SciDataCon 2025

Monday 13 October 2025 - Thursday 16 October 2025

Brisbane Convention & Exhibition Centre

Book of Abstracts

Contents

Presentations Session 1: CAREful Indigenous Data Governance / 5

Developing a framework for enhancing data literacy in the Indigenous Australian community

Authors: Becki Cook[1]; Kerrie Mengersen[1]; Stephen Corporal[2]

[1] *QUT Centre for Data Science, Queensland University of Technology*

[2] *Indigenous Data Network, University of Melbourne*

Corresponding Authors: stephen.corporal@unimelb.edu.au, k.mengersen@qut.edu.au, r24.cook@qut.edu.au

Enhancing data literacy is important for Aboriginal and Torres Strait Islander Peoples because it essential for effective engagement and communication with health services, self-determination, personal health management and Indigenous Data Sovereignty. Data literacy incorporates accessing, using and sharing data in order to make decisions within personal and professional spheres and it is a critical skill for participation in contemporary society. However, while there has been a movement toward ensuring good Indigenous Data Governance processes, for example the CARE Principles developed by the (Global Indigenous Data Alliance, 2018), and the National Indigenous Australians Agency's Framework for Governance of Indigenous Data (2024) and improvements around data infrastructure and research capability within Australia, as led by the (Australian Research Data Commons, 2021), there has not been any specific focus on improving data literacy within Indigenous Australian communities. This work also aligns with the National Agreement on Closing the Gap Priority Reform 4, "Shared Access to Data and information at a regional level"which highlights (a) the need for communities to have access to data to make decisions about their future; (b) data and information sharing; (c); transparency and capacity building so communities have access and ability to collect and use data (Department of Prime Minister & Cabinet, 2020) to assist with closing the gap between Indigenous and non-Indigenous Australians.

There is currently no mechanism for determining data literacy in Indigenous communities or any framework to enhance Aboriginal and Torres Strait Islander Peoples'data literacy. The overall aim of this research is to support the enhancement of data literacy within the Aboriginal and Torres Strait Islander community by developing and evaluating a community informed and culturally appropriate data literacy enhancement framework that can be used by organisations.

This research was designed in collaboration with the Aboriginal and Torres Strait Islander Community Health Service Brisbane (ATSICHS). This co-design approach is important to ensure that research is Indigenous led and meets the needs of the Aboriginal and Torres Strait Islander community. Furthermore, this approach aligns with developing data literacy initiatives as suggested by Komosar, et al. (2024) data literacy practices need to be conceptualised through the group it is about. This presentation will highlight key considerations around Indigenous research and data collection as well as explore data literacy from an Indigenous perspective. It will also explain how the framework has been developed and how it will be implemented.

This research adopted an Indigenous research methodology and incorporated a case study approach using mixed methods. Within this study qualitative data has been collected through interviews or focus groups and yarning (Bessarab & Ng'andu, 2010) with community members and ATSICHS staff, and quantitative data collected through surveys. Indigenous Standpoint Theory (Moreton-Robinson, 2013) and Cultural Interface Theory (Nakata, 2007) were incorporated as theoretical and conceptual frameworks to assist interpret findings and theorise knowledge from an Indigenous perspective. This research upholds the Australian Institute for Aboriginal and Torres Strait Islander Studies (AIATSIS) Code of Ethics for Aboriginal and Torres Strait Islander Research (2020).

Data collected through interviews underwent thematic analysis using the six steps outlined by Braun and Clark (2006). Responses were also compared quantitatively, for example, where participants suggested their confidence level when using data. All data were considered through an Indigenous lens to ensure that Indigenous Australian voices were privileged, and Indigenous perspectives are considered. Following the analysis a number of factors were identified as being important for inclusion in the data literacy framework. This included community perceptions, understanding of and engagement with data, as well as how ATSICHS staff were using data in their interactions with clients and

how they could improve data literacy. The incorporation of the identified factors into the framework will be discussed in this presentation.

It is expected that this framework will assist ATSICHS staff engage with clients about their data, and data held and collected at ATSICHS. This research has potential to improve data literacy Indigenous communities by supporting staff to empower community members regarding data. Establishing a data literacy framework that encompasses best practices in data access, collection and utilisation, can enhance self-determination and data management within Indigenous Australian communities, thereby fostering improved socio-economic outcomes. This research could be used to support other community organisations and Indigenous communities to enhance data literacy.

References

Australian Government. (2024). Framework for Governance of Indigenous Data. Commonwealth of Australia. Retrieved from https://www.niaa.gov.au/resource-centre/framework-governance-indigenous-data

Australian Institute of Aboriginal and Torres Strait Islander Studies. (2020). AIATSIS Code of Ethics for Aboriginal and Torres Strait Islander Research. Canberra: AIATSIS. Retrieved from http://aiatsis.gov.au/ethics

Australian Research Data Commons. (2021). HASS and Indigenous Research Data Commons. https://ardc.edu.au/hass-and-indigenous-research-data-commons/

Bessarab, D., & Ng'andu, B. (2010). Yarning about yarning as a legitimate method in Indigenous research. International Journal of Critical Indigenous Studies, 3(1), 37-50. https://search.informit.org/doi/10.3316/informi

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.

Department of Prime Minister & Cabinet. (2020). National Agreement on Closing the Gap. Retrieved from https://www.closingthegap.gov.au/national-agreement

Global Indigenous Data Alliance. (2018). CARE Principles for Indigenous Data Governance. https://www.gida-global.org/care

Komosar, A., Koprivica, S., Taibi, D., Ivic, A., Stefanović, D., & Kijanovic, S. (2024). Data literacy measurements: a systematic literature review. https://doi.org/10.33965/IS2024_202401C008

Moreton-Robinson, A. (2013). Towards an Australian Indigenous women's standpoint theory: A methodological tool. Australian Feminist Studies, 28(78), 331-347. https://www.tandfonline.com/doi/pdf/10.1080/0816

Nakata, M. (2007). Disciplining the savages, savaging the disciplines. Aboriginal Studies Press.

Presentations Session 2: Data and Research & Data Science and Data Analysis / 7

Beehive Health Patterns Using Multi-modal Data Analysis and Unsupervised Machine Learning

Authors: Arbind Agrahari Baniya[1]; Jiqiang (John) Chen[2]

Co-authors: Kieran Murphy [1]; Thabo Thayalakumaran [1]

[1] *Agriculture Victoria Research, Department of Energy, Environment and Climate Action, Victorian State Government*

[2] *University of Melbourne*

Corresponding Authors: thabo.kumaran@agriculture.vic.gov.au, jiqiang.chen@student.unimelb.edu.au, arbind.agraharibaniya@agr kieran.murphy@agriculture.vic.gov.au

Agriculture Victoria Research (AVR) undertook a project for multimodal data analysis and anomaly detection for beehive health to address the critical challenge of determining hive health comprehensively in a pollination environment where diverse modalities of data are at interplay. Honeybee populations play an essential role in pollination within Victoria's horticulture industry, making the monitoring and maintenance of hive health crucial for agricultural productivity and pollination sustainability. However, existing methods for monitoring hive health often rely on commercial sensor companies that typically provide inferred hive health matrices using proprietary modelling and algorithms, often lacking transparency and scientific validation. At the same time, commercial monitoring systems heavily rely on IoT sensors alone for measurements and do not take into account other environmental variables at play that cannot be measured via these sensors. This highlights a

gap and opportunity for assessing hive health using multimodal data streams and the application of scientifically validated data-driven methodologies for health pattern recognition.

The project collected multimodal data from—IoT sensors (e.g., temperature, humidity, acoustic activity), environmental DNA results, agrochemical records and manual inspections—to demystify interconnected insights into the health of a hive. The data was collected across three different almond farms in the Sunraysia region during 2024 almond pollination. The project aimed to develop a system capable of detecting abnormalities in hive behaviour and determining the overall health status of the hive by cross-checking with chemical sprays and human inspection to demonstrate functional relationships using a data-driven approach.

To address this, the project conducted a comprehensive data analysis on the high-dimensional dataset, including data pre-processing, feature engineering, visualisation and modelling. For modelling, data-driven unsupervised machine learning alongside classical statistical methods was used for anomaly detection. The models, which include Gaussian Mixture Models (GMM), Local Outlier Factor (LOF), and Isolation Forest (IF), were used to flag unusual hive behaviours over the pollination period.

An interactive data dashboard was then created to visualise the results of anomaly detection and cross-validate the identified anomalous behaviours in a particular hive/site with corresponding spray records and eDNA results. The anomaly detection system successfully identified periods of abnormal colony behaviour, which were cross-referenced with chemical records and validated through human inspection data.

These initial findings confirm the scientific base for future research and development in this new domain. For instance, incorporating eDNA data directly into model training would enable the model to capture additional ecological variables that may improve its pattern recognition capability. Similarly, establishing a standardised rubric for human inspections with high accuracy could serve as labels for supervised learning models, enabling the exploration of a broader range of machine-learning approaches. Scaling data collection across more sites and longer periods would further improve model robustness by capturing seasonal trends and environmental influences on hive health that could be made available for industry as a generalisable service.

Presentations Session 9: Empowering the global data community for impact, equity, and inclusion / Education / 9

The Effect of Data Journalism Knowledge to the Good Graphing Practice in Indonesian Media

Authors: Arran Ridley[1]; Utami Diah Kusumawati[2]

Co-author: Ambang P [3]

[1] *Monash University*

[2] *RMIT*

[3] *Universitas Multimedia Nusantara*

Corresponding Authors: utamid.kusumawati@gmail.com, ambang@umn.ac.id, arran.ridley@gmail.com

The availability of more data in the public sphere has driven the rise of data journalism practices in media worldwide, including in Asia (Mutsvairo, 2019). Consequently, audiences are now more exposed to data-based products when consuming news. Data journalism serves several functions, from explaining to the audience more accurately because they are presented with numbers and data to helping the audience save time from reading too many words (Wang &Li, 2019). Meanwhile, journalists believe that using data journalism can improve the quality of their products and the journalists themselves (Heravi & Lorenz, 2020). Through data journalism, journalists collect, clean, analyze, visualize, and narrate data in addition to reporting and publishing their stories, rather than just traditional reporting. To undertake these data-driven steps, journalists are beginning to produce new technologies such as applications and interactive journalistic products where users can gain insights from the data (Howard, 2017).

Among the data journalism practices found in media, we can see data visualization emerging as a component of this practice. Data visualization, as a product of data journalism, transforms sorted and analyzed data into graphic form. Graphics become a communication tool for the eye that has the function to 'store, understand, and communicate' useful information. Cairo (2013) explains that graphics refer to the "aggregate of data"and help us find many patterns among the data. The shift from data to visual representation aims to enhance cognitive capabilities (Card, S. et al., Mackinlay, J. & Shneiderman, B., 1999). This enhancement includes "improving memory and processing resources available to users", saving time in searching for information, increasing the ability to detect patterns, "enabling perceptual inference operations", applying mechanisms related to perceptual attention to monitor something, and encoding information.

Despite having benefits in transforming complex data into simple and easily understood visual information, data visualization still has the potential to mislead. Inaccurate graphics found during pandemic reporting and going viral, for example, have caused misinformation among the public. Cairo (2019) mentions inadequate design and labeling, exaggerated perspectives of scale and proportion, biased data visualization, vague and incomplete data, as some factors contributing to the inaccuracy of graphics. Misleading graphics can then lead to the manipulation of truth or that graphics can lie (Cairo, A., 2019).

The fact is not all these misleading graphics are made intentionally; rather, they are poorly executed (Crisan, A., 2022). In other words, literacy about data and visualization is needed to increase knowledge to detect miscalculated and deceptive graphics. Research conducted by Lee et al. (2021) mentions that literacy in data visualization is related to the ability to understand and interpret graphics. The definition of visual literacy, as stated in previous research 'VLAT: Development of a Visualization Literacy Assessment Test' has almost the same meaning as it refers to the ability to 'read and interpret' data visualizations and obtain information from data visualizations. Besides the ability to understand graphics, literacy also refers to the ability where graphic creators and designers have good graphic creation practices, including data integrity. Data integrity, coined by Jacques Bertin, occurs when the graphics are 'clear, simple, and easy to understand' (Olande, O., 2013). Having good graphic creation practices is crucial because journalism ethics and graphic design mention that both hiding the truth (about data and graphics) and portraying graphics in twisted truth are 'highly unacceptable' (Cairo, 2020).

Given the above phenomena, this research attempts to identify the effect of data journalism knowledge among Indonesian journalists to the good graphing practices in Indonesian media. Using a mixed- methods approach, we are going to examine data visualization products collected from the submissions of participants for the Indonesian Data Journalism Award (IDJA) over two consecutive years (2023 and 2024). IDJA, as quoted from its website, is the first data journalism competition held in Indonesia by the nonprofit journalism organization Indonesian Data Journalism Network (IDJN) and was created in 2023. This competition is joined by hundreds of media in Indonesia, both local and national, as narrated in an interview with the Executive Director of the organization, Wan Ulfa Nur Zuhra (2023).

To measure the level of good graphic creation practices, the researchers will use an index measured from a set of criteria combining several theories related to statistics, graphics, design, and ethics from experts in the field and applying this set of good graphic creation criteria to the IDJA data visualization submissions to determine the level of good graphic creation practices. Meanwhile, to measure the level of data journalism knowledge among Indonesian journalists, the researcher will distribute survey to journalists who participate in the IDJA award (2023 and 2024). At last, the researcher will use Pearson correlation method to measure the relationship and effect between data journalism literacy towards the good graphing practice among Indonesian journalists.

Furthermore, to understand the data more comprehensively, this study will take a qualitative approach by interviewing data journalists and graphic designers as follow-up to help answer 'the why and the how'of the analysed quantitative data.

Poster Session / 12

Relationship building in CARE

Authors: Kristen Smith[1]; Stephen Corporal[2]

[1] *The University of Melbourne*

[2] *University of Melbourne*

Corresponding Authors: kristens@unimelb.edu.au, stephen.corporal@unimelb.edu.au

The presentation will discuss the importance of relationship building between the university researchers and local Aboriginal community organisations as part of a national research project. The project from the University of Melbourne is led by Distinguished Prof Marica Langton included the current presenters who are members of the research team.
The research project was codesigned by university researchers and three community controlled organisations of the Aboriginal and Torres Strait Islander Community Health Service Brisbane, Southeast Queensland, Binarri Binyja Yarrawoo (BBY) East Kimberley, Western Australia and Ngaanyatjarra Pitjantjatjara Yankunytjatjara NPY Women's Council Alice Springs, NT.
Of the four priority reforms for action from the National Agreement on Closing the Gap 2020 (National Agreement) this research project was developed in relation to priority 4. The sharing data and information with Aboriginal and Torres Strait Islander people to ensure Aboriginal and Torres Strait Islander people have more power to determine their own development. The project codesign involved preliminary meetings between the researchers and members of each community controlled organisation to put community agreements in place. These preliminary meetings are where much of the relationship building is located to establish the ongoing relationships throughout the project. Relationship is more important than the task. The protocol of providing information about yourself and your cultural location so people can connect and establish a relationship with you culturally, socially and politically. This was for non-Indigenous people as well so that they could connect with the Indigenous individual, family and community too (Corporal 2017).
The community priorities of Indigenous data governance and data ecosystems as well as technical infrastructure and Indigenous data capacity building were the main aims of the interview questions that were agreed on
The process of building relationships with each of the Aboriginal and Torres Strait Islander community controlled organisations was an important part of the process to have a connection with the local community organisations, families and individuals. Building relationships takes time and one of the most appropriate ways of building relationships is through listening to stories. Many Aboriginal and Torres Strait Islander people, especially Elders, tell the stories or yarning and if you listen closely, you will hear what they want to tell you (Corporal 2017) p:221
This project has the potential to produce outcomes from a grassroots focus on Indigenous data stewardship and data governance, with potential results that can be upscaled across community-controlled organisations nationally.

13

Building an Open Data Collaborative Network in Asia-Oceania

Author: Masaki Kanao[1]

[1] *Research Organization of Information and Systems*

Corresponding Author: kanao@nipr.ac.jp

Session Title:
Building an Open Data Collaborative Network in Asia-Oceania

Session Contents:
One of the primary objectives in the current Open Science movement is to initiate new research fields and technologies leveraging the vast amounts of data now being generated across numerous scientific domains. Adherence to the FAIR Principles for data has been recognized as a standard for data-oriented activities, driving the development of Open Data infrastructures. These include enhanced and integrated metadata catalogues, persistent identifiers (PIDs), metadata standards for research data management, and certification of data repositories to ensure the long-term stewardship and management of quality-assessed data. Additionally, many academic institutions are adopting

persistent identifiers for people, places, and other entities as a best practice for preserving and providing access to data generated by their research activities.
Despite these advancements, significant efforts are still required to address the challenges surrounding scientific research data, particularly regarding its sharing and reuse. While the importance of multidisciplinary data integration is widely acknowledged, data reuse by scientists—whether within their own discipline or across others—remains challenging in terms of the FAIR Principles. Issues such as difficulties in discovering and accessing data or insufficient metadata to enable seamless analysis are common. Furthermore, when sharing research data with the public domain, including policymakers, additional contextual information is often required to ensure proper understanding.
To address these challenges, it is crucial to foster collaboration among researchers and scientists across various disciplines and countries. Establishing systems that facilitate interaction between research data users and providers is key to improving data sharing, reuse, and interdisciplinary integration.

Building on Open Data and Open Science principles, this session invites "**Lightning Talks**" on topics related to data-oriented activities in Asia and Oceania within the context of a globally developing society. For example, data-sharing platforms have been developed as a result of discussions held during previous WDS and CODATA conferences. An interactive discussion session will also be convened following the set of talks.
The purpose of this session is to build consensus among stakeholders on various aspects of research data management, aligning with open research policies and the FAIR principles. The session will explore new approaches for promoting interdisciplinary and collaborative research, advance data management solutions, and facilitate efficient data reuse across diverse scientific disciplines. These efforts will be supported by evidence and feedback from communities across Asia and Oceania.

This session is related to the following **SciDataCon 2025 conference themes**.
- Rigorous, responsible, and reproducible science in the era of FAIR data and AI
- Open research through Interconnected, Interoperable, and Interdisciplinary Data
- Empowering the global data community for impact, equity, and inclusion
- Infrastructures to Support Data-Intensive Research - Local to Global

Lightning Talks: (each 5 min.) (Presenters were fixed)

- Johnathan Kool (Australian Antarctic Division), "International Collaborative Data Management –Challenges and Opportunities"
- Noorsaadah Abd Rahman (Malaysian Open Science Alliance), "Malaysian Open Science Platform"
- Juanle Wang (China Academy of Science), "Challenges of Open science data governance and China's practice"
- Pei-shan Liao (Research Center for Humanities and Social Sciences, Academia Sinica), "Sharing Social Science Data: The Importance and Challenges"
- Sa-kwang Song (Korea Institute of Science and Technology Information), "Research Data Management Platform at KISTI: DataON and Beyond"
- Madiareni Sulaiman (University College London), "Navigating the Research Data Management Compliance Maze as an Indonesian Early-Career Researcher: Unravelling Policies Across Borders"
- Chandra Shekhar Roy (Bangladesh Bureau of Statistics), "Open Government Data (OGD) at Bangladesh NSO"
- Chantelle Verhey (World Data System International Technology Office）, "Polar data Discovery and Mobilization Pathways"
- Masaki Kanao (Research Organization of Information and Systems), "International Symposium on Data Science (DSWS); Asia & Oceania Collaboration"

(Presenters were fixed)

Structured Discussion: (30 minutes)

Following the Lightning Talks (5 minutes each) by presenters from individual centers and organizations across Asia and Oceania, the session conveners will facilitate a **Structured Discussion** (30

minutes) involving both the speakers and the audience. The main topics for the panel discussion include:
- What types of scientific data can be shared under the FAIR Principles using existing platforms within the Asia and Oceania community?
- How can advanced technologies for creating and sharing data, developed by individual centers and organizations, be effectively shared across the region?
- What policy arrangements are needed to facilitate data sharing on a regional basis, and who can enact them?
- What are the practical challenges in increasing data sharing within the Asia and Oceania community, and what realistic solutions can be proposed in alignment with global strategies and regulations?

The lightning talks and structured discussion will provide valuable insights, sparking discussion on how to establish international collaborative networks related to open data in the Asia and Oceania region, and the development of concrete international cooperative frameworks. The primary goal of the session is to build consensus among stakeholders on various aspects of research data management, ensuring alignment with open research policies and the FAIR principles in the region.

Conveners:
- Masaki Kanao (Research Organization of Information and Systems)
- Johnathan Kool (Australian Antarctic Division)
- Juanle Wang (China Academy of Science)

Suggested session types
• Lightning Talks and Structured Discussion
(Lightning talks are through invitation only)

Poster Session / 15

International collaboration activities for open sciences and data sciences by the Polar Environment Data Science Center, Tokyo, Japan

Author: Masaki Kanao[1]

[1] *Research Organization of Information and Systems*

Corresponding Author: kanao@nipr.ac.jp

Polar Environment Data Science Center (PEDSC) of the Joint Support-Center for Data Science Research (DS), the Research Organization of Information and Systems (ROIS) aimed to promote opening and sharing scientific data obtained by research activities in polar regions. Its purpose is to strengthen collaboration with universities and other communities, and to support creation of further scientific outputs and advancement of polar research. PEDSC is also expected to play a role of the national data center for polar science in Japan. In this presentation, several international collaborative activities regarding open and data sciences conducted by PEDSC are introduced.

1) Invitation of the International Strategic Advisors
A total of five famous researchers / data managers from UK, China and Australia were invited by PEDSC to serve as the ROIS International Strategy Advisor in FY2022, 2023 and 2024 (+2025 as planned). They stayed at ROIS and had fruitful meetings and discussions with the staff of ROIS-DS and gave precious advice and suggestions in terms of data management and operation of ROIS and PEDSC. They also made related research presentations at institutional seminars and international symposiums such as DSWS. These invitations made strength the collaboration not only with two institutions where they are belongings and ROIS, but also among open and data science communities in the areas of Asia and Oceania, as well as global data initiatives such as WDS and CODATA.

2) Organizing International Data Science Symposium
PEDSC organized several international conferences hosted by Japan, with many participants form

overseas. 1) International Workshop on Sharing, Citation and Publication of Scientific Data across Disciplines (DSWS-2017, December 2017, Tachikawa, Tokyo), 2) International Workshop on Data Science - Present & Future of Open Data & Open Science - (DSWS-2018, November 2018, Mishima, Shizuoka), 3) International Symposium on Data Science - Global Collaboration on Data beyond Disciplines - (DSWS-2020, September 2020, online), 4) International Symposium on Data Science 2023 - Building an Open-Data Collaborative Network in the Asia-Oceania Area - (DSWS-2023, December 2023, Science Council of Japan (hybrid conference), https://ds.rois.ac.jp/article/dsws_2023). Special issues for 3) and 4) are published as the Special Collection in the CODATA Data Science Journal (https://datascience.codata.org/collections/open-data-collaborative-network).

3) SCAR - Standing Committee on Antarctic Data Management (SCADM)
The Scientific Committee on Antarctic Research (SCAR) under the International Science Council (ISC) has established the Standing Committee on Antarctic Data Management (SCADM) to discuss data management and publication in the Antarctic and exchange information regarding data activities in polar regions. SCADM has developed a data policy called the Data and Information Management Strategy of SCAR. SCADM also requests that each country involved in Antarctic observations should establish a National Antarctic Data Center (NADC). In addition to the in-person meetings once a year, including the SCAR General Meeting every two years, SCADM conducts monthly online meetings to facilitate close information exchange.

PEDSC has been served as the NADC in Japan and participates in the SCADM related activities. As a part of NADC activities, PEDSC publishes metadata and data obtained from polar regions through a Polar Science Database (http://scidbase.nipr.ac.jp/). The information registered is also forwarded to the Antarctic Master Directory (AMD) within the Global Change Master Directory (GCMD) of NASA. The GCMD aggregates metadata from SCADM countries.

As related activities, SCADM collaborates with the International Arctic Science Committee (IASC) under ISC and WDS, CODATA and has been continuously holding international symposiums (Polar Data Forum; PDF) since 2013. PEDSC has been involved in organizing the PDFs from the first symposium (2013 at the National Museum of Nature and Science in Tokyo).

Presentations Session 4: Data Stewardship / 16

Geoscience Sample Management and Discovery through Best Practices and Digital Solutions –CSIRO Mineral Resources (Discovery) Case Study

Author: Anusuriya Devaraju[1]

Co-authors: Anusree Ramachandran Menon [1]; Jacob Walmsley [1]; Kirsten Fenselau [1]; Tenten Pinchand [1]; Tina Shelton [1]

[1] *CSIRO*

Corresponding Authors: tenten.pinchand@csiro.au, anu.ramachandranmenon@csiro.au, kirsten.fenselau@csiro.au, jacob.walmsley@csiro.au, tina.shelton@csiro.au, anusuriya.devaraju@csiro.au

Physical samples are essential research assets. Systematic curation ensures their accessibility and reusability for future scientific studies. CSIRO Mineral Resources (CMR) scientists collect diverse physical samples—including rock, regolith, water, and vegetation—for research and mineral exploration projects. These samples are costly to obtain, irreplaceable, and critical for generating and validating downstream research data. In mineral exploration, sample analyses allow geologists to compare newly discovered deposits with previously studied ones, helping identify similarities and new exploration targets. However, the lack of standardised curation practices and digital solutions has resulted in fragmented, ad-hoc sample management. Metadata is often recorded manually and inconsistently, increasing the risk of data loss, inefficiencies, and missed opportunities for future analysis. This disorganisation also poses safety hazards, including trip risks, dust accumulation, and storing unidentified or mislabelled samples. Over time, labels on archived samples have faded or detached, and pallet cardboard boxes in storage facilities have deteriorated, making it difficult for researchers to locate and access samples beyond their immediate projects. As a result, researchers

frequently rely on project leads or laboratory technicians for assistance, adding to operational inefficiencies. To address these challenges, it is essential to enhance sample accessibility while reducing costs and staffing efforts associated with their management and tracking.
This presentation outlines best practices developed and implemented within CMR Discovery to enhance the physical and digital curation of samples. These practices include standardised procedures for sample identification, labelling, and packaging, as well as the deployment of a comprehensive sample management system to ensure integrity, traceability, and efficient curation. Built on Specify, an open-source biological collections management platform, this system streamlines the recording and discovery of research samples that were previously challenging to track. We highlight the system's key functionalities, including an adapted data model tailored for geological samples and its integration with FAIR-enabling services to enhance sample discoverability and reuse in future mineral exploration projects. Finally, we share key lessons from implementing these practices and insights gained from system adoption, demonstrating an effective integration of technical and non-technical components.

Presentations Session 8: Policy and Practice of Data in Research; Data, Society, Ethics and Politics / 17

Governing Sensitive Personal Data Access and Data Publication in Intra-, Inter- and Transdisciplinary Research

Author: Olga Churakova[1]

Co-authors: Christine Krebs [1]; Dirk Verdicchio [1]

[1] *University of Bern*

Corresponding Authors: olga.churakova@unibe.ch, christine.krebs@unibe.ch, dirk.verdicchio@unibe.ch

Open Research Data (ORD) is fundamental to the transparency and reproducibility of scientific results. It fosters scientific exchange and networking. In line with ORD strategies and funding agencies' requirements, research data should be published as openly as possible. Despite the advantages ORD brings to research, however, the publication of research data can be subject to restrictions. This applies to sensitive personal data, health-related data, synthetic data, and copyright protected data (e.g., software, protected works of literature, art and related works). Additionally, the rapidly increasing research on and with Artificial Intelligence (AI) in intra-, inter- and transdisciplinary studies, particularly in biomedicine and engineering, health and environmental sciences, poses certain risks when working with sensitive data. Therefore, researchers need best practices for their work with sensitive data considering ethical questions before data collection and/or developing of AI as well as throughout the research data life cycle; they also need to be aware of the lack of control and the potential risks when processing unpublished and sensitive research data with AI.

The Data Stewards of the Open Science Team at the University Library of Bern support researchers across the University of Bern and Insel Hospital by developing guidelines on sharing and publishing sensitive and personal data. These guidelines are intended to ensure that the management and publication of sensitive, personal and protected data comply with ethical as well as legal requirements and FAIR principles.

In our presentation, we will address key aspects of these guidelines on managing and publishing sensitive, personal and protected data across different disciplines. Firstly, we will show how to manage access to sensitive personal data, health-related data, synthetic data, and non-personal protected research data, which are subject of the Swiss cantonal and federal regulations and recommendations aligning with General Data Protection Regulations, which are relevant for research projects supported by the European Commission. Secondly, we will draw on how to publish unprocessed (raw) data, pseudonymized and synthetic data, anonymized and copyright-protected data to facilitate transparency and excellence of research data in intra-, inter- and transdisciplinary research projects. The third aspect concerns data protection, security and privacy when personal data is being processed with AI. Finally, we will discuss how regional and national heterogeneity in guiding researchers working with sensitive data can be governed and aligned in research data management globally.

Poster Session / 18

From Cloud to Clarity: Architecting Resilient AI Systems in Enterprise Environments

Author: Weiying Chen[None]

Corresponding Author: wchen@costco.com

Delivering AI systems in enterprise settings requires more than model optimization —it demands infrastructure clarity, cross-functional orchestration, and the ability to navigate complexity over time. This poster highlights real-world lessons from deploying intelligent systems within a cloud-native architecture, combining applied technical experience with strategic delivery in organizational environments undergoing transformation.

The work draws from two significant experiences: the completion of the **UC Berkeley Professional Certificate in Machine Learning and AI**, which provided a strong foundation in applied AI practices; and active participation in the **2025 NVIDIA GTC Hackathon**, where real-time problem-solving and innovation under pressure shaped the architecture of resilient, scalable ML pipelines.

The poster explores the intersection of infrastructure, AI delivery, and organizational complexity through a practitioner lens. Topics include:

1. **Infrastructure design** using GCP, Kubernetes, and hardened microservices
2. Techniques for building AI systems that can scale across teams and environments
3. Patterns for **security**, **observability**, and operational clarity at deployment
4. Navigating delivery friction in systems where multiple priorities compete

Many teams attempt machine learning projects, but few sustain them to completion. This poster examines why some AI efforts quietly stall, even with strong talent and intent, and what's required to close that gap.

Drawing from the personal experience of completing a rigorous internal nanodegree, this work surfaces the often-invisible factors that determine whether AI systems succeed —including follow-through, detachment from noise, and **infrastructure maturity**.

In addition to modeling static user baselines, the system leverages **sequential modeling** techniques to capture time-ordered patterns in user activity. By understanding not only what a user does but when and in what sequence, the model can detect context-aware anomalies—such as unusual login behavior or compromised session flow—that would go unnoticed with traditional static rules. This approach supports long-term adaptability and improves detection of subtle shifts in behavior across evolving enterprise systems.

This poster will be of interest to engineers, data scientists, and research leaders who seek to operationalize AI in large organizations —particularly those undergoing cloud migration or seeking to create lasting research-ready infrastructure. It connects theory to delivery and highlights the critical infrastructure layer where sustainable AI impact begins.

19

Units, Symbols and Terminology for Data Driven Science: Philosophy to Pragmatism

Authors: Aileen Day[1]; Blair Hall[None]; Jeremy Frey[1]; Max Gruber[2]; Samantha Pearman-Kanza[1]; Stuart Chalk[3]; Vanessa Seifert[4]

[1] *University of Southampton*

[2] *PTB*

[3] *UNF*

[4] *University of Athens*

Corresponding Authors: s.pearman-kanza@soton.ac.uk, maximilian.gruber@ptb.de, vs14902@bristol.ac.uk, schalk@unf.edu, a.e.day@soton.ac.uk, blair.hall@measurement.govt.nz, j.g.frey@soton.ac.uk

The rapid changes in the way we undertake scientific research driven by digitalisation and artificial intelligence means we need to look again at the basis of scientific methods, requiring the input from the philosophy of science and one end while being pragmatic about the uses of technology at the other. This approach will provide a way to ensure that our research and teaching is relevant, effective and efficient.

More specifically, this means revisiting the topics of quality and provenance frameworks as developed by certain disciplines; exploring the role of metadata and semantics in such frameworks and in the context of use of data by AI systems; exploring how the digital representation of units is best managed in such frameworks.

Consequently, the session will address the following questions:

1. How should we best interact with AI systems, how can they contribute to scientific discovery?
2. How can we ensure that sensible provenance is provided with data that is consumed and produced by AI systems? Why is it so hard to get quality metadata?
3. How do we ensure that the metadata and semantic framework is useful and used by researchers and not seen as yet another barrier to work?
4. The role of metadata standards and ontologies. How do we evolve the publication and dissemination framework to provide these details?
5. What software tools are needed to support researchers to use digital terminology and units? What support can be built into programming languages?

This session builds on the agenda set by a successful small conference held at the Royal Society of Chemistry in London in March 2025, and convened by the UK Physical Science Data Infrastructure (PSDI, www.psdi.ac.uk), the International Union for Pure and Applied Chemistry (IUPAC, Green and Gold Book projects) and CODATA (DRUM Task Group).

Proposed speakers:
We appreciate that this list would be too many for a 90 min session but we wanted to show that we had a wider group of speakers to call upon to support the session as we will need to raise funding for some of them to be able to attend. We will seek funding to bring several speakers to Brisbane and take advantage of the hybrid format to enable those who can not travel to participate online or via recorded talks if the time zones are an issue.

1. Philosophy of science in the AI age –Vanessa Seifert (Greece), Will McNeil (UK)
2. IUPAC –Digital Transformation of IUPAC –Stuart Chalk. (USA) , Jeremy Frey (UK)
3. Digital SI –Max Gruber (Germany) Pragmatic Semantics –Samantha Pearman-Kanza (UK) 4.Usable Metadata –Cerys Willoughby (UK)
4. Definition of units, symbols and terminology in PSDI - Aileen Day (UK) (presenting/participating remotely)
5. M-Layer –Blair Hall, (New Zealand)
6. CODATA DRUM Task Group members

.

21

Double Identification to Support Equitable Participation in Global Open Research

Authors: Xiaoli Chen[1]; QI XU[2]; Satoko Fujisawa[3]; JIA LIU[None]

Co-authors: xiaolei xia [4]; shu wang [4]; Lijuan Wang [4]

[1] *DataCite*

[2] 中国科学院国家空间科学中心

[3] *Japan Link Center*

[4] *Computer Network Information Center (CNIC) of the Chinese Academy of Sciences (CAS)*

Corresponding Authors: xiaoli.chen@datacite.org, xlxiao@cnic.cn, liujia@cnic.cn, xuqi@nssc.ac.cn, wlj@cnic.cn, satoko.fujisawa@jst.go.jp, wangshu@cnic.cn

The exponential growth of data-intensive research demands robust, interconnected infrastructure that can seamlessly translate local scientific efforts into global knowledge resources. This session directly addresses the critical challenge of developing scalable, interoperable research data infrastructure that supports the complex journey of research data from local repositories to international discovery platforms.

Research data increasingly represents a complex, dynamic ecosystem, and data repositories continue to shape and implement strategies to ensure local scientific outputs transcend institutional and national boundaries to realize their full potential. Persistent identifier (PID) systems emerge as a crucial facilitator, serving as the connective tissue that enables data discoverability, accessibility, and reusability across diverse research contexts.

Our session will present a comprehensive exploration of how national-level data centers can strategically develop infrastructure that supports data-intensive research. Drawing on concrete case studies from the Chinese National Space Science Data Center (parallel implementation of CSTR and DataCite DOI) and Japan Link Center (JaLC, offering JaLC DOI and DataCite DOI), we will demonstrate practical approaches to transforming local research data repositories into globally accessible scientific resources.

Session Structure and Approach: The 90-minute session will be structured into three integrated segments, each designed to provide insights into infrastructure development for data-intensive research. The first segment will provide theoretical and conceptual foundations, examining the current landscape of research data infrastructure. Presentations from DataCite, CSTR, Chinese National Space Science Data Center and Japan Link Center will offer nuanced perspectives on identifier implementation strategies.

Proposed Session Agenda:
Introduction (15 minutes): Framing the global challenges in research data infrastructure
Presentations (40 minutes):
DataCite: Global perspectives on persistent identifier adoption
CSTR: PIDs infrastructure for open data sharing
Chinese National Space Science Data Center: Chinese National Space Science Data Center User Case
Japan Link Center: Technological and community-based approaches to metadata infrastructure for research data
Interactive Discussion and Q&A (25 minutes): Collaborative exploration of challenges and opportunities

The session will examine how persistent identifier systems can be strategically deployed to:
Create seamless pathways between local and global research data ecosystems
Enhance the discoverability and impact of scientific research outputs
Support interdisciplinary and international research collaboration
Develop more equitable and accessible research infrastructure

Participants will gain actionable insights into developing infrastructure that supports data-intensive research, with a particular focus on technological, policy, and collaborative strategies that enable effective research data sharing.

By bringing together perspectives from national data centers, identifier service providers, and research infrastructure experts, this session will offer a comprehensive exploration of the complex journey from local data generation to global scientific discovery.

Proposed speakers and presentations:
CSTR-Liu Jia
Chinese National Space Science Data Center –Qi Xu
DataCite-Xiaoli Chen
JaLC - Satoko Fujisawa

Discussion prompts (work in progress)
As a data center or institution manager, how do you think identifiers can be better used in the context of resource sharing?
As an identifier service provider, what services do you think are most important to offer in the context of resource sharing?

Poster Session / **22**

The Establishment of the National Bushfire Data Commons Dashboard

Author: Richard Sinnott[1]

Co-author: Luca Morandini [1]

[1] *The University of Melbourne*

Corresponding Authors: luca.morandini@unimelb.edu.au, rsinnott@unimelb.edu.au

The Australian Research Data Commons (ARDC) funded Bushfire Data Commons (BDC) established a range of projects focused on an ever increasing problem to Australia: bushfires. These projects are highlighted in https://ardc.edu.au/program/bushfire-data-challenges/. One of the ARDC funded projects focused on establishment of a front end dashboard to showcase the data sets and tools arising from the various BDC projects. This talk will provide an overview of the Cloud-based dashboard that was established and the technologies used. We explore the various data sets stemming from the different ARDC funded projects as well as other data sets of relevance to the bushfire research community. This included: national air quality data; under-insured houses; population statistics and the distribution of the indigenous population; hospital admissions data related to pulmonary disorders; distributions of rare and threatened species; the historic fires of Australia including bushfires and prescribed burns; fuel load data, and satellite and raster data including data from the Sentinel-2 satellite.

The BDC platform also offered an API providing programmatic access to the data, e.g. through Jupyter notebooks. The talk will showcase a range of examples of analytics around the data, e.g. the number and size of prescribed burns vs bushfires.

Poster Session / **23**

Introducing the Australian Internet Observatory Research Dashboard

Author: Richard Sinnott[1]

Co-author: Luca Morandini [1]

[1] *The University of Melbourne*

Corresponding Authors: rsinnott@unimelb.edu.au, luca.morandini@unimelb.edu.au

The Australian Internet Observatory (AIO - https://internetobservatory.org.au/) has been funded by the Australian Research Data Commons (ARDC –www.ardc.edu.au) in 2024 to support large-scale access to and use of social media and digital data more broadly by Australian researchers. A core part of the AIO is the Australian Internet observatory Research Dashboard (AIReD - https://www.aio.eresearch.unimelb.edu.au established at The University of Melbourne. The AIReD platform includes large-scale Cloud-based data harvesting from Reddit, Mastodon, BlueSky, FlickR, Foursquare, YouTube, and historic data from Twitter. The platform supports sentiment analysis on the posts as well as topic modelling to help understand the daily pulse of what citizens and organisations across Australia are talking about online. Most importantly, the platform supports discovery and access to data by the broader research community. This includes targeted search interfaces as well as use of large language model interfaces to support natural language querying of the resources.

The AIReD platform underpins many research and teaching efforts across Australia. This talk will cover some of the stories AIReD has been used to tell by researchers and students across Australia requiring access to social media at a scale that would otherwise be impossible without such a national capability.

The talk will also explore some of the technical challenges related to the Cloud infrastructure upon which AIReD depends and the ramifications of dealing with truly big data.

The Australian Internet Observatory is national research infrastructure supporting digital platform and smart data research. AIO received investment from the Australian Research Data Commons (ARDC) through the National Collaborative Infrastructure Strategy (NCRIS) in partnership with RMIT University, Queensland University of Technology, The University of Queensland, The University of Melbourne, Swinburne University of Technology and Deakin University.
DOI https://doi.org/10.3565/hjrp-b141

Poster Session / **26**

Navigating Dataspaces in Australasia: Challenges and Opportunities for Innovation

Author: Kheeran Dharmawardena[1]

Co-authors: Andrew White [1]; Jonathan Smillie [1]; Muhammad Ali [1]; Rob Clemens [1]; Shannon Callaghan [1]

[1] *Australian Research Data Commons*

Corresponding Authors: jonathan.smillie@ardc.edu.au, andrew.white@ardc.edu.au, shannon.callaghan@ardc.edu.au, kheeran.dharmawardena@ardc.edu.au, muhammad.ali@ardc.edu.au, rob.clemens@ardc.edu.au

Dataspaces can unlock the potential of data-intensive research by enabling trusted data sharing. This session explores the challenges and opportunities facing organisations and researchers as they navigate the adoption of trusted sharing using dataspaces.

Bringing together key stakeholders, including researchers, policymakers, and infrastructure providers, this forum will identify barriers, share best practices, and contribute to a more coordinated and inclusive global dataspace landscape.

With a focus on the Australasian region, this session will have a mix of presentations and a panel discussion examining practical considerations for building and connecting dataspaces. Topics will include infrastructure readiness, data governance frameworks, interoperability concerns, and the need for skilled personnel.
We will discuss the role of coordination mechanisms –such as hub facilitators –in fostering knowledge sharing and supporting the development of robust, interoperable dataspaces. The session will also explore requirements for regional infrastructure and supporting services to accelerate adoption.

Poster Session / 28

Design And Application of Experimental Schemes for Thermoelectricity Dataset

Author: Junmin Fang[1]

Co-authors: Chunjiang Liu [2]; Shuying Li [1]; Yiran Cai [1]

[1] *National Science Library (Chengdu), Chinese Academy of Sciences*

[2] *National Science Library (Chengdu), CAS*

Corresponding Authors: lisy@clas.ac.cn, caiyiran23@mails.ucas.ac.cn, liucj@clas.ac.cn, fjm@clas.ac.cn

This study focuses on intelligent data mining and knowledge discovery services, addressing the critical challenges researchers face in processing massive datasets and optimizing experimental schemes. We propose an innovative solution integrating knowledge graph and artificial intelligence technologies, with a specific application to thermoelectric domain through the development of a three-dimensional experimental scheme coordinate system.
Methodologically, this research first employs ontology modeling techniques to systematically construct a knowledge graph for thermoelectric experimental schemes, enabling structured representation of multivariate relationships among experimental entities. We then innovatively introduce a three-dimensional coordinate analysis model, establishing a visual evaluation framework based on key performance metrics including power conversion efficiency (PCE), short-circuit current density (Jsc), open-circuit voltage (Voc), and fill factor (FF).
The principal innovations of this work include: (1) proposing a novel research paradigm combining knowledge graph with multidimensional coordinate analysis; (2) developing a quantitative evaluation system specifically for thermoelectric experimental schemes; and (3) achieving data-driven decision support for scheme optimization. The research outcomes not only provide new analytical tools for experimental design in thermoelectric research but also offer a transferable methodological framework for intelligent data mining and knowledge discovery across other scientific disciplines.
This study holds practical value for advancing AI for Science initiatives and enhancing research innovation efficiency, demonstrating particular technical advantages and practical guidance in the field of energy materials development.

Poster Session / 30

Data Management and Utilization of National Metrology Institutes (NMIs) in the Republic of Korea

Author: HEE KYEOM YOO[1]

Co-author: Se-Hyun Shim [1]

[1] *Korea Research Institute of Standards and Science*

Corresponding Authors: s2h@kriss.re.kr, hky@kriss.re.kr

National metrology institutes (NMIs) serve to establish measurement standards for their respective nations and disseminate these standards to end users, including various industries. Through this process, NMIs facilitate freedom in economic activities by overcoming technical trade barriers through compliance with the International Committee for Weights and Measures Mutual Recognition Arrangement (CIPM MRA). Countries participating in the MRA recognize each other's calibration and measurement capabilities, enabling the NMIs of these nations to gain international recognition for their measurement results based on mutual trust. Consequently, NMIs play a crucial role in producing and disseminating credible research data grounded in traceability, high accuracy, and reliability.

The Korea Research Institute of Standards and Science (KRISS), the NMI of the Republic of Korea, categorizes and manages its research data into four distinct types. First, KRISS produces research data

through its research and development pursuits that it undertakes as a government-funded research institute. These data are managed on KRISS's data management platform (DMP) and are also linked to Korea's national research data platform, 'DataOn'to be made publicly available as open data for nationwide utilization.

The second category involves measurement data used to establish national measurement standards, ensuring reliability and transparency through international key comparisons between countries. Calibration and measurement capabilities (CMCs) of individual countries are registered in the CIPM MRA database (KCDB) of the International Bureau of Weights and Measures (BIPM). Additionally, these measurement standards are managed as legally recognized research outputs of national R&D projects.

Third, national reference data are certified data derived from scientific analyses that evaluate the accuracy and reliability of measurement data and information. These data are produced by relevant organizations and data centers in various fields, such as physics, chemistry, biology, medicine, and materials science. Managed as research data for continuous and repeated use, national reference data are made available through the national research data platform.

Lastly, measurement data used to verify measured values on calibration certificates are integrated into a system for issuing digital calibration certificates (DCCs). A dedicated platform has been established to ensure that relevant measurement data are provided alongside calibration certificates when issued.

This study examines the characteristics of the four aforementioned research data types and explores methods of collecting, storing, and managing each type of data. Additionally, this study reviews case studies of research data utilization, namely the application of the Korean standard time as national measurement standard data for terrestrial navigation systems and the implementation of brain magnetic resonance imaging data as standard reference data.

Presentations Session 2: Data and Research & Data Science and Data Analysis / 31

Mining Meaning: How SMU Libraries Use NLP and AI Tools to Uncover Strategic Insights

Author: Danping Dong[1]

Co-authors: Aaron Tay [1]; Pin Pin YEO [1]; Bella Ratmelia [1]

[1] *Singapore Management University*

Corresponding Authors: ppyeo@smu.edu.sg, dpdong@smu.edu.sg, aarontay@smu.edu.sg, bellar@smu.edu.sg

Academic institutions often hold large volumes of unstructured text data—such as chat transcripts, research publications, and strategic documents—but may lack accessible methods to analyze and interpret these resources effectively. This presentation shares how Singapore Management University (SMU) Libraries leveraged BERTopic, an AI-driven topic modeling tool for text clustering, along with generative AI tools like ChatGPT and Deepseek, to extract meaningful insights from institutional data—supporting both service enhancement and strategic planning, and illustrating the growing potential of generative AI in supporting institutional goals and strategies.

In the first case, we applied BERTopic to anonymized library chat transcripts to identify recurring topics in user queries. This approach allowed us to efficiently analyze thousands of transcripts and uncover common types of enquiries received through the library's chat service. The findings provide insights that may inform future service improvements, staff training, and chatbot development —areas of growing interest for many academic libraries. By applying modern topic modeling to a traditionally underused dataset, we demonstrated how unstructured service data can support evidence-based decision-making.

In the second case, we used BERTopic to cluster and analyze a collection of publications by university faculty. The goal was to identify thematic groupings that reflect our university's research strengths

and to respond to senior leadership inquiries about research trends and institutional output. To aid interpretation of the keyword-based topic representations generated by BERTopic, we used generative AI tools such as ChatGPT and Claude to produce easily understandable summaries.

Additionally, we will briefly share a small-scale application of Deepseek API for zero-shot classification to categorize faculty publications by the university's strategic priorities—demonstrating the value of generative AI for institutional insights.

Across these cases, we reflect on the strengths and limitations of using BERTopic and generative AI tools. Challenges included interpreting noisy topic groups, tuning model parameters to improve results, and balancing automation with human judgment. We discuss how data preparation, prompt refinement, and iterative experimentation influenced the outcomes. Importantly, these tools are now significantly more accessible than in the past, making it feasible for staff without deep technical expertise to conduct advanced text analysis. We also note the privacy advantages of using BERTopic, as it can be deployed locally without transmitting sensitive data to external servers. These reflections aim to provide a grounded view of how AI-based tools can be responsibly and effectively applied in institutional settings, while remaining mindful of their constraints.

These projects illustrate how librarians and institutional research staff, equipped with accessible NLP and AI tools, can contribute meaningfully to both operational and strategic initiatives. We highlight the evolving role of libraries and research support units in advancing institutional data intelligence, improving workflows, and enhancing decision-making in the age of AI.

Presentations Session 4: Data Stewardship / 33

Sustainability and findability of important global geoscience information standards

Author: Mark Rattenbury[1]

[1] *GNS Science*

Corresponding Author: m.rattenbury@gns.cri.nz

Important global geoscience information standards are developed, managed and governed by the Commission for the Management and Application of Geoscience Information (CGI), a commission under the auspices of the International Union of Geological Sciences (IUGS). CGI's standards include logical data models such as GeoSciML and EarthResourceML. These data models are supported by controlled vocabularies, currently numbering more than 50 with another 50 in various stages of preparation. Much of the development of the GeoSciML model and attendant vocabularies was influenced and showcased by the international OneGeology initiative. Similarly the EarthResourceML model and supporting vocabularies development occurred with the European Union's Minerals4Eu project. Without supporting initiatives and project like these, CGI struggles to maintain, let alone develop geoscience information standards yet these standards are influencing geoscience data management around the world. CGI is heavily reliant on the participation of individuals whose host organisations support their involvement. Without iconic cooperative projects, this support can erode. CGI is also reliant on server infrastructure provided by Geoscience Australia to publish and serve its geoscience vocabularies. This currently works well for both parties but nevertheless represents a single point of failure. The sustainability of CGI standards is open to several points of vulnerability.

Another challenge for CGI relates to the findability of its standards. CGI vocabularies are strongly FAIR-compliant, including scoring nominally very well in terms of being Findable. Yet this has not translated into visibility in a data-saturated world. There are many alternative geoscience vocabularies available for more localised needs, and while many of these reference CGI vocabulary terms and sources, locating truly international standards is difficult to the uninitiated.

CGI, through a strategic thought process, is realising that promotion of its standards is an increasingly important part its business. The leadership of the International Science Council and CODATA working with various international unions may be able to help with sustainability and findability

of international standards, potentially providing information resources that improve visibility and findability.

Poster Session / 35

Plans and challenges for FAIR and open data and an enhanced transparency of the IPCC Seventh Assessment Report

Author: Martina Stockhause[1]

Co-authors: Lina Sitz [2]; Azra Alikadic [3]; April Lamb [4]; Charlotte Pascoe [5]; Xiaoshi Xing [6]

[1] *IPCC Data Distribution Centre (DDC)*

[2] *Instituto de Física de Cantabria (CSIC-UC) and Intergovernmental Panel on Climate Change (IPCC), WGI-TSU, Université Paris-Saclay*

[3] *Deltares / IPCC Working Group II Technical Support Unit (WGII TSU)*

[4] *North Carolina Institute for Climate Studies, North Carolina State University*

[5] *Centre for Environmental Data Analysis (CEDA), STFC*

[6] *CIESIN / IPCC Data Distribution Centre (DDC), Columbia Climate School, Columbia University*

Corresponding Authors: azra.alikadic@deltares.nl, sitzl@ifca.unican.es, charlotte.pascoe@stfc.ac.uk, xxiaoshi@ciesin.columbia.edu, alamb@cicsnc.org, stockhause@dkrz.de

The Intergovernmental Panel on Climate Change (IPCC) has been providing climate Assessment Reports (ARs) since 1988, which document the state of climate change and future projections under various options for action. These ARs form the basis of international agreements and actions. UN Secretary-General Guterres described climate action as "the 21st century's greatest opportunity to drive forward all the Sustainable Development Goals." (Guterres, 2023). Data is an important basis for action and the transparency of data generation is a key contribution to trust in AR results.

Within the IPCC, the Task Group on Data Support for Climate Change Assessments (TG-Data) provides guidance on the curation, traceability, stability, availability and transparency of data and facilitates the availability and consistent use of climate change-related data through the work of the Data Distribution Centre (DDC) and the Working Group (WG) Technical Support Units (TSUs).

DDC Partners and WGI TSU made a concerted effort to improve the transparency of the IPCC's Sixth Assessment Report (AR6) by documenting the generation of figures and archiving figure datasets, figure generation software, key intermediate data products and key input data collections (Pirani et al., 2022; Stockhause et al., 2024a). This work and the experiences of the WGI partners and the AR6 WGI authors were incorporated into the formulation of the IPCC TG-Data recommendations for AR7, involving WGII and WGIII (IPCC, 2023; Stockhause et al., 2024b). The recommendations include:

- Exhaustive treatment of figure generation for all reports including data and software preservation in the DDC;
- Support of authors with training and tools in providing figure data and documenting the figure generation process;
- Intensified collaboration with sibling platforms like the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) and with related scientific and technical organizations like the Coupled Model Intercomparison Project (CMIP) and Coordinated Regional Climate Downscaling Experiment (CORDEX) of the World Climate Research Programme (WCRP) and the Research Data Alliance (RDA);
- Improved documentation and accessibility of input data citations and the provenance of figure generation in the AR7.

The AR6 WGI concept of open and FAIR data and traceable results was presented at the International Data Week (IDW) 2018 in Gaborone (Stockhause et al., 2019) and its implementation in AR6 at the IDW 2023 in Salzburg (Stockhause et al., 2023). The focus at IDW 2025 in Brisbane lies on the ongoing work for an improved and more comprehensive implementation:

Initial work on the workflow, metadata schema updates and Complex Citation/provenance was carried out under the condition of an uncertain funding situation of several DDC Partners. Therefore, every implementation plan includes a fallback solution in case no funding can be found for its (full) implementation. The Complex Citation/provenance plan is presented as an example for the ongoing implementations of TG-Data's recommendations for AR7. The DDC has reached out to developers of key community frameworks (ESMValtool and CMIP Rapid Evaluation Framework), providers of common interactive data science tools (Jupyter and Notebooks Now! projects) and the RDA Complex Citation WG (Agarwal et al., 2025) to align IPCC requirements with these groups. Several technical challenges have already been identified, but the main challenge lies in the coordination and collaboration across different and sometimes very large and diverse groups (IPCC, infrastructure, RDA) and the required acceptance by different stakeholders (including IPCC authors, DDC Partners, IPCC WGs, funders, indexers and publishers). Furthermore, the scale of IPCC AR7 brings additional challenges. This is relevant both in terms of number of authors, but also in the number of figures and the associated number of digital objects with complex interrelationships as well as the need for a stepwise implementation under multiple uncertainties (funding, national/institutional support, availability of technical solutions, etc.).

References:

Guterres, A. (2023). Secretary-General's briefing to the General Assembly on Priorities for 2023. New York. 06 February 2023. https://www.un.org/sg/en/content/sg/statement/2023-02-06/secretary-generals-briefing-the-general-assembly-priorities-for-2023-scroll-down-for-bilingual-delivered-all-english-and-all-french-versions

Pirani, A., Alegria, A., Khourdajie, A. A., Gunawan, W., Gutiérrez, J. M., Holsman, K., Huard, D., Juckes, M., Kawamiya, M., Klutse, N., Krey, V., Matthews, R., Milward, A., Pascoe, C., Van Der Shrier, G., Spinuso, A., Stockhause, M., and Xiaoshi Xing. (2022). The implementation of FAIR data principles in the IPCC AR6 assessment process. https://doi.org/10.5281/ZENODO.6504469

Stockhause, M., Pascoe, C., Sitz, L. and Pirani, A. (2024a). IPCC FAIR data approach. Zenodo. https://doi.org/10.5281/ZENODO.10821975

Intergovernmental Panel on Climate Change. (2023). TG-Data Recommendations for AR7 (1.0). Zenodo. https://doi.org/10.5281/ZENODO.10059282

Stockhause, M., Huard, D., Al Khourdajie, A., Gutiérrez, J. M., Kawamiya, M., Klutse, N. A. B., Krey, V., Milward, D., Okem, A. E., Pirani, A., Sitz, L. E., Solman, S. A., Spinuso, A. and Xing, X. (2024b). Implementing FAIR data principles in the IPCC seventh assessment cycle: Lessons learned and future prospects. In J. A. Añel (Ed.), PLOS Climate (Vol. 3, Number 12, e0000533. p.). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pclm.0000533

Stockhause, M., Juckes, M., Chen, R., Moufouma Okia, W., Pirani, A., Waterfield, T., Xing, X. and Edmunds, R. (2019). Data Distribution Centre Support for the IPCC Sixth Assessment. In Data Science Journal (Vol. 18). Ubiquity Press, Ltd. https://doi.org/10.5334/dsj-2019-020

Stockhause, M., Pirani, A., Sitz, L., Krüss, B., Pascoe, C., MacRae, M., Anderson, E. and Fisher, E. (2023). Implementation of the IPCC FAIR Guidelines into the Sixth Assessment Report (AR6): benefit, challenges and recommendations for AR7. Zenodo. https://doi.org/10.5281/ZENODO.10039597

Agarwal, D., Ayliffe, J., J. H. Buck, J., Damerow, J., Parton, G., Stall, S., Stockhause, M. and Wyborn, L. (2025). Complex Citation Working Group Recommendation. Zenodo. https://doi.org/10.5281/ZENODO.14106602

36

Certified Data Repositories in Asia-Pacific and Africa: towards Sustainable Science, Education, and Development

Authors: Bapon Fakhruddin[1]; Guoqing LI[2]; Jing Zhao[3]; Olivier Rouchon[4]; Daisy Selematsela[5]; Juanle Wang[6]

[1] *Green Climate Fund*

[2] *AIRCAS*

[3] *China National Satellite Meteorological Centre*

[4] *Centre National de la Recherche Scientifique*

[5] *University of Witwatersrand*

[6] *Institute of Geographic Sciences and Natural Resources Research*

Corresponding Authors: wangjl@igsnrr.ac.cn, olivier.rouchon@cnrs.fr, ligq@aircas.ac.cn, bfakhruddin@gcfund.org, zhaoj@cma.gov.cn, daisy.selematsela@wits.ac.za

This 90-minute session, featuring five 10-minute presentations followed by a 40-minute panel discussion, aims to tackle critical challenges in scientific data repository certification across the Asia-Pacific and Africa (APA) regions. The session seeks to establish a collaborative framework for improving repository management standards and fostering cross-regional synergies. By convening experts from academia, policymaking, and technical fields, it will explore how certification mechanisms can enhance data interoperability, service quality, and ethical compliance, ultimately supporting the United Nations Sustainable Development Goals (SDGs).

Current challenges in APA regions include fragmented certification criteria, uneven infrastructure development, and insufficient multilingual support, collectively hindering the adoption of globally recognized standards such as the FAIR (Findable, Accessible, Interoperable, Reusable) principles and cross-domain interoperability. For instance, a 2024 UNESCO report indicates that only 18% of repositories in these regions meet basic metadata certification requirements, compared to 42% in Europe and North America. The session will highlight case studies demonstrating how certification protocols can bridge gaps in data accessibility and governance while respecting regional socio-technical contexts.

This proposal is jointly supported by CoreTrustSeal, World Data System, and CODATA, with additional collaboration from early-career researchers via WDS-ECR and WDS China Group. Potential participants include leading experts from China, Singapore, Thailand, Japan, South Africa, Australia and New Zealand. Their contributions will emphasize practical strategies for familiarizing themselves with existing certification systems, such as CoreTrustSeal and Re3Data. Additionally, the session will demonstrate how certified data repositories can work in sector such as higher education, ocean, disaster risk reduction, scientific research and poverty reduction.

The panel discussion will concentrate on solutions that can be implemented to overcome the barriers encountered in the certification process, with a particular focus on indigenous language and capacity-building initiatives. The anticipated outcomes encompass the promotion of mutual assistance, the cultivation of cooperation, and the training of experts on certification among data repository institutions in these regions. The session will achieve this by aligning technical rigor with cultural inclusivity, thus catalyzing long-term advancements in data-driven sustainability efforts across the Global South.

Session Team:

Li Guoqing, Director of China National Earth Observation Data Center, and the Board member of CoreTrustSeal, co-chair of CODATA FAIR-DRR TG (onsite)

Dr. Zhao Jing, China National Satellite Meteorological Centre, and the co-Chair of WDS ECR (online)

Yuyun Wirawati, Nanyang Technological University Library, and leader of Community of Practice (CoP) for Southeast-Asia repositories (online)

Daisy Selematsela, University of Witwatersrand, and the Vice President of CODATA (onsite)

Olivier Rouchon, Centre National de la Recherche Scientifique, and chair of the CoreTrustSeal board (onsite)

Wang Juanle, Institute of Geographic Sciences and Natural Resources Research, and Scientific Committee member of WDS (onsite)

Noriko Kaneshima, Japan National Institute of Informatics (online)

Bapon Fakhruddin, Green Climate Fund, and co-chair of CODATA FAIR-DRR TG(onsite)

Liu Chuang, Institute of Geographic Sciences and Natural Resources Researchs, and Director of WDS center on Global Change Research Data Publishing Repository, and co-chair of CODATA GIES TG.(onsite)

Jun Shen, School of Computing and Information Technology, University of Wollongong, Australia.(onsite)

Potential Speakers (5 x 10 mins):

(1)CoreTrustSeal update, challenges in Asia-Pacific region (by Olivier Rouchon & Guoqing LI, in-person)

(2)Certified repositories development in Earth Science and their value added potential in APA regions (by Juanle Wang, in-person)
(3)Asia Pacific case study on World FAIR (by Bapon Fakhruddin, in-person)
(4)Conceptualisation of an RDM framework for the Wits Graduate Online Learning and Development (GOLD) Programme, South Africa (Daisy Selematsela and Lazarus Matizirofa, in-person)
(5)Community of Practice: Collaborating Across Certified Repositories (by Yuyun Wirawati, online)

Panel Discussion (40 mins)
Panelist (TBD): Jun Shen, Olivier Rouchon,Liu Chuang, Daisy Selematsela, Bapon Fakhruddin
Panel Discussion Topics (draft):
(1)How to select a right certification for scientific data repository?
(2)Lessons and Learened: preparation, writing, and communication during with certification procedure
(3)The benefits from certification of data repositories
(4)Regional cooperation due to the certification of data repositories

Poster Session / 38

When policies meet practices, research data governance at the university of Lille, France

Author: Julien ROCHE[1]

[1] *university of Lille*

Corresponding Author: julien.roche@univ-lille.fr

In a context where the amount of data is doubling every three years, there is an urgent need to develop and define policies and harmonised practices for research data at an institutional level, connected to national strategies and international frameworks. The aim of this paper will be to demonstrate how international, national and institutional levels can be connected, through the example of the University of Lille, in France.
At a national level in France, the awareness has grown quite recently in the past five years, leading to the vision that every ministry needs to adopt a roadmap regarding the data produced by institutions belonging to their perimeter, under the umbrella of a national coordinator from the Office of the Prime Minister. For Higher Education and Research, a national data officer has been appointed in 2021, being in charge to coordinate the policy regarding both administrative and research data, with three overarching goals : to facilitate the re-use of public (research) data, to develop transparency for public institutions as data producers, to simplify and foster efficiency of the public sector thanks to a well-reasoned and relevant use of their data.
As a result, French universities as well as other research-performing institutions were advised to appoint a chief officer for research data in order to propose and prepare a policy for the whole life cycle of research data.
The university of Lille, a comprehensive research-performing university in France, has created a framework for this research data produced and managed by the university, consistent with the national strategy in France. This framework is including :
• the creation of a data cluster, as a one-stop shop for the challenges the researchers are facing regarding their data,
• the development of tailored courses, in-person and online, to help researchers at all stages during the data lifecycle,
• the coordination of all the research-performing organisations in the area inside a network and a common collaborative-based structure,
• the creation of a data repository to leave no researcher with no solution for their data storage, curation and dissemination,
• the appointment of a chief officer for research data,
• the coordination of a unit including all the helpful bodies and competencies inside the university to prepare guidelines and solve complex issues raised by researchers, including relevant stakeholders in research-support services, Data Protection Office, libraries, archives, legal department, ethics committee, research integrity office, IT department, security office and so on,
• the participation in the national network to share and spread good practices and guidelines,

• the adoption of a strategic framework at the University of Lille for the governance of research data, also including codes and algorithms.
The framework, called « principes for the governance of research data, algorithms and codes at the university of Lille » and adopted by the research council of the university (March 2025), has defined several key principles :
• the respect of key values (transparence, scientific integrity⋯),
• the importance of the FAIRness for data,
• the need to describe the data, regarless of their openness or closeness,
• the need to take into account the specificity of algorithms and codes, that are different from research data,
• the importance of energy sobriety for data.
The framework has also defined the role of every stakeholders (researcher, research unit, faculties, university level) and is progressively including practical guidelines and documents that are relevant either directly for the researchers of indirectly for the staff supporting research.

Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global / 39

PANGAEA –30 years of publishing data for Earth & Environmental Science

Authors: Frank Oliver Glöckner[1]; Janine Felden[2]; Uwe Schindler[3]

[1] *Alfred Wegener Institute - Helmholtz Center for Polar- and Marine Research & MARUM - Center for Marine Environmental Sciences University of Bremen*

[2] *Alfred Wegener Institute - Helmholtz Center for Polar- and Marine Research, PANGAEA*

[3] *PANGAEA, MARUM/University of Bremen, Germany*

Corresponding Authors: uschindler@pangaea.de, frank.oliver.gloeckner@awi.de, janine.felden@awi.de

PANGAEA –Data Publisher for Earth & Environmental Science is a worldwide recognised digital data repository that plays a pivotal role in archiving, publishing, and disseminating scientific data related to earth and environmental sciences. As a publicly accessible information system, PANGAEA ensures that high-quality, well-structured, and interoperable datasets are preserved and made available to the scientific community. The platform fosters collaborations across various scientific disciplines, including geology, oceanography, climatology, ecology, and biodiversity by allowing scientists to archive and share georeferenced observational and experimental data. Each dataset is assigned a Digital Object Identifier (DOI), ensuring persistent citation and long-term accessibility. Its commitment to the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles ensures that data curation, management and dissemination align with international best practices, enhancing scientific transparency and reproducibility. In this way, PANGAEA has grown into a widely respected platform serving a diverse range of research communities since its establishment in the early 1990s by the Alfred Wegener Institute –Helmholtz Centre for Polar and Marine Research (AWI) and the Center for Marine Environmental Sciences (MARUM) at the University of Bremen.
PANGAEA plays a vital role in supporting large-scale international research projects and initiatives such as the Intergovernmental Panel on Climate Change (IPCC), the International Ocean Discovery Program (IODP), and the World Data System (WDS). holds a mandate from the World Meteorological Organization (WMO), to host the World Radiation Monitoring Center (WRMC). It is accredited as a World Data Center by the International Council for Science (ICS) since 2001 and has been certified as a trustworthy long-term data archive by Core Trust Seal. The repository integrates seamlessly with other global data infrastructures, ensuring compatibility with frameworks such as the Global Earth Observation System of Systems (GEOSS) and the European Open Science Cloud (EOSC). By linking datasets with scientific publications and ensuring proper attribution, PANGAEA strengthens the credibility and impact of research findings.
The platform's manual data curation process adheres to rigorous standards, involving expert review and validation before datasets are published. Researchers uploading data are required to provide comprehensive metadata compliant to ISO 19115, including descriptions of methodologies, instrumentation, and data provenance for each data set. This meticulous approach minimizes errors and enhances the reliability of published datasets. Additionally, PANGAEA supports a wide variety of data formats, including numerical, textual, image, and geospatial datasets, facilitating diverse applications in scientific research. All data and metadata are compiled in close collaboration between

the scientists and trained field experts acting as data editors. Both, data and metadata are checked for completeness and plausibility, ensuring high quality standards according to the FAIR data principles. Semantic interoperability during data curation is ensured through strict application and dynamic evolution of terminologies according to international protocols and standards. All published datasets carry a licence information (CC0 or CC-BY). The structured metadata accompanying each dataset enhances discoverability and usability, allowing researchers to effectively integrate PANGAEA's resources into their work. Currently, PANGAEA provides access to over 434,000 datasets containing over 31 billion individual measurements, including those collected through over 889 national and international projects.

Beyond its function as a repository, PANGAEA offers numerous tools and services. In addition to the classic access to data via the website, an integrative use of data in the form of a DataWarehouse and a set of tools for programmatic data processing are available for this purpose. The two applications written for the scripting languages Python and R, pangaeapy and pangaear, respectively, make use of the well-developed interoperability framework of PANGAEA. This framework allows most effective dissemination of metadata and data to all major internet search-engine registries, library catalogs, data portals, and other service providers, and ensures the optimal findability of data hosted by PANGAEA. The respective web services entail SOAP and REST APIs, a Schema.org/Dataset compliant metadata endpoint and OAI-PMH for various metadata content standards like DataCite, Dublin Core, DIF and ISO 19115 for harvesting. These technical capabilities make PANGAEA an essential resource for interdisciplinary studies addressing complex environmental challenges such as climate change, biodiversity loss, and natural resource management.

Looking forward, PANGAEA aims to further strengthen its role as a cornerstone of global earth and environmental data infrastructure. The rapid increase in data volume and complexity requires to extend PANGAEAs front-office model with trained data stewards all over the world. Efforts to integrate artificial intelligence and machine learning techniques into data curation and retrieval processes are ongoing, promising to further enhance the efficiency and scalability of PANGAEA's operations. By collaborating with established data initiatives and delivering data products for data portals as well as fostering international partnerships, PANGAEA will continue to facilitate innovative research in the earth and environmental sciences. As scientific data management evolves, PANGAEA remains at the forefront, providing researchers with the tools and resources necessary to address some of the most pressing environmental challenges of our time.

The presentation will provide an overview of the current status and further perspectives of PANGAEA - data publisher for earth & environmental science.

Poster Session / 40

Promoting Open Data Sharing through Scientific Data Publishing: Innovations from the Global Change Research Data Publishing & Repository

Author: ZHAOCAI JIANG[1]

[1] *Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences*

Corresponding Author: jiangzhaocai@igsnrr.ac.cn

Against the global backdrop of "data sharing and reuse", publishing scientific data as a formal research output—compared to traditional data submission or repository archiving—proves more effective in advancing open data sharing and academic recognition. Currently, data publishing remains in its early exploratory stages in China and worldwide, with dedicated scientific data publishing journals and platforms still relatively limited. This study examines the Global Change Research Data Publishing & Repository (GCdataPR), a World Data Center and China's first and only platform supporting bilingual (Chinese-English) publishing of integrated "dataset + data paper" outputs. We present its innovative practices in data publishing models, sharing policies, workflows, platform functionalities, and current achievements. The system consists of two core components: Digital Journal of Global Change Data Repository and the Journal of Global Change Data & Discovery. To date, it has published over 1,400 datasets and 550 data papers, while curating thematic collections aligned with the UN Sustainable Development Goals (SDGs), such as Geographical Indications Environment & Sustainability (GIES), Global Island Data, and Global Oasis Data, etc.

By synthesizing GCdataPR's practical experience, this study aims to provide insights for the global

development of scientific data publishing, fostering its broader adoption in scholarly evaluation and open science initiatives.

Poster Session / 41

Publishing the Scientific Data of Chinese Academic Journal: a Case Study on Global Change Data Publishing and Repository System

Authors: Junhua MA[1]; Chuang Liu[1]; Ruixiang Shi[1]

[1] *Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences*

Corresponding Authors: shirx@igsnrr.ac.cn, majh@igsnrr.ac.cn, lchuang@igsnrr.ac.cn

Since 2015, China has introduced policies and regulations to encourage the sharing of scientific data. Among them, the publication of scientific data is an important part of scientific data sharing. The publication of scientific data includes two parts: data paper and dataset, in which the data paper is published as journal paper, and the dataset is published by registering DOI or e-journal on Internet. Data publishing in the field of global change science has been going on for 10 years. The Global Change Science Research Data Publishing and Repository System (GCdataPR) was launched in June 2014, which is jointly sponsored by the Institute of Geographic Sciences and Natural Resources Research (IGSNRR), Chinese Academy of Sciences (CAS) and Geographical Society of China (GSC). GCdataPR is a regular member of the World Data System and the Data Publishing Center of China Integrated Earth Observation System (China GEOSS). GCdataPR includes two main parts. One is the Journal of Global Change Data & Discovery (a bilingual journal), which is mainly used for the publication of data papers, and the other is the Digital Journal of Global Change Data Repository, mainly used for the publication of dataset. In order to better promote the publication of scientific data, GCdataPR has been a data repository for a number of academic journals to promote the data publication of these journal papers. Until now there are about 400 dataset published from above journals. This paper summarizes and analyzes the publication of related data of academic journals to provide reference for the sharing of scientific data. Taking the published data results in GCdataPR from June 2014 to December 2024 as an example, the related journals, data authors and foundations were analyzed. Through the analysis of the publication of the association data of academic journals, we found that: (1) the publication of the related data of Chinese journals is significantly more than that of English journals; (2) more than half of the dataset authors came from the same institution; (3) as for the funding for data development, more than 48% of the dataset were founded by National Natural Science Foundation of China and more than 23% from Ministry of Science and Technology of China; (4) the journals with the most published related data are Acta Geographica Sinica, Geographical Research, Journal of Natural Resources, Acta Ecologica Sinica, Resources Science. Compared with the annual number of papers published in academic journals, the number of related data published is obviously much lower, which is less than 1% of the annual publication. There is still great potential to share and publish the related data of academic journals.

Poster Session / 43

Data education for researchers: Designing for learner empowerment at a data skills Summer School

Authors: Adeline Wong[1]; Kathryn Greenhill[1]

[1] *Australian Research Data Commons*

Corresponding Authors: adeline.wong@ardc.edu.au, kit.greenhill@ardc.edu.au

This presentation explores ways of improving researchers'data skills by creating an environment that engages learners, helps them form networks and gives them greater control over what happens.

It draws on three years'experience facilitating a national research data skills summer school for the Australian Research Data Commons (ARDC). The presentation is suitable for anyone who mentors, guides or trains researchers about data in any discipline.

The presenters are Kit Greenhill, Skills Development Lead (HASS and Indigenous) and Adeline Wong, Skills Development Lead (Learning Design) at the ARDC. With decades of experience in university teaching and learning design they will share strategies to design and deliver data skills training for researchers.

The ARDC provides Australian researchers with competitive advantage through data.
Since 2023, the HASS and Indigenous Research Data Commons of the ARDC has run an annual three day Summer School to help researchers learn digital skills, network, and create new research outcomes.

Skills in data collection, analysis, governance and management are often not included in any depth in formal research training for Humanities Arts and Social Sciences (HASS) researchers in Australia. When data-intensive research becomes part of a project, new HASS researchers can feel isolated. It can be difficult to locate disciplinary mentors, discuss data issues with peers and learn data skills in an environment tailored to HASS, rather than a Science, Technology Engineering and Mathematics (STEM) background. Any data skills training for HASS needs to focus on building supportive relationships for future networking, not just transferring knowledge.

Designing the 2025 Summer School, Kit and Adeline reviewed participant feedback from previous years, addressing issues in levels of material, information available at registration, the structure of the day and opportunities for attendees to share their research. Potential attendees and presenters were more involved in curriculum design through online consultation before the programme was decided.The 2025 event had foundational 101 sessions for everyone, then workshop streams that progressed from introductory to more advanced. Learners determined the topics of several sessions on the same day, and collaborated on content. More time was scheduled during the event for attendees and presenters to network and learn about each others'research. These improvements to Summer School resulted in a 93.5% satisfaction rating from attendees.

Although the Summer School content was designed for HASS researchers, the presentation will invite the audience to consider whether this more social and learner-centred approach would improve data skills learning for researchers across all disciplines.

Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global / 44

SOOSmap: Empowering Southern Ocean Research Through Data Access

Authors: Michaela Miller[1]; Petra ten Hoopen[2]; Antonio Novellino[3]; Alyce Hancock[1]; SOOS Data Management Sub-Committee[1]

[1] *Southern Ocean Observing System*

[2] *British Antarctic Survey*

[3] *ETT Solutions*

Corresponding Author: miller@soos.aq

The Southern Ocean's influence on global climate and marine ecosystems underscores the urgent need for comprehensive and accessible observational data. Here we present a tool for discovery of, and access to, existing Southern Ocean data –SOOSmap, Version 2. SOOSmap is a collaborative effort between the Southern Ocean Observing System and European Marine Observations and Data Network Physics, free to use for anyone, from ocean science experts to classroom students. Currently hosting over 50,000 observations across multiple disciplines, SOOSmap is a gateway to physical, biogeochemical and ecological open access data. We describe the platform's architecture, user interface, and its role in facilitating data discovery, visualisation, and download. Furthermore, we contextualise SOOSmap within the broader polar data ecosystem and emphasise the importance

of community engagement and data contribution for its continued growth and utility. By fostering open access and data sharing according to FAIR principles, SOOSmap aims to empower a wide range of stakeholders and ultimately improve our capacity to understand, predict, and respond to changes occurring in the Southern Ocean.

Presentations Session 9: Empowering the global data community for impact, equity, and inclusion / Education / 46

Strengthening Global Training and Skills Development Partnerships: The ARDC-Alliance Staff Exchange Initiative

Authors: Catherine Di Vita[1]; Kathryn Unsworth[2]

[1] *Research Data Alliance of Canada*

[2] *Australian Research Data Commons (ARDC)*

Corresponding Authors: kathryn.unsworth@ardc.edu.au, catherine.divita@alliancecan.ca

In early 2024, a proposal to advance bilateral collaboration around national training strategies and frameworks between the Alliance and the Australian Research Data Commons (ARDC) was introduced. The agreement was formally signed in February 2025, effective through to December 31, 2026, launching the start of a two-year knowledge and staff exchange pilot.

This pilot is in the form of a series of bilateral staff exchanges both in Canada and Australia and involves each host country showcasing their respective work in-country. The objective of the pilot is to evaluate the efficacy and impact of on-site international staff exchanges, documenting the lessons learned along the way. By comparing and contrasting the skills development landscapes of ARDC and the Alliance, the pilot aims to identify points of convergence and opportunities for collaboration. Additionally, the pilot leads will share and document their experiential knowledge about training, skills, and workforce development from their respective countries (research sectors) and leverage the knowledge and understandings from their local training communities.

In this presentation we will provide an overview of each organisation, an outline of the areas of priority identified and work initiated during the first and second exchanges, along with an update of the progress made thus far. We will then discuss how we plan to develop and further the work with a focus on international stakeholder relationships to expand skills and workforce development in our two jurisdictions.

This presentation is designed for a general audience but is particularly valuable for sectoral skills development leaders and training providers. It offers insights into innovative ways for exchanging skills and knowledge among research infrastructure organisations and their staff, while also using international expertise and practices to broaden and enhance skills development and training strategies.

Finally, we will invite all stakeholders and interested audience members to provide any feedback on this exciting journey of international collaboration. Your insights and experiences are invaluable to the success of this pilot. We will share an interactive poll to capture your thoughts on skills-related topics that currently resonate with you. Additionally, we welcome questions during Q&A time and/or post presentation.

Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI / 47

FAIR-by-design Pipelines to ensure reproducibility and transparency of innovative remote-sensing Data Products

Author: Fernando Aguilar Gómez[1]

Co-authors: Aina García-Espriu [2]; Cristina González-Haro [2]; Daniel García-Díaz [1]

[1] *IFCA-CSIC*

[2] *ICM-CSIC*

Corresponding Authors: cgharo@icm.csic.es, aguilarf@ifca.unican.es, garciad@ifca.unican.es

The rapid growth in global data production, particularly from remote sensing and Earth observation, has created significant opportunities to address pressing global challenges such as climate change, biodiversity loss, and sustainable resource management. Open data from satellites, drones, and environmental sensors, although increasingly available, often require complex integration due to diverse formats, large data volumes, and heterogeneous sources. Effectively exploiting these datasets needs automated processes that rely heavily on detailed metadata. This proposed presentation will detail the advances and developments of the GOYAS project and its platform, which emphasizes adopting metadata standards and automated processing pipelines to ensure a FAIR (Findable, Accessible, Interoperable, and Reusable) data lifecycle, generating novel remote-sensing-based products.

Introduction

Earth Observation and environmental sciences are experiencing significant growth in available data, not only in terms of volume but also in terms of quality at various resolutions (spatial, spectral/radiometric, temporal) and through new instruments and sensors. Moreover, recent developments in techniques, methods, and Artificial Intelligence (AI) algorithms facilitate the estimation and production of new physical, chemical, or biological georreferenced variables. Addressing global issues such as climate change and ecosystem management requires integrating diverse datasets. For instance, AI-based algorithms for estimating chlorophyll concentration in freshwater reservoirs depend on in-situ data to train and validate new models. Remote-sensing data pose challenges due to their format variability, large volumes, and complexity. Global initiatives such as Copernicus and Digital Earth Australia standardize data products derived from Earth observation missions, facilitating user access. However, incorporating new algorithms or variables into these platforms can be complex, particularly for variables that cannot be produced globally. Nevertheless, innovative and experimental AI-derived data products are beneficial for various stakeholders. Ensuring FAIR compliance of these datasets is crucial as it enables their reuse, validation, and reproducibility. Achieving this requires robust data lifecycle management supported by comprehensive metadata standards.

FAIR Data Life Cycle in GOYAS project

The OSCARS project aims to enable Open Science services adhering to FAIR principles, interconnected with Research Infrastructures and the European Open Science Cloud (EOSC). Under the OSCARS umbrella, the GOYAS project is developing an Open Science service linked to ENVRI (European Environmental Research Infrastructures). The GOYAS platform employs the ISO19139 metadata standard, comprehensively addressing descriptive, administrative, and structural requirements for remote sensing data.

- `Descriptive metadata ensures datasets are discoverable by incorporating unique persistent identifi`
- `Administrative metadata provides essential context regarding data creation, quality assurance, lic`
- `Structural metadata, including format details, encoding, and logical attributes, enable interopera`

GOYAS systematically documents these metadata categories, ensuring data are ready for integration, interpretation, and validation, both manually and through automated pipelines. Metadata also includes quality details such as accuracy and expected errors, enhancing transparency and helping users in evaluating data suitability.

FAIR-by-design Pipeline

The GOYAS infrastructure incorporates data products from four Spanish research institutes belonging to CSIC (Spanish National Research Council): Doñana Biological Station, Institute of Marine Sciences of Andalusia, Institute of Marine Science, and the Physics Institute of Cantabria. Product types are diverse, including flood data from Doñana National Park, satellite-derived bathymetries, oceanic data (temperature, salinity), and freshwater quality indicators, among others. The lifecycle of these products involves multiple complex steps—data collection, corrections, preprocessing, curation, processing, and ingestion. The FAIR-by-design approach adopted by GOYAS documents all components necessary for reproducibility. This pipeline systematically records each dataset action, from initial collection through final outputs, ensuring transparency and reproducibility. It

references preprocessing algorithms using persistent identifiers and explicitly describes performed actions. Additionally, the pipeline integrates FAIR assessment tools (such as FAIR EVA) to conduct tests based on Research Data Alliance FAIR indicators (RDA FAIR Maturity Group). Consequently, new data products are published automatically only after all preceding steps have achieved FAIR compliance.

The primary users of the GOYAS platform include researchers developing innovative, experimental, or very localized remote sensing data products not suited for integration into large-scale portals like Copernicus. Additionally, the produced data are valuable to environmental scientists, policymakers, and public administrations for informed decision-making on environmental issues. This structured, FAIR-compliant pipeline significantly supports interdisciplinary research, environmental monitoring, resource management, and decision-making.

The proposed presentation will provide an overview of the GOYAS project, showcasing the FAIR-by-design workflow, including metadata management, automated data integration, processing, and FAIR validation. It will also include technical details of the underlying technologies and illustrate practical applications of the platform using available data products. This transparent and automation-supported approach is crucial for effectively adopting FAIR principles, and the GOYAS platform could serve as a model for similar initiatives.

Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data / 48

Local Expertise, Global Impact: The Growing Role of Institutional Data Repositories in Research Infrastructure

Author: Mikala Narlock[1]

Co-authors: Aundria Parkman ; Rachel Preisman Márquez ; Scout Calvert ; Shawna Taylor

[1] *Indiana University*

Corresponding Author: mnarlock@iu.edu

This presentation will examine the critical role of Institutional Repositories (IRs) and Institutional Data Repositories (IDRs) as foundational knowledge infrastructures supporting data-intensive research across academic institutions. As data sharing mandates and standards from funding agencies, publishers, and disciplinary societies continue to evolve, understanding the role of institutional data services, in particular IRs and IDRs, as complex knowledge infrastructures become increasingly important. This has been thrown into sharp relief as the United States grapples with unprecedented loss and manipulation of government data and federally funded research repositories.

Based on a comprehensive longitudinal study spanning 2017-2023, we will share findings on the significant growth of institutional data sharing solutions among Association of Research Libraries (ARL) member institutions. Our analysis reveals that by 2023, over 54% of ARL academic libraries maintain dedicated IDRs alongside traditional IRs, marking a milestone where more institutions now have dedicated data repositories than those without. Additionally, the number of datasets shared in institutionally managed solutions grew exponentially, with IDRs seeing a 199% increase from 2020 to 2023.

Beyond quantitative growth, our presentation will explore how these repositories function as complex adaptive systems that seamlessly integrate technical infrastructure with human expertise, institutional policy, and social practices. Through an "infrastructural inversion" lens, we demonstrate how IRs and IDRs provide unique advantages in meeting federal data sharing requirements while facilitating local-to-global data interoperability. We will demonstrate how research libraries function as an "installed base" that provides critical organizational sustainability, curation expertise, and integration with institutional systems.

Our research demonstrates how IRs and IDRs serve as critical connective tissue between researchers, institutions, and broader scholarly communities. These repositories provide not only technical infras-

tructure for data storage but also facilitate human-mediated curation, standardization, and interoperability that enables data to move from local contexts to global discovery and reuse. We will discuss how institutional repositories facilitate interoperability through persistent identifiers, standardized metadata, and integration with campus research systems, enabling local data to be discoverable and reusable globally.

Our presentation will analyze how institutional repositories are uniquely positioned to meet the requirements and expectations of funding agencies and publishers through their integration of technical systems with human expertise. Specifically, we will highlight how the "articulation work" performed by data stewards and curators—often invisible but essential labor—transforms repositories from mere storage platforms into true knowledge infrastructures capable of supporting the entire research data lifecycle.

Our findings suggest several key advantages that institutional repositories provide over generalist or disciplinary alternatives. First, their integration into the university context enables better authentication and validation of researcher identities, increasingly important for research integrity and compliance with security requirements like NPSM-33. Second, local administration facilitates integration with other campus systems, from library catalogs to grant management workflows, creating a more seamless experience for researchers and administrators alike. Third, the presence of local data curation expertise helps researchers navigate complex requirements around sensitive data, human subjects protections, and intellectual property considerations.

Despite the development of robust, free, self-upload generalist repositories over the past decade, interest in institutionally managed repositories continues to grow. Our presentation will explore this phenomenon and suggest that the human layers of repositories—the expertise, relationships, and institutional knowledge that facilitate effective data stewardship—may explain this continued investment. We will conclude by discussing opportunities for further development of institutional data infrastructure, including deeper integration with research information management systems and machine-actionable data management plans.

This presentation will be valuable for institutions developing or refining their data infrastructure strategies as federal mandates for research data sharing continue to evolve, offering both empirical evidence of current trends and a theoretical framework for understanding repositories as complex sociotechnical systems rather than merely technical solutions.

Presentations Session 2: Data and Research & Data Science and Data Analysis / 49

Introducing FJORD: a framework for FAIRly Jointed Open Research Data

Author: Federico Grasso Toro[1]

[1] *University of Bern*

Corresponding Author: federico.grasso@unibe.ch

The increasing digitalization of science, coupled with the push for Open Science and FAIR data (Findable, Accessible, Interoperable, Reusable), presents significant challenges for managing diverse research outputs effectively throughout their lifecycle. Traditional Data Management Plans (DMPs) often lack the detail and machine-actionability needed for dynamic research processes, while Current Research Information Systems (CRIS) struggle to capture the complexity of modern research workflows and assets beyond publications.
Furthermore, coordinating data management efforts and information flow between key stakeholders –researchers, research management offices (RMOs), infrastructure providers, university management (HEIs), and funders –remains a major hurdle, hindering efficient resource allocation, reproducibility, and robust research assessment.
This presentation introduces FJORD (FAIRly Jointed Open Research Data), a novel framework designed to address these challenges by creating an integrated ecosystem for managing diverse intellectual assets FAIRly by design and Jointed by interdisciplinarity.
FJORD builds upon previous work on "enhanced DMPs" as input for machine-actionable, tailored

to specific research contexts and SMART (Specific, Measurable, Achievable, Relevant, Time-bound) metrics for FAIR assessment, developed for both research software engineers and infrastructure managers.
The core of FJORD comprises:
(1) A suite of enhanced DMP templates specifically designed for diverse intellectual asset types, including i. publications, ii. research workflows, iii. models, iv. code/software, and v. datasets;
(2) "Fjordie," a prototype bot acting as a user-friendly frontend; and
(3) A vector metadata-database backend leveraging "knowledgement" –an ontology-driven approach to knowledge base(s) management –to process information from multiple sources and disciplines, while constructing internal customised knowledge graphs.
FJORD facilitates three crucial information pipelines:
(Pipeline 1) Streamlining reporting and compliance from researchers via RMOs to funders;
(Pipeline 2) Enabling better infrastructure planning and investment by connecting researcher needs through infrastructure managers to university management;
(Pipeline 3) Enhancing institutional research intelligence by feeding enriched, structured information (data and metadata) from researchers via RMOs to university management, functioning as an advanced, asset-aware CRIS 2.0.
The framework is currently being designed and validated through three distinct Proof-of-Concept (PoC) case studies within the context of my work as Data Steward for Data Science and IT, where I act as an embedded Open Research Data expert on each team.
These PoCs deliberately target different intellectual asset types and research domains, typically ignored and not harmonized on other international efforts (e.g., EOSC):
Workflow Focus (Cell Biology Laboratory Case): Applying enhanced DMPs to map and manage the entire lifecycle of life science, from instrument data at laboratory, by the generation pipeline through processing, curation, analysis, and plotting, to ultimately publishing standard operating procedures of FAIR datasets in local open data repositories.
Code Focus (Phenomics Case): Utilizing enhanced DMPs and FAIR-by-design principles during the coding of development and deployment of the lifecycle of a domain-specific repository for phenomics research.
Model Focus (Digital Humanities Case): Drafting enhanced DMPs combined with Behavior-Driven Development (BDD) using Cucumber/Gherkin syntax to define FAIR requirements for managing multiple complex metadata models and a unifying meta-metadata-model for digital editions.
Although these case studies are in their early stages (started during 2025), initial findings demonstrate the feasibility and utility of the FJORD approach.
Preliminary results indicate that asset-specific enhanced DMPs provide valuable structure for planning and tracking diverse research outputs.
The BDD approach proved effective in translating FAIR principles into actionable requirements for complex metadata models.
Early observations suggest the potential for improved coordination between researchers and support units, and more proactive FAIR implementation when integrated early in the research lifecycle via the FJORD templates.
This presentation will detail the FJORD framework's design, its theoretical underpinnings, and the drafted architecture supporting the key stakeholder pipelines. And I will share practical experiences, challenges encountered, and initial findings from the ongoing, diverse PoCs, emphasizing their relevance to the Swiss context, since they can be explicitly connected to current Swiss Open Research Data initiatives.
Attendees will gain insights into a novel, integrated approach for operationalizing FAIR principles across various research assets and fostering better alignment between researchers, institutions, and funders in the evolving digital research landscape.

Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global / 52

Research Data Ecosystem: Innovating Infrastructure for the Social Sciences in the 21st Century through Building a Modernized Software Platform, Data Description Framework, and Tools for the Research Data Community

Authors: Aalap Doshi[1]; Margaret Levenstein[1]

[1] *ICPSR University of Michigan*

Corresponding Authors: maggiel@umich.edu, aalapd@umich.edu

ICPSR, one of the world's largest social science data archives, located at the Institute for Social Research at the University of Michigan, is leading a $38M National Science Foundation project, the Research Data Ecosystem (RDE), to modernize research data infrastructure and support efficient, cutting-edge, and reproducible data-driven science.

In this presentation, we will briefly discuss the current state of data infrastructure, outline the infrastructure needed to support 21st century social science, and show how RDE will help close the gap between the two states. The presentation will be 60 minutes allowing 30 minutes for Q&A.

The extant data infrastructure cannot adequately support 21st century social science. Diverse types of data enable path-breaking analyses into human behavior but also present challenges of scale, sensitivity, and structure, requiring new research approaches. There is an urgent need for new modes of access, confidentiality protection, methodological approaches, and tools so that research using a variety of data types meets accepted scientific standards of transparency, reproducibility, efficiency, and ethics. Current barriers include multiple incompatible standards for data, lack of interoperability, and the inherent difficulty of managing big and often-sensitive data.

Data types driving research across the social science disciplines include social media, administrative, commercial transaction, streaming, audio, video and photo, satellite imaging, and biological. These data enable path-breaking analyses into human behavior as innovative projects combine data from different sources to allow for timely analysis with unprecedented granularity. New data types present challenges of scale, sensitivity, and structure, requiring new approaches to collection, privacy, analysis, storage, and preservation. A consensus exists in the scientific community on the urgent need for new modes of access, confidentiality protection, methodological approaches, and tools so that research using a variety of data types meets accepted scientific and ethical standards.

Frontier social science, relying on new categories of data, needs convergent standards and interoperable research infrastructure for producing, managing, and analyzing research data that includes non-designed data and "big"data. All stages of the data lifecycle need infrastructural support. RDE is building infrastructure that spans across all stages of the research lifecycle (ensuring research data are FAIR), with standards and methods making the tools for using the infrastructure interoperable. RDE, encompassing the full data life cycle, will make scientific analyses using that data more rigorous, transparent, and reproducible.

We will share how the RDE infrastructure is being built to enable social and data scientists across disciplines to conduct their work more efficiently and to create, organize, archive, access, and analyze data in ways that they cannot with existing infrastructure through the specific components outlined below:

Research Data Description Framework: a flexible system of metadata standards
Research Document Registry: a registry for digital documents including pre-registered research designs and hypotheses, data management plans, participant consent statements, and data use agreements
Tubocurator: software for harmonizing data and generating appropriate metadata to assure FAIRness.Turbocurator facilitates the curation and sharing of high quality, discoverable, and re-usable data by reducing the cost of preparing data and metadata. It will help harmonize data across studies, make it easy for researchers to attach appropriate metadata, maintain provenance, prepare data for re-use and re-discovery, and check for confidentiality issues.
Explore Data: interactive tools to preview, explore, and discover data
Video Data Tools: tools for facilitating data discovery and integration of video data
Geospatial Data Tools: tools for facilitating data discovery and integration, including confidential and geospatial data
Researcher Passport: credentialing system for researcher access to confidential data
COBRE: cloud-based platforms for analyzing confidential or large, complex, social science data

This session will be relevant to any person or organization involved in modernizing their cyberinfrastructure across the research data lifecycle to ensure FAIR data.

The speakers will be Dr. Maggie Levenstein and Aalap Doshi.

Levenstein is Director of ICPSR, Professor in the School of Information, Research Professor, Institute for Social Research, at the University of Michigan. She is the Principal Investigator of the NSF infrastructure project, RDE, and the NIH's Social, Behavioral, and Economic COVID-19 Consortium Coordinating Center. She is Co-Director of the Michigan Federal Statistical Research Data Center. She serves on the boards of the Social Science Research Council; World Data System; the Data Documentation Initiative (DDI); National Internet Observatory; Data Archiving and Access Requirements Working Group (DAARWG) of the NOAA Science Advisory Board; Criminal Justice Administrative Records System (CJARS); and the Wealth and Mobility Study, Stone Center for Inequality Dynamics. She received her PhD in economics from Yale University and BA in economics from Barnard College, Columbia University. She is the author of Accounting for Growth: Information Systems and the Creation of the Large Corporation and a fellow of the American Association for the Advancement of Science. Her research examines the production, dissemination, and confidentiality protection of novel, non-designed data for social and economic measurement.

Aalap Doshi is the Director of Technology for ICPSR at the Institute for Social Research and Lecturer in the School of Information at the University of Michigan where he teaches "Navigating Ambiguity in User Experience." He holds a Master of Science in Information (Human-Computer Interaction) from the University of Michigan, and has over two decades of experience at the intersection of design, technology, and strategy. Previously, he helped establish and scale human-centered design and innovation at Michigan Medicine, where he led the design of UMHealthResearch, an award-winning health research recruitment platform. He is also the co-founder of Findcare, a nonprofit connecting low-income communities to affordable healthcare.

53

Stronger together: Advancing the data repository ecosystem through strategic coopetition

Author: Ana Van Gulick[1]

Co-authors: Ishwar Chandramouliswaran [2]; John Chodacki [3]; Kristi Holmes [4]; Mark Hahnel [1]; Matt Buys [5]; Stefano Iacus [6]; Traci Snowden [7]

[1] *Figshare, Digital Science*

[2] *NIH Office of Data Science Strategy*

[3] *California Digital Library*

[4] *Northwestern University*

[5] *DataCite*

[6] *Harvard University*

[7] *Mendeley Data, Elsevier*

Corresponding Authors: siacus@iq.harvard.edu, ishwar.chandramouliswaran@nih.gov, m.hahnel@digital-science.com, t.snowden@elsevier.com, ana@figshare.com, mattbuys@datacite.org, john.chodacki@ucop.edu, kristi.holmes@northwestern.edu

Overview
The NIH Office of Data Science Strategy launched the Generalist Repository Ecosystem Initiative (GREI) in February 2022, in recognition of the key role generalist repositories play in the NIH data sharing landscape to support the FAIR sharing of data and other research outputs. The GREI program represents a groundbreaking collaborative model that brings together seven repositories (Dataverse, Dryad, Figshare, Mendeley Data, Open Science Framework, Vivli, and Zenodo) to enhance data sharing and reuse in alignment with NIH's mission. By prioritizing user needs, adopting shared standards, and creating flexible governance, GREI has modeled an innovative approach to balancing competition and cooperation among diverse repository partners.

In this session, we will build off a previous GREI presentation at SciDataCon in Salzburg in 2023 highlighting work in progress at the midpoint of the program. As we approach the next phase of the program, we look forward to sharing the impact of the work we have accomplished together, as well as the challenges and successes of the coopetition model itself.

Over the last four years, the initiative has focused on improving data accessibility, standardizing metadata, and facilitating comprehensive metrics, all in an effort to create a more interoperable research data ecosystem, both within and beyond generalist repositories. By working collaboratively across repositories and through strategic partnerships with organizations and community initiatives like DataCite, the Carpentries, and ROR, GREI has managed to create collective progress that would not have been possible through individual repository efforts. Critically, GREI's success lies in its ability to leverage each repository's unique strengths while maintaining a shared vision of open and FAIR data sharing infrastructure.

As we look to the future of our initiative, we are seeking feedback from the research data community on our work thus far and suggestions for where we might collectively turn our focus next. In this session, we will seek audience input to gather feedback on the work we have done on interoperable metadata, consistent metrics, and community engagement, and will mediate a forward-looking audience discussion on future GREI priorities to serve the research data ecosystem.

Session Agenda
Our proposed session agenda includes short presentations from GREI team members highlighting GREI accomplishments, impact, and future directions. We will thread interactive polls throughout the session, and follow this with a facilitated discussion (~30 minutes) and Q&A focused on the SciDataCon audience's suggestions for future GREI efforts.

Overview of the GREI program - the NIH perspective
Proposed speaker: Ishwar Chandramouliswaran, NIH Office of Data Science Strategy
This presentation will expand on the program history and overview shared in the Session Overview above, and give the NIH perspective on both the role of generalist repositories in the NIH data sharing landscape and the rationale behind the GREI program.

Coopetition way of working - challenges & successes
Proposed speaker: Kristi Holmes, Northwestern, Zenodo
This presentation will describe how GREI demonstrates the potential of coopetition to advance open science and collectively address complex challenges to create a more interoperable research data infrastructure. By working together, the seven GREI repositories have produced collective improvements that would have been unattainable by siloed individual repositories.

Highlights & impacts from 4 years of GREI

Metadata & Search
Proposed speaker: Mark Hahnel, Figshare
This presentation will showcase the development of our refined core metadata recommendation based on the DataCite Metadata Schema, through which we have strategically enhanced data discoverability and compliance with NIH data sharing policies. We will detail how we have standardized our handling of optional fields, introduced controlled vocabularies, and integrated persistent identifiers to streamline research data workflows and improve data interoperability across repositories.

Common Metrics
Proposed speaker: John Chodacki, California Digital Library, Zenodo
This presentation will explore how GREI has developed and continues to implement common metrics through the Make Data Count principles, creating a standardized approach to tracking data usage, citation, and impact across repositories. We will show how these harmonized metrics provide actionable insights for researchers, funders, and institutions, ultimately recognizing and reinforcing the value of open data sharing practices.

Community Outreach
Proposed speaker: Traci Snowden, Mendeley Data
This presentation will trace the evolution of GREI's extensive community engagement efforts, revealing how we have built capacity and support for researchers navigating data management and sharing requirements via targeted webinars, workshops, training materials, and collaborative partnerships. Our proactive approach to outreach has ensured we address emerging needs, engage diverse research communities, and foster trust through transparent feedback mechanisms.

Real-world user stories highlighting program impact
Proposed speaker: Ana Van Gulick, Figshare

This presentation will share compelling real-world user stories that illustrate how GREI's collaborative efforts have made NIH data sharing and reuse more impactful. By highlighting concrete examples of researchers benefiting from improved metadata standards, enhanced discoverability, and streamlined repository workflows, we will demonstrate the tangible value of GREI's innovative approach to open science.

Future directions for collaboration
Proposed speaker: Ana Van Gulick, Figshare
This presentation will explore future directions for GREI, focusing on improving data value and interoperability across repositories through enhanced metadata, connected digital objects, and thoughtful AI integration for data curation and metadata enhancement. We will also prioritize user engagement and incentives for transparent data sharing. Potential future tasks include developing a repository maturity model, facilitating federated search, and fostering a culture of data sharing and reuse to advance the research data ecosystem.

Facilitated audience discussion and Q&A

54

"So much going on!" How to best coordinate international efforts for data management —a polar to global case study and discussion

Authors: Chantelle Verhey[None]; Mark Parsons[None]; Michaela Miller[None]

Corresponding Authors: chantelle.verhey@gmail.com, miller@soos.aq, parsonsm.work@icloud.com

As data managers and practitioners, we've all felt it: "I want the data I handle to be broadly discoverable,interoperable, and preserved according to described best practice, but I first have to address the demands of my data providers and users who have their own special (and well-loved) way of doing things. I want to make sure I'm plugged into the leading practices of the global data management community while also making sure my domain/community is represented and heard. I need to solve my local problems in a way that fits into a global data ecosystem. I need to be in multiple places and spaces at once!"

Meanwhile, we are living in the world of big data and open science. The Fourth Paradigm is here. The possibilities for machines and algorithms to process and help interpret massive amounts of diverse data continues to accelerate while the array of scientific and operational initiatives collecting that data continues to grow. All this is testing human organizational capacity.

The polar science community has a long history of international collaboration and has been coordinating data since the first International Polar Year in 1882. Major data coordination initiatives continue today including the recently concluded MOSAiC (Multidisciplinary drifting Observatory for the Study of Arctic Climate) expedition, the Antarctica InSync (Antarctica International Science & Infrastructure for Synchronous Observation) initiative, as well as the next IPY in 2032 (IPY5). Correspondingly, the polar data community seeks to learn from the experiences of the global community as well as other regions and disciplines.

Using IPY5 as a milestone, the polar data community seeks to foster more effective and more efficient data collaboration at all levels. In this session, we will introduce current large-scale data coordination activities and challenges. We will then conduct a structured discussion around "integrated autonomy"to begin to identify "How is it that we can be more integrated and more autonomous at the same time?"The goal is to identify several focus areas for international collaborations (i.e. standardization) that also address local needs (i.e customization).

Session Speakers and Structure:

Introduction and Context (10 min) Mark Parsons CODATA Polar Data Advocate
- Overview of major data coordination activities and challenges in the Antarctic (10 min) Michaela Miller Data Officer, Southern Ocean Observing System
- Overview of major data coordination activities and challenges in the Arctic (10 min) Chantelle

Verhey Co-chair, Arctic Data Committee
- Structured Discussion (55 min)
– Introduction to the exercise "How is it that we can be more integrated and more autonomous at the same time when addressing the needs of large-scale science initiatives?"(5 min)

– Small groups work to:
— Identify tensions between a desire to standardize and the request for more customizing or autonomy?"(10 min)
— Pick an activity from the list above and ask "What is the rationale for standardizing? What is the rationale for customizing?"Then develop action steps that achieve standardization and action steps that achieve customization. (10 min)
— Identify actions that boost both standardization and customization and what modifications can be made to some actions so that they boost both standardization and customization. (10 min)
— Prioritize the most promising actions that promote both integration and autonomy. (10 min.)
– Report out (10 min)

– Closing and follow up (5 min)

Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research / 55

Challenges and Strategies for Ensuring the Quality and Reliability of Scientific Data at BrCris

Authors: MARCEL SOUZA[1]; Thiago Rodrigues[2]; Washington Segundo[2]

[1] *IBICT*

[2] *Ibict*

Corresponding Authors: marcelsouza@ibict.br, thiagomagela@gmail.com, washingtonsegundo@ibict.br

Introduction
The understanding of the development of scientific disciplines, knowledge dissemination, and technological evolution is predominantly informed by analyzing scientific publications, collaborative networks, and patent records. Brazil holds a prominent position in Latin America's scientific production, becoming a key regional player and talent attractor. The growing digitization of academic output has resulted in the availability of extensive datasets for scientific analysis, leading to the establishment of systems known as Current Research Information Systems (CRIS). CRIS aggregates scientific information, facilitating comprehensive data analysis on research ecosystems at various institutional and national levels.
BrCris, modeled after CRIS standards, consolidates information related to Brazilian scientific output —researchers, institutions, publications, and projects—and enables comprehensive bibliometric and scientometric analyses. However, integrating data from diverse sources generates considerable challenges related to harmonization and consistency, compromising analysis accuracy.
Persistent identifiers (e.g., DOI, ORCID) are vital to data accuracy, yet records often lack these identifiers or contain errors, complicating accurate data linkage and author attribution. Therefore, strategies for data disambiguation, deduplication, and metadata validation are paramount for improving BrCris reliability.
Common Data Quality Issues in BrCris
Integrating diverse data sources into BrCris presents several recurrent issues:
1. Record Duplication:
Duplicate records occur when identical authors or publications appear repeatedly due to absent or inconsistent identifiers and metadata discrepancies. This redundancy distorts co-authorship networks and artificially inflates productivity statistics, impairing scientific analyses.
2. Inconsistent Persistent Identifiers:
Incorrect or missing DOI, ORCID, and ISSN identifiers complicate normalization processes, author recognition, and citation attribution. Approximately 7% of Brazilian researchers on Plataforma Lattes have registered ORCID IDs, limiting the interoperability and accuracy of author-publication associations.

3. Entry and Standardization Errors:
Manually entered textual fields (authors'names and institutions) contain variations, inconsistent abbreviations, and typographical errors. For example, Universidade Federal de Minas Gerais can appear in multiple formats, complicating aggregation and consistency. These variations significantly impair accurate network analyses and institutional evaluations.
4. Outdated and Incomplete Records:
Many records become outdated due to researchers' negligence in updating profiles or linking publications to identifiers. Incompleteness in repository metadata indexing further reduces visibility, impacts accurate performance indicators, and weakens data-driven policy formulation.
These challenges profoundly influence the reliability and accuracy of scientific performance assessments, negatively affecting policymaking and resource allocation processes based on BrCris data.
Impact on Bibliometric and Scientometric Analysis
Data quality directly impacts bibliometric and scientometric accuracy, crucial for scientific evaluation, institutional assessments, and policy formulation. Duplicate records inflate productivity indicators, while inconsistent standardization underestimates scientific output, distorting accurate assessments of research impact.
Mismanaged co-authorship networks, caused by inadequate disambiguation, lead to fragmentation or artificial collaborations, misrepresenting institutional performance. Inconsistent article indexing impacts accurate citation attribution, compromising bibliometric analyses, such as citation counts and impact factor measurements.
The integration of multiple sources, each with varying metadata standards, exacerbates interoperability issues, complicating accurate identification of international collaborations. Such inaccuracies can lead to misguided resource allocation, skewing policy decisions and affecting institutional funding equitably.
Strategies for Improving Data Quality
To mitigate these challenges, several technical solutions have been implemented in BrCris:
• Record Deduplication:
Applying machine learning algorithms, heuristic rules, and clustering techniques efficiently identifies and merges duplicate records. OpenRefine and similar tools help standardize metadata, improving database accuracy.
• Researcher and Institutional Disambiguation:
Utilizing persistent identifiers (ORCID for researchers and ROR for institutions) substantially reduces ambiguity and enhances tracking capabilities. Integrations between Lattes and OpenAlex demonstrate effective correlation between author identities and institutional affiliations.
• Data Validation and Certification:
Implementing international metadata standards (CERIF) and automated API-based validation procedures contributes significantly to enhancing metadata consistency. Cross-referencing data from trusted platforms (Oasis.Br, OpenAIRE Research Graph) supports automatic correction, ensuring reliability.
• Integration with Bibliometric Tools:
Strengthening integration with analytical tools (VOSviewer, Gephi, Visão) enables systematic anomaly detection and data quality monitoring through interactive dashboards, identifying inconsistencies for proactive resolution within BrCris.
The continuous adoption and refinement of these strategies significantly bolster BrCris's credibility and enhance the precision and robustness of its bibliometric and scientometric analyses.
Concluding Remarks
Data quality is fundamental for BrCris effectiveness, reliability, and usefulness in scientific evaluation. Implementing robust strategies for deduplication, disambiguation, and metadata validation is essential to ensuring data integrity.
BrCris has significant potential as a reliable scientific information ecosystem if systemic challenges are adequately addressed. Adopting international metadata standards and persistent identifiers enhances global interoperability, data synchronization, and credibility.
Recent studies demonstrate data quality directly affects metrics such as productivity, scientific impact, and research collaboration. Innovative technological solutions such as machine learning and semantic analysis provide viable pathways for further enhancing BrCris capabilities.
Beyond data management, BrCris represents an initiative to strengthen Brazil's scientific information infrastructure, promoting greater transparency and data accessibility. Improvements in data quality not only enhance domestic scientific credibility but also facilitate international collaboration and recognition.
Despite significant advancements, persistent challenges remain, especially concerning intentional or unintentional errors and misuse of identifiers. Issues like deliberate duplicate entries and incorrect authorship attribution underscore the importance of complementing automated solutions with active community involvement and institutional policies promoting transparency and responsible data management.
Ultimately, maintaining data quality within BrCris demands continuous technological and method-

ological improvements combined with community participation, ensuring the accuracy, reliability, and global competitiveness of Brazilian scientific information.

Poster Session / 56

Mapping permafrost in the Northern Hemisphere

Author: Youhua Ran[1]

[1] *Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences*

Corresponding Author: ranyh@lzb.ac.cn

Due to the permafrost develops at certain subsurface depths, it cannot be directly observed by remote sensing, and ground-based surveys are costly. As a result, there remains considerable uncertainty in our current understanding of permafrost distribution. This study employs an ensemble simulation approach using multiple machine learning models, integrating the most comprehensive international ground borehole observations with environmental parameters derived from remote sensing, such as freezing and thawing indices. We have developed high-accuracy data products for the Northern Hemisphere at a spatial resolution of 1 km, including mean annual ground temperature, active layer thickness, and subsurface ice content. From the perspectives of formation conditions and vulnerability, we classified permafrost zones into five types: climate-driven, climate-driven with ecosystem regulation, climate-driven with ecosystem protection, ecosystem-driven, and ecosystem-protected. This classification provides a more detailed representation of the vulnerability of permafrost across the Northern Hemisphere.
Specifically, the novel permafrost datasets for the Northern Hemisphere, including the mean annual ground temperature (MAGT) at the depth of zero annual amplitude (DZAA) (approximately 3 m to 25 m) and active layer thickness (ALT) with 1-km resolution for the period of 2000–2016, as well as the probability of permafrost occurrence and the biophysical zonation. These datasets integrate unprecedentedly large amounts of field data (1,002 boreholes for MAGT and 452 sites for ALT) and multisource geospatial data, especially remote sensing data, using statistical learning modelling with an ensemble strategy. Thus, the resulting data are more accurate than those of previous circumpolar maps (bias=0.02 ℃, RMSE=1.32 ℃ for MAGT; bias=2.71 cm, RMSE=86.93 cm for ALT). The datasets suggest that the areal extent of permafrost (MAGT≤0 ℃) in the Northern Hemisphere, excluding glaciers and lakes, is approximately 14.77 (13.60–18.97) million square kilometers and that the areal extent of permafrost regions (permafrost probability>0) is approximately 19.82 million square kilometers. We developed a rule-based decision framework to delineate the biophysical permafrost zones in the Northern Hemisphere at 1-km resolution that incorporates the interactions among biophysical factors on permafrost vulnerability. The permafrost regions was classified into five types: climate-driven (19%), climate-driven/ecosystem-modified (41%), climate-driven/ecosystem protected (3%), ecosystem-driven (29%), and ecosystem-protected (8%). This map indicate that the 81% of the permafrost regions in the Northern Hemisphere are affected by ecosystems, indicating the dominant role that ecological processes have in controlling permafrost stability. The finding highlights the importance of reducing ecosystem disturbances (natural and human activity) to help slow permafrost degradation and lower the related risks from a warming climate. The map is potentially useful for predicting permafrost degradation and ecological transitions, and for assessing the future risks to infrastructure and society from climate warming, as well as for planning the mitigative strategies and measures of engineering infrastructure in cold regions.

Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research / 57

Integrating Machine Learning Standards in Disseminating Machine Learning Research

Authors: Scott Edmunds[1]; Nicole Nogoy[2]; Qing Lan[None]; Hongfang Zhang[None]; Yannan Fan[None]; Hongling Zhou[None]; Chris Hunter[None]; Chris Armit[None]

[1] *GigaScience Press, BGI Hong Kong*

[2] *GigaScience Press*

Corresponding Authors: scott@gigasciencejournal.com, nicole@gigasciencejournal.com

The increasing use of AI-based approaches such as machine learning (ML) across diverse scientific fields presents challenges for reproducibly disseminating and assessing research. As ML becomes increasingly integral to clinical applications, there is also a critical need for transparent reporting methods to ensure both comprehensibility and the reproducibility of pre-clinical research and clinical trials supporting them. To address this issue there are a growing number of standards, checklists and guidelines enabling more standardized reporting of ML research, but the proliferation and complexity of these make them challenging to use in the process of disseminating research. Particularly in assessment and peer review of scientific papers which has to date been an ad hoc process that has struggled to throw light on increasingly complicated computational supporting methods that are otherwise unintelligible to other researchers. Taking the publication process beyond these black boxes, GigaScience Press has experimented with integrating many of these ML-standards into their assessment and publication workflows to make the outputs more FAIR. But having a broad-scope has necessitated going beyond the many field specific and standards to look for more generalist and automated approaches. In this talk we will introduce and map the landscape of different field-specific (predominantly clinical) reporting guidelines alongside some new broader-scope generalist ML-standards and checklists that have been released. Outlining the rationale for our eventual adoption of the DOME recommendations for machine learning in biology. The DOME recommendations initially formulated by the ELIXIR-Machine Learning Focus Group, DOME being an acronym for: Data, Optimisation, Model, and Evaluation. Initially use of these guidelines has been to help screen which ML publications are suitable to send out to peer-review. Through further collaborating with the DOME-community, we have now integrated their DOME Data Stewardship Wizard (DOME-DSW) and DOME Registry tools into our peer-review and publication process. At the end of the review process process a DOME Registry persistent identifier is included in the manuscript to increase the visibility and discoverability of the annotations and different ML components (data, models, methods, etc.) making up the study to the research community. Carrying out these experiments we have found a more practical approach for assessing and sharing ML-research is “Trust and Transparency”rather than trying to test de facto reproducibility. Presented here these efforts provide a useful case study of approaches, workflows and strategies to more logically handle the peer review and dissemination of data intensive ML research. We emphasise the need for continued dialogue and collaboration among various ML communities to create unified, comprehensive standards, ultimately enhancing the credibility and impact of ML-based scientific research. With the hope of others testing this approach in outlets with different scopes and publication volumes to see if it remains practical and can become a wider standard for sharing of ML research in a rigorous, reproducible and FAIR manner.

Presentations Session 2: Data and Research & Data Science and Data Analysis / 59

Study on Handling Dark Data in HPCI Shared Storage System using the WHEEL Workflow Tool

Author: HIdetomo Kaneyama[1]

Co-authors: Tomohiro Kawanabe [1]; Hiroshi Harada [1]

[1] *RIKEN R-CCS*

Corresponding Authors: tkawanabe@riken.jp, hidetomo.kaneyama@riken.jp, hiroshi.harada@riken.jp

Overview

Understanding unnecessary data, known as Dark Data is a major operational challenge in large-scale shared storage. We propose an alternative approach that leverages HPC workflow tools to collect extended metadata at each stage of job execution, minimizing the need for fundamental system changes.

Background

The High Performance Computing Infrastructure (HPCI) project was established to provide a unified environment for accessing supercomputers deployed at universities and research institutes on Japan. This project provides network-storage service to HPCI users for the shared research data between supercomputing center among HPCI, called HPCI Shared Storage(HPSS), jointly operated by The University of Tokyo and RIKEN Center for Computational Science. HPSS began service in 2012 and will transition to 3rd generation system by 2025, to serve a total logical capacity of 95 petabytes(PB).HPSS is built upon the network-based file system known as Gfarm 1.

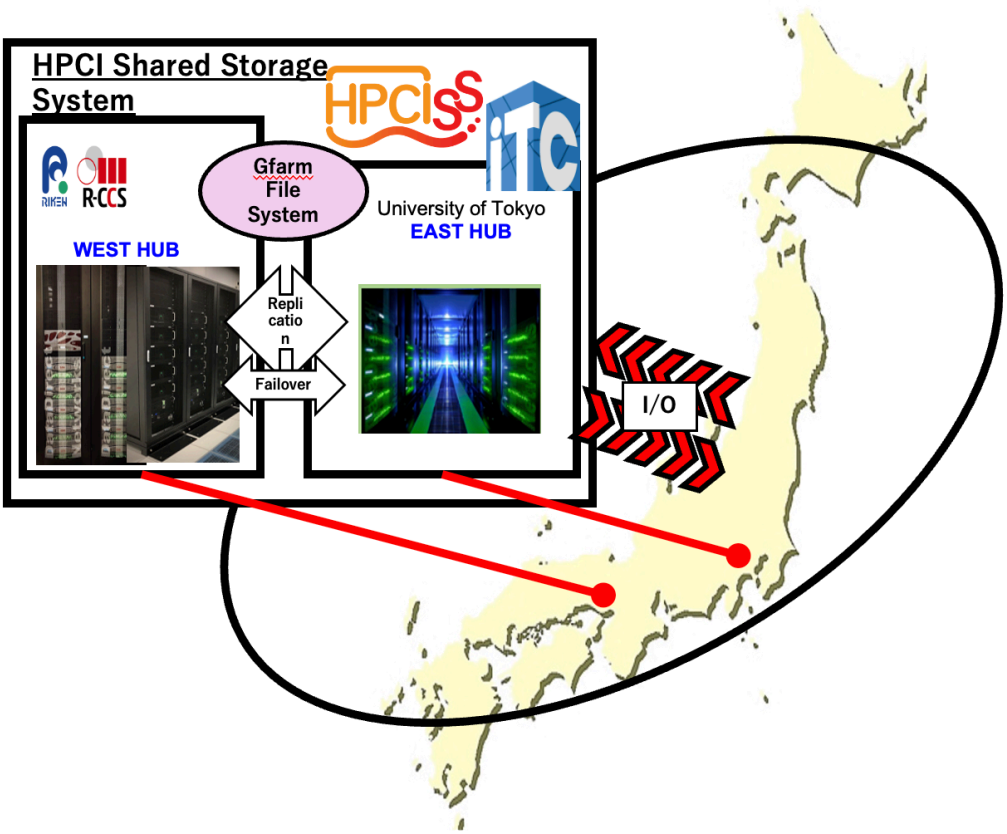

Figure 1: HPCI Shared Storage Overview

By 2024, user demand for 2nd generation HPSS surpassed the 50 PB limit, to operate under an overcommitted state. An analysis of metadata via Gfarm, revealed that much of the preserved data was Cold Data, rarely accessed despite occupying large capacity.

Discussion

Since HPSS is a free service provided to researchers with HPCI projects, maximizing its public utility is essential. Accordingly, two policies have been established

1. Prioritize users who do important research and share their data openly
2. Help users improve data management by showing infrequently used data.

To support the second policy, metadata such as access times are collected automatically and shown in dashboards like Grafana to highlight rarely used data. However, Gfarm and many HPC filesystem, such as Lustre and BeeGFS, do not provide functions to annotation custom tags or store detailed data meta information (e.g. when research data was generated). Without such extended metadata, users are unsure if certain old data and left by inactive project members, can be safely removed. This

uncertainty fuels the accumulation of Dark Data, defined as data whose purpose or ownership is unclear 2. Because users receive storage at free, there is no incentive to remove Dark Data, further contributing to Dark Data growth. A more advance data management framework is therefore critical. Ideally, expanded metadata should be automatically recorded at the time of data generation or stored storage, and should accompany the data lifecycle, from creation during computation to HPSS. However, many HPC filesystem and job schedulers (e.g. PBS) lack automated extended metadata tagging. Introducing new services (e.g. Globus, Starfish) or adding such functionality directly to Gfarm must demand extensive redevelopment and collaboration among supercomputing centers.

Proposed Method

To solve these challenges, we propose an alternative approach that gathers extended metadata at each stage of job execution by leveraging HPC workflow tools, thereby minimizing the need for fundamental system changes. By coordinating job submissions and data storage within a unified workflow framework, we can automatically capture computational resources used, JobID, software name and version, user who executed job, and the time of execution. This metadata is annotated for all result data.
We utilize WHEEL workflow tool 3, which already supports I/O operations to HPSS.

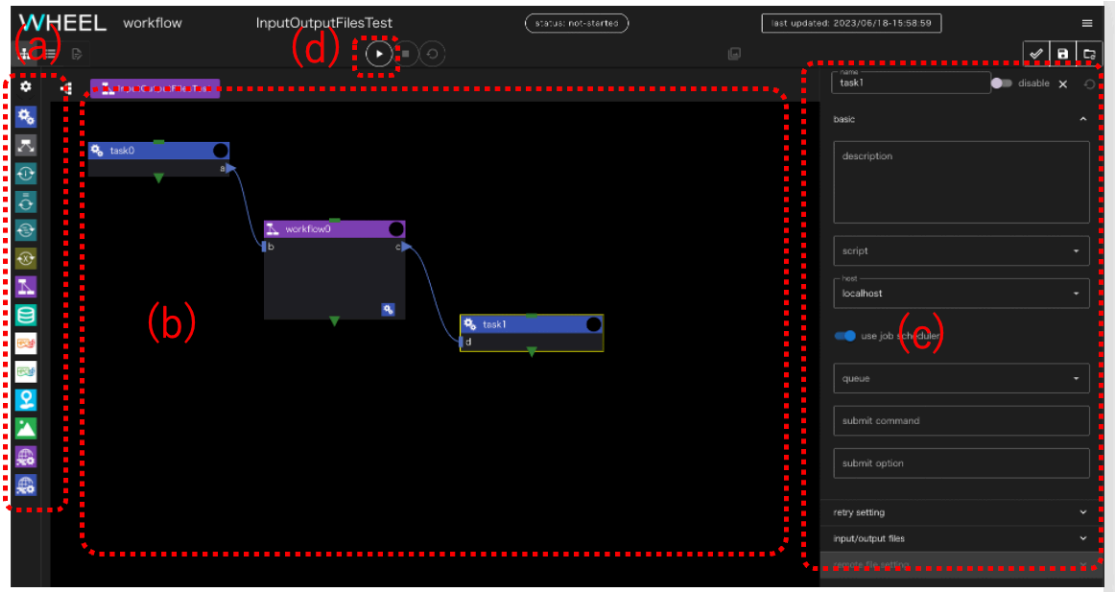

Figure 2: WHEEL's editor screen. (a): Component list, (b): Graph view, (c): Property pane, (d): Run-proj. btn.

By extending WHEEL to incorporate metadata assignment, we can automatically annotated expended metadata, such as supercomputer names, Job IDs, software name and versions, create usernames. Existing research on workflow tools sometimes focuses on annotating extended metadata to entire workflows 4, but our goal is to directly annotate each data for clearer provenance and future integration with research data management (RDM) services. In practice, WHEEL automatically collects expanded metadata, storing these alongside file paths in NoSQL database. Because Gfarm has data archive function, so it's possible to annotate extended metadata to each packaged of data, allowing users to detailed information about the package-data later. This approach mitigates the risk of changing to Dark Data. Currently, our proof of concept writes extended metadata to a NoSQL database, however we are considering the following for future development.

1. Automated data publication, with digital object identifiers (DOI) assigned.
2. Extending Gfarm to embed extended metadata.
3. Integrating with external RDM services. (as explored in projects such as HOMER 5)
4. Extend Metadata visualization.

While this paper focuses on HPSS and WHEEL, our broader aim is to develop metadata management framework applicable across diverse HPC environments. Automating metadata annotation can simplify future data management for researchers and support practices such as RDM and open science. In particular, it can help reduce the accumulation of Dark Data in large-scale storage systems. Such efforts are expected to promote more equitable resource allocation and more efficient, transparent use of research data.

Reference

1 Tatebe, O., et al. (2010). Gfarm grid file system. New Generation Computing, 28(3), 257–275. https://dl.acm.org/doi/10.1007/s00354-009-0089-5
2 Bauer, D., et al. (2022). Revisiting data lakes: The metadata lake. In Proceedings of the 23rd International Middleware Conference Industrial Track (pp. 8–14). ACM. https://doi.org/10.1145/3564695.3564773
3 Kawanabe, T., et al. (2024). Introduction of WHEEL: An analysis workflow tool for industrial users and its use case on supercomputer Fugaku. In Proceedings of the 2024 IEEE International Conference on Cluster Computing Workshops (pp. 180–181). IEEE. https://www.computer.org/csdl/proceedings-article/cluster-workshops/2024/834500a180/21EtRZFluLu
4 Jain, A., et al. (2015). FireWorks: A dynamic workflow system designed for high-throughput applications. Computational Materials Science, 96, 118–124. https://doi.org/10.1016/j.commatsci.2014.10.037
5 Chiapparino, G., et al. (2024). From ontology to metadata: A crawler for script-based workflows. INGGRid. https://www.inggrid.org/article/3983/galley/3912/download/

Presentations Session 6: The Transformative Role of Data in SDGs and Disaster Resilience / 62

A Collaborative Data Network for the Asia Oceania Region Enabled by Emerging Technologies to Foster Innovation in a Secure and Open Environment.

Author: Alison Specht[1]

Co-authors: Kim Byrceson [2]; Margaret O'Brien [3]; Michelle Waycott [4]; Pedro Correa [5]; Shoufeng Cao [2]; Siddeswara Guru [1]

[1] *TERN, University of Queensland*

[2] *School of Agriculture & Food Sustainability, the University of Queensland*

[3] *University of California, Santa Barbara*

[4] *Univ. Adelaide*

[5] *University of Sao Paulo, Brazil*

Corresponding Authors: margaret.obrien@ucsb.edu, s.guru@uq.edu.au, s.cao@uq.edu.au, pedro.correa@usp.br, k.bryceson@uq.edu.au, michelle.waycott@adelaide.edu.au, a.specht@uq.edu.au

A discoverable inventory of items of value and importance to a community, country or region has many benefits. These benefits can be for reporting (state of the environment, achievement of Sustainable Development Goals, conservation objectives), for research (exploration of international phylogenetics, development of new pharmaceuticals), or for economic reasons (tracking the status of fish stocks, timber trade and other food products). As clear from the principles of Open Science, confidence in the value of data is best achieved if the information used to underpin statements and conclusions is open to scrutiny.
Active participation in conserving data in a temporal and spatially relevant manner is desirable. This can result in better knowledge mobilisation and dissemination, and increased community benefits through the capacity building of community partners and their greater involvement in knowledge production and mobilisation activities. Making resources available for the community means ensuring that data (and related materials) are findable and accessible on the Web, and that they comply with adopted international standards making them interoperable and reusable by others (David et al., 2020).
This presentation presents the 'Collaborative Regional Data Access Network' (CREDAN) for the Asia-Oceania region, using the progress made by the Pacific Environment Data Portal (pacific-data.sprep.org)

as an exemplar. CREDAN blends the attributes of centralised and decentralised networked databases for the management and discovery of data to comply with (a) FAIR principles, uses criteria to ensure (b) CARE principles are followed, while (c) proposing that the information is held in repositories that follow the TRUST principles. Using Local Context (localcontexts.org/) labels, data owners determine the sovereignty and sharing permissions for their data. To ensure ownership is always clear, and subsequent use of data is tracked, we propose harnessing blockchain technology alongside smart contracts and Decentralised Identifiers (DIDs) to allow for secure discovery and sharing. The use of blockchain, with its features of immutability, security and transparency (data anchoring), will ensure the information about each dataset (metadata, provenance and access rights) is immutably connected to each dataset. Using DIDs, each party in a transformation or movement of data can be recorded on the distributed blockchain ledger. This creates an auditable trail of data provenance that is transparent and tamper-proof, allowing for easy tracing of data back to its origin. The use of smart contracts can further automate and enforce the establishment of common standards for data types, metadata requirements, and vocabulary descriptions, promoting consistency, transparency, and ease of discovery in data management. Smart contracts can be used to enable access permissions to a person or entity, record a data request, and communicate transactions to the primary data owner.
To ensure security of the data gathered by the community CREDAN employs a combination of on-chain and off-chain components: (i) the management of the original data and metadata (off-chain), (ii) data management including establishment of the DIDs and smart contracts (on-chain), and (iii) user access and verification (bridging).
We recommend a strong governance structure to ensure best practices are aligned with international criteria and those set by the component communities. Management of the eventual large volume of data should be anticipated, and partnerships may need to be considered. We propose a seven-step pathway to creation of such a network. A seven-step implementation pathway is suggested, beginning with establishing a common intent, through data collection, and using blockchain technology and unique DiDs to enable secure labelling and tracking of the data.
We consider this work to be of wide interest to data engineers, repository managers, those involved in community-driven data security among others.
This presentation is based on a recent publication (doi: 10.5334/dsj-2025-001).

64

Implications of new access and benefit sharing regimes for global research using genetic data

Authors: Scarlett Sett[1]; Yiming Bao[2]; Eizadora Yu[3]; Michelle Rourke[4]

[1] *CSIRO Australian Centre for Disease Preparedness*

[2] *National Genomics Data Center, China National Center for Bioinformation*

[3] *Marine Science Institute, University of the Philippines Diliman*

[4] *Law Futures Centre, Griffith University*

Corresponding Authors: m.rourke@griffith.edu.au, etyu@up.edu.ph, scarlett.sett@csiro.au, baoym@big.ac.cn

Fair and equitable sharing of the benefits generated by using genomic sequence data and other related digital data, referred to in policy circles as "Digital Sequence Information"(DSI), is a hot topic in several international fora. It builds off similar discussions related to the sharing of benefits from using *physical* genetic resources, which led to the creation of the Nagoya Protocol in 2010. The Protocol has shaped and impacted how research is done across the globe. As governments discuss how to create benefit sharing mechanisms for *digital* genetic resources, it is key that the scientific community be aware of the debates and their possible implications.

This session will provide an overview of the debates under the Convention on Biological Diversity (CBD), under the High Seas Treaty agreement on Biodiversity in Areas Beyond National Jurisdiction (BBNJ), under the WHO Pathogen Access and Benefit Sharing system (PABS), and the International Treaty on Plant Genetic Resources (IT-PGRFA).

Speakers will discuss the implications and opportunities of these different and possibly overlapping agreements, considering how they could improve data management to address concerns about fair-

ness and ethics, while also presenting possible challenges to the open science infrastructure. Panelists will outline some of the implications for researchers and data managers in terms of compliance, legal certainty and capacity building.

Format: this will be a panel discussions, with four presenters. The proposed agenda is as follow:

5 min - Welcome and introduction to what "DSI"means - Session chair
10 min - Presentation 1 - Scarlett Sett
10 min - Presentation 2 - Michelle Rourke
15 min - Q&A
10 min - Presentation 3 - Yiming Bao
10 min- Presentation 4 - Eizadora T. Yu
20 min - Q&A
5 min - Closing remarks - Session chair

Panellists:
- **Scarlett Sett, CSIRO Australian Centre for Disease Preparedness, Australia** –What is going on with DSI? An overview of current negotiations under the Convention on Biological Diversity and implications for research
- **Michelle Rourke, Law Futures Centre, Griffith University, Australia** –Exploring tensions between fairness, equitable access and open science: Possible implications of the WHO Pathogen Access and Benefit Sharing system
- **Yiming Bao, National Genomics Data Center, China National Center for Bioinformation, China** –Data governance: what can databases do to improve fairness and equity in support of open science collaborations?
- **Eizadora T. Yu, Marine Science Institute, University of the Philippines Diliman, The Philippines**: How can the new agreements on benefit sharing be channeled to reduce inequities in the research community and empower researchers from LMICs?

Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global / 65

Breaking the Silos in Environmental Science One Infrastructure at a Time

Author: Anca Hienola[1]

Co-authors: Alex Vermeulen [2]; Andreas Petzold [3]; Claudio Dema [4]; Daniele Bailo [5]; Dario de Nart [6]; Federico Drago [7]; Magdalena Brus [7]; Marta Gutierrez David ; Ulrich Bundke [8]; Zhiming Zhao [9]

[1] *Finnish Meteorological Institute*

[2] *ICOS*

[3] *FZ Juelich*

[4] *CNR*

[5] *Istituto Nazionale di Geofisica e Vulcanologia (INGV)*

[6] *CREA*

[7] *EGI*

[8] *FZ Julich*

[9] *UVA*

Corresponding Authors: u.bundke@fz-juelich.de, dario.denart@crea.gov.it, a.petzold@fz-juelich.de, anca.hienola@fmi.fi, claudio.dema@cnr.it, z.zhao@uva.nl, marta.gutierrez@egi.eu, magdalena.brus@egi.eu, federico.drago@egi.eu, alex.vermeulen@icos-ri.eu, daniele.bailo@ingv.it

Despite major investments in Open Science infrastructures, environmental research remains fragmented across domains, systems, and borders. We have no shortage of data. What we lack are better ways to turn it into science. ENVRI-Hub NEXT project tackles this by breaking silos and building real bridges—one infrastructure at a time.

The ENVRI-Hub NEXT advances the ENVRI-Hub, a platform that integrates European environmental Research Infrastructures (RIs) across atmosphere, marine, terrestrial, and biodiversity domains. It provides harmonized discovery, access, and use of multidisciplinary data, services, and training resources. But ENVRI-Hub is not just another portal. It is a functioning federation, connecting European research infrastructures with global policy objectives, including the Green Deal and the SDGs. This hub moves beyond isolated silos and embraces true interoperability—not only at the metadata level, but at the service, computation, and knowledge layers. Researchers can compute Essential Environmental Variables, run cross-domain analyses, and access services mapped directly to policy-relevant frameworks—supporting science that matters, and science that leads to action.
In this practice presentation, we will present:
• How ENVRI-Hub structures cross-domain data services to support environmental research from the local to the global scale.
• How the Hub advances FAIR principles across diverse infrastructures, while maintaining domain-specific richness.
• How ENVRI-Hub connects scientific infrastructures with global policy agendas, creating an infrastructure that informs governance, not just research.
We will also reflect critically on key challenges:
• Why many current Open Science infrastructures underperform in interdisciplinary data access, and how ENVRI-Hub addresses this gap.
• Why sustainable governance models matter more than just technical integration.
• How future infrastructures must embed transnational and interdisciplinary workflows by design —not as an afterthought.
ENVRI-Hub is a case study in building infrastructure that is scientifically relevant, policy-relevant, and user-driven. It shows that environmental research infrastructures can work across domains and borders without sacrificing scientific depth or usability. It also shows that federation across RIs is not only possible—it's necessary if we are serious about supporting the global research community to address climate change, biodiversity loss, and sustainable resource management.
The work presented will be particularly relevant to developers and managers of research infrastructures, policymakers and funders designing Open Science and data-sharing strategies, researchers and research communities seeking better ways to access and use interdisciplinary environmental data.
In a landscape crowded with infrastructures that promise connection but deliver isolation, ENVRI-Hub NEXT proves that true federation is not only possible —it is essential if environmental science is to meet the global challenges it faces.

Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research / 67

Supporting dataset curation through automation at KU Leuven

Authors: Dieuwertje Bloemen[1]; Ozgur Karadeniz[1]

[1] *KU Leuven*

Corresponding Author: dieuwertje.bloemen@kuleuven.be

KU Leuven RDR is the CoreTrustSeal certified institutional data repository of KU Leuven, where curation plays an important role in data FAIRification and ensuring the quality of published datasets. The curation phase is not only crucial to have some quality control on the FAIRness of the data by ensuring correct metadata input, the presence of documentation and a choice of license, but to also ensure that the researchers are fully informed and supported in their efforts to publish their data.
After the repository's launch in 2022, the monthly number of datasets published slowly increased overtime, and with that the number of dataset reviews to be carried out. As these numbers increased, it became clear that there was a need to better track the reviews and who is picking each review up, as well as a need to streamline this process in general. This would not only prevent unnecessary duplication of work, but would also potentially free up more time for support rather than the evaluation itself. To streamline the curation process, the RDR team developed an open-source review dashboard that plugs in to a Dataverse instance and automates different parts of the review process.
In the initial iteration of the dashboard, the automation focused on the administrative side of the reviews. For example, in the dashboard, reviewers can easily track who reviews what dataset, can

add notes to any review and look back at the review history of said dataset. On top of that, the effort to streamline the feedback process resulted in the implementation of simple checklist in the review dashboard they can use to autogenerate feedback. This ensures uniformity in reviews, while still allowing for customizations, and prevents reviewers having to type the same feedback over and over again. This initial version of the dashboard was key to processing more datasets ready for publication and enabled reviewers to focus on the reviews themselves and not the administrative mess that previously came with it.
A second version of the review dashboard goes even a step further in its automation efforts. As the reviews were being carried out, some frequently made mistakes were flagged as having potential to be automatically found. With this idea an initial exploration began of what curation elements could all be automatically checked and how. From exploration, we found a lot of potential, such as indicating when a README file is likely missing, or when a README file is present, but empty. The list of potential automated checks was longer than expected and were easier to implement than we had anticipated. A bigger challenge, however, was to balance this automation with the human effort and input that is key in data curation. Some brainstorming on how to visualize this automation and how to always allow for human overwrites were necessary to ensure that the review supports human curation through automation and doesn't replace it.
In this presentation, we'll share our road to the creation of the review dashboard and a look at our UI, but also provide an insight into the logic of the automated checks. We hope to spark conversation on how to further support the human task of curation through tools and technology without losing the important human touch and interpretation that is so valuable to making a dataset as FAIR as possible.

Presentations Session 4: Data Stewardship / 68

Reexamination of historical secondary data given federal funding cuts?

Author: Shannon Farrell[1]

Co-authors: Julia Kelly ; Kristen Mastel [1]; Lois Hendrickson [1]; Nikki Galloway [2]; Stephanie Sparrow [1]

[1] *University of Minnesota Libraries*

[2] *Old Dominion University*

Corresponding Authors: sfarrell@umn.edu, l-hend@umn.edu, ngallowa@odu.edu, jkelly@umn.edu, ssparrow@umn.edu, meye0539@umn.edu

Older data in paper or analog format (e.g., field/lab notebooks, photos, maps) held in labs, offices, and archives across research institutions are an often overlooked resource for potential reuse in new scientific studies. However, with the uncertainty around federal funding and research administration in the United States, there has been speculation that scientists across multiple disciplines will be utilizing more secondary data and that historical data will become more valuable. Reuse of historical data is particularly important in studies of biodiversity and climate change.

We have been examining the landscape of historical scientific data use and scientific researcher perspectives on their use of historical analog data. We have also been investigating scalable institutional workflows for organizing, describing and digitizing paper data at a large research University. When transitioning historical data that originates in paper to digital, we identified several unique challenges, including the presence of sticky notes, difficulty in interpreting handwritten notes, and unclear provenance and dates.

Our research tells us that scientific researchers care about this data, consider it valuable, and that these datasets exist in large amounts and are a potentially large, untapped resource. We explore how these datasets could be unearthed and shared to wide benefit. Currently, there are few mechanisms to help researchers find existing historical analog data in order to reuse it. This is a persistent problem, as evidenced by several large projects that have tried to address this over the last few decades. Large-scale solutions have not been identified to date and researchers are siloed by discipline. We are seeking solutions that can work at different scales, across disciplines, and for both researchers and data stewards.

We will discuss example projects involving fruit breeding data, and methodology including how data curators and researchers can work together to balance repository criteria (authenticity, data integrity, reliability and persistence of service) and expectations of the scientists, curators and the research community. The work to organize and convert these datasets from paper to digital formats and the time spent with data producers in improving the metadata and description increased the likelihood that this data would be more in alignment with the FAIR principles.
This presentation will be beneficial to data curators working with contemporary and historical data and scientists interested in exploring legacy data that could be applicable to current and future research.

Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research / 71

Research Infrastructure for Solid Earth Sciences: The Case for International Collaboration

Authors: Carmela Freda[1]; Daniela Mercurio[1]; Daniele Bailo[2]; Elisabetta D'Anastasio[3]; Elizabeth Abbott[3]; Federica Tanlongo[1]; Gaetano Festa[4]; Helen Glaves[5]; Jan Michalek[6]; Jonathan Hanson[3]; Otto Lange[7]; Rebecca Bendick[8]; Rebecca Farrington[9]; Rossana Paciello[2]

[1] *EPOS ERIC*

[2] *INGV*

[3] *GNS NZ*

[4] *University of Naples "Federico II"*

[5] *BGS*

[6] *University of Bergen*

[7] *University of Utrecht*

[8] *Earthscope Consortium*

[9] *AuScope*

Corresponding Authors: rebecca.bendick@earthscope.org, rebecca@auscope.org.au, jan.michalek@uib.no, o.a.lange@uu.nl, federica.tanlongo@epos-eric.eu, daniela.mercurio@epos-eric.eu, hmg@bgs.ac.uk, e.danastasio@gns.cri.nz, executive.director@epos-eric.eu, j.hanson@gns.cri.nz, e.abbott@gns.cri.nz, rossana.paciello@ingv.it, gaetano.festa@unina.it, daniele.bailo@ingv.it, tim@auscope.org.au

Solid Earth science seeks to understand the complex chemical and physical processes shaping our planet. This knowledge is essential for addressing key societal challenges, from mitigating natural hazards to managing vital resources. Yet, the scale and nature of the data required for such research —spanning petabytes and crossing geographical, disciplinary, and temporal boundaries—necessitate a global response supported by robust, interoperable Research Infrastructures (RIs).
Research Infrastructures play a central role in enabling this science by providing high-quality, open, and standardized data and services. They foster scientific excellence, promote equitable access to resources, and underpin the collaborative frameworks necessary for large-scale, multidisciplinary research. In the field of solid Earth science, international collaboration among RIs is not just beneficial—it is essential.
Recognizing this, EPOS (Europe), AuScope (Australia), and EarthScope (United States)—the leading solid Earth RIs in their respective regions—have formally committed to working together through a Memorandum of Understanding signed in September 2024. Their collaboration is rooted in shared principles: advancing Open Science; upholding FAIR (Findable, Accessible, Interoperable, Reusable) data practices; and striving toward global equity in scientific research. Recently, this initiative has expanded with the participation of GNS Science, New Zealand's national RI for geoscience, which is in the process of joining the agreement.
The grand vision underpinning this collaboration is the creation of a globally federated research infrastructure—one that interlinks regional platforms into a cohesive, interoperable system. Such a federation would allow scientists worldwide to access and contribute to high-quality, multidisciplinary data and services, accelerating discovery and innovation in solid Earth sciences on a planetary scale.

To realise this vision, these organizations aim to federate and harmonize their platforms, enabling seamless access to multidisciplinary data and services across continents. This vision demands coordinated action to align technical standards, protocols, and vocabularies, and to overcome significant barriers—legal, political, economic, and infrastructural.
Key challenges include:
Interoperability: Aligning technical frameworks and data standards across regions to enable seamless data sharing and integration. A key challenge lies in the varying levels of maturity across geoscience subdomains when it comes to data and metadata standards. While some scientific communities, such as seismology and geodesy, have a strong tradition of global collaboration and well-established protocols, others remain fragmented and operate with divergent standards or limited interoperability. This heterogeneous landscape complicates the integration of multidisciplinary datasets, which is essential for addressing complex solid Earth processes. Overcoming this hurdle requires fostering cross-domain dialogues, supporting the development and adoption of shared standards, and creating incentives for communities to converge around interoperable solutions that respect disciplinary specificities.
Legal and regulatory hurdles: Navigating diverse national policies governing data access and cross-border collaboration. While the European Union benefits from a common regulatory framework that generally facilitates alignment across member states, many other regions lack such harmonization. This disparity can hinder international cooperation and the free exchange of scientific data. Bridging this gap will require multilateral dialogues to develop shared legal principles, the adoption of interoperable data governance frameworks, and capacity-building efforts to support regions'more fragmented regulatory systems. International organizations and policy fora can play a vital role in fostering convergence while respecting national sovereignty and data rights.
Sustainability: While long-term funding remains a cornerstone, it represents only one facet of the broader sustainability challenge. Equally critical are open and accessible technical infrastructure, high-performance computing, data storage, and robust connectivity—resources still lacking in many regions. Disparities in expertise and limited training opportunities further constrain participation and innovation.
. Ensuring sustainability therefore demands a holistic approach that couples investment in digital infrastructures with the widespread development of skills and knowledge through inclusive education and training programs, workforce mobility and capacity building.
An integral part of the sustainability effort is also long-term data preservation—particularly for geospatial data, whose volume and multidimensionality pose significant challenges. Ensuring the longevity and usability of such data demands coordinated strategies that optimise storage, adapt to evolving formats, and support reproducibility and future reuse.

Inclusivity and equity: Promoting research autonomy in lower-resourced regions and avoiding scientific dominance by wealthier nations. This includes upholding the CARE (Collective benefit, Authority to control, Responsibility, and Ethics) principles, particularly in relation to Indigenous data governance. Research Infrastructures must recognize and support the rights, knowledge systems, and agency of Indigenous People by integrating Indigenous knowledge respectfully, ensuring community-led data stewardship, and enabling meaningful participation in research and infrastructure development.
Addressing these challenges requires not only scientific and technical coordination but also sustained policy support at national and international levels. Governments have a critical role to play in fostering an enabling environment for global scientific collaboration. This includes aligning national research infrastructure roadmaps, simplifying regulatory frameworks, supporting staff development and mobility, and committing to long-term investment in globally shared infrastructure.
Global RIs contribute directly to international policy goals, including the United Nations Sustainable Development Goals, by delivering the knowledge and capacity needed to sustain a resilient, habitable planet. Their role in supporting global scientific diplomacy and cooperation is vital in an era of growing geopolitical complexity and environmental urgency.
The partnership among EPOS, AuScope, EarthScope, and GNS-NZ exemplifies a growing global movement to build coordinated, inclusive, and impactful research infrastructures and underlines the need for policymakers to support global scientific collaboration not as a luxury but as a necessity to confront humanity's grand challenges.

Presentations Session 4: Data Stewardship / 72

Establishing a Data Culture using Data Frameworks to Navigate the Waves of Marine Data

Authors: Andrew Conway[1]; David Currie[1]; Eoin O'Grady[1]; Sarah Flynn[1]; Tara Keena[1]

[1] *Marine Institute*

Corresponding Authors: david.currie@marine.ie, eoin.ogrady@marine.ie, sarah.flynn@marine.ie, tara.keena@marine.ie, andrew.conway@marine.ie

Advancing sustainable, high-quality, long-term data stewardship and management is fundamental to ensuring that marine data remains a valuable resource for research, policy, and environmental monitoring. This paper explores the practicalities of establishing organisational policies and methodological approaches that govern how operational and research data are collected, processed, stored, shared, and preserved for long-term sustainability. By engaging a multidisciplinary team of scientists, data managers, and IT specialists, the Marine Institute has demonstrated how structured frameworks evolve to meet the needs of diverse stakeholders while maintaining core principles of high-quality data stewardship; sharing the lessons learned along the way.

Since 2017, the Marine Institute has focused on implementing best practices, organisational strategies, and institutional frameworks that align with the FAIR principles. By adopting internationally recognised accreditation and certification processes - such as the Data Management Quality Management Framework (DM-QMF) (Leadbetter, Carr, et al., 2020) and the CoreTrustSeal Certification - the Institute has established a foundation for responsible data stewardship within a global digital ecosystem.

Good practices in data stewardship require a balance between technical standards and effective collaboration with data producers. The Marine Institute has developed structured workflows to facilitate seamless data deposition, ensuring that researchers and stakeholders can contribute data in a way that meets high-quality standards while also respecting ethical considerations and long-term usability. These practices have been integral to the Marine Spatial Planning (MSP) process in Ireland (Flynn et al., 2020, 2023), where multi-disciplinary and multi-stakeholder data are managed in a way that supports transparency, accessibility, and interoperability.

Long-term data stewardship is not only about maintaining technical infrastructure but also about fostering trust between data providers, users, and governing institutions. Certifications and frameworks provide a strong foundation, but successful implementation relies on the human element - ensuring that data producers are supported with clear guidelines, training, and incentives to deposit their data in a way that meets sustainability and accessibility goals. Expectation management, investment in capacity building, and continuous process improvement are key to advancing data stewardship at an institutional level.

Aligning marine data management with international best practices ensures that oceanographic and environmental data remain usable, interoperable, and impactful for global research efforts. The integration of FAIR principles supports data governance, transparency, and long-term preservation. Sustainable data stewardship also enhances collaborative research on pressing environmental challenges such as climate change, marine biodiversity conservation, and ocean resource management.

Marine data management requires a comprehensive approach that bridges scientific research, technological advancements, and policy frameworks. It involves rigorous processes for data validation, standardisation, and long-term curation, alongside the use of advanced sensing technologies, remote monitoring tools, and scalable data storage platforms. The Marine Institute's commitment to data frameworks and international certifications reflects its ongoing efforts to enhance trust, accessibility, and sustainability in marine data stewardship.

Ultimately, while structured frameworks set the foundation for best practices, it is the commitment to continuous improvement, stakeholder engagement, and investment in skilled personnel that elevates marine data management from a procedural necessity to a strategic advantage for long-term environmental and scientific impact.

Leadbetter, A., Carr, R., Flynn, S., Meaney, W., Moran, S., Bogan, Y., ... & Thomas, R. (2019). Implementation of a Data Management Quality Management Framework at the Marine Institute, Ireland. Earth Science Informatics, 1-13. DOI: 10.1007/s12145-019-00432-w

Flynn, S., Meaney, W., Leadbetter, A., Fisher, J. P. & Nic Aonghusa, C. (2020). Lessons from a Marine Spatial Planning data management process for Ireland, International Journal of Digital Earth, DOI: 10.1080/17538947.2020.1808720

Flynn, S., Tray, E., Woolley, T., Leadbetter, A., Nic Aonghusa, C., … Conway, A. (2023). Management of spatial data integrity including stakeholder feedback in Maritime Spatial Planning, Marine Policy, DOI: 10.1016/j.marpol.2023.105799

Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI / 73

FAIR for Now and into the Future: Building Blocks for Long-Term Data Stewardship in a Shared Data Repository Service

Author: Julie Shi[1]

[1] *Scholars Portal*

Corresponding Author: juli.shi@utoronto.ca

Funding bodies and publishing venues increasingly require researchers to deposit and share their data in order to support rigorous, responsible, and reproducible science. Rising to the occasion, libraries have been expanding their scope to support research data as a scholarly resource and are increasingly recognized as providers of research data management and repository services. These trends are driven by the FAIR principles of findability, accessibility, interoperability, and reusability, which aim to enhance data quality to improve discovery and reuse. Data that is FAIR today may not be FAIR tomorrow, however, and ensuring that research data can be found, accessed, understood, and used into the future requires stable infrastructure and ongoing and informed interventions by the stewards of that data. Preserving research data to support these principles for the long term reveals several challenges, however, including developing policies and procedures, determining the various stakeholders that should be involved and the roles that they have in these processes, considering needs across the heterogenous and evolving landscape of research data (e.g. big data, sensitive data, disciplinary practices and protocols) and, by extension, research communities, as well as managing the associated costs and capacity required for curation, repository management, storage, ongoing preservation, and more.

This presentation will outline efforts at Borealis to develop technical building blocks and foster community-driven initiatives in support of research data preservation in Canada. Borealis, the Canadian Dataverse Repository, is a multi-disciplinary, bilingual, national data repository service provided by Scholars Portal at the University of Toronto Libraries in partnership with regional academic library consortia and the Digital Research Alliance of Canada. With almost 24,000 published datasets, the service supports over 80 institutions and research organizations across Canada, and each institution or research organization manages their own collection and provides local support to their researchers. Borealis is library-led, community-based infrastructure, and the preservation of deposited data is a responsibility shared between Borealis as the repository service and each participating institution as an expert on their institutional context, research communities, and data collections.

To lighten the load for minimum preservation, Borealis has developed technical building blocks for the repository as a whole, including ongoing backup processes and monthly fixity checks for all deposited files, as well as fixity remediation workflows developed in line with best practices. These building blocks are complemented by preservation-friendly features provided in the repository software. Export and integration options are also available for further preservation processing and management beyond Borealis for institutions interested in engaging in additional research data preservation activities. A 2022 survey of institutional administrators further identified support for preservation planning and workflows as a primary gap in the research data services landscape. Responding to these needs, Borealis relaunched a community initiative to update and develop resources and documentation related to policy and service modelling. This building block is based on CoreTrustSeal requirements to directly support participating institutions seeking to apply for CoreTrustSeal certification as well as those interested in using the requirements to benchmark and plan their services. Bridging the technical and community, Borealis is also undertaking a format analysis project to support institutions with file format management and preservation planning.

By providing an overview of the various preservation building blocks developed for Borealis, this presentation will demonstrate the extensive work and coordination by the service provider, collection administrators and depositors required to ensure the long-term stewardship of research data. Collaborating on projects with the community to draw on our collective knowledge and experiences also provides extensive communal benefits for building local capacity and ensuring data can continue to be FAIR into the future.

Presentations Session 2: Data and Research & Data Science and Data Analysis / 74

Accelerating Research with Data Commons and Data Meshes

Author: Robert Grossman[1]

Co-authors: Bernie Pope [2]; Claire Rye [3]; Steven Manos [4]

[1] *University of Chicago*

[2] *University of Melbourne*

[3] *New Zealand eScience Infrastructure*

[4] *Australian BioCommons*

Corresponding Authors: steven@biocommons.org.au, rgrossman1@uchicago.edu, bernie@biocommons.org.au, claire.rye@nesi.org.nz

About Data Commons and Data Meshes

A data commons is a cloud-based software platform with a governance framework that enables a research community to manage, analyze and share its data. A data mesh is a collection of two or more data commons, cloud-based computational resources, and other cloud-based resources that interoperate using a common set of core software services and a hybrid governance model. Data commons and data meshes are two important types of data platforms supporting the research community.

Reason for a Session

Today, there are over 50 data commons and similar platforms around the world and a growing number of data meshes. The primary goal of this session is to bring together IDW/RDA participants interested in data commons and data meshes and in sharing lessons learned about building, operating and using data commons and data meshes. Data commons and data meshes are emerging as important platforms for building, distributing and federating AI models. A secondary goal of the session is to bring together IDW participants interested in building AI models over data commons and data meshes.

Relevance to the conference

Data commons support open reproducible research and provide FAIR access to the data they manage, both of which are important themes for the conference. Data meshes provide a mechanism for interoperating and federated data commons and supporting interdisciplinary research, which are also important themes for the conference. The session will include talks on CAREful Indigenous Data Governance and all of the talks will provide regional and international perspectives.

Structure of Session

Four 15 minute talks about data commons and data meshes (see below). One 30 minute panel discussion with audience participation about using data commons to support reproducible scientific research and discovery.

Four 15 Minute Talks

Robert L. Grossman - From Data Commons to Data Meshes

The talk will provide a brief introduction and overview of data commons and data meshes and how they are beginning to support AI. He will discuss some of the lessons learned over the past decade

that contribute to a successful data commons and data mesh. Finally, he will provide an introduction to some of the approaches that are emerging for integrating AI models and frameworks with data commons and data meshes.

About Robert L. Grossman. Robert Grossman is a Professor of Medicine and Computer Science at the University of Chicago and the Director of the Center for Translational Data Science at the University of Chicago. He is also the Director of the Open Commons Consortium. He is the lead for the open Gen3 data platform which has been used to build over 20 data commons and data meshes.

Claire Rye - Building a genomic data repository for taonga species in Aotearoa New Zealand

The Aotearoa Genomic Data Repository (AGDR) is an Aotearoa-based resource that enables researchers and Māori communities to fulfil their obligations relating to the guardianship, management, sharing and use of genomic data from biological samples that are taonga (precious or treasured). The AGDR was jointly developed by Genomics Aotearoa and the New Zealand eScience Infrastructure, with funding from the Ministry of Business Innovation and Employment. Its design is based on the principles of Māori data sovereignty, enabling kaitiaki (Māori guardianship) centric control over who can access genomic data, and for what purposes. AGDR has been developed based on gen3 data commons and Globus, and is in line with the globally-relevant CARE and FAIR Principles, ensuring data is findable, interoperable, and re-usable in cases where there is prior and informed consent, and access is provided by kaitiaki.

About Claire Rye. Claire Rye is a Product Manager at New Zealand eScience Infrastructure (NeSI) based out of the University of Auckland. She works across the Aotearoa Genomics Data Repository and Rakeiora Pathfinder projects and generally looks at research data management and the data lifecycle across NeSI and nationally as part of the Research Data Culture Conversation.

Steven Manos - Some Best Practices for Building Data Commons: The Australian BioCommons Experience

The talk will provide an overview of how the Australian BioCommons develops and operates data commons and discuss some of the the related services supporting research for data managed by data commons.

About Steven Manos. Steven Manos is the Associate Director of Cyberinfrastructure for the Australian BioCommons. His interests lie in building digital platforms designed specifically for researchers, which includes developing partnerships and community building.

Bernie Pope - establishing data commons for human omics data in Australia

Australian BioCommons is working with several partner organisations in Australia to build human omics data commons tailored to the particular needs of the respective research communities. This work has received national funding from the National Collaborative Research Infrastructure Strategy (NCRIS), supporting the GUARDIANS program, and from the Medical Research Future Fund (MRFF), supporting the Australian Cardiovascular Disease Data Commons (ACDC) and the OMIX3 program. In this presentation we will describe the steps taken to establish these data commons, including critical aspects of data management and governance.

About Bernie Pope. Bernie is a Professor in the Faculty of Medicine, Dentistry and Health Sciences at The University of Melbourne. He works in the fields of computer science and bioinformatics, where he applies scalable and robust computational methodologies to key challenges in human health, particularly in genomics and cancer. He is Associate Director at Australian BioCommons and leads the Human Genome Informatics division.

Presentations Session 1: CAREful Indigenous Data Governance / 76

Enhancing African Data Sovereignty and Representation: The Role of the Africa PID Alliance in Ensuring Ownership and Recognition of African Indigenous Knowledge

Author: Joy Owango[1]

Co-author: Nabil Ksibi [2]

[1] *Training Centre in Communication*

[2] *Africa PID Alliance*

Corresponding Authors: joy.owango@tcc-africa.org, nabi.ksibi@africapidalliance.or

Introduction to the Africa PID Alliance and the Digital Object Container Identifier (DOCiD ™)
The Africa PID Alliance is an Open Infrastructure program of the Training Centre In Communication that is focused on producing African African-originated persistent Identifiers designed to enhance the representation, sovereignty, and visibility of African research output globally. This program aims to bridge the gap in access to Digital Object Identifiers (DOIs) across African institutions, ensuring greater visibility and recognition of African research, including grey literature, indigenous knowledge, cultural heritage, and patents, in global scholarly ecosystems.

Why is this important?
The Lack of digitizing research outputs is a real challenge in Africa. Through the Africa PID Alliance innovative projects, through reliable open research infrastructure services, we will make access to knowledge and metadata about digital objects closer to the wider communities, including indigenous knowledge and patent metadata, starting from Africa. We intend to solve these challenges by producing Persistent, Accessible Affordable Persistent Identifiers (PIDs) in Africa.

The System: Digital Object Container Identifier (DOCiD ™)
The Digital Object Container Identifier (DOCiD ™) is a persistent identifier (PID) developed by the Africa PID Alliance. DOCiD started as a vision to create a unified platform for managing Digital Object Identifiers in Africa, focusing on making research output on indigenous knowledge, cultural heritage and Patents more discoverable and reusable.The system is structured to be multilinear in nature, enabling it to accommodate different types of persistent identifiers such as,CSTR, Handles, DOIs, Ark Keys, ROR and RAID. It seamlessly integrates the entire DOI ecosystem by allowing multiple persistent identifiers to be assigned to various outputs linked to the original object.

DOCiD ™ System Interface is made up of the following
• Resource Types: Crucial information or systems that are focus areas for DOCiD™ .
• Publications: Integration of various categories of African Research Outputs.
• Documents: This includes the integration of and is not limited to datasets, audio, video and images .
• Creators: Integration of tools that help in identifying and Tracking researchers and their academic output
• Organizations: Integration of systems that assign PIDs to organizations involved in research
• Funders: Integration of research funding organizations
• Projects: Integrations of systems that assign ID to research projects

What stage are we in?
DOCiD™ 1.0 is live and we are working with use cases in the following personas : **1)Researcher,2) Librarian,3)Museum Collections Manager/Curator** from the following respective institutes, Jomo Kenyatta University of Agriculture and Technology, Thomas Sankara University, Burkina Faso, North West Maphikeng University South Africa, National Museum of Tanzania and Kigulu Cultural Centre, Uganda.

Conclusion
Africa PID Alliance is transforming how African knowledge is owned and represented globally by enhancing research visibility, promoting data sovereignty, and securing Africa's intellectual assets.
Call to Action: Stakeholders are invited to join the initiative and prioritize African data sovereignty.
https://africapidalliance.org/sign-apa/
For more information visit
docid.africapidalliance.org, africapidalliance.org

Poster Session / 80

Practices and perceptions in research data management: a cross-sectional study based in a Brazilian university hospital

Authors: Daniel Umpierre[1]; ROBERTO DA SILVA[2]

Co-authors: Cibeli Fernandes [1]; Evelin Michels [1]; Jessica Carvalho ; Raphaella Scherer [3]; Vania Hirakata [2]

[1] *Federal University of Rio Grande do Sul*

[2] *Hospital de Clínicas de Porto Alegre*

[3] *Universidade Luterana do Brasil*

Corresponding Authors: vhirakata@hcpa.edu.br, danielumpierre@hcpa.edu.br, jpscarvalho@hcpa.edu.br, cofernandes@hcpa.edu.br, pietracarvalho.silva2@gmail.com, rpscherer@hcpa.edu.br, robertosilva@hcpa.edu.br

Introduction: Proper data management is essential to ensure the integrity, transparency, and reuse of data in scientific research. **Objectives:** This study aimed to investigate the practices and perceptions related to data management among researchers at a university hospital in southern Brazil. **Methods:** This was a cross-sectional, exploratory study. Consecutive sampling was used to invite researchers with at least one active human research project. Data collection was conducted between April and November 2024 through an electronic questionnaire. The questionnaire was divided into two sections: Practices (37 items across six dimensions) and Perception (10 items). Document analysis (14 items) was carried out through duplicate review of research protocols. Additionally, a project structure assessment was performed in the REDCap platform (11 items). **Results:** Of the 375 researchers invited, 184 agreed to participate; 166 responded to the section on data management practices, and 160 completed the entire questionnaire, answering both the practices and perception sections. The analysis of practices revealed significant use of spreadsheets (62.7%), electronic forms (78.9%), and paper forms (43.4%). The main tools reported for data entry included platforms such as Google (68.7%), MS Excel (57.2%), and REDCap (39.8%). Although 69.9% of participants reported preserving their research data, only 10.2% planned to use public repositories. Researchers'perceptions were positive in 9 out of 10 items on a 5-point scale. However, non-response rates for items such as interoperability (13%) and data transparency and accessibility (7%) highlight gaps that need to be better understood. Additionally, 544 research projects registered at the institution were analyzed through document review, of which 103 (17.6%) had databases registered on the REDCap platform. The results revealed shortcomings in data management planning, with 52.8% of projects not specifying tools for this purpose, as well as variations in storage and security practices. Despite the growing use of platforms such as REDCap, many researchers still lack adequate training. **Conclusion:** It is concluded that strengthening institutional policies and investing in researcher training can improve research quality and integrity, promoting a more transparent and collaborative scientific culture.

Presentations Session 1: CAREful Indigenous Data Governance / 81

Open science and the management of traditional and scientific knowledge: A case study of the Takinahakỹ Center for Indigenous Higher Education at the Federal University of Goiás, Brazil.

Authors: Carlos Abs da Cruz Bianchi[1]; Cassia Oliveira[1]; Geisa Müller de Campos Ribeiro[1]; Larissa Bárbara Borges Drumond[1]; Laura Rezende[1]; Maria das Graças Monteiro Castro[1]

[1] *Federal University of Goias (UFG)*

Corresponding Authors: laura_rezende@ufg.br, larissa.barbara@ufg.br, gracamcastro@ufg.br, cassiaoliveira@ufg.br, geisamuller@ufg.br, cbianchi@ufg.br

This study presents the first findings of the project: **Open science and the management of traditional and scientific knowledge: A case study of the Takinahakỹ Center for Indigenous Higher Education at the Federal University of Goiás, Brazil.** The main aim is to improve the capacity of indigenous students to effectively manage and safeguard the records of traditional and

scientific knowledge they generate during their work on the undergraduate degree course in Intercultural Education at the Takinahakỹ Center for Indigenous Higher Education at the Federal University of Goiás (UFG).
Specifically, the aim is to:
• Propose strategies that encourage indigenous students to design protocols for the management of their data and information guided by CARE principles in an integrated manner;
• Enable the use of innovative technologies related to data sharing and traceability within the Research Data Repository of the Federal University of Goiás (UFG);
• To build a conceptual modeling and technological implementation of the collection of records generated within the scope of the Takinahakỹ Center of UFG considering the FAIR principles and replicability requirements.

In the degree course of Intercultural Education at UFG, the research carried out by indigenous students during the course begins in the basic training matrix (pedagogical and transdisciplinary principles of intercultural education) and continues in the specific matrices (Natural Sciences and Mathematics, Cultural Sciences and Language Sciences). Their purpose is to support dialogue between the specific knowledge produced by indigenous groups and the so-called scientific or universal knowledge, thus favoring the realization, in practice, of transdisciplinarity and interculturality. This scientific practice also supports language policies, the struggle for citizenship, political participation in various intercultural contexts, the production of teaching materials and the construction of pedagogical projects and school management, as well as encouraging the entry of indigenous intellectuals into the scientific scene.
The thematic axes of the course to be worked on will be guided by five lines of research, namely:
• Indigenous Education and School Education;
• Environment and Self-Sustainability;
• Language Policies and Bilingual Education;
• Art, Tradition and the Market;
• Indigenous Policies, Interculturality and Indigenous Movements.

At the end of the course, each "new"indigenous teacher, based on their research and the area they have chosen to specialize in, will present an alternative project to improve life for their community. This is not a monograph, but an extension project aimed at quality teaching in indigenous schools, linked to the projects of the communities in which they live. There are two presentations of this final project: one in the University and another for their own group, with the elder evaluating according to their own approach.
The proper management of traditional and scientific knowledge generated by indigenous students depends, first of all on the terms agreed between the interested parties (researchers/students, scientific institutions, government, project team, among others). The training of the indigenous students is essential in order to align the understanding of the decisions to be taken and the modus operandi of the entire process of registering the knowledge. These terms between the parties, also known as protocols, must include:
• previously established conditions for handling information/material during the project;
• appropriate data management approach;
• local governance structures to support the project;
• consensus on the implementation of the project;
• Terms of Consent previously established and clarified from the students considering the context and self-designation;
• Gender issues as a priority for women's needs and opportunities.

In order to guide the drafting of agreements and protocols, which can bring benefits and generate ethical and responsible partnerships in the context of scientific research carried out by indigenous students, the CARE principles provide a set of guidelines to be considered so that these actors effectively hold the governance of their data, basically establishing data standards on traditional and scientific knowledge; relationships and research practices defined in detail.
This is a mixed-methods study, which will rely on initial quantitative surveys and a qualitative approach. The methodology for implementing the Center´s collection in the research data repository follows the bottom-up approach with multiple case studies, considering that the students come from more than 30 different ethnic groups. Besides that, a multidisciplinary working group involving researchers and students will be responsible for carrying out the proof of concept for the collection of traditional and scientific knowledge in UFG's digital research data repository, that uses dataverse software.
We are planning to use TK labels (Traditional Knowledge labels) and BC labels (Biocultural labels),

that are digital labels developed through definitions established by the indigenous communities that hold their data and are been used in various countries, enabling them to publicize local and specific conditions for sharing data and information, as well as detailing how involvement in future research and relationships should be consistent with rules, governance and community protocols signed for the use, sharing and circulation of knowledge and data.
This is an innovative and pioneering effort in the state of Goiás and the central-western Brazilian region, as it aims to systematize the traditional and scientific knowledge generated by indigenous students in order to strengthen their groups, generating a secure and reliable base of traditional and scientific knowledge considering previously defined protocols (agreements) guided by the CARE and FAIR principles, which will enable the construction of the conceptual modeling and technological implementation of the collection of records generated within the scope of the center that can be replicated in other experiences of this type of knowledge registry.

Presentations Session 9: Empowering the global data community for impact, equity, and inclusion / Education / 84

Ten Simple Rules for Researchers Training the Rapidly Evolving Workforce

Authors: Amany Gouda-Vossos[1]; Meirian Lovelace-Tozer[2]

Co-authors: Adeline Wong [1]; Ellen Lyrtzis [2]; Kathryn Greenhill [1]; Kathryn Unsworth [2]; Liz Stokes [1]; Rob Clemens [1]

[1] *Australian Research Data Commons*

[2] *Australian Research Data Commons (ARDC)*

Corresponding Authors: amany.gouda-vossos@ardc.edu.au, ellen.lyrtzis@ardc.edu.au, liz.stokes@ardc.edu.au, meirian.lovelace-tozer@ardc.edu.au, adeline.wong@ardc.edu.au, kit.greenhill@ardc.edu.au, kathryn.unsworth@ardc.edu.au, rob.clemens@ardc.edu.au

We are entering the fourth research paradigm following the digital revolution, which is evidenced by rapid advancements in the scientific methodology of data-intensive practices. Upskilling the next generation of the research workforce is pivotal. Further, discipline-specific advancements highlight the importance of researchers acquiring new skills to meet evolving demands.

Researchers are likely to assist others in learning new skills, from computational tools to different data analysis methods. This knowledge transfer often happens organically and informally through mentorship and collegiality, but a more structured approach would be vastly beneficial and impactful to those needing to upskill. When organising a skills event for others, researchers may require guidance on where to begin.

Short-format training (SFT) provides a quick and efficient medium to address needs and gaps in data skills, especially for the modern workforce. An SFT refers to a non-formal workshop, short course, boot camp, or similar, that teaches skills and knowledge over a brief period. A study led by Williams (2023) called to make SFT more reliable, effective, inclusive, and career-spanning in the face of rapid technological changes.

Recommendations for effective SFT development were inspired by productive discussions at the Australian Research Data Commons (ARDC) Digital Research Skills Summit. The Skills Summit brought together researchers, learning designers, skills trainers, and librarians. Attendees collaboratively formulated innovative, effective, and transferable strategies to increase data literacy in researchers, applicable to all research disciplines. These recommendations were curated into ten simple rules by the Skilled Workforce Development Team at ARDC. For the full list visit: https://eresear.ch/10sr.

Our ten simple rules provide a streamlined workflow to assist in developing SFT for researchers who train the research workforce. These rules outline how to think about skills learning for researchers, plan training sessions, and efficiently maximise learning. We offer recommendations on how to design and develop learner-centered training programs, foster outreach, and connect with trainer communities. We then provide tips to manage and optimise training, and conclude with valuable

insights on preparing for uncertainty and the importance of post-training operations and continued learning.

85

Building data platforms to reduce inequities

Authors: Aiden Price[1]; Aswi Aswi[2]; Daminda Solangaarachchi[3]; Helen Thompson[1]; Jessica Cameron[4]; Susanna Cramb[1]

[1] *Queensland University of Technology*

[2] *Universitas negeri Makassar*

[3] *Australian Institute of Health and Welfare*

[4] *Cancer Council Queensland*

Corresponding Authors: aswi@unm.ac.id, daminda.solangaarachchi@aihw.gov.au, jessicacameron@cancerqld.org.au, susanna.cramb@qut.edu.au, helen.thompson@qut.edu.au, a11.price@qut.edu.au

Increasing amounts of data are routinely collected by governments, but providing insights from this data is often challenging. This can be due to restrictions in accessing the data and/ or data sparsity.

This session showcases innovative, interactive health and environmental data platforms that are providing inequity-focused data insights, ranging from platforms generated using low-cost solutions like R Shiny to more expensive, bespoke platforms. Some focus on visualising raw data; others on providing reliable estimates through advanced modelling. Most are freely available online, and Indigenous-specific data platforms are featured.

The session format will comprise five speakers from a range of government, non-government and academic institutions, each presenting for 10 minutes, followed by a 30 minute speaker panel session, inviting audience interaction.

Outcomes include a greater awareness and understanding of:
• Available data platforms
• Options and capabilities when building data platforms in different settings
• Protecting sensitive data when releasing estimates
• Using data platforms in decision making
• Communicating data insights to decision makers for impact.

The speakers and moderator are:

Dr Jessica Cameron, Senior Research Fellow and Group Lead of Understanding Cancer Inequalities, Cancer Council Queensland, Australia

'**The Australian Cancer Atlas and geographic inequalities in cancer**'
• The Australian Cancer Atlas is an online, freely available digital platform that visualises spatial differences in cancer outcomes.
• With ethics approvals and in collaboration with Data Custodians, statistical modelling enables the release of estimates of cancer disparities at a fine geographical granularity, making information accessible to the public without risking the release of sensitive health information, using the FAIR principles.
• Visualisations and clear communication make complex statistical and epidemiological concepts accessible to a broad audience, empowering diverse users to employ Atlas estimates.
• This has resulted in a range of stakeholders using the Atlas estimates to make data-driven positive change to clinical practice, government policy, service provision and community engagement.
• The Atlas also provides an infrastructure to support further research.

A/Prof Aswi, Head of Master's Program in Statistics, Statistics Study Program, Universitas negeri Makassar, Indonesia

'**From Model to Map: R Shiny Platforms for Identifying Health Disparities through Spatial Disease Modelling**'
• A user-friendly R Shiny web application was developed, integrating the shiny and CARBayes packages to implement multiple Bayesian spatial Conditional Autoregressive (CAR) models, making them accessible to users without advanced statistical programming skills.
• Users can input count data, population, and covariates. The app fits mutliple CAR models using different hyperpriors, evaluates model fit, checks convergence, and helps users choose the most appropriate model.
• The core output is an interactive thematic map displaying relative risk across regions, allowing easy identification of high-risk areas and supporting targeted public health interventions.
• The tool has been applied to Indonesian COVID-19 and stunting data, highlighting its flexibility in different epidemiological contexts.

Dr Aiden Price, Senior Research Associate, Centre for Data Science, School of Mathematics, Queensland University of Technology, Australia
Environmental Health Domain Specialist, Australian Urban Research Infrastructure Network (AURIN)

'**Extracting insights from data through the Australian Environmental Health (AusEnHealth) project**'
• The AusEnHealth Project creates accessible, interpretable indicators of environmental health vulnerability across Australia, covering heat, cold, and air pollution.
• These indices offer rapid insight into complex data, helping users explore patterns without needing advanced technical expertise.
• By using open data and producing publicly available and FAIR indicators and indices across all of Australia, AusEnHealth improves equitable access to decision-relevant environmental health information.
• The platform lowers technical barriers through open-source infrastructure and visual tools designed for government, research, and community use.

Dr Daminda Solangaarachchi, Senior Project Manager, Regional Insights Unit, First Nations Health and Welfare Group, Australian Institute of Health and Welfare

'**Regional Insights for Indigenous Communities (RIFIC): A platform for Indigenous data that empowers decision-making**'
• RIFIC is an online data visualisation platform developed by the Australian Institute of Health and Welfare. It serves as a one-stop shop for data and statistics about First Nations Australians, focusing on their health and wellbeing.
• Data is available at the lowest possible geographic level, and varies by underlying source.
• The prototype website was presented to government and in three workshops seeking feedback from key stakeholders, including representatives from Primary Health Networks, Aboriginal Community Controlled Health Organisations, and Empowered Communities.
• RIFIC is a powerful tool for decision-making.

A/Prof Susanna Cramb, Principal Research Fellow, Australian Centre for Health Services Innovation, School of Public Health and Social Work, Queensland University of Technology, Australia
Biostatistician, Jamieson Trauma Institute, Metro North Health, Australia

'**Identifying inequities in trauma care: the Queensland Injury Atlas**'
• The Queensland Injury Atlas aims to comprehensively visualise injury cases and costs across Queensland for clinical and government stakeholders.
• Underpinned by large linked datasets: all Queensland hospital injury admissions were linked to Queensland ambulance, emergency department and insurance compensation scheme data.
• Spatial data included patient residential area, health facility locations, and ambulance pick-up geocoordinates.
• An optimized database schema was developed and rigorously tested for performance, scalability, and reliability. Visualisations used HTML5, CSS3, and JavaScript frameworks like React.
• Information on injuries can be filtered by diagnosis or procedure codes, region, facilities, time point, or patient age, down to the individual patient journey, and is visualised through maps and graphs.
• This resource can aid in planning, auditing care and preventing injuries.

Session chair and panel moderator: **A/Prof Helen Thompson**, Associate Professor of Statistics, Centre for Data Science, School of Mathematics, Queensland University of Technology, Australia.

86

Approaches to Indigenous Data Governance in the HASS and Indigenous Research Data Commons

Author: Gavin Stanbrook[1]

[1] *ARDC*

In 2020, the HASS and Indigenous Research Data Commons (HASS and Indigenous RDC) was established to create a comprehensive digital HASS and Indigenous research infrastructure capability as part of the Australian Research Data Commons (ARDC). The HASS and Indigenous RDC is developing infrastructure across a range of focus areas to serve the needs of diverse HASS and Indigenous research communities with shared approaches to common needs.

All research disciplines served by the HASS and Indigenous RDC use Indigenous data, making Indigenous Data Governance a central shared consideration across the entire RDC. Indigenous data governance refers to the right of Indigenous peoples to autonomously decide what, how and why Indigenous Data are collected, accessed and used. The HASS and I RDC has utilised strength-based practices by centring Indigenous leadership, worldviews, knowledge, and aspirations. Through a critical analysis of current practices and national research infrastructure priorities, it has highlighted the importance of Indigenous-led approaches to implementing and Indigenous Data Governance.

The HASS and Indigenous RDC approach to Indigenous Data Governance centers around five guidelines:

1. Recognising the data assets owned by Indigenous people and entities and observing their legal rights to use, share, store and report such data and comply with their decisions while also supporting their status as data custodians.
2. Partnering with Aboriginal and Torres Strait Islander people: Ensuring collective benefit, meaningful representation of, and collaboration with, Indigenous communities, respecting their perspectives and complying with their rights.
3. Building and enabling data-related capabilities: Empowering Indigenous data custodians and stakeholders with the skills and resources necessary for effective data management and utilisation, and especially data ecosystem maturity including cybersecurity measures to enable their qualification for data access. Building and enabling data-related capabilities: Empowering Indigenous data custodians and stakeholders with the skills and resources necessary for effective data management and utilisation, and especially data ecosystem maturity including cybersecurity measures to enable their qualification for data access.
4. Providing knowledge, accessibility, and usability of data assets: Ensuring Indigenous communities have access to and benefit from data resources, emphasising transparency and inclusivity.
5. Building an inclusive data ecology: Fostering collaboration and inclusivity within the research data landscape to promote greater accessibility, collective growth and understanding.

This panel will bring together Indigenous team members from across the different focus areas that make up the HASS and Indigenous RDC, as well as the ARDC's Program Manager, Indigenous Data Governance and Indigenous Intern. Together they will discuss how each focus area is addressing Indigenous Data Governance, and the importance of embedding Indigenous Data Governance into the design and provision of research data infrastructure.

Format:
60 minutes structured panel discussion
30 minutes panel Q&A with the audience

Speakers:
(note: speakers to be confirmed closer to the date)

Distinguished Professor Marcia Langton (University of Melbourne) is Project Lead of the Improving Indigenous Research Capability focus area. This focus area is working to enable Aboriginal and Torres Strait Islander peoples and researchers at the interface of research data science and Indigenous knowledge systems to have access to effective research data tools.

Robert McLellan (University of Queensland) is Program Manager of the Language Data Commons of Australia focus area. This focus area is working to secure vulnerable and dispersed language collections of written, spoken, multimodal, and signed text, and to link these with improved analysis environments for new research outcomes.

Neenah Gray (UNSW Sydney) is Indigenous Research Fellow in the Australian Creative Histories and Futures focus area. This focus area will improve interoperability of cultural data sets about the arts consistent with FAIR and CARE principles.

Dr Danielle Armour (University of Queensland) is an Activity Lead in the Social Sciences Research Infrastructure Network focus area. This focus area aims to improve discoverability, accessibility and usability of social science data (particularly linked government data assets).

Gavin Stanbrook is Program Manager (Indigenous Data Governance) at the Australian Research Data Commons.

Casey Haseloff is Indigenous Intern at the Australian Research Data Commons, located at the Indigenous Data Network at the University of Melbourne.

Moderator: Grant Sarra has spent over 40 years working hands-on with clients from many different urban, rural and remote Aboriginal and Torres Strait Islander communities. Grant uses his own experiences as an Aboriginal man to anchor meaningful conversations, foster respectful relationships and create accountability as individuals.

Poster Session / 88

Towards Integrated Monitoring of Antimicrobial Resistance and Usage in horticulture, water, and wine sectors in Australia

Authors: Cherry Green[1]; Noorul Amin[1]; Ricardo Soares Magalhaes[1]; Sahil Arora[1]; Tatiana Proboste Ibertti[1]

[1] *The University of Queensland*

Corresponding Authors: cherry.green@uq.edu.au, r.magalhaes@uq.edu.au, sahil.arora@uq.edu.au, noorul.amin@uq.edu.au, t.probosteibertti@uq.edu.au

Background: Antimicrobial resistance (AMR) is a growing concern in agribusiness sectors with serious consequences to productivity and public health. A data centric approach is needed to support Australian agribusinesses and water sectors to understand the impact of antimicrobial usage on the emergence of resistance for diseases that farmers are faced on a daily basis. The SAAFE CRC Analytics Program has partnered with the Australian Research Data Commons (ARDC) to co-design the SAAFE Data Code that promotes responsible and ethical data sharing across sectors and SAAFE CRC project partners in managing the AMR data. Furthermore, to alleviate the problem of siloed data, we developed SAAFE Data Dictionary, an ontology-based and relational schema-based framework to standardise and enable FAIR (Findable, Accessible, Interoperable, Reusable) data practices for AMR data across the agribusiness and water sector in Australia.

Method: This study will lay the foundation for data centric approach to AMR monitoring in the agribusiness and water sectors and has the following distinct yet connected milestones:

- SAAFE Data Code (governance): To co-design a framework for responsible and trusted data access between sectors, project partners, key stakeholders and collaborators. The Code promotes and encourages best practices on data sharing and accessibility to ensure emerging challenges,

such as AMR data, are managed in an ethical manner. In this regard, we have conducted workshops with partnered sectors based on specific "what-if"scenarios and have elicited from the participants the risks involved and then the guiding principles to mitigate those risks. The risks and guiding principles covered, existing data, future and new data that will be generated, and cross-sectoral data.

- SAAFE Data Dictionary: To address heterogeneity and promote FAIR (Findable, Accessible, Interoperable, and Reusable) data principles of antimicrobials in Agribusiness across Australia, we developed a standardised ontology-based framework. In the first phase, we compiled a dictionary of domain-specific terms related to antimicrobials, pathogens, sample types, and more, across concerned sectors such as water, wine, and horticulture. These terms were sourced from authoritative international bodies like the FAO and WOAH, which monitor antimicrobial usage and resistance globally. Building on this, we designed a relational data schema to capture laboratory test results, focusing on four key data types: antimicrobial usage, AST testing, antimicrobial residue, and minimum inhibitory concentration (MIC). This schema supports standardised recording and interoperability of laboratory findings across organisations. Finally, we developed a user-friendly standardised data submission framework that allows partner sectors to upload their datasets and respond to a guided questionnaire. This input is automatically mapped through a backend ontology, enabling consistent data integration across the horticulture, water, and wine sectors.

Results: Overall, the SAAFE Data code and Data dictionary provide for ethical, interoperable and standardised data practices, and strengthen cross-sectoral collaboration in order to establish effective surveillance integrated AMR monitoring in agribusiness and water sectors in Australia. We will show the SAAFE data Code and the working AMR standardisation framework for the horticulture, water and wine sectors.

Presentations Session 2: Data and Research & Data Science and Data Analysis / 89

The Australian Reference Genome Atlas: supercharged exploratory infrastructure for national-scale genomic data discovery

Authors: Kathryn Hall[1]; Jack Brinkman[1]; Keeva Connolly[2]; Christopher Mangion[1]; Winnie Mok[None]; Goran Sterjov[1]

Co-authors: Matt Andrews [1]; Peter Brenton [1]; Simon Checksfield [1]; Jeff Christiansen [2]; Nick dos Remedios [1]; Hamish Holewa [3]; Yasmina Kankanamge [1]; Vikas Nagaraju [1]; Lars Nauheimer [1]; Caitlin Ramsay [1]; Sarah Richmond [4]; Nigel Ward [2]

[1] *Atlas of Living Australia, CSIRO*

[2] *Australian BioCommons*

[3] *Australian Research Data Commons (ARDC)*

[4] *Bioplatforms Australia*

Corresponding Authors: kathryn.hall@csiro.au, matt.andrews@csiro.au, nick.dosremedios@csiro.au, lars.nauheimer@csiro.au, jack.brinkman@csiro.au, srichmond@bioplatforms.com, keeva.connolly@qcif.edu.au, nigel@biocommons.org.au, goran.sterjov@csiro.au, christopher.mangion@csiro.au, mok@biocommons.org.au, yasima.kankanamge@csiro.au, peter.brenton@csiro.au, hamish.holewa@ardc.edu.au, jeff@biocommons.org.au

The Australian Reference Genome Atlas (ARGA) is a next-generation platform designed to index, connect and expose genomic data for Australia's mega-biodiversity. It sits at the interface between the twin problems of genomic data discovery in an age where data are rapidly proliferating, and the crisis in documenting and understanding Australia's vulnerable biodiversity. More than 80% of Australia's estimated 500,000 species are endemic, including diverse lineages of marsupials, reptiles, flowering plants, fungi, and marine invertebrates. This exceptional biodiversity presents both an opportunity and a challenge for genomic science: the need to coordinate, contextualise, and make accessible a growing body of data across highly diverse taxa and ecosystems.

Despite the rapid proliferation of genomics datasets, researchers face persistent obstacles to discovery and reuse. Genomic data are scattered across disconnected repositories, stored under inconsistent taxonomies, and often lack sufficient provenance metadata to support informed reuse. No single source captures the full range of sequence types, methods, and specimen contexts relevant to a given taxon. This fragmentation significantly hampers the biosciences community's ability to conduct comprehensive, comparative, or ecologically contextualised research.

ARGA provides a concrete response to these challenges, combining rigorous data stewardship with practical infrastructure for researchers. For example, a conservation biologist studying threatened plants can use ARGA to locate and compare available genome assemblies for Critically Endangered plant species, trace sample provenance from herbarium vouchers through to public sequence repositories, and identify key taxonomic gaps where genomic data are still lacking. As a national infrastructure platform, ARGA has been purpose-built to bring FAIR and TRUST principles to life, making genomic data for Australian biodiversity taxa not only findable, accessible, interoperable, and reusable, but also transparent, contextualised, and trusted.

With the foundations now in place following an ambitious two-year development pilot, ARGA is ready to be commended to the scientific community for integration into research workflows. It represents the culmination of technical innovation, collaborative platform design, and principled data stewardship. Developed by the Atlas of Living Australia, Bioplatforms Australia, and the Australian BioCommons, with investment from the Australian Research Data Commons, ARGA offers researchers a unified platform (https://app.arga.org.au) for exploring genome assemblies, annotations, barcodes and marker sequences, and linked specimen metadata, contextualised through taxonomic, geographic, and traits filters.

ARGA's architecture harmonises Darwin Core standards with MIxS checklists (Genomic Standards Consortium) via a custom event model that traces the provenance of genomic data derived from biological samples. At the core of ARGA is a belief in transparent infrastructure: every datum indexed in ARGA is traceable to its source. A specimen-to-sequence timeline allows researchers to interrogate data quality, completeness, and methodology.

ARGA's technical architecture is purposefully lightweight and independent. A React-based frontend supports intuitive exploration of taxonomically indexed data, while a GraphQL layer provides fine-grained control over queries. Underneath, PostgreSQL serves as the backbone for structured metadata, supported by a custom Rust-based resolver layer optimised for speed and stability. Harmonisation of data across external sources (including NCBI GenBank, Barcode of Life Data Systems, and Bioplatforms Australia Data Portal) is achieved through ingest pipelines that map records to a shared, extensible event model aligned with Darwin Core concepts. These architectural choices are deliberate: to ensure flexibility, and to support an Open Source, Open Science ecosystem. All code is maintained in public repositories under a copyleft licence, and the platform is structured to support community reuse, extension, and review.

ARGA was co-designed with users to prioritise clarity over complexity. Researchers can navigate by systematic groupings, explore ecological traits, and expose under-sequenced lineages. Key features of the ARGA platform include:
- rich metadata and visualisations of **genomic data** for species, with integrated download functionality and evidenced taxonomic histories;
- **taxon dashboards** showing genome coverage, gaps, and sequencing progress by systematic rank;
- **specimen-to-sequence timelines** that visualise provenance from original collection to data reuse;
- **trait-based filtering** for ecological and management attributes (*e.g.* bushfire vulnerability, invasive species);
- **curated species lists** drawn from authoritative sources to guide strategic data use;
- **linked specimen metadata** from museums, herbaria, and biobanks;
- **persistent identifiers and transparent mappings** to support reproducibility and trust.

FAIR and TRUST principles are foundational. From transparent mappings of openly available vocabularies to fully citable and reproducible data downloads, ARGA is engineered to be not just functional, but credible. It is a place where the absence of data is as visible as its presence — where researchers can engage critically with the structure, lineage, and limitations of the data they use.

SciDataCon 2025 at International Data Week marks the full product launch of ARGA. Here we demonstrate the platform's functionality, share technical and governance lessons, and discuss fu-

ture product direction and planned integrations. We will showcase key product features, including Genome Tracker, a newly developed visual tool to assess genomic coverage across Australia's biota, which we see as having utility as a strategic planning aid and gap analysis tool for both research and policy sectors. Tools like Genome Tracker have been made possible through key data architecture decisions made early in the conceptualisation of ARGA, and demonstrate the breadth of data insights and dividends that can be actualised from a core commitment to data provenance principles.

Presentations Session 6: The Transformative Role of Data in SDGs and Disaster Resilience / 90

Marine knowledge value chain: How the European Marine Observation and Data Network supports international marine policy against marine pollution.

Author: Chiara Altobelli[1]

Co-authors: Alessandra Giorgetti [1]; Ann Kristin Østrem [2]; Athanasia Iona [3]; Charles Troupin [4]; Dick M.A. Schaap [5]; Hans M. Jensen [6]; Julie Gatti [7]; Karin Wesslander [8]; Lotta Fyrberg [8]; Luminita Buga [9]; Marilena Tsompanou [3]; Martin M. Larsen [10]; Neil Holdsworth [6]; Reiner Schlitzer [11]

[1] *National Institute of Oceanography and Applied Geophysics - OGS, Trieste, Italy*

[2] *Institute of Marine Research - IMR, Norway*

[3] *Hellenic Centre for Marine Research/Hellenic National Oceanographic Data Centre - HCMR/HNODC, Greece*

[4] *University of Liège (ULiege), Belgium*

[5] *MARIS, Nootdorp, Netherlands*

[6] *International Council for the Exploration of the Sea (ICES), Denmark*

[7] *French Research Institute for Exploitation of the Sea - IFREMER, France*

[8] *Swedish Meteorological and Hydrological Institute - SMHI, Sweden*

[9] *National Institute for Marine Research and Development "Grigore Antipa" - NIMRD, Romania*

[10] *Aarhus University, Danish Centre For Environment And Energy, Denmark*

[11] *Alfred Wegener Institute - AWI, Germany*

Corresponding Authors: lbuga@alpha.rmri.ro, dick@maris.nl, neilh@ices.dk, ctroupin@uliege.be, caltobelli@ogs.it, karin.wesslander@smhi.se, reiner.schlitzer@awi.de, lotta.fyrberg@smhi.se, ann.kristin.ostrem@hi.no, sissy@hnodc.hcmr.gr, mml@ecos.au.dk, julie.gatti@ifremer.fr, mtsompanou@hcmr.gr, agiorgetti@ogs.it, hans.jensen@ices.dk

This contribution gives an overview of the European Marine Observation and Data Network (EMODnet) and focuses on the use cases of EMODnet managed data on marine pollution.
EMODnet has been funded by the European Commission for more than 15 years and is a trusted source for marine in-situ data and data products in Europe. It is a data infrastructure supporting data-intensive research and evidence-based policy making. Its data and products are accessible through a single entry point: the EMODnet portal at emodnet.ec.europa.eu. EMODnet is highly multidisciplinary and covers the entire marine environment for seven thematic areas: bathymetry, geology, physics, chemistry, biology, seabed habitats, and human activities. Over the years, each thematic consortium has succeeded in building a specific marine knowledge value chain. The starting point is a network of data providers who supply data to national data aggregators. They are responsible for aggregating and harmonising data and metadata to make them Findable, Accessible, Interoperable, Reusable, and reproducible (FAIR) according to European and global vocabularies and standards. Furthermore, EMODnet fosters international partnerships to support the interoperability of marine data in Europe and beyond.
The final link in this value chain is the extensive EMODnet user community, which includes national authorities and administrations, academia, non-governmental organisations and companies.
EMODnet Chemistry is the thematic consortium that supports the development of evidence-based knowledge on eutrophication, ocean acidification, and pollutants, including marine litter. Its work draws on the experience of a network of 66 organisations in 32 countries. Most are part of the UNESCO/IOC/IODE network of National Oceanographic Data Centres (NODCs) and a growing number have been officially recognised by the UNESCO-IODE Committee as accredited NODCs or are ISO-certified.

The EMODnet Chemistry value chain relies on SeaDataNet: the pan-European marine data management infrastructure with 110 organizations (NODCs, marine research institutes and international bodies) that has developed consolidated services, standards, and best practices. EMODnet Chemistry has harvested nearly 1.3 million metadata records and associated data from more than 500 different data providers.
EMODnet Chemistry has accumulated dozens of success stories across various types of data providers and users. For example, regarding data users, the European Environment Agency, the European Commission's Joint Research Centre, and most regional sea conventions in Europe have used EMODnet's chemical data to implement the European Union's marine framework policies. Researchers and Copernicus Marine Service use these data to develop tools, data products, and models for assessing environmental status and trends. Recently, partners in the Horizon Europe Blue Cloud 2026 project, which supports the implementation of the European Open Science Cloud, used EMODnet Chemistry data together with data from Copernicus Marine Service and the World Ocean Database (WOD). The goal is to develop a toolbox for creating customizable, validated datasets on key ocean variables of eutrophication and assessing the consistency of the information. Finally, EMODnet together with Copernicus Marine Service will form the data backbone for the European digital twin of the ocean. Although EMODnet focuses on European data sources, it is increasingly contributing to global data systems, for example to share data on marine litter and ocean acidification.
Finally, a call to action: As recent European and global environmental reporting shows we can no longer ignore the transformative role of data for the UN sustainable development goals and disaster resilience. Now is the time for all countries to act quickly and collectively to strengthen the weakest link in the marine knowledge value chain –from the local to the global level –so that evidence-based adaptation and mitigation measures can be defined and implemented to tackle the looming environmental crisis.

91

Open science actions toward achieving the SDGs: an infrastructure dialogue with the Global South

Authors: Jianhui Li[1]; Tshiamo Motshegwa[2]; LILI ZHANG[3]; Francis P. Crawley[4]; Simon Hodson[5]

[1] *Nanjing U*

[2] *African Open Science Platform*

[3] *COMPUTER NETWORK INFORMATION CENTER, CAS*

[4] *CODATA IDPC*

[5] *CODATA*

Corresponding Authors: simon@codata.org, fpc@gcpalliance.org, motshegwat@ub.ac.bw, lijh@cnic.cn, zhll@cnic.cn

The UN 2030 Agenda has become a cornerstone of global scientific and policy efforts, calling for urgent, collective action to address poverty, inequality, environmental degradation, climate change, and other systemic challenges. The Global South is at the frontline of the risks and the opportunities inherent in the SDG framework. These regions face the most acute development challenges while also being home to a wealth of untapped knowledge systems, scientific talent, and emerging infrastructures.
This workshop addresses this question by fostering a critical dialogue on how trusted, cross-national, and cross-regional e-infrastructures can drive SDG-oriented science and innovation. The session responds directly to the IDW 2025 theme 'Trusted Research Data Driving Transformation'by focusing on the enabling role of open science infrastructures in supporting development goals. It also supports the implementation of the UNESCO Recommendation on Open Science. It aligns with priorities articulated in the UN Pact for the Future, emphasizing the use of digital infrastructure to accelerate global cooperation in scientific knowledge sharing.
The GOSC initiative is an international effort to build interconnected, interoperable science clouds. These infrastructures aim to provide reliable, open, and secure access to essential research tools and data, particularly by strengthening scientific collaboration between the Global South and the broader international research community. GOSC is supported by its CSTCloud, AAI, and CSTNet backbone, which enable secure, federated, and policy-compliant access across institutions and borders.

The session is grounded in practical engagement with global partners, translating policy into research infrastructure development. High-level GOSC dialogues with African Academy of Sciences leadership in early 2025 and others have helped align science policies and address regional capacity needs. GOSC has also advanced data governance and AI readiness through a series of workshops on open data methods and metrics in 2024, a workshop on data quality, and a follow-up FSCI Satellite Workshop in 2024. Participation in the international and regional Internet community dialogues underscored GOSC's role in global Internet governance and cross-border data sharing. Regionally, GOSC co-hosted a disaster management training workshop in Mongolia and led a session at the 2024 FBAS Forum in Africa to assess infrastructure needs and federated governance. At the 2024 EGI Conference, GOSC strengthened collaboration with European partners around shared infrastructure architecture. These research and practice initiatives are brought together in this session.
This session will consist of a 90-minute interactive workshop featuring short thematic presentations, followed by a moderated panel discussion with active audience engagement.
Session co-chairs
• Prof. Jianhui Li (Principal Investigator, CAS GOSC Initiative)
• Dr. Tshiamo Motshegwa (Coordinator, African Open Science Platform).
• Mr. Francis P. Crawley (Chair, CODATA IDPC)
• Simon Hodson (Executive Director, CODATA)
• Dr. Lili Zhang (Coordinator, CNIC,CAS & GOSC IPO)

Agenda
1.Opening Remarks and Framing the Dialogue (5 minutes)
By the session chairs, introducing the goals of the workshop and how it fits into IDW 2025 and SciDataCon priorities.
2.Thematic Presentations (4-6 speakers: approx. 10 minutes each)
Each presentation presents concrete experiences, infrastructure models, and challenges related to open science for the SDGs in the Global South. Confirmed and invited speakers include:
A.Prof. Lise Korsten (African Academy of Sciences) TBD
"Open science and data governance for sustainable development in Africa"
Drawing on agricultural and environmental sciences, this talk will reflect on regional infrastructure and policy needs in Africa.
B.Dr. Tshiamo Motshegwa (University of Botswana & AOSP) TBD
"Linking national and regional science clouds for SDG impact"
Presentation on how the AOSP is creating a federated and policy-aware research environment in Africa.
C.Dr. Rania Elsayed Ibrahim (National Authority for Remote Sensing & Space Sciences, Egypt, online)
"Remote sensing and open data for urban resilience"
Use cases in Earth observation for disaster risk reduction and climate-resilient urban planning.
D. Francis Agamah (H3Africa & DS-I Africa)online
"Open data systems for health and genomic research in Africa"
Focus on bioinformatics, health data systems, and infrastructure for local, regional, and national contributions to global science.
E.Nicky Mulder TBD
F.Agnes Kiragga, APHR TBD
G.Dr. Lili Zhang and Jianhui Li (CNIC, CAS & GOSC IPO, Nanjing U)
"Building shared infrastructure for SDG research with the Global South"
Insights from CSTCloud development and GOSC coordination in enabling platform interoperability and trusted access.

H.Panel discussion and open dialogue (30 minutes)
Moderated by Francis P. Crawley, this portion will invite interaction between speakers and the audience. Key questions will include:
I. What are the immediate infrastructure needs to support SDG-related research in the Global South?
II. How can we ensure ethical, equitable, and sovereign data access?
III. How can AI and advanced analytics be responsibly integrated into SDG science platforms?
I. Closing remarks and session summary (5 minutes)
Key takeaways, recommendations, and pathways for collaboration beyond IDW 2025.
Outcomes
• A synthesis of infrastructure needs and opportunities for the Global South
• Practical models for federated access and responsible data governance

• Recommendations for integrating open science infrastructures into national and regional SDG strategies
• Opportunities for future collaboration within the GOSC and broader CODATA communities

This session focuses on infrastructure coordination through research and practice and highlights pathways to making open science a technical and policy enabler for addressing urgent global challenges.

Presentations Session 6: The Transformative Role of Data in SDGs and Disaster Resilience / 93

Remote Sensing Data for large lakes to asses Water Availability for accelerating SDG 6 implementation and enhance human well-being

Authors: Ioana Popescu[1]; Carmen Cillero[2]; Andrea Jonoski[None]; Harriet Wilson[None]; Krishna Patil[None]

[1] *IHE Delft*

[2] *Estonian University of Life Sciences*

Corresponding Authors: i.popescu@un-ihe.org, carmen.cillero@emu.ee

Freshwater is one of our most precious resources, essential for drinking water, agriculture, inland fisheries, and recreation. However, both its availability and quality face significant challenges from human activities, with these pressures potentially intensifying due to environmental changes.
Large lakes, while covering only 3% of Earth's surface, hold approximately 87% of surface freshwater, making them crucial water reservoirs. The urgency to understand and protect these water bodies is immediate: an estimated 4.4 billion people globally depend on unmonitored water sources. This knowledge is vital for mitigating risks associated with the planetary crisis of climate change, pollution, and biodiversity loss. The United Nations' 2030 Agenda, endorsed by 193 member states, includes 17 Sustainable Development Goals, with SDG 6 specifically targeting sustainable water and sanitation for all. The water sector's significance is further emphasized by the fact that 60% of climate adaptation strategies involve water management, establishing it as a global priority.
Achieving SDG 6 requires innovative strategies to overcome major data gaps, especially in vulnerable regions. The latest reports (2024) highlight insufficient water quality data where it is most needed. Earth Observation (EO) technologies, including satellite remote sensing, combined with smart in situ networks, modelling and forecasting tools, offer effective solutions to bridge these gaps. When used together, they enable a proactive, data-driven approach to global water management—allowing for informed decision-making
We focus our work is on 2 large lakes in Africa (Tanganyika and Turkana).We will highlight the power of available remote sensing (RS) data to tackle the gaps in data collected for water resources management and SDG6 monitoring and the potential of the synergy between RS and other data streams to asses, and predict water availability and improve this through decision making. The main focus is on proving the potential for improving the SDG 6 implementation, as well as improving social safety in regard to water availability.
We will describe how data/knowledge tools can be the missing link to increase social engagement and how they can be used in favour of water quality management, peace and security. Through the results of 2 projects based in Lake Turkana and Lake Tanganyika we will describe the design of a new water quality indicator based on EO data; tools for water and conflict mapping, and the development of a dashboard for planning and policymaking process.

Poster Session / 94

Context and Provenance for FAIR Health Data - The who, what, why, where, when and how

Author: Esmond Urwin[1]

Co-authors: Andy Rae [1]; Tim Beck [1]; Grazziela Figueredo [1]; Phil Quinlan [1]

[1] *University of Nottingham*

Corresponding Authors: svzpq@exmail.nottingham.ac.uk, esmond.urwin@nottingham.ac.uk, msztjb@exmail.nottingham.ac.uk, pmzgf@exmail.nottingham.ac.uk, uizar1@exmail.nottingham.ac.uk

The COVID-19 pandemic has altered how health data is regarded and was a distinct driver for change. The need for rapid analysis and assessment of health data at scale brought sharp focus to the challenges, highlighting the importance of Findable, Accessible, Interoperable and Reuseable (FAIR) data. The heterogeneous nature of health data, together with the wide array of systems and associated formats that record and represent data is a fundamental impediment to this. This is compounded by the representation of data, i.e., how it is coded and thus understood. This aspect alone varies widely, from localised efforts to encode data through to the use and application of international standardised controlled vocabularies (e.g. SNOMED).

One approach used to address these issues is the application of Common Data Models (CDM). A CDM that is gaining traction internationally is the Observation Health Data Sciences and Informatics (OHDSI) Outcomes Medical Observation Partnership (OMOP) CDM. Its aim is to harmonise the representation of heterogeneous health datasets to enable the FAIR assessment and analysis of converted datasets. The approach to creating OMOP CDM datasets utilises a process called Extract, Transform and Load (ETL). The Transform process converts source data representation to a target representation using OHDSI's own OMOP controlled vocabulary concepts by way of the OMOP CDM. Currently, how the Transform process is performed is ambiguous because the decisions necessary to select the target vocabulary terms to represent source data terms are not explicit. Thus supporting knowledge cannot be used to aid the Transform process. In effect, the whole approach to transforming data is variable, open to interpretation and thus subjective.

To be able to make better informed decisions when transforming data, a potential answer lies in the ability to represent and use context and provenance. In other disciplines, data provenance is used to record and represent the origin of the data, how it has been moved, processed and who has performed these. Yet, for OMOP datasets such approaches have yet to be made available. The more amorphous properties of context are somewhat problematic to be able to represent, but they are in part key to understanding why decisions are made and the factors that influenced them.

In November 2024, a workshop was run to focus upon 'How to be FAIR with data standards'. Its focus was the conversion of datasets to the OMOP CDM, such conversions can be lossy –in that data granularity may not be translated fully. 55 people from around the United Kingdom and Europe took part in the workshop. These constituted academics, clinicians and industrial representatives. Activities were designed to elicit perspectives and thoughts from attendees by posing questions intended to derive a response. Two themes were set out and groups were assigned accordingly. The main question posed was 'All OMOP datasets are unreliable'. Currently there is no way to prove whether or not they are reliable.

The workshop produced a rich set of perspectives focused upon current transformation challenges. Utilising these, a context and provenance data model has been developed. It represents the pertinent attributes that potentially affect decision making and the provenance of the data conversion. The data model is composed of several concepts, they are: OMOP Datasets, Quality, Standards, Sharing, Circumstances, Provenance, Version, Transformation, Decisions, Context, Datasets, People, Bias and Training.

The central concept of the model is OMOP Rulesets, i.e., the rules that have been created to transform source data to the OMOP CDM. The concepts of Quality and Standards support this, Quality representing the quality of the rulesets and Standards, the associated standards. Sharing represents FAIR data approaches to sharing rulesets.

The next key concept is Circumstances which relates to OMOP Rulesets. It models the who, why, what, where, when and how (bringing together the concepts of Provenance, Decisions and Context). It coalesces the factors that influence data transformation. Provenance relates to the OMOP Ruleset concept, and Circumstances. It represents the factors of data origin, data transportation, changes made to the data, when and by whom. This is supported by Version to evidence changes made. The concept of Decisions models reasons why decisions were made and the transformation aim. Context concerns environmental aspects, where was the transformation performed, under what conditions

and time pressures. Transformation represents the processes undertaken to convert the data to the OMOP CDM, this in turn relates to the concept of Datasets (the source data). Additionally, the concepts of People, Training and Bias represent the people that perform the decisions, their expertise and training, plus any potential bias.

The context and provenance data model seeks to make explicit the factors surrounding CDM data transformations. The conversion of data to the OMOP CDM is a time consuming and complex undertaking. Hence, there is a need to understand how better to transform data in an ever-efficient manner by providing methods and systems to support this. Additionally, the growth in federated data approaches underlines the need to create reliable OMOP datasets, to which the data model can provide provenance and context information for such transformations.

Further work will entail (i) running the workshop again at the Elixir All-hands meeting in June 2025 and to gain additional insight; (ii) development of a software tool to instantiate the data model to capture the attributes of data provenance and context for end users. The objectives of these are to further support OMOP CDM transformations and improve their reliability.

95

Bridging the FAIR gap: transforming the long tail of supplementary data & generalist repositories into FAIR datasets

Authors: Julien Gobeill[1]; Melissa Harrison[2]; Patrick RUCH[1]; Wolmar Nyberg Åkerström[None]

[1] *SIB Text Mining group, Swiss Institute of Bioinformatics*

[2] *EMBL's European Bioinformatics Institute (EMBL-EBI)*

Corresponding Authors: patrick.ruch@sib.swiss, mharrison@ebi.ac.uk, julien.gobeill@hesge.ch, wolmar.n.akerstrom@uu.se

The rapid rise in adoption of open science practices, coupled with growing mandates from publishers and funders for data to be published, has led to a dramatic increase in supplementary data files published alongside articles and generalist repository uploads. Supplementary data are now submitted with approximately 80% of publications, a substantial increase from about 40% in the early 2000s, and this "long tail of data"signifies a vast and under-exploited source of scientific information—a potential gold mine. However, the inherent heterogeneity, lack of standardisation, and often limited metadata associated with these files pose significant barriers to their discovery and reuse.

This session addresses the challenges associated with achieving increased Findability, Accessibility, Interoperability, and Reusability (FAIRness) for this long tail of data, focusing on textual and image-based information. Building on the report "FAIRness of shared data in life sciences, and opportunities to improve"1, it brings together generalist repositories, data curators and publishers in a cross-disciplinary discussion and interactive workshop to showcase innovative solutions for implementing the FAIR principles and incrementally "FAIR-ifying" these data. Ultimately, the session aims to identify recommendations for improved workflows and foster new collaborations between researchers and practitioners to tackle complex challenges, such as extracting meaningful information from images and figures, which are often crucial components of supplementary data.

The session is facilitated by members of ELIXIR2—a European research infrastructure, bringing together life science resources across over 20 member countries. Part of ELIXIR's activities3 aims at improving the exchange of knowledge, best practices, and technologies to ultimately strengthen global efforts to make life science data more useful. In this context, ELIXIR's data platform4 builds on the broader efforts of the biological data curation (biocuration) community and a network of representatives of repositories and publishers to enable "scalable curation support from the long tail of biological data".

The following agenda designed to support a dynamic and interactive session:

- *Welcome and introduction to the topic (10 min)*

- *Showcase Innovative Solutions (30 min):* Feature presentations from leading experts developing tools and methodologies for automated metadata extraction, data harmonization, and image analysis from supplementary files and generalist repositories.
- *Facilitated Cross-Disciplinary Dialogue (20 min):* Bring together representatives from generalist repositories (e.g., Zenodo, Dryad, BioStudies), data scientists, publishers, and researchers to discuss the practical challenges and potential solutions for enhancing FAIRness.
- *Synthesis of Actionable Recommendations (20 min):* Engage participants in collaborative discussions to identify key bottlenecks and develop concrete recommendations for improving data curation workflows, metadata standards, and repository policies.
- *Next steps to Promote Community Building (10 min):* Foster a network of researchers and practitioners dedicated to addressing the challenges of the long tail of data, facilitating ongoing collaboration and knowledge sharing.

Speakers and panelists will be recruited from an international network of collaborators, and will include:

- ELIXIR Data Platform to present and expand on the findings and recommendations from their report.
- Leading generalist repositories (Zenodo, Dryad, BioStudies) to discuss their efforts in supporting FAIR data.
- Research groups developing tools for automated metadata extraction and image analysis.
- Representatives from publishers who are working on better ways to handle supplementary data.

By focusing on practical solutions and fostering a collaborative environment, this workshop will contribute to the development of a more robust and accessible ecosystem for scientific data, ultimately accelerating discovery and innovation.

References

1 FAIRness of published data in life sciences, and opportunities to improve (Elixir Data Platform D4.1). https://doi.org/10.5281/zenodo.15007096. In press. Copy at https://docs.google.com/document/d/19c7CrCE2sILHnbI7iD-Q/edit?usp=sharing
2 About us | ELIXIR, https://elixir-europe.org/about-us
3 ELIXIR Scientific Programme 2024–28 | ELIXIR, https://elixir-europe.org/how-we-work/scientific-programme
4 Data Platform | ELIXIR, https://elixir-europe.org/platforms/data

96

Emerging technologies in the global context: challenges and opportunities for the long-term environmental data management lifecycle.

Authors: Alison Specht[1]; Gretchen Stahlman[2]

[1] *TERN, University of Queensland*

[2] *Florida State University School of Information*

Corresponding Authors: a.specht@uq.edu.au, gs23j@fsu.edu

Conversations around the data lifecycle from creation to re-use frequently revolve around the challenges of limited resources, lack of understanding by parties at various points of the process of what is involved in making it successful, and lack of interest from funding and re-use communities (e.g. Borgman & Groth, 2025, Specht et al., 2025, Stahlman, 2022). This is exacerbated when the data collected and preserved for reuse is to be sustained over the long term. Maintaining the integrity and quality of data collection, data deposition, curation and discovery while technology, funding, and expectations from those data evolve presents persistent challenges. Moreover, questions increasingly

surround the provision of sufficient (super)computer access to properly curate data holdings while allowing effective augmentation in the data stocks, as well as open access to those holdings.
This 90-minute session will attempt to identify critical factors affecting the data lifecycle over the long term. These include the inherent expansion of data over time across domains and knowledge contexts, the effect of emerging technologies, the increasing energy and financial cost of handling, curating, and using data, and who takes responsibility for the curation of data for the common good. We shall use environmental data for research as our domain focus.
Invited panellists from various stages of the data lifecycle (see details below) will use their experiences and how they have overcome the challenges of maintaining good practice against adversity to answer the three questions listed below. Each panellist will present pre-prepared short statements (not all panellists will necessarily have comments against every question) followed by a short Q & A with the audience.
How do we balance the promise of emerging technologies with the practical risks of data loss and preservation challenges across the long-term data lifecycle?
What does true democratization of environmental data look like—and who might be left out?
How can we ensure that our environmental data management practices remain sustainable—both environmentally and ethically—in a rapidly shifting global context?
The session will close with audience discussion guided by various data lifecycle stages in relation to the panel discussion (20 minutes). This will help formulate key themes from the session, and participants (panellists and attendees) will be invited to contribute to a paper aimed to describe a clear understanding of current status and provide guidance for future work, such as (a) the Identification of possible roles of new technologies, (b) the quantification of the cost of high quality, persistent and abundant data, and (c) the identification of risks to data when supported by big corporations, and mechanisms to reduce risk.
The session will be supported by a questionnaire of the audience and an on-line space for future collaboration.
References
Borgman CL, Groth P (2025) Harvard Data Science Review 7, doi:10.1162/99608f92.35d32cfc
Specht A, et al. (2025) Data Science Journal 24, 1. doi:10.5334/dsj-2025-001
Stahlman GR (2022) Journal of the Association for Information Science and Technology 73, 1692–1705. doi:10.1002/asi.24687
Panellists
Our assembled experts represent various components of the data lifecycle. Insights into the challenges of data retention will be provided by Wim Hugo, a member of the EOSC long-term data retention task force, while experience of the data demands imposed on an observatory and repository will be provided by Siddeswara Guru. Greg Maurer of the US-LTER can provide practical ways to achieve consistency in data acquisition across a distributed network of research sites, while Gretchen Stahlman brings expertise in data curation education and legacy data integration. Stephen Bird (or colleague from the Queensland Cyber Infrastructure Foundation, part of an Australia-wide network that provides cloud facilities for research organisations) will bring understanding of the options and limitations of cyber infrastructure support for the present and the future. Shelley Stall is a pre-eminent global thinker and promoter of Open Data practices for researchers and an expert in scholarly publishing, while Tavita Su'a provides fundamental understanding of the requirements for building and supporting an emerging data network and repository. As chair, Alison Specht brings theory and practice in field (including transdisciplinary), experimental, and synthesis centre research, data management and education, with an eye to timeliness.
Stephen Bird: Queensland Cyber Infrastructure Foundation (QCIF), 0009-0001-9846-0990; Siddeswara Guru: TERN, University of Queensland, 0000-0002-3903-254X; Wim Hugo: Chief Technology Officer for DANS, an institute of the KNAW (Royal Dutch Academy of Science)) 0000-0002-0255-5101; Greg Maurer: US Long Term Ecological Research Network, State University of New Mexico, 0000-0002-3007-8058; Gretchen Stahlman: Florida State University School of Information, 0000-0001-8814-863X; Shelley Stall: Data Vice-President, American Geophysical Union, 0000-0003-2926-8353; Alison Specht: TERN, University of Queensland, 0000-0002-2623-0854; Tavita Su'a: South Pacific Regional Environmental Portal and Data Hub.

97

The highs and lows of providing digital infrastructure to enable safe access to sensitive data for research

Author: Tim Beck[1]

Co-authors: Alain-Dominique Gorse [2]; Clair Sullivan [2]; Jason Ferris [2]; Neerja Karnani [3]; Simon Thompson [4]; Sumir Panji [5]; Xing Yi Woo [3]; Philip Quinlan

[1] *University of Nottingham*

[2] *University of Queensland*

[3] *Bioinformatics Institute, A*STAR, Singapore*

[4] *Swansea University*

[5] *University of Cape Town*

Corresponding Authors: neerja_karnani@bii.a-star.edu.sg, j.ferris@uq.edu.au, d.gorse@uq.edu.au, philip.quinlan@nottingham.ac.u simon@chi.swan.ac.uk, sumir.panji@uct.ac.za, tim.beck@nottingham.ac.uk, woo_xing_yi@bii.a-star.edu.sg, clair.sullivan@health.qld.

Significance of the issues to be tackled in the session

A Trusted Research Environment (TRE) is a highly secure computer system where sensitive data is stored. TREs are designed to be safe, allowing only authorised individuals to access the data. Data cannot be added or removed without proper permissions, ensuring transparency and accountability. Multiple sources of data can be combined in a TRE to create a comprehensive dataset for research. The technical and legal demands on these environments are considerable, resulting in challenges for these systems to interoperate.

International standards-setting organisations such as the Global Alliance for Genomics and Health (GA4GH) and ELIXIR provide resources that lower the barrier for TRE providers to align the development of their infrastructures. The HEALTHWISE consortium brings together providers of TREs, and other digital infrastructure for safe access to sensitive data, to tackle challenges around transnational and transcontinental federation. During this session, HEALTHWISE members will present the state-of-the-art in digital infrastructure to enable safe access to sensitive data. During an expert panel discussion, the challenges and opportunities for federation of these infrastructures will be discussed in conversation with the audience.

Description of the approach, structure, format, and suggested agenda for the session

The aim of the session is to set out the international landscape of digital infrastructures that enable safe access to sensitive data for research. Audience questions about the highs and lows of providing such infrastructure will be addressed by a panel of experts in conversation. The session will begin with brief presentations from invited speakers on the theme of providing secure digital infrastructure to access sensitive data. This will be followed by an interactive panel discussion with infrastructure providers, led by a moderator who will guide the conversation, explore the topic from multiple perspectives, and facilitate audience questions.

Draft agenda

Part 1: Welcome and introduction to the aims of the session (10 mins)
Tim Beck, University of Nottingham

Part 2: Setting the scene - 10 min presentations (50 mins)
Chair: TBC
Five speakers:

1. Phil Quinlan, University of Nottingham, Federation of UK TREs & NHS Secure Data Environments (SDEs)
2. Dom Gorse, University of Queensland, Making Queensland Health Data Accessible for Research: The Role of SMART Hub and UQ TRE
3. Sumir Panji, University of Cape Town, Making African Data more Discoverable
4. Simon Thompson, Swansea University, Large Scale TRE Provision & Operation
5. Neerja Karnani, A*STAR, Making Data Science Secure and Connected: Insights from Singapore

Part 3: Panel discussion with infrastructure providers and audience questions (30 mins)
Moderator: TBC
Panellists: Phil Quinlan, Dom Gorse, Sumir Panji, Simon Thompson and Neerja Karnani.

Proposed speakers and the subject of their talks

Phil Quinlan, University of Nottingham, UK
The UK has invested in federated programmes, such as in Health Data Research UK (HDR UK) and in Data and Analytics Research Environments UK (DARE UK). The team in Nottingham are leaders in these programmes and during this talk we will present the current state of the possible and the international collaborations that are flourishing via open source and open standards methodologies.

Dom Gorse, University of Queensland, Australia
Queensland's Integrated Electronic Medical Record (iEMR) is a single system deployed across the state, providing a unified platform for health data management. Access to Queensland Health data for research is facilitated through the University of Queensland's SMART Hub and Trusted Research Environment (UQ TRE). The SMART Hub connects researchers with health data, ensuring secure and compliant access, while the UQ TRE offers a highly secure environment for analysing sensitive health data. This talk will also explore the principles of TREs and the CARE principles, which guide the ethical and responsible use of data.

Sumir Panji, University of Cape Town, South Africa
The eLwazi platform is developed by the University of Cape Town and is an African-led Open Data Science Platform (ODSP) designed to support health and biomedical research across the continent. It is a key component of the DS-I Africa initiative, which aims to harness data science for health discovery and innovation in Africa. The eLwazi platform actively incorporates Global Alliance for Genomics and Health (GA4GH) standards to enhance data interoperability, security, and accessibility. This talk will explore how these initiatives are making African data more discoverable.

Simon Thompson, Swansea University, UK
Swansea University makes available SeRP (Secure e-Research Platform), which is a secure data platform that provides TREs for researchers to access and analyse sensitive data, particularly health and administrative data, while ensuring data privacy and security. It is used by academic institutions, government bodies, and health organisations. A prominent example is SAIL Databank, which uses UK SeRP to provide access to anonymised health and administrative data from Wales. This talk will explore the diversity and demands of running multiple large-scale TREs in the UK and Internationally.

*Neerja Karnani, Bioinformatics Institute, A*STAR, Singapore**
This talk will provide insights into our cross-institutional efforts and collaborations with national platforms, industry, and international data communities to advance secure data science in Singapore. These initiatives are shaping the future of federated analytics by aligning technological innovation with strong data governance. It will also highlight emerging use cases, the adoption of open standards, and expanding opportunities for global collaboration.

Poster Session / 98

A Practice of Science Data Bank on Promoting Data Sharing in China

Authors: Lulu Jiang[1]; Pengyao Wang[1]; Chengzan Li[1]; Yuanchun ZHOU[1]

[1] *Computer Network Information Center, Chinese Academy of Sciences*

Corresponding Authors: zyc@cnic.cn, jianglulu@cnic.cn, lichengzan@cnic.cn, wangpengyao@cnic.cn

Data sharing is considered one of the effective practical means to enhance research transparency. Data repositories are pivotal e-infrastructure in fostering this. As a generalist data repository, Science Data Bank (ScienceDB) offers free services to the global community for sharing and dissemination of non-traditional research outputs, such as datasets and codes. It has been built and maintained by Computer Network Information Center, Chinese Academy of Sciences since 2015 and shared more than 8 million datasets by now.

But the recognition and practice of data sharing among researchers are generally insufficient. This requires multiple efforts to jointly promote policies, standards, digital infrastructure service capabilities, and incentive mechanisms.

To foster the practice of open data, the repository integrates with a variety of external systems including preprint servers, manuscript submission systems and published paper systems. Through these integrations, Science Data Bank streamlines the data sharing process across diverse research outputs, making data sharing more convenient at the system operation level, supporting the data review of the articles'peer-review, and assisting policy makers such as journals and publishers encourage the data sharing behaviors. Moreover, the integrations establish the links between data and literature not only in the repository but in the external systems. ScienceDB pulls all these links to DataCite, Scholix, Dimensions and CSCD database（a famous paper database in Chinese）, and ensure that enriched output can be easily retrieved from both the data and paper ends.

To facilitate institutional data management and data sharing, ScienceDB integrates with the institutional data management tools. This work effectively reduces the repetitive operation of researchers uploading data files multiple times under requirements from different policies. By establishing an institutional Data Community on ScienceDB, it provides a better display and unified service portal for the data sharing of an institution.

To incentive data sharing among researchers, ScienceDB integrates with Intellectual Property Registration System in China. In order to encourage data management practices, China's National Intellectual Property Administration has launched a pilot project to register data intellectual property rights. As the only pilot platform for scientific data supporting property rights registration, ScienceDB has successfully completed multiple data intellectual property registration tasks for research data shared by Chinese scientists. The brand-new attempt to enable researchers to share data while achieving recognition for more efficient intellectual contributions.

By integrate with various external system, ScienceDB connects multiple scenarios of data sharing with many positive impacts on disseminate the best practice of data sharing and foster the research transparency.

Poster Session / 100

From Principles to Practice: Designing Researcher-Centred Solutions for Open Science

Author: Graham Smith[1]

[1] *Springer Nature*

Corresponding Author: graham.smith@springernature.com

Over the past decade, open science has moved from the margins to the mainstream. Yet for many researchers, putting open science into practice remains challenging —due to disciplinary norms, fragmented support systems, and tools that prioritize policy over usability. At Springer Nature, we ar evolving our approach to open science support by embedding FAIR principles into product design and workflows, making open practices easier, more attractive, and more aligned with researchers' day-to-day needs. We are committed to ensuring researchers can do this in a safe and secure way.

This session will present a set of practical interventions developed to support researchers in sharing data and adopting open practices. These tools have been shaped by ongoing user research and iterative development cycles, ensuring they meet real needs rather than just compliance goals. We will share examples of how we've shifted from a policy-driven to a researcher-centred mindset, with solutions that guide users through complexity, offer meaningful feedback, and demonstrate value through social proof and community engagement.

A core evidence base informing this work is The State of Open Data project, in partnership with Figshare and Digital Science, that is the longest running survey and analysis on open data. Cele-

brating the State of Open Data's tenth year in 2025,this year's report will introduce a deeper focus on disciplinary trends and feature new expert commentary on what's next for open science —insights that we will preview and build upon in this session. Drawing on longitudinal survey data from tens of thousands of researchers globally, The State of Open Data has revealed not only enduring challenges in data sharing, but also shifting attitudes and emerging norms, with growing discipline-specific differences.

By combining fresh findings from the newest The State of Open Data survey with practical case studies from our ongoing initiatives, this presentation offers both a reflection on progress and a look ahead. It will highlight what it takes to make responsible and reproducible science not just possible, but preferable, through support that researchers actively choose to use.

Poster Session / 102

Leveraging Open Science for Geographical Indications Environment & Sustainability Study Case

Author: Limin Li[1]

[1] *Global Change Research Data Publishing and Repository,World Data System,Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences*

Corresponding Author: lilimin@igsnrr.ac.cn

Introduction：China has rich geographical and biological diversity, and cultural resources, which have enriched people's lives.Its diverse and complex geographical environment has given rise to a wealth of geographical products.Geographical Indications (GIs) products hold significant importance in promoting agriculture, enhancing product quality, preserving cultural heritage, and driving sustainable development. According to the latest data from the China National Intellectual Property Administration, China has been certified a total of 2,547 GIs products by March 2025, demonstrating their critical role in the country. Taking Baoshan Coffee as an example, it has gradually become a vital cash crop in the region. Listed in the inaugural China-EU Geographical Indications Agreement in 2020, Baoshan Coffee has spurred local economic development, improved agricultural quality, and increased farmers' incomes. However, challenges remain in balancing economic benefits with environmental sustainability:1.Unclear delineation of geographical production areas with insufficient scientific basis for geographical boundaries; 2.Low brand awareness and limited consumer understanding of GIs products; 3.Inadequate traceability systems leading to inconsistent product quality and reliability; 4.Weak intellectual property rights protection;5.Incomplete standard. To address this issue, we propose the "Geographical Indications Environment & Sustainability" (GIES) initiative.The initiative focuses on GIs products, Geographical Specific Product, and Geographical Traditional civilized Product. The initiative started from six dimensions: variety, quality, appearance, brand, morality, and taste. By utilizing open science, geographical big data, and IoT technologies, and through open science data publishing, the initiative integrated science, technology, engineering, standards, and culture. It established a case traceability and intellectual property rights protection system.The GIES Initiative through collaborative efforts among multiple stakeholders and partners.Ultimately, it aims to bolster regional ecological protection and foster local sustainable development.
Methods: (1) Open science on the geographical information system, remote sensing, big data, and IoT technologies of physical and human geography, (2) standard management of agricultural products, (3)The traceability of products enabled by IoT technologies. （4） GIES case linking producers the market and consumers to promote economic growth.
Findings:(1)The GIES use of geographical science to define the production boundaries of key GI products and employs remote sensing, big data technology, and IoT technologies to network the product chains for better quality control.
(2） Physical geographic data, including air quality, soil composition, water quality, terrain features, geographical boundaries, land use patterns, and natural ecology data, are comprehensively gathered through remote sensing.This wealth of data provides a scientific foundation for understanding the unique environmental conditions that contribute to the distinct characteristics of GIs products.
（3） Leveraging Open Science, the GIES case datasets, and relevant research papers of each of these cases have been published in the Digital Journal of Global Change Data Repository and Journal of Global Change Data & Discovery.
（4） Participating in product expos, forums, festivals, and training workshops has generated high

demand for GIES products and significantly improved local farmers' livelihoods.
Conclusion:The GIES initiative has achieved a balance between economic benefits and environmental sustainability, fostering regional ecological protection and local sustainable development.
Keywords: Geographical Indications;Open Science;Data Publishing;environmental sustainability;economic growth

Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research / 103

Sustainable Open Data Infrastructure to Protect Government Data during Regime Changes: African Examples

Author: Lynn Woolfrey[1]

[1] *University of Cape Town, DataFirst*

Corresponding Author: lynn.woolfrey@uct.ac.za

1. Introduction and Background Recent policies and laws in the US that threaten data access have highlighted the need for infrastructure to ensure important government data is not lost during regime changes. US federal government information has been in the public domain since the 1895 Printing Act which prohibited any copyright on federal government publications (United States Congress, 1895: 608). However, ideologically motivated political interference can threaten government data even in the most robust data ecosystems. Webb and Kurtz (2022: 68) reveal data removal by administrators at 12 federal agencies during 2017-2020, and (Nost et al., 2021: 1) calculate that about 20% the EPA's website content was removed between 2016 and 2020. In 2025 executive orders of the (The White House, 2025a; 2025b) that targeted federal programmes related to diversity, equity and environmental protection led to many departments and offices removing datasets related to gender, race and climate change {Choo, 2025 #1362@1;Mallapaty, 2025 #1358. A further executive order of the (The White House, 2025c) to pause foreign aid obligations and disbursements ended many US programmes to co-collect and disseminate data with LMICs (Offord, Cohen and Enserink, 2025) and shuttered agency data sites (Mallapaty, 2025).

Regime changes in African countries have also impacted access to government data. The government of Tanzania withdrew from the (Open Government Partnership, 2017) and reneged on their open government data (OGD) commitments in 2017 after a change of government. (Mutuku and Idriss, 2019: 557) note that early Nigerian and Ivorian OGD initiatives also did not survive changes of leadership. Supportive OGD infrastructures such as open data repositories can help data to endure through such crises (Jarvenpaa and Essen, 2023: 11).
2. Data Rescue at DataFirst
The team at DataFirst, an African research data service, has stepped in when African data is threatened with erasure. Apart from regime changes, threats to data originate from the absence of robust data infrastructure in many African countries, which hamper ongoing preservation of and access to both government data and research datasets (Willoughby, 2019: 239). DataFirst (2025) is an African research data service and internationally certified repository based at the University of Cape Town (UCT) The DataFirst team have saved data from South Africa's first income panel survey when the survey site was closed during institutional changes. Data sustainability can also face risks other than political interference, such as media deterioration and loss of contextual knowledge (Mayernik et al., 2020: 5). This was the case with the 2015-2016 rescue of data in hard copy from two South African surveys documenting outcomes for victims of early and later forced removals. These rich historical datasets are available on our data site.

DataFirst's US data rescue efforts focus on two valuable data series. First, USAID-funded DHS program data, which includes African demographic and health data collected since the 1980s (DHS Program, 2025). The discontinuation of the DHS program will have negative consequences for SDG progress monitoring by African governments, as DHS data has been used heavily for this purpose (Woolfrey, 2020). After the shuttering of USAID, it was no longer possible to register projects with

the DHS Program, but an API enabled bulk downloads of the data files by those with already registered projects. DataFirst downloaded all African datasets, and we are currently creating metadata records and organising data sharing permissions with African government agencies, and will apply for further permissions to share the rest if DHS Program access is not restored. The second data series targeted for data rescue is education data from USAID funded projects. Access to this data was lost with the shuttering of the USAID's Development Data Library. DataFirst's AFLEARN project, which focuses on foundational learning, offered project implementers in African countries the option to upload their data via a Nextcloud file hosting service to a secure server at UCT. With agreements from implementers DataFirst will provide long-time hosting and secure and sustained access to these threatened datasets.

3. Open Government Data as A Risk Reduction Strategy

The 2025 data rescue efforts exposed weaknesses in funding agencies'data management policies and procedures. USAID's data policy fell short of implementing Obama-era policy and legislation to make openness the default for government data, including data from federally funded scientific research (United States Agency for International Development, 2020: 10-12, 16-17) As a result, Africa-based USAID project implementers, who had in some cases effectively signed over their data to USAID did not have sharing permissions and were therefore hesitant to place the data in rescue repositories. A policy fully aligned with federal open data legislation could be a strong data sustainability mechanism. Data rescue obstacles relating to ownership demonstrate that building OGD infrastructure can be a risk reduction strategy to preserve data as a societal good (Dodds and Wells, 2019: 260-267).

Poster Session / 104

From the people, for the community: Using a Residents'Assembly to build the Liverpool City Region Data and AI Innovation Charter

Author: Emily Rempel[1]

[1] *University of Liverpool*

Corresponding Author: emily.rempel@liverpool.ac.uk

Introduction

Citizens' juries and assemblies are increasingly popular public participation methodologies for deliberation on data and artificial intelligence. They are formal, top-down exercises that aim to address power imbalances in the design, application, or regulation of data and AI processes through allowing residents to debate and provide recommendations on specific questions. In March 2025, we invited over 60 residents in the Liverpool City Region (LCR) in the United Kingdom to take part in a Residents'Assembly on Data and AI Innovation. We specifically asked residents what trustworthy and beneficial data and AI practice looks like for the City Region to build a Charter on Data and AI Innovation. In this paper, we critically reflect on the design and outcomes of the Assembly and whether it was able to better represent public voice in the development of regional data and AI policy and practice in the UK.

Assembly Methodology and Data Sources

Designing the Assembly

The Assembly was designed and run by the LCR Civic Data Cooperative, a local government-funded research project. It was convened jointly by local health, government, and University partners. We aimed to recruit 60 residents via randomised postcode sampling. Residents who signed up were then stratified to represent seven demographic characteristics in the region including knowledge of AI. Residents took part in four full days of activities as well as two online or phone sessions. This included group induction, two days of learning on data and AI, two days of interactive deliberation, and a final debrief session. The core activity of the latter two days was a ranked choice vote on resident suggested Charter principles.

Around this process, we specifically sought to address inclusivity and ensure diverse representation. Residents were compensated for their time, travel, and meals. All recruitment was completed via post to address digital exclusion. Some of the additional services advertised to residents included letters of support for employers, translated materials, live translation, childcare, and 1-2-1 debrief sessions.

Data Sources

Throughout the sessions, facilitators took notes and select recordings to report on resident perspectives, key questions, and reflections. Notes and transcribed recordings will be thematically analysed for common themes related to these topics. In addition, we report the process and validation of the principles put forward by residents through the two-days of deliberation, the ranked choice voting exercises, and the debrief sessions. Voting was fully anonymous using Poll Everywhere software. Residents ranked the principles from most to least importance. A formal independent evaluation was also conducted and is forthcoming.
Early Results and Discussion
Principles and Resident Perspectives
59 residents took part in the four in-person Assembly sessions. 60 principles were put forward by residents. These principles were reduced by the Assembly team for repetition and semantic similarity. These 22 unique principles were then ranked by residents and the top 12 were taken forward for further review. 55 total votes were recorded during the ranked choice vote. The final set of principles will be available in May 2025. They currently include four key concept areas including benefits and harms, inclusivity and transparency, oversight and accountability, and security and legality. Findings on key resident hopes and concerns are forthcoming as full data collection will be complete in April 2025.
Process
Early reflections on the process of designing the Assembly include three core concepts: representivity, literacy challenges, and positive bias.
First that the nature of a Jury or Assembly process is exclusionary by design. 60 residents cannot, and are not, expected to be representative of the full range of perspectives on data and AI. Four full days is a significant investment in time for residents and inevitably causes a self-selection in who takes part. While childcare and translation services were offered, they were not taken up by any residents. This means exercises must, themselves, challenge participants to think about different perspectives on data and AI in order to represent diversity in an Assembly setting.
Second, the necessity of accommodating a variety of levels of data and AI literacy meant that some activities were too complicated and some too simple for the residents who took part. Thus, some were more and less successful an eliciting critical reflection on the contents of a potential data and AI charter. Embracing reflexivity was as important as robust planning for learning and deliberation activities. However, we found an Assembly methodology was not well suited to representing the voices of those with no knowledge of data and AI.
And third, the nature of positive bias both in how materials were presented and what residents then subsequently shared back about their perspective on data and AI. While we continuously worked with our external evaluator to ensure we adapted materials to maintain balance in what was presented, there were challenges in ensuring all residents fully reflected on the harms and risks of data and AI technologies.
Conclusion
The Assembly was successful in developing regional data and AI policy, namely a Charter. Similar deliberative events can learn from our experiences and challenges to ensure residents are fairly represented in the design of regional data and AI policy and practice.

Poster Session / 105

World Data System Early Career Researcher Network: what it is and why you should join.

Authors: Claire Rye[1]; Daniela Santos Oliveria[2]; Jing Zhao[3]; Yuya Shibuya[4]

[1] *University of Auckland/WDS ECR co-chair*

[2] *World Data System International Program Office*

[3] *National Satellite Meterological Center (NSMC), China Meteorological Administration (CMA)*

[4] *University of Tokyo*

Corresponding Authors: yuya-shibuya@iii.u-tokyo.ac.jp, zhao.j@aircas.ac.cn, claire.rye@auckland.ac.nz, doliveir@utk.edu

The World Data System (WDS) is an affiliated body of the International Science Council (ISC) that supports research data repositories and data service providers worldwide. The WDS Early Career Researcher (WDS-ECR) Network is dedicated to nurturing, advancing, and strengthening the capacities of early career scientists within data-centric fields. The network has ten overarching goals

as part of its charter (https://worlddatasystem.org/charter-wdsecr/), and through fostering connections between communities and across disciplines, the WDS-ECR Network encourages cooperative projects and proactive participation.

This vision of an inclusive environment is oriented towards innovation and advancement in research data stewardship. In recent years, the WDS-ECR membership has grown significantly, encompassing a wide range of academic fields from Social Sciences to Genomics. WDS is committed to strengthening its support for early career data scientists and data stewards by promoting an interconnected community with access to opportunities and resources that will ensure the long-term stewardship and provision of quality-assured data and data services to the international science community.

Our proposed poster will showcase some of the resources offered by the network, including recognition through the WDS Data Stewardship Award, guidance via mentorship programs, skill-building through specialised webinars, and opportunities for worldwide collaboration during networking events. The presentation's objective is not only to illustrate these benefits but also to demonstrate how affiliation with the WDS-ECR Network can empower researchers at their career inception. We endeavour to build a dynamic community that upholds the highest standards in data management such as FAIR and CARE principles and supports open science initiatives. We invite dynamic data science professionals from any discipline or region to join us!

Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data / 106

Bridging the Gap: A Research-Ready AI/ML Infrastructure on the Nectar Research Cloud

Author: Glen Charlton[1]

Co-authors: Long Le [1]; Jiaxin Fan [1]; Anastasios Papaioannou [2]; Andy Botting [3]; Meirian Lovelace-Tozer [3]; Ben Chiu [3]

[1] *Intersect Australia*

[2] *University of Technology Sydney*

[3] *Australian Research Data Commons*

Corresponding Authors: andy.botting@ardc.edu.au, jiaxin@intersect.org.au, long@intersect.org.au, glen@intersect.org.au, ben.chiu@ardc.edu.au, anastasios.papaioannou@uts.edu.au, meirian.lovelace-tozer@ardc.edu.au

Introduction: Researchers and scientists are increasingly using programming languages for data processing, visualisation, and analysis. Advancement in machine learning (ML) and artificial intelligence (AI) helps accelerate the process of analysing complex research data and conducting experiments, leading to the discovery of hidden patterns in the data. However, installing and configuring the environment to begin researching using AI/ML or transitioning from local scripting to utilising cloud computing to accelerate research is often hindered by the need for extensive knowledge or expertise in the use of command line and back-end operating systems. In collaboration with the Australian Research Data Commons (ARDC), the Advanced Analytics and AI (3AI) Platform at Intersect has developed a solution to address this barrier by developing an application that provides a research-ready AI/ML infrastructure on the Nectar Research Cloud. The goal is to reduce the technical barrier of utilising cloud infrastructure for AI/ML-enabled research, empowering researchers to focus on science.

Methods: The application is featured with a modular design, offering an AI-ready environment that covers the foundational aspects of data-intensive research, including a pre-installed and configured Python environment, Data Operations (DataOps), Machine Learning Operations (MLOps), and real-time system monitoring. The pre-installed and configured Python environment is the core of the application with all the necessary libraries for users to begin their AI journey or to go further with optional additional layers focusing on specific AI topics including; Computer Vision, Large Language

Models and Generative AI. This is enabled by the pre-installed and configured state-of-the-art open-source tools (i.e. Apache Airflow for DataOps, MLflow for MLOps, and Prometheus and Grafana for system monitoring). Additionally, interactive software like RStudio and JupyterLab are also included to offer user-friendly development environments. Researchers can customise their environment with specialised Python installations optimised for tasks like computer vision or generative AI. Guidance on hardware specifications (vCPUs, RAM, and storage) is provided. The comprehensive documentation, covering installation, setup, and usage, further supports researchers in utilising the platform effectively. For more advanced users, the infrastructure could be customised from ARDC' s open-source image repository to create their unique platform to meet their needs.

Results: The initiative provides pre-configured, optimised environments that enable researchers to focus on their scientific inquiry rather than complex system administration. The infrastructure simplifies the entry to cloud computing for AI/ML, empowering researchers to quickly and efficiently begin their AI/ML journey. The platform offers essential tools for data analysis, workflow orchestration, and model development within a user-friendly framework. The guidance on hardware specifications also ensures optimal performance for research workloads.

Conclusion: This collaborative effort between Intersect and ARDC aims to empower researchers, regardless of their system administration expertise, to leverage cutting-edge AI/ML tools. By providing a pre-configured, documented, and user-friendly environment, the project removes a significant hurdle in the research workflow. The initiative aims to significantly accelerate AI/ML research on the Nectar Research Cloud, enabling researchers to focus on innovation and discovery and enhancing the Australian research landscape by providing accessible and powerful computational resources.

Presentations Session 9: Empowering the global data community for impact, equity, and inclusion / Education / 107

Higher Learning's Next Era: Enabling Innovation and Interoperability through a Sector-Aligned Data Standard

Authors: Charlsey Pearce[1]; Greg Sawyer[2]

[1] *CEO, MortarCAPS Higher Learning Data Standard*

[2] *CEO, CAUDIT*

Corresponding Authors: charlsey.p@mortarcaps.com, greg.sawyer@caudit.edu.au

Data fragmentation is a persistent barrier to innovation, interoperability, and student mobility in the global post secondary ecosystem. This session presents the MortarCAPS Higher Learning Data Standard (MCDS) —a sector-aligned, data standard co-developed by post secondary institutions, technology leaders, and policy advocates in Australia and Canada.

Charlsey Pearce (CEO, MortarCAPS) and Greg Sawyer (CEO, CAUDIT –the Council of Australasian University Directors of Information Technology) will explore how MCDS is enabling a new paradigm of data portability and collaboration across institutions. Drawing on real-world implementations and government engagement, this session will highlight how MCDS serves as a blueprint for education data ecosystems globally —promoting efficiency, transparency, and interoperability without sacrificing institutional autonomy.

The MortarCAPS Higher Learning Data Standard (MCDS) provides a common language and structure for how data is captured, exchanged, and interpreted across the student lifecycle —from enquiry and enrolment to graduation and beyond. It simplifies system integrations, reduces duplication, and allows institutions to work more seamlessly with government agencies, pathway providers, and each other. Built with extensibility in mind, MCDS is designed to support both public and private providers, enabling scalable digital infrastructure that grows with evolving education models. At its core, MCDS is about empowering institutions to own and use their data more effectively —while reducing friction in cross-institutional collaboration and service delivery.

The session will be structured as a hybrid presentation and discussion, combining insights into the creation and architecture of the standard, use cases from leading universities, and broader implications for the global research and education data community. We will dedicate the final 25 minutes to an open discussion with attendees around adoption challenges, opportunities for cross-sector alignment, and lessons for other geographies seeking to develop or implement similar frameworks.

The conversation will connect directly to SciDataCon's themes of data policy, stewardship, and ecosystem development. By showcasing a concrete solution already adopted across a national education system, we hope to inspire practical strategies for building sustainable, interoperable data infrastructures across disciplines and borders.

Poster Session / 109

From Bureaucracy to Usability - How OSTrails Simplifies Open Science

Authors: Anca Hienola[1]; Elli Papadopoulou[2]; Tassos Stavropoulos[3]

[1] *Finnish Meteorological Institute*

[2] *Athena*

[3] *OpenAire*

Corresponding Authors: tassos.stavropoulos@openaire.eu, elli.p@athenarc.gr, anca.hienola@fmi.fi

The Open Science landscape today is full of good intentions, grand infrastructures, and endless new portals. Yet for many researchers, the daily reality has changed very little. Managing data, publications, and workflows remains fragmented, bureaucratic, and painfully slow. FAIR principles are widely endorsed, but in practice, they are often difficult and time-consuming to implement. The distance between Open Science policy and everyday research practice is still wide.
OSTrails tackles this problem head-on. Rather than adding another layer of complexity, OSTrails builds practical, machine-actionable pathways that connect scientific workflows directly to Open Science best practices. It makes Open Science easier, not harder, by embedding automation and guidance into the tools and systems researchers already use.
By integrating Scientific Knowledge Graphs (SKG), machine-actionable Data Management Plans (maDMPs), and FAIRness assessment tools, OSTrails helps researchers move from scattered documents and disconnected systems to streamlined, reproducible, and responsible science. Researchers no longer have to treat Open Science as an extra administrative burden —it becomes a natural part of their research process.
At the heart of OSTrails is a simple but disruptive idea: researchers should not have to choose between doing good science and doing Open Science. Good workflows should make FAIR, reproducible research automatic, seamless, and intuitive.
In this poster, we present how OSTrails:
• Connects research outputs and datasets automatically through Scientific Knowledge Graphs, improving traceability and reuse.
• Simplifies data management planning with machine-actionable DMPs that integrate directly into research workflows,updating dynamically as research evolves.
• Embeds FAIRness assessment into the research lifecycle without adding extra bureaucracy, helping researchers check and improve the FAIRness of their outputs in real time.
• Supports researchers in producing better, more reproducible science with less administrative overhead, freeing up time for discovery and innovation.
OSTrails is a call to rethink how we operationalize Open Science: not by building yet another system that researchers must learn separately, but by embedding FAIRness, reproducibility, and responsibility into the systems, tools, and practices they already know and use.
Ultimately, Open Science will succeed not because researchers are forced to comply, but because Open Science becomes the easiest and most natural way to work. OSTrails is showing how to make that shift real by designing Open Science for researchers, not around them.

Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data / 112

Metadata Meets Standardization: Leveraging a Staging Database to Integrate African Longitudinal Mental Health Data into OMOP CDM 5.4

Authors: Agnes Kiragga[1]; Bylhah Mugotitsa[1]

Co-authors: Dorothy Mailosi [2]; Evans Omondi [1]; Jay Greenfield ; Jim Todd [3]; Pauline Andeso [1]; Reinpeter Momanyi

[1] *African Population And Health Research Center*

[2] *Committee on Data of the International Science Council- CODATA*

[3] *Catholic University of Health and Allied Sciences*

Corresponding Authors: eomondi@aphrc.org, pandeso@aphrc.org, rmomanyi@aphrc.org, akiragga@aphrc.org, dorothy@codata.org, jim.todd@lshtm.ac.uk, jay@codata.org, bmugotitsa@aphrc.org

Background: Longitudinal studies are necessary for tracking the progression of mental health disorders such as depression, anxiety, and psychosis. However, the integration of diverse mental health data from different sources and waves—especially in low- and middle-income countries—remains complex due to variability in instruments, socio-cultural expressions, and data structural formats. This study introduces a novel staging database framework developed under the INSPIRE Mental Health (INSPIRE MH) project, which bridges standard OMOP CDM concepts and non-standard INSPIRE concepts, using a harmonized metadata-driven approach. The database supports longitudinal, multi-source mental health datasets while preserving socio-contextual detail that would otherwise be lost in direct OMOP CDM conversions.

Methods: The staging database is structured using the DDILifecycle metadata standard and follows a snowflake schema, enabling seamless documentation of datawaves, instruments, instrumentitems, and social context variables across individual and household levels. This schema allows data from HDSS populations (Kilifi, Iganga Mayuge, Kagando) and secondary data accessed from African researchers to be captured uniformly. A dynamic ETL pipeline was created using R to map the data from the source to the staging database, in preparation for mapping to the OMOP CDM while preserving variable traceability. Importantly, the database accommodates non-standard vocabularies (INSPIRE Concepts) such as religion, income sources, and household size—variables essential for contextualizing mental health in African settings.

Results: Out of 14 datasets targeted for migration, 10 have been successfully transformed from the source to the staging database, representing over 163,000 individuals. The staging database allows for a preliminary understanding of the trends of the three outcomes of interest. The SD facilitates standardized ETL, has helped reduce data loss, and provides an intermediate layer for data visualization and harmonization quality checks. Automated quality assessments were run using the Data Quality Dashboard (DQD) in R, ensuring fidelity between source and staging database layers. This architecture also allows experimentation with HDSS-linked datasets and non-coded mental health concepts.

Conclusion: The INSPIREMH staging database represents a significant advancement in the standardization and integration of longitudinal mental health data across heterogeneous African contexts. It is more than a pre-processing step; it acts as a living metadata repository and research-ready platform that captures both clinical and socio-contextual variables. By combining the strengths of DDI Lifecycle, OMOP CDM, and a novel staging model, this architecture sets a precedent for future mental health data science pipelines in LMICs, ensuring both interoperability and local relevance.

Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI / 113

EcoCommons: Advancing Reproducible and Scalable Ecological Modelling with FAIR Data

Author: jenna wraith[1]

Co-authors: Abhimanyu Raj Singh [1]; Jo Morris [2]; Ryan Newis [1]; Xiang Zhao [1]

[1] *QCIF*

[2] *ARDC*

Corresponding Authors: r.newis@qcif.edu.au, abhimanyuraj.singh@qcif.edu.au, xiang.zhao@qcif.edu.au, j.wraith@qcif.edu.au, jo.morris@ardc.edu.au

EcoCommons is a national digital infrastructure purpose-built to advance responsible, reproducible, and scalable ecological modelling in the era of FAIR data. It enables researchers, policymakers, and environmental managers to access integrated datasets, run validated modelling workflows, and share reproducible outputs that can inform biodiversity conservation, climate adaptation, and land-use planning. Built through close collaboration between ecological scientists, infrastructure providers, and technical experts, EcoCommons addresses long-standing barriers in environmental modelling. Many researchers face challenges in accessing high-quality data, configuring reproducible workflows, or building interoperable models at scale. EcoCommons responds to these challenges with a robust, cloud-based platform that brings together open data, scalable
compute environments, and a curated suite of Jupyter notebooks. These notebooks, developed by domain experts, allow users to undertake complex tasks such as species distribution modelling, climate projections, and ecological forecasting without requiring extensive programming skills.
The platform is designed to serve a broad user base—from advanced coders to those with limited technical experience—by offering guided workflows alongside flexible, customisable modelling environments. This dual approach supports both accessibility and sophistication, enabling individual researchers, government agencies, NGOs, and university educators to conduct modelling that is scientifically rigorous, transparent, and replicable. EcoCommons strongly aligns with FAIR data principles, embedding metadata standards and promoting data provenance throughout its workflows. The platform supports integration with external data repositories, including the Atlas of Living Australia, enabling researchers to connect multiple datasets from different sources and disciplines. By linking data from diverse origins—species observations, climate projections, environmental layers, and more—EcoCommons promotes interdisciplinary collaboration and enhances the quality and scope of ecological insights.
In addition to its technical capabilities, EcoCommons is committed to building a national community of practice around ecological modelling. It provides training resources, reusable code templates, best-practice guidance, and collaborative forums. This helps users improve not only their technical skills but also their understanding of the assumptions, limitations, and implications of different modelling approaches. EcoCommons is more than just a modelling tool—it is a research enabler that supports data-intensive, policy-relevant ecological science. By fostering FAIR data use, reproducible workflows, and cross-sector collaboration, EcoCommons is helping to future-proof Australia's environmental decision-making processes. As pressures on ecosystems intensify, platforms like EcoCommons will be critical for equipping the next generation of researchers and decision-makers with the infrastructure needed to generate robust, timely, and transparent ecological insights.

Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global / 114

Wildlife Observatory of Australia (WildObs): First National Infrastructure for Automated Wildlife Image Analysis

Author: Jenna Wraith[1]

Co-authors: Daraka Hewa Vithanage [1]; Jo Morris [2]; Matthew Luskin [3]; Renee Piccolo [3]; Zachary Amir [4]

[1] *QCIF*

[2] *ARDC*

[3] *University of Queensland*

[4] *TERN*

Corresponding Authors: daraka.hewavithanage@qcif.edu.au, z.amir@uq.edu.au, m.luskin@uq.edu.au, r.piccolo@uq.edu.au, j.wraith@qcif.edu.au, jo.morris@ardc.edu.au

The Wildlife Observatory of Australia (WildObs) is building the Australia's first national infrastructure dedicated to automated wildlife image analysis. Designed to process large volumes of camera trap data using artificial intelligence, WildObs provides the tools and infrastructure necessary to support scalable, standardised, and reproducible biodiversity monitoring across Australia's varied ecosystems. Camera traps are widely used across research institutions, government agencies, and conservation groups to monitor wildlife presence and behaviour. However, processing this data manually is labour-intensive and often inconsistent, limiting the utility of these datasets for broad-scale or repeatable analysis. WildObs addresses this challenge by applying advanced computer vision models to automate species identification from images, reducing bottlenecks in data processing and improving accuracy, consistency, and speed. The platform leverages cloud computing and reusable AI model pipelines to support the high-throughput processing of images at local, regional, and national scales.

WildObs also plays a critical role in connecting wildlife observation data with other national infrastructures. It is designed to integrate with repositories like the Terrestrial Ecosystem Research Infrastructure (TERN), the Atlas of Living Australia (ALA) and analytical platforms such as EcoCommons, allowing seamless flow from image capture to species detection, and onward to modelling and decision support. This end-to-end interoperability enables researchers and policymakers to derive greater value from ecological data, supporting initiatives like the Threatened Species Index and national reporting. Importantly, WildObs is a collaborative infrastructure that serves the broader environmental community. By engaging research institutions, conservation NGOs, Indigenous rangers, and land managers, the initiative promotes knowledge-sharing and encourages participation in the co-development of training datasets and detection models. Its open and extensible design supports customisation to regional contexts and species groups, enhancing its relevance and scalability.

WildObs is a foundational infrastructure that transforms fragmented and underutilised image data into a national resource for science and policy. It improves the visibility, usability, and consistency of wildlife data across jurisdictions, enabling evidence-based decisions in biodiversity conservation and land management. By connecting on-ground monitoring efforts with digital analysis pipelines and national-scale reporting, WildObs exemplifies the kind of research infrastructure needed to support data-intensive science—locally and globally. It represents a significant step forward in the use of AI and interoperable systems for environmental resilience and ecological knowledge generation.

Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data / 115

Interoperability in Practice: Integrating Natural History Collections with Modern Ecological Data Streams

Authors: Andrew Young[1]; Owen Forbes[1]; Peter Thrall[1]

[1] *CSIRO*

Corresponding Authors: peter.thrall@csiro.au, owen.forbes@csiro.au, andrew.young@csiro.au

Natural history collections are increasingly recognised as critical infrastructure for addressing complex ecological challenges, yet their full potential can only be realised through strategic integration with other ecological data streams. While FAIR principles and the Digital Extended Specimen concept provide theoretical frameworks for data integration, practical implementation requires navigating real-world constraints and trade-offs. This presentation explores two case studies demonstrating pragmatic approaches to collections data integration, highlighting both opportunities and challenges in building interoperable biodiversity data assets.

The first case study demonstrates integration of historical specimen records with climate projection data to forecast flowering phenology responses in Australian Acacia species. By combining specimen data, historical climate records, and phylogenetic information with CMIP6 climate projections,

we model potential shifts in flowering patterns under different climate scenarios (high- and low-emissions Shared Socioeconomic Pathways) at 2050 and 2100 horizons. This integration of specimen-based spatiotemporal data with climate projections provides insights into potential ecosystem-level impacts of phenological changes, while highlighting technical challenges in integrating collections datasets and other ecological data layers with projected outputs from modern climate models.

The second case study examines the integration of specimen data from Australian herbaria with the AusTraits database, focusing on practical strategies for linking specimen records with trait measurements. Rather than pursuing universal completeness in metadata and vocabulary overlap, this work emphasises feasible approaches that maximise utility for diverse research applications while maintaining manageable data standards. We discuss key decision points in balancing comprehensive coverage against practical constraints, and present solutions for creating flexible, interlinked data resources that serve multiple research needs.

These case studies illustrate how thoughtful integration of collections data with other ecological data streams can unlock new research capabilities while adhering to FAIR principles. We present practical lessons learned about balancing ideal standards against operational realities, and share suggestions for similar integrative studies. Our experiences demonstrate that successful data integration need not achieve perfect universality to deliver significant scientific value, suggesting pragmatic pathways for expanding the utility of natural history collections in modern ecological research.

Presentations Session 8: Policy and Practice of Data in Research; Data, Society, Ethics and Politics / 117

Connecting researchers to the Australian data linkage landscape through institutional investment and communities of practice

Author: Nadine Andrew[1]

Co-authors: Dianne Brown [1]; Komathy Padmanabhan [1]

[1] *Monash University*

Corresponding Authors: dianne.brown@monash.edu, nadine.andrew@monash.edu, komathy.padmanabhan@monash.edu

Situation
Recent technological progress has significantly enhanced our capacity to link person-level data across diverse sectors for research in Australia. Key advancements include: legislation to facilitate data access and availability, streamlined governance processes, enriched metadata, more efficient data linkage, and enhanced statistical methods and training. Collectively these advancements have created significant opportunities to maximise data sharing and expand the utilisation of existing research data across multiple domains but not without creating additional complexities for researchers.

Task
Monash University has long been recognised for its leadership in Health, Social, Epidemiological, and Translational research. Monash supports over 43 registries, 700 clinical trials, and 47 cohort studies, each leveraging complex, multimodal, large, and often sensitive datasets with many linked to administrative health datasets that exist at multiple levels of government in Australia. However, many of these research activities have been occurring in isolation. In response to the need for internal capacity building and increased potential to leverage national and global advancements in population data linkage, Monash University has significantly invested in data infrastructure, networks and highly skilled people to support best practice in data linkage and population research.

Action
In 2019, Monash University established M-Link - a community of practice to advance data linkage capabilities across all schools and faculties. An initial multidisciplinary working group was established –covering expertise in data management, data governance, privacy, data engineering, statistics and epidemiology. The overarching objective of the group was to develop a comprehensive roadmap for

expansion to a university-wide community of practice. In parallel, Monash invested in infrastructure support through the establishment of a Monash hosted Trusted Research Environment.
As these activities progressed it became increasingly clear that the many challenges could not be addressed solely at an institutional level and that collaborative approaches across institutions and with government funded bodies were required. Advocacy for data linkage included submissions to government agencies on availability of their data and culminated in the establishment in 2023 of a Monash DATA (Data Availability and Transparency Act) scheme accreditation working group to provide collective leadership for Monash's accreditation as a data user
for Australian Government data, resulting in Monash being one of the first institutions to be accredited under the scheme . In addition, events have been organised in response to member surveys every six months with leading researchers and data custodians to share knowledge and increase exposure of data linkage capacity at the University.

Results
Over the past six years, M-Link has successfully cultivated a thriving community of practice at the institutional level, growing from a working group of 10 volunteers to a community of over 130 researchers. Activities undertaken by Monash at the institutional and national level supported Monash to be one of the first academic institutions to become an accredited data user under the DATA scheme. The expansion of M-Link into the national landscape led to 143 researchers from multiple institutions registering for a national co-hosted event. Activities revealed that while government data custodians, infrastructure providers, and researchers acknowledged the value of advancements in linkage technology and data availability, these came with new challenges such as navigating multiple TREs, rising data provision costs and the enduring issue of timeliness.
The workshop highlighted that by fostering knowledge exchange and promoting best practices in data linkage across Australia, initiatives like M-Link serve as crucial connectors between researchers and government entities providing valuable insights into the research community's perspectives on service delivery models. This has positioned M-Link to drive the conversation within the University to promote collaborative approaches to extend data use and fully leverage research data's value.

Presentations Session 2: Data and Research & Data Science and Data Analysis / 118

Creating an integrated Electronic Health Record data platform for revolutionising healthcare research

Author: Nadine Andrew[1]

Co-authors: Richard Beare [1]; Velandai Srikanth [1]

[1] *Monash University*

Corresponding Authors: richard.beare@monash.edu, nadine.andrew@monash.edu, velandai.srikanth@monash.edu

Background
The digitalisation of Electronic Health Record (EHR) data has unlocked unique opportunities for research. Unlike administrative datasets, EHRs provide granular clinical data, real-time updates within systems, and access to detailed clinical notes. Despite these advantages, EHR data—primarily collected for operational purposes—remains siloed, lacks standardisation between systems, suffers from poor interoperability, and contains large amounts of unstructured text. Consequently, EHR data is often not curated to the standard required for research, which hampers its optimal use in healthcare studies. The National Centre for Healthy Ageing (NCHA), a partnership between Monash University and Peninsula Health (comprising four hospitals and over ten outpatient and community services), has developed a unique EHR-derived Data Platform. Its primary objective is to integrate multi-site healthcare data from an entire geographic region to support translational research focused on healthy ageing across the life course.

Methods
Our approach involved establishing a core set of EHR data suitable for research from the sole public health provider within a defined geographic region. Relevant items were identified based on published literature and consensus processes, then curated within a specialised research data warehouse.

The curation process included data validation, quality assessment, internal linkage using a patient data spine, and data harmonisation/merging. Semi-automated data extraction processes were developed for approved research projects. End-users were engaged in defining the platform's content, and consumer workshops were conducted to understand community perspectives on data management and governance. Further efforts to expand the platform's content included implementing an AI pipeline to extract concepts from clinical notes, routinely collecting Patient Reported Outcome Measures, and linking to a variety of local, state, and national datasets (e.g., primary care, medication, aged care, hospital, and mortality datasets). Publicly available datasets were scoped for inclusion, and collaborations with local councils were established to incorporate community data.

Results
The platform's research data warehouse contains curated data for over one million patients, collected over ten years and updated weekly. A total of 131 core data items from 11 research-relevant datasets across the four hospitals and ten community/outpatient services have been identified for inclusion. Data access, extraction, and release processes are guided by the Five Safes Framework. A natural language processing (NLP) pipeline has been implemented and trained to detect dementia in clinical notes. A framework for the routine collection and integration of patient-reported outcome measures has been developed. Linked state and commonwealth data for 179,089 residents aged >60 years (January 2010–May 2021) has been obtained, achieving a linkage accuracy of 98.4%. Environmental (greenery, air pollution, walkability) and Census data have been incorporated at the neighbourhood level (Statistical Area 1). Over 50 use-case projects have tested data access, extraction, and release into a Monash-hosted Trusted Research Environment, covering topics such as dementia, residential aged care, medication use, homelessness, environmental impacts on health, ageing, and health service redesign.

Implications
The NCHA Data Platform provides an international exemplar for leveraging linked EHR data to advance population health research. In addition to delivering high-quality, research-grade data to clinicians and researchers, the platform serves as critical infrastructure to underpin data-driven innovation across multiple domains.

120

Digital Research Infrastructure Supporting FAIR, Reproducible and Impactful Research: A Global Ecosystem of Tools, Resources and Skills

Authors: Jeff Christiansen[1]; Susanna-Assunta Sansone[2]; Wolmar Nyberg Åkerström[3]; Ishwar Chandramouliswaran[4]; Fabio Liberante[5]

[1] *Australian BioCommons*

[2] *ELIXIR Interoperability Platform; ELIXIR-UK; University of Oxford*

[3] *NBIS - National Bioinformatics Infrastructure Sweden, SciLifeLab, Uppsala university*

[4] *NIH Office of Data Science Strategy*

[5] *ELIXIR Hub*

Corresponding Authors: ishwar.chandramouliswaran@nih.gov, fabio.liberante@elixir-europe.org, susanna-assunta.sansone@oerc.ox.ac.uk, wolmar.n.akerstrom@uu.se, jeff@biocommons.org.au

As the volume and complexity of research data continue to grow, researchers increasingly rely on robust digital research infrastructure and interoperability across regional and disciplinary boundaries to achieve FAIR and reproducible data, and impactful results. Building new collaborations and extending the reach of digital research infrastructure initiatives will be crucial in fostering a global ecosystem of tools, resources, and skills that supports the inherently international and collaborative nature of research. This session presents examples from infrastructure providers, their value and use in practice, and also serves as an invitation and call to action for researchers, data managers, librarians, and IT professionals involved in supporting research data management and infrastructure across any discipline to consider their use and adoption, and/or engage in their development.

This session aims to showcase how some core resources and services assist research and professional communities to effectively manage, share, and reuse research data and other digital research assets, such as computational workflow descriptions and training materials, in a FAIR-enabling manner. The session will also present a showcase of successful collaborative efforts between Australian, European, and North American groups, highlighting some of the core challenges that researchers face, and corresponding digital research infrastructure systems designed to address them. The systems covered also illustrate how cross-disciplinary adoption have served to increase the utility and sustainability of solutions that began their journeys in support of biological research and have since been successfully redeployed, expanded and utilised in support of a diverse spectrum of research domains and disciplines.

The core challenges these resources and services aim to address are the lack of standardised data management practices, the fragmented landscape of informatics and data science training resources, the difficulty of navigating the thousands data and metadata standards essential to data stewardship, and the complexity of capturing and sharing computational workflows in a way that ensures reproducibility and discoverability. The session highlights four systems and their role in addressing the above-mentioned challenges:

- (1) The *Data Stewardship Wizard (DSW)* –a tool that facilitates the creation of machine-actionable FAIR-enabling data management plans.
- (2) *FAIRsharing* –a registry, a service and an educational resource to discover and use the right standards and the appropriate repositories, enabling a number of data management tasks.
- (3) The *Training eSupport System (TeSS)* –a comprehensive platform for discovering and accessing bioinformatics training events and materials, fostering skills development and knowledge dissemination.
- (4) *WorkflowHub* –a system for describing and sharing rich descriptions of computational workflows, enabling researchers to document, share, and reuse complex analysis pipelines, enhancing transparency and reproducibility.

The session combines a series of short presentations with questions and answers leading into a structured discussion as follows:

- *Introduction: Challenges and Collaborations* (10 minutes):
  An overview of the challenges in biological data management and workflow sharing, emphasising the global context and the importance of globally interoperable Digital Research Infrastructure, and efforts and collaborations across Europe, Australia and North America to address these challenges.
- *Showcase: Systems, Case Studies and Global Impact* (4 x 15 minutes):
  An introduction to the DSW, FAIRsharing, TeSS, and WorkflowHub, illustrating their features, functionalities, current and potential applications, as well as connections among these resources that, where relevant, are powered by each other's content. Highlighting real-world applications of these tools and infrastructure, showcasing their impact on research outcomes and data management practices across the globe.
- *Discussion: Future Directions and Wider Adoption* (20 minutes):
  A facilitated discussion on the future development and adoption of digital research infrastructure, with a focus on fostering collaboration and expanding the user community.

Speakers from partner institutions in Australia, Europe, and North America, each a leading expert on their respective topic, will present and participate in the discussions:

- The implementation of the DSW for machine-actionable data management plans within the European research landscape.
- The development and utilisation of FAIRsharing by humans and machines.
- The development and utilisation of TeSS for bioinformatics training resource discovery across Europe and how it has been modified to develop Digital Research Skills Australasia (DReSA), to support data science training across all disciplines.
- The development of WorkflowHub and its role in enhancing computational reproducibility.

We are excited to present our work, build new collaborations and engage with the broader data community at International Data Week 2025.

Presentations Session 9: Empowering the global data community for impact, equity, and inclusion / Education / 121

Data Science Education Across Academic Disciplines: A Comprehensive Approach to Campus-wide Integration

Author: Daphne Raban[1]

Co-author: Niv Ahituv [2]

[1] *University of Haifa, CODATA Israel NC*

[2] *Tel Aviv University*

Corresponding Authors: ahituv@tauex.tau.ac.il, draban@univ.haifa.ac.il

Data Science—encompassing data collection, storage, analysis, visualization, and interpretation—has become a foundational pillar of modern research, decision-making, and innovation. Its influence extends across scientific, social, industrial, and governmental domains, transforming both methods and outcomes. This contribution discusses the importance of integrating Data Science education across all academic disciplines, recognizing its multidisciplinary nature and universal relevance. While core Data Science programs remain essential, a broader educational approach is needed to equip all students—of every academic discipline—with basic data literacy, data search and retrieval, analytical skills, and ethical awareness. The Israeli National Academy of Sciences and Humanities has established a committee to evaluate and promote this vision, in coordination with national academic policy bodies. Central to this initiative is the "data cycle"model, a generic framework for data-informed inquiry and decision-making, applicable across disciplines. The report proposes differentiated teaching objectives for three student groups: Data Science majors, students in data-adjacent fields, and the broader student body (where, in addition to a fundamental section of Data Science course, a discipline-oriented section will be offered). This framework aims to foster informed citizenship, critical thinking, and workforce readiness in an increasingly data-driven world. Our findings and recommendations contribute to global discussions on embedding Data Science into higher education systems.

The National Academy report suggests a universal approach to data handling entitled The Data Cycle. The data cycle offers a structured framework for addressing data-driven questions, spanning six key stages: problem definition, data collection, cleaning and integration, analysis, visualization, and drawing conclusions. Each stage requires distinct computational and ethical competencies, from sourcing and validating data to applying statistical, AI, and machine learning methods. Technological tools—from qualitative design software to advanced programming environments—support this process, with differentiated access for students based on background and field. The cycle's iterative nature reinforces the evolving relationship between data, knowledge, and decision-making. Ethical and legal considerations, including privacy, bias, transparency, and replicability, are integrated throughout, emphasizing critical thinking as a core component of data education. This presentation outlines a pedagogical model designed to equip all students—not only Data Science majors—with the skills and sensitivity needed for responsible and effective data use across academic and professional domains in an immersive information world.

This initiative outlines a modular and interdisciplinary framework for integrating Data Science education across all academic disciplines. Recognizing the varying needs of fields such as humanities, law, engineering, exact sciences, and life sciences, the proposed model begins with a universal core course introducing the data cycle, ethical considerations, basic tools, and types of digital data. This is followed by discipline-specific modules that align Data Science concepts and tools with the methodologies and priorities of each field. Content delivery can take the form of standalone introductory courses, enriched existing courses, or a hybrid of both. Emphasis is placed on experiential learning through real-world data, collaborative team projects, and critical thinking. Partnerships with university Data Science centers and academic libraries are central to the program's success, as they provide

both technical expertise and infrastructure, including access to domain-specific datasets. Furthermore, the initiative recommends developing national data resources and supporting pre-university education to instill data literacy from an early age. (In fact, the Ministry of Education has established a three-year program entitled Data and Information in high schools to precede the academic studies in Data Science). This comprehensive approach not only equips students with essential skills for research and employment but also fosters cross-disciplinary communication and data-informed thinking—crucial for addressing today's complex societal challenges.

Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research / 122

OntoPortal –An Open Technology for Discipline-Specific Terminology Repositories

Author: Rafael S. Gonçalves[1]

Co-authors: Jennifer L. Vendetti [1]; Alex Skrenchuk [1]; Michael Dorf [1]; Syphax Bouazzouni [2]; Clement Jonquet [3]; Mark A. Musen [1]; OntoPortal Alliance members

[1] *Stanford Center for Biomedical Informatics Research (BMIR), School of Medicine, Stanford University, Stanford, USA*

[2] *Laboratory of Computer Science, Robotics and Microelectronics of Montpellier (LIRMM), Montpellier, France*

[3] *French National Research Institute for Agriculture, Food and Environment (INRAE), Paris, France*

Corresponding Authors: alex.skrenchuk@stanford.edu, mdorf@stanford.edu, gs_bouazzouni@esi.dz, vendetti@stanford.edu, rafael.goncalves@stanford.edu, musen@stanford.edu, clement.jonquet@inrae.fr

Background and Motivation

Controlled vocabularies and ontologies are essential for enabling data interoperability, discovery, and integration across domains. Repositories that host and expose these artifacts play a critical role in implementing the FAIR (Findable, Accessible, Interoperable, Reusable) principles. The OntoPortal Alliance, a collaboration of academic and commercial partners, supports the development and deployment of terminology repositories through a shared open-source platform. Originally derived from the National Center for Biomedical Ontology's BioPortal resource (https://bioportal.bioontology.org), OntoPortal has evolved into a modular, extensible, domain-independent solution for managing collections of discipline-specific controlled terminologies. The system has been adopted by diverse research groups internationally to archive and manage terminologies in biomedicine, agronomy, ecology, biodiversity, earth science, materials science, and other domains. The groups that have deployed OntoPortal to provide terminology services in these disciplines form a network of linked repositories, collaborating as part of an affiliation called the *OntoPortal Alliance*.

Despite the increasing importance of ontologies and controlled terminologies in implementing FAIR data practices, many communities lack sustainable, adaptable infrastructure to support the lifecycle management of such resources. Most existing solutions are either overfitted to other research areas, not reusable, or fail to interoperate effectively. OntoPortal addresses this gap by offering a unified, open-source foundation for building and federating terminology repositories across diverse scientific disciplines.

System Overview

OntoPortal provides a full-featured repository infrastructure for publishing, discovering, and interacting with ontologies—both interactively and programmatically (Figure 1). The system supports uploading and curating terminologies and ontologies, searching across the uploaded content, browsing terms, and annotating data using terminology-based services. OntoPortal handles multiple representation formats (viz., OWL, SKOS, RDF, OBO) and it provides multilingual support, metadata enrichment, version tracking, mappings among terms, and usage metrics. The platform exposes RESTful APIs and a web-based interface for both end users and terminology developers.

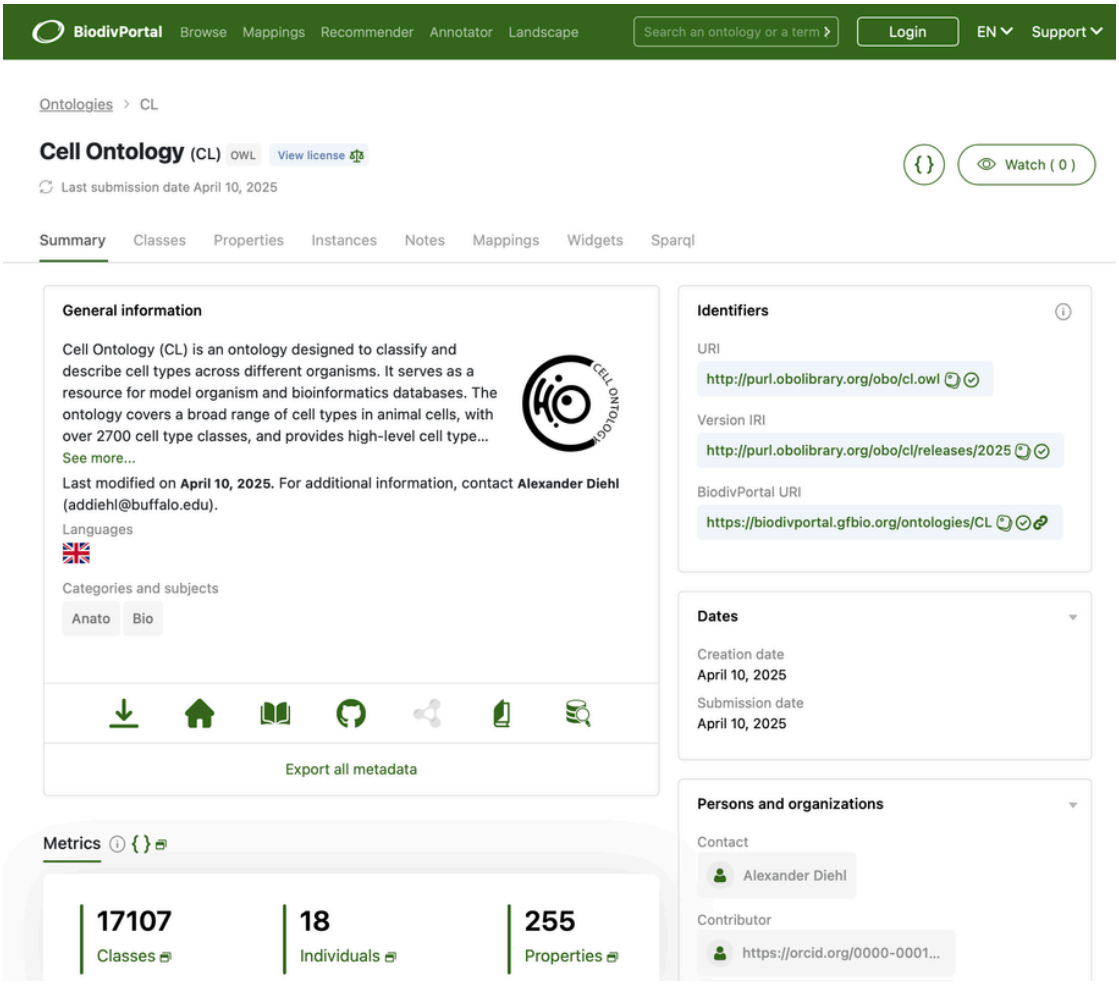

Figure 3: Figure1

Figure 1: Landing page of the Cell Ontology in the BiodivPortal implementation of OntoPortal.

OntoPortal's architecture includes backend services supporting RDF triple-store integration and indexing, and frontend components for user interaction. Management of terminology metadata is a central feature, and OntoPortal supports configurable metadata schemas including MOD (Metadata for Ontology Description). Term mappings, annotation services, change tracking, and user-generated content are integral to the system. Each deployment can adopt a flexible editorial policy, ranging from community-contributed catalogues to curated domain-specific repositories.

Recent work by OntoPortal Alliance members led to a major milestone: a federated release between AgroPortal, EcoPortal, EarthPortal, and BiodivPortal. This release will feature shared browsing across the four platforms, integrated search, and metadata harmonization using a common classification system (the UNESCO Thesaurus), serving as a foundation for further technical alignment.

In other recent work, BioPortal upgraded its platform stack, improved multilingual search, and added API functionality for remote terminology pulls. AgroPortal introduced features such as a URI management service and embedded SPARQL editor. EcoPortal implemented a multilingual UI overhaul and structured documentation, while EarthPortal deepened connections with research infrastructures. BiodivPortal expanded its terminology collection, prototyped version diff views, and began exploring LLM integration for enhanced annotation. The OntoPortal Alliance will promote the sharing of these new features across all repositories in the consortium.

OntoPortal provides a shared documentation platform, accessible via ontoportal.github.io. It offers a modular system allowing each portal to present local variations while maintaining a common documentation structure.

For the data community, OntoPortal offers a proven model for scalable terminology repository de-

velopment, extensible tooling, and metadata interoperability grounded in real-world deployments. The work contributes a blueprint for how FAIR-aligned infrastructure can be adapted and federated across scientific domains.

Deployment and Applications

OntoPortal powers a growing number of public repositories across disciplines. These include BioPortal for biomedical ontologies, AgroPortal for agri-food data, EcoPortal for ecological sciences, EarthPortal for Earth science, and BiodivPortal for biodiversity research. Additional deployments, such as MatPortal and IndustryPortal, demonstrate the software's adaptability. Newer initiatives include OntoPortal-Astro for astronomy, CHPortal for cultural heritage, and LovPortal for Semantic Web vocabularies. Each portal deployment group tailors the software to its community's needs while contributing to shared development across the OntoPortal Alliance.

Key Features

OntoPortal enables users to publish and explore controlled terminologies with rich metadata, track terminology evolution, and reuse terminological content across applications. Mappings of terms across vocabularies help users to identify conceptual relationships, while an Annotator service supports terminology-based text processing. The system provides usage statistics, API access, and community interaction features such as user notes and change proposals. Federation capabilities allow OntoPortal instances to interoperate, sharing artifacts and metadata while retaining local autonomy.

The OntoPortal codebase is open-source and maintained on GitHub. Researchers, developers, and infrastructure projects are invited to deploy new instances, contribute enhancements, or join the OntoPortal Alliance. Our goal is to make ontology repository development as accessible and collaborative as the web itself.

Conclusion

OntoPortal provides an adaptable platform for managing controlled terminologies across disciplines, supported by a growing community of engaged investigators who are tailoring the system for use in a wide range of research domains. The OntoPortal Alliance is reducing the technical and organizational barriers to building sustainable, FAIR-aligned terminology infrastructure. By supporting a federated model with shared tooling and community governance, the OntoPortal platform enables interoperability across domains and the broader Semantic Web ecosystem. As more communities adopt OntoPortal, its role as shared infrastructure continues to expand—lowering barriers to terminology reuse, accelerating adoption of FAIR principles across diverse domains, and offering a practical foundation for semantic services that support open science, linked data, and AI integration.

Presentations Session 2: Data and Research & Data Science and Data Analysis / 124

Data-Driven Risk Identification in Supervision Reports of the Ministry of Health

Author: Avital Zadok[1]

Co-author: Daphne Raban [2]

[1] *University of Haifa*

[2] *University of Haifa, CODATA Israel NC*

Corresponding Authors: draban@univ.haifa.ac.il, avitalz21@gmail.com

In response to inefficiencies in governmental regulation, such as excessive regulation, an overabundance of laws and procedures, lack of flexibility, and disregard for the costs, countries around the

world began efforts to optimize regulation, in part by privatization. The trend toward privatization of social services necessitated substantial development of governmental supervision practices. Risk management in regulation emerged as an efficient approach to supervising public sector services, assisting regulators in deciding the extent of intervention necessary to prevent harm to the public interest and to ensure that service recipients are protected and safe. Today, risk management in supervision is a critical component of decision-making processes under conditions of uncertainty and is recognized as one of eleven principles of best practices in supervision and enforcement by the OECD.

This study explores the potential of artificial intelligence (AI) in identifying and categorizing risks from unstructured open text, using advanced natural language processing (NLP) architectures such as Dicta and HeBERT. The research aimed to develop a methodology for analyzing supervision reports from the healthcare sector, enabling risk detection and classification into predefined categories. The study's results indicate high performance of the Dicta model in identifying and classifying risks from unstructured text, achieving an accuracy of 93.3%, a recall of 85.9%, and an F1 score of 92.3%. In comparison, the HeBERT model yielded lower results across all metrics. In the multi-class classification task, Dicta also outperformed HeBERT, with an accuracy of 74.4% versus 65.1%, respectively. These differences were statistically significant ($p < 0.05$), underscoring the advantages of using Hebrew-adapted models, particularly those tailored to the healthcare domain.

The study highlights the critical role of semantic features and keywords in risk identification. It also addresses challenges associated with ambiguous sentences and overlapping categories, emphasizing the need for future research to develop multi-category classification algorithms. While Dicta showed superior performance in identifying key categories such as "Infrastructure, Equipment, and Logistics"and "Medical Services and Quality of Care,"HeBERT exhibited limitations in distinguishing mid-range categories, resulting in higher error rates.

The findings suggest practical applications for regulatory bodies, such as optimizing resource allocation, enhancing decision-making through data-driven insights, and improving transparency and service quality. Despite its promising results, the study acknowledges limitations, including the reliance on a single corpus of healthcare supervision reports and the constrained sample size. Future research should expand the corpus and explore AI techniques for less structured texts.

This research provides a foundational framework for applying AI to risk detection in healthcare and other domains, offering valuable insights for improving supervision, monitoring, and service delivery.

Poster Session / **127**

Data Management Plan (DMP) –From FAIR to FAIRER

Authors: Su Nee Goh[1]; Willie Koh[2]

[1] *Nanyang Technological University*

[2] *Nanyang Technological University, Singapore*

Corresponding Authors: williekoh@ntu.edu.sg, sunee@ntu.edu.sg

Many universities have adopted the use of Data Management Plans (DMPs) for research teams to outline how their research data will be handled both during and after a project. DMPs support responsible data management in accordance with the **FAIR principles**: Findable, Accessible, Interoperable, and Reusable. The objectives are to contribute towards research integrity, reproducibility, and efficient reuse.

At the Nanyang Technological University (NTU), Singapore, the requirement for a DMP was introduced in 2016. This was purposefully integrated into the university's research grant management system, so that Principal investigators (PIs) must complete and submit their project-specific DMPs within the system. Completion of their DMP is a prerequisite for financial account creation before PIs can access their grant funding.

NTU's DMP initiative began with the aim of supporting open science by encouraging researchers to plan for the sharing of their research data, especially as this becomes an increasingly common requirement by funders and journals. This includes making non-sensitive research data available on NTU's open access research data repository, thereby enhancing research transparency and increasing impact. Researchers who demonstrated exemplary practices of making their scientific contents,

tools, and processes open and accessible are eligible to receive the NTU / Singapore Open Research Awards, which were presented in 2022 and 2024.

However, the research data management landscape within NTU has evolved significantly in recent years. This shift was driven primarily by internal factors such as internal audits, as well as external forces including geo-political developments and increased scrutiny on data security. Data protection issues are particularly pressing in collaborative projects involving commercial companies or governmental agencies. These have led researchers to resist data sharing, citing data protection concerns as a reason to withhold data sharing. This growing tension between the imperative to make research data open and reusable, and the need to protect sensitive data, highlights the need for a balanced and strategic response.

In light of this, NTU is revamping its DMP template to better support researchers in making their data FAIRER, where '**E**'**stands for** '**Ethical**' and '**R**'**stand for** '**Responsible**'. The new template will not only guide researchers in meeting FAIR data principles, but also to help them fulfil their ethical and institutional responsibilities under the University's new institutional data security framework. For example, the revised DMP questionnaire will direct researchers to using appropriate University-supported research data storage solutions for sensitive research data.

We present here our university's journey towards implementing a **FAIRER DMP** that reflects both openness and the evolving responsibilities of data stewardship.

Presentations Session 6: The Transformative Role of Data in SDGs and Disaster Resilience / 128

Utilizing Health Data for Malaria Surveillance and Prompt Response: Experience from Karenga District, North-Eastern Uganda

Author: Mallo Paul Lokiru[1]

[1] *Karenga District Local Government*

Corresponding Author: paullokirumallo@gmail.com

Introduction: The transformative role of health data in achieving SDG 3 cannot be overstated, as data remains "the lifeblood of public health"(Ghebreyesus, 2019). The significant role of data in promoting the Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-being), and improving disaster resilience continues to gain attention worldwide. Globally, in 2023, the number of malaria cases was estimated at 263 million, with an incidence of 60.4 cases per 1,000 population at risk, and the number of deaths was estimated at 597,000, with a mortality rate of 13.7 per 100,000 (World Malaria Report, 2024). Additionally, the same report indicates Uganda as the third country in Africa with the third-heaviest estimated burden of malaria cases at 5%. In a district like Karenga in the Karamoja Sub-Region of northeastern Uganda, where malaria is the top cause of morbidity and mortality, having access to timely and accurate data can significantly enhance health outcomes by enabling early detection.
Methodology: This Quality Improvement (QI) project started, May 2024, focused on a district-wide initiative aimed at collecting and analyzing malaria data from all public health facilities in Karenga District through the Health Management Information System (HMIS). Working closely with the Health Facility Incharges (HFIs) and the Health Information Assistants (HIAs) to ensure timely and complete submission of reliable data, the data gathered included weekly outpatient attendance, suspected malaria cases, malaria test positivity rates, and essential medicines and supplies. Data analysis was carried out with the aid of the inbuilt capabilities within the DHIS2, and geo-spatial maps indicating low- and high-risk areas were created or developed.
Results: The value of data lies not just in its collection but in its application for real-time decision-making (Gawande, 2010), as this case study shows. This initiative has resulted in several important outcomes:
Improved Reporting Timeliness and Accuracy: Since early 2024, the percentage of timely HMIS

reporting has improved from 40% in May 2024 to 82.5% in March 2025, thanks to the tailored mentorship and feedback based on data, conducted by the District Health Team (DHT).
Identification of Malaria-Concentration Areas: Hotspot mapping using GIS technology has shown that three Sub-Counties of Karenga Town Council, Lobalangit Sub County, Kakwanga Sub County, and Sangar Sub County registered between 83 and 133 confirmed malaria cases, while the remaining seven Sub-Counties or Town Councils registered between 13 and 84 malaria cases. This finding led to the targeted implementation of community-level outreach for mass malaria testing and treatment in these areas.
Community-Driven Initiatives: The access to real-time data spurred quick community actions, such as health education, door-to-door distribution of mosquito nets, and routine entomological surveillance efforts organized by the district and some of the implementing partners within the district, including Doctors With Africa (CUAMM). This data-driven strategy enabled the prioritization of interventions based on incidence levels and transmission risk.
Advocacy for Resource Allocation: The District Health Team began using data visualizations in their discussions with district and national stakeholders to lobby for more resources and funding for malaria control and elimination efforts. This initiative resulted in Karenga being identified as a high-priority district to host a national event, "Walk Against Malaria,"on 10th April 2025, including other control activities such as malaria vaccination.

Enhanced Disaster Resilience: Ensuring local ownership of data is critical for sustainability and accountability in rural contexts (Alkire, 2015). By having the community and district-level teams consistently map health and environmental data, the district was able to proactively plan, budget, and direct resources to highly vulnerable areas during times of heavy rainfall, effectively preventing potential malaria upsurges. This integration of health and weather data marked a significant step forward in our disaster risk management.
Conclusion: This case study from Karenga District illustrates how localized, democratized, and contextualized data can serve as a significant catalyst for change, even in underserved areas. Key takeaways highlight the necessity of engaging the community, enhancing the data literacy of frontline health workers, and utilizing geospatial mapping for effective disease surveillance. Innovative concepts emerging from this study include: integrating community-level data from the Village Health Teams (VHTs) into the data value chain to improve data credibility and foster trust; testing a real-time malaria alert system via SMS to inform health teams when malaria hotspots exceed critical levels; and combining health and climate data dashboards at the district level to facilitate proactive measures. The value of data lies not just in its collection but in its application for real-time decision-making (Gawande, 2010), a key lesson adopted by district health teams and political leaders in Karenga. To sustain this initiatives, it is essential for government and development or implementing partners to invest in the infrastructure needed for data management, encourage multi-sector collaborations, and cultivate a culture of data utilization across all levels of administration and management.

Presentations Session 4: Data Stewardship / 129

A FAIR compliance review of a major open, biological data repository in Korea

Authors: Wonsik Shim[1]; Haeyoung Jeong[2]

[1] *Sungkyunkwan University*

[2] *Korea Bioinformation Center (KOBIC)*

Corresponding Authors: wonsik.shim@gmail.com, hyjeong@kribb.re.kr

We can state the following on the sharing of research data. Policy shapes awareness, repositories promote practices and culture shapes practice. The existence of reliable and robust data repositories across disciplines, institutions, and geographical areas is a crucial step in the data sharing process. One of the world leaders in research investment and production is South Korea, which ranks second globally in terms of R&D investment as a percentage of GDP in 2023 (4.96%) behind Israel. In 2022, Korea accounted for 3.6% of global research article production, placing it at number 12. It is not doing enough to promote open science and the sharing of research data, despite the fact that it actively performs research. South Korea has just 13 repositories listed by re3data.org, a global register of research data repositories, while Australia has 107, the Netherlands has 82, and Japan has 65. But a

recent event indicates that the nation is moving in a different direction.
A number of revisions to the Act on the Acquisition, Management, and Utilization of Biological Research Resources, passed by the Korean government in 2009, led to the launch of the Korea BioData Station (K-BDS), operated by the Korea Bioinformation Center (KOBIC), in December 2022. KOBIC is a part of the Korea Research Institute of Bioscience and Biotechnology, which is run by the government. A wide variety of biological datasets, such as those related to genomics, proteomics, metabolomics, imaging, chemical compounds, and clinical trials, are currently stored and shared throughout the nine open, reliable, full-scale data repositories that make up K-BDS. While more than two million records are still accessible to the public, private access is permitted in accordance with security and privacy requirements. K-BDS offers standardized data submission, curation services, and limited analytical capabilities. K-BDS databases provide an OpenAPI based on XML for external organizations looking to search and get its metadata and research data. Korea Research Institute of Science and Technology Information (KISTI) provides large-scale computing resources and a venue for exchange to promote collaboration among researchers as part of K-BDS. K-BDS and related data repositories are a major step forward in the sharing of biological data in Korea and around the world, and they are driving the rise of data repositories, which are becoming crucial elements for data-powered knowledge discovery and innovation.
The establishment of K-BDS represents a significant milestone in Korea's journey toward open science and data-driven research. This initiative addresses a critical gap in Korea's research infrastructure, particularly considering the country's substantial investments in R&D and its significant contribution to global scientific literature. By creating a centralized platform for biological data sharing, Korea is positioning itself to maximize the impact of its research outputs and foster greater international collaboration.
What makes K-BDS particularly noteworthy is its comprehensive approach to data management. Beyond mere storage, the platform incorporates sophisticated mechanisms for ensuring data quality, interoperability, and usability. The standardized submission protocols and curation services help maintain consistency across diverse datasets, while the OpenAPI functionality enables seamless integration with external systems and applications. This technical infrastructure is crucial for realizing the full potential of shared research data.
The two main parts of the proposed presentation will be providing an overview of K-BDS as open research data repositories and reporting on the FAIR compliance evaluation study.
The first part of the presentation will cover the following:
·the legal and policy frameworks of the K-BDS establishment;
·data registration, quality control, search & retrieval, data use and analytics;
·data standardization and metadata standards; and
·continuous enhancement and the launch of Korea's National Bio Big Data Project, a five-year initiative (2024-2028)
The K-BDS's FAIR compliance assessment will be the main topic of the presentation's second section. The FAIR principles served as a general guide for the development of K-BDS; a rigorous compliance evaluation may reveal both its strengths and weaknesses. The FAIRplus DataSet Maturity Model, designed for the biological sciences, serves as the foundation for the assessment. In close contact with the KOBIC staff in charge of running K-BDS, an external evaluation team carries out the compliance review. We will discuss the evaluation's implications for Korea's data repositories' future development.
This evaluation is particularly timely as Korea embarks on its National Bio Big Data Project. The findings will not only inform the ongoing development of K-BDS but also shape the broader strategy for research data management in Korea. By identifying areas of strength and opportunities for improvement, the assessment will help ensure that Korea's investments in data infrastructure yield maximum scientific and societal benefits. Moreover, the lessons learned from this evaluation could provide valuable insights for other countries seeking to enhance their own research data repositories, particularly those with similar profiles of high research output but limited data sharing infrastructure.

Presentations Session 8: Policy and Practice of Data in Research; Data, Society, Ethics and Politics / 130

Public Trust, Literacy and Health Data Foundations in Canada

Author: Kim McGrail[1]

Co-authors: Catherine Street [2]; Jannath Naveed [2]; Julia Burt [2]

[1] *UBC*

[2] *Memorial University of Newfoundland*

Corresponding Authors: cstreet@mun.ca, kim.mcgrail@ubc.ca, juliaburt@mun.ca, jnaveed@mun.ca

Public trust in governments, organizations and institutions that collect, protect, share and use health data is critical. Health data refers to information that describes a person's health, their health care or anything about their health status or condition. It can be about individuals (personal health data or information) or about populations (population health data). Access to timely and reliable health data, by health care providers, systems and patients, is the foundation of providing high quality health care services to individuals and public health services to populations. Access to health data is also critical for health system planning, management, public health, evaluation, research and innovation. These uses often involve combining or linking data sets, and in the case of research and innovation, includes interests from the public, non-profit and private sectors.

Canada's constitutional federation gives power over large parts of public health and health care to provinces and territories. This creates a fragmented health data ecosystem. New federal funds for provinces and territories were negotiated with provinces and territories, in exchange for commitments to modernize health systems with standardized health data and digital tools. Bi-lateral agreements with provinces and territories were aligned with the Pan-Canadian Health Data Charter which outlines a shared vision for health data in Canada. The 10 principles of the Charter put people and populations at the centre of health data management. The Charter starts with principles of person-centred data design, inclusive data governance, and distinctions-based approaches to data sovereignty as defined by First Nations, Inuit and Métis Peoples, their governments, representatives and expert/technical organizations.

Building on the Charter, this paper on public trust and health data expands on these principles. It highlights important considerations for public trust and identifies the risks and benefits of data sharing as well as the protections interest holders identify to maximize benefits and minimize risks (noting that data sharing does not necessarily imply moving data, and that we use the term interest holders as an alternative to stakeholders). It includes considerations around the uses of health data in clinical care and patients'access to their own health data (often referred to as "primary uses"), as well secondary uses that include but are not limited to health system planning, assessment, evaluation, improvement, research and innovation.

This work is meant to help different levels of government and health system organizations work together to earn public trust in and understanding of health data. It can also be useful to other organizations, both inside and outside Canada, that hold and use health data.

We reviewed relevant grey (e.g., policy reports, working papers, guidelines) and peer-reviewed literature, held focus groups and interviewed key informants. Focus groups and interviews add perspectives from individuals who have expertise related to public trust, public engagement and health data protections, and those who have experience as users of healthcare and public health services. Opportunities for review and comment on key paper materials were included at several different points in the process, including two rounds of public consultations on the health data glossary.

We find that trust is complex, and public trust in health data collection, sharing and use can fluctuate depending on many factors, including the broader political and social context in which health systems operate. Data literacy should be viewed as one of several foundational elements that create the possibility of trust, alongside other important elements, such as transparency and public benefit. Trust comes from trustworthy practice and requires reciprocity. Organizations and institutions that collect, use and share health data need to demonstrate trustworthy practices that are guided by well-developed principles, including those identified in the Pan-Canadian Health Data Charter.

We identify six recommendations for trustworthy practice that include: ongoing, inclusive public engagement; reconciliation that includes respect and support for Indigenous data sovereignty; the alignment of health data use with public benefit; clear rules and supports for data sharing, access and use; technology standards for safe and seamless data sharing; and transparency, communications and reciprocal learning.

The multitude of interests involved in both primary and secondary uses of health data make clear that there is no simple one-size-fits-all solution for trustworthy practices. The processes used to govern data collection, protection, sharing and use must align with their intended purpose. They

will need to evolve over time, as data, technology, analytic approaches, and public expectations also evolve. Public trust must be earned and can easily be lost. Trust is based on relationships that are mutual and require ongoing attention.

Trustworthy practices should be transparent, and part of that transparency is showing accountability for both successes and failures in meeting the intended practices. Trust itself should be measured and monitored over time, as part of an ongoing dialogue between the public and the governments, institutions and organizations that have responsibility for health data.

Earning and keeping public trust is both a laudable and achievable goal. It will require changes in practice, including adopting new models of data stewardship and creating more interoperability in technologies, policies, standards, and relationships across organizations and geography. The principles of the Health Data Charter provide a common goal, and trustworthy practices can help earn the public trust that will be essential for getting there.

Presentations Session 6: The Transformative Role of Data in SDGs and Disaster Resilience / 131

Beyond Data: Leveraging LowCost Sensors for Policy Impact and Regulatory Acceptance

Author: Nangila Wafula[1]

Co-author: Libby Hepburn [1]

[1] *CSGP*

Corresponding Authors: libhepburn@gmail.com, faithwafula4@gmail.com

Low-cost sensors are revolutionizing air quality monitoring, especially in under-resourced regions like Kenya. These devices enable affordable, localized data collection, which allows communities to identify pollution hotspots, raise awareness, and advocate for action. However, their integration into formal regulatory frameworks remains limited due to concerns over data reliability and perceived shortcomings compared to traditional systems (Lewis et al., 2016).

Many African cities face mounting air pollution challenges but lack consistent urban air quality monitoring. Although Kenya enacted Air Quality Regulations in 2014, data on particulate pollutants in Nairobi remains scarce. This gap is common across many African nations, hindering efforts to assess pollution impacts, inform policy, and respond effectively to deteriorating air quality. The global Air Quality Community of Practice (CoP) of the Citizen Science Global Partnership (CSGP) actively addresses these challenges by working to scale up air quality monitoring in under-resourced regions and demonstrating evidence-based policymaking through citizen science.

This presentation proposes actionable strategies to enhance the credibility and impact of low-cost sensors in policymaking and regulatory contexts in such regions. First, establishing universal calibration and validation protocols in collaboration with academic and industry stakeholders can significantly bolster the credibility of sensor data by ensuring alignment with regulatory standards (Crilley et al., 2018). The Air Quality CoP is collaborating with the WorldFAIR+ Project and the CitiObs project to create interoperability frameworks based on FAIR principles for citizen science air quality monitoring. Second, creating effective data communication strategies can maximize the visibility and impact of sensor-derived insights. Platforms that transform complex datasets into accessible visualizations and narratives can engage policymakers and the public, fostering broader support (Kumar et al., 2022).

Integrating citizen science into policy through multi-stakeholder collaborations institutionalizes community-driven data collection. Open-access platforms, such as OpenAQ, bridge local monitoring efforts with policy-level interventions, building stakeholder trust and cooperation. Finally, advocating for adaptive regulatory systems that position low-cost sensors as complementary tools to traditional monitoring methods and not replacements can drive innovation and amplify impact.

Drawing from case studies within the CoP and successful implementations, this session explores how these solutions can bridge the gap between citizen-driven data and institutional action. By tackling technical, communication, and policy challenges, low-cost sensors can be repositioned as essential tools for community agencies, filling data gaps, raising awareness, and impactful policy change.

References:
Lewis, A., et al. (2016). Evaluating low-cost sensors for air quality monitoring. Faraday Discussions. https://doi.org/10.1039/C5FD00201J
Crilley, L., et al. (2018). Calibration of low-cost air quality sensors for urban environments. Atmospheric Measurement Techniques. https://doi.org/10.5194/amt117092018
Kumar, P., et al. (2022). Bridging the gap between citizen science and policymaking. Environmental Monitoring and Assessment. https://doi.org/10.xxxxxx
Code for Africa. (2019, October 21). Measuring Nairobi's air quality using locally assembled low-cost sensors. Medium. https://medium.com/code-for-africa/measuring-nairobis-air-quality-using-locally-assembled-low-cost-sensors-94a356885120
Hasenkopf C. et al (2024) Energy Policy Institute at Chicago. The Case for Filling Air Quality Data Gaps with Local Actors: A Golden Opportunity for the Philanthropic Community

Poster Session / 132

Life Cycle of Metagenomic Research Data Management

Authors: Charlie Pauvert[1]; Maja Magel[2]; Thomas Clavel[1]

Co-author: Catherine Gonzalez [3]

[1] *RWTH/UKA/NFDI4Microbiota*

[2] *University Hospital RWTH Aachen*

[3] *RWTH/UKA*

Corresponding Authors: c.gonzalez@itc.rwth-aachen.de, tclavel@ukaachende, cpauvert@ukaachen.de, mmagel@ukaachen.de

Effective research data management (RDM) is essential for ensuring that data adheres to the FAIR principles of Findability, Accessibility, Interoperability, and reusability. In this session we will examine how these principles drive the life cycle of metagenomic data at the Uniklinikum University in Aachen, from data generation to long-term storage and reuse.

The process begins when biological samples are sent for sequencing, generating raw data that is then processed using a reproducible Snakemake workflow developed by the Clavel Lab. This workflow is documented and publicly available in GitHub: GitHub - ClavelLab/genome-assembly: A Snakemake workflow assembling bacterial genomes according to the standard operating procedure in the Clavel Lab.
The output of the workflow includes the assembled genomes (in gzipped FASTA format), plasmid sequences (results/plasmids/isolate.plasmids.fa.gz), and a structured metadata table in CSV format. This metadata captures key information about the genome assemblies and conforms to the minimal metadata standards recommended by NFDI4Microbiota. The associated metadata profile is created using the Metadata Profile Generator, ensuring that each field is linked to appropriate ontology terms for improved interoperability. The profile is then made available for use in Coscine, the platform used at the UKA for storing and archiving research data along with its linked metadata profile.

In the next stage, the metagenomic data and its metadata are prepared for storage and reuse. This process is facilitated by the data steward using Python to automate the extraction of metadata from the CSV file. The extracted metadata is added to the metadata form and then both the file and metadata form are uploaded to specified resources in Coscine.

Each resource in Coscine has a unique persistent identifier, ensuring the findability of the data. Data stored in Coscine remains accessible for at least ten years following the conclusion of the research project, in accordance with good scientific practice. Researchers and collaborators can access the

data using institutional credentials or ORCiD. Project-based permissions enable secure sharing of data and collaboration.

By walking through each stage of this workflow—from data generation to archiving—this presentation demonstrates how the FAIR principles are applied in practice to support transparent, sustainable, and reusable metagenomic research at UKA.

Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI / 134

Promoting the Use of Discipline-Specific Metadata for Data FAIRness

Authors: Christine Kirkpatrick[1]; Mark Musen[2]

Co-authors: Erik Schultes [3]; Minfang Wu [4]; Wonsik Shim [5]

[1] *San Diego Supercomputer Center / CODATA*

[2] *Stanford University*

[3] *GO FAIR Foundation*

[4] *Austrailian Research Data Commons*

[5] *Sungkyunkwan University*

Corresponding Authors: mingfang.wu@ardc.edu.au, musen@stanford.edu, wonsik.shim@gmail.com, erik@gofair.foundation, christine@sdsc.edu

The FAIR guiding principles indicate that scientific datasets should be annotated with "rich"metadata that adhere to relevant community standards. Those standards include metadata reporting guidelines, which enumerate the attributes needed to describe the features of the experiments that led to the corresponding data, and the controlled terms that standardize the values of those metadata attributes. Some scientific communities have been developing their own metadata standards for nearly three decades, while others are only now beginning. The functional genomics community, for example, released the Minimal Information About a Microarray Experiment (MIAME) back in 2001—offering a reporting guideline that has been widely adopted by researchers, by publishers, and by the discipline-specific repositories where microarray data are archived, enhancing data discoverability, interpretability, and reusability. Yet the majority of scientific communities still lack metadata standards, seriously impeding the FAIRness of their datasets.

This panel session will highlight the importance of developing discipline-specific metadata reporting guidelines to support the FAIRness of data both within and across domains.

The panel will involve brief presentations by a group of international experts, with considerable time for discussion among the panelists and with the audience. The panel is being organized by Christine Kirkpatrick, Founder and Head of GO FAIR US and Secretary General of CODATA, and by Mark Musen, Stanford Medicine Professor of Biomedical Informatics Research at Stanford University (USA). The sequence of presentations will be as follows:

Christine Kirkpatrick will introduce the problem and the panel. She will survey the use of discipline-specific metadata standards in science and discuss efforts on the part of GO FAIR US to educate research communities in the development and use of rich metadata standards that are tailored for different classes of experiments within their particular disciplines.

Mingfang Wu, Product Manager at the Australian Research Data Commons, has performed studies of researchers in clinical investigation, social science, and ecology, and she has identified how metadata granularity directly affects their ability to perform dataset search and reuse. She will offer evidence for how discipline-specific metadata can enhance data FAIRness.

Wonsik "Jeff"Shim, Professor of Library and Information Science at Sungkyunkwan University in Seoul, KR, will discuss the use of rich, discipline-specific metadata within the Korean Research Memory—a comprehensive repository of research output from all publicly funded projects supported by

the National Research Foundation of Korea. Prof. Shim will describe the repository's rich, hierarchical organization for metadata, particularly in the humanities and social sciences, and how such discipline-specific metadata facilitate data search, data understanding, and data reuse.

Erik Schultes, FAIR Implementation Lead for the GO FAIR Foundation, NL, will discuss the Metadata for Machines workshops that his organization has pioneered over the past eight years. He will present the structure of these 2–3 day events that lead subject matter experts through the process of constructing relevant discipline-specific reporting guidelines and sets of controlled terms to offer new community standards for detailed metadata authoring. He will offer an assessment of the strengths and limitations of intensive workshop environments to help scientists to create and document relevant and meaningful metadata standards and to deploy those standards more routinely.

Mark Musen will provide perspective on the preceding presentations. He then will describe the use of the CEDAR system for managing and disseminating discipline-specific metadata standards. He will discuss how CEDAR encodes a research community's preferences for discipline-specific standards as metadata templates, and how those templates can become shared representations for use by a variety of data-management technologies—for metadata authoring, for metadata correction, and for metadata harmonization.

Overall, the panel will overview the status of discipline-specific metadata, discussing the feasibility and utility of incorporating such annotation in real-world repositories. The panel will discuss strategies for creating and disseminating new discipline-specific, community-based metadata standards, and for encoding those standards in a way that makes them maximally reusable across a variety of applications that rely on those standards.

The FAIR guiding principles mandate the use of "rich,"community-relevant metadata standards, but they do so without offering guidance for how research groups might actually accommodate such standards. We now have considerable experience within the data-management community that offers practical solutions for the creation, promulgation, and use of discipline-specific metadata—demonstrating that the creation of truly FAIR data is not only desirable, but also manifestly achievable.

137

Building Earth and Environmental Science Data Repository Ecosystems: actioning locally - operationalising globally.

Authors: Lesley Wyborn[1]; Rebecca Farrington[2]; Kelsey Druken[3]; Donald Hobern[4]; David Lescinsky[5]; Peggy Newman[6]; Andrew Robinson[7]

[1] *Australian National University*

[2] *AuScope*

[3] *ACCESS-NRI*

[4] *Australian Plant Phenomics Facility*

[5] *Geoscience Australia*

[6] *Atlas of Living Australia*

[7] *Australian National Computational Infrastructure*

Corresponding Authors: david.lescinsky@ga.gov.au, peggy.newman@csiro.au, donald.hobern@adelaide.edu.au, kelsey.drucken@anu.edu.au, andrew.robinson1@anu.edu.au, rebecca@auscope.org.au, beryl.morris@uq.edu.au, lesley.wyborn@anu.edu.au, tim@auscope.org.au

Significance of the issues to be tackled:
Earth and environmental (E&E) datasets covering the six spheres of Earth System Science (geosphere, hydrosphere, biosphere, cryosphere, atmosphere, and anthroposphere) have been collected over centuries. Properly curated and preserved over time, E&E datasets can provide evidence-based inputs into longitudinal monitoring of changes over decades (e.g., desertification, sea level rise, anthropogenic contamination, climate variability, groundwater quality). When integrated with data from

Health, Social Science and Humanities, E&E datasets make valuable contributions to the UN Sustainable Development Goals.

Early collections of E&E datasets were dominated by ground- or marine-based human observations and/or measurements by manual instruments, either in situ or laboratory-based. Airborne platforms emerged in the 1930s, followed by satellites (1960s), and Uncrewed Aerial Vehicles (UAVs, 2000s). Data acquisition underwent a paradigm shift in the 1950's with the computerisation of instruments, which quickly evolved to a capability for born-digital data outputs. Initially, most raw Primary Observational Datasets (PODs) were stored and managed locally within the institution that collected them; some PODs remained in this local state, others evolved into large-scale, internationally managed community resources. Additionally, non-observational datasets, including Climate/weather reanalysis and modelling, have emerged as fundamental datasets and are exposed to the same tensions as PODs.

Longitudinal requirements of E&E research necessitate that datasets be managed over decades and accommodate changes in instrumentation, hardware, standards, software, compute power, etc. Over time, the resolution and scale of PODs have increased: some data volumes from individual surveys are now in petabytes and require specialised HPC-D for storage and curation. Simultaneously, PODs collected as small-scale measurements are in megabytes and can be stored locally or on the cloud, yet their complexity and richness require specialised repositories to sustainably curate (meta)data to community-agreed domain standards. For both large and small volume PODs, the application of successive levels of processing throughout the data life cycle creates an incredible diversity of downstream datasets and products. In many cases, derivative products from very large volume datasets can be megabytes or smaller, no longer needing HPC-D for storage.

Apart from this diversity of data types and volume, there is a range in researcher capability, from highly skilled users who expertly use PODs and minimally processed reference datasets at any scale, to those who depend on pre-processed datasets to answer their research questions or to fit the needs of their software applications (and sometimes the capacity of the hardware and bandwidth they have access to).

While actioning systems locally or nationally within a single research community can be achieved, systems need to be operationalised within the global context and compatible with international strategies, including:

1. 2007 OECD Principles and Guidelines for Access to Research Data from Public Funding;
2. 2019 Beijing Declaration on Research Data: 'publicly funded research data should be interoperable, and preferably without further manipulation or conversion, to facilitate their broad reuse in scientific research';
3. 2021 UNESCO Recommendations on Open Science for open access to data, both raw and processed, and the accompanying metadata, analysis code and work flows;
4. FAIR Guiding Principles for Scientific Data Management and Stewardship;
5. TRUST Principles for Digital Repositories;
6. CARE Principles of Indigenous Data Governance.

It is nearly impossible for a single repository to meet all these requirements for every E&E data type and to serve all users. A 'Repository Ecosystem'is needed, one that balances and emphasises resources across the full data cycle and meets multiple considerations including:

1. Curation and sustainable preservation of raw full-resolution PODs captured directly off the instruments, under the expectation of limited need for ongoing access;
2. Calibration and conversion of the raw PODs into full-resolution reference datasets using community-agreed formats, data standards, etc. and their annotation with rich, FAIR- and CARE-compliant metadata. Many E&E PODs are dynamic and require regular updates with new data, calibrations, standards and technology. Once standardised, these high-resolution datasets can be aggregated into national/global datasets. These upstream datasets will mainly be accessed by power users;
3. Systematic reprocessing of full-resolution reference datasets into reusable downstream analysis-ready products, including mapping to uniform space-time grids, model outputs, syntheses and

subsampling products. These can be accessed from distributed data platforms, virtual laboratories, portals, dashboards, etc. that are customised for specific user communities.

Separating the PODs'curation, preservation, calibration and conversion in 1 and 2 from the short-term and ever-changing distribution environments of analysis-ready products in 3, highlights the need for curation and preservation of raw full-resolution PODs to enable sustainable reuse over time, thus ensuring a capability to generate new knowledge in future scientific research.

But sustainable management of E&E data needs to also consider geopolitical changes that can abruptly impact repositories caring for key datasets. Securing these datasets for future use will depend on multiple countries proactively collaborating to ensure adequate redundancy and risk management.

Approach, structure, format, and suggested agenda:
This session will explore national actioning approaches for E&E Data Ecosystems that are operationalised globally around the above considerations.

1. Introduction (10 Minutes)
2. Lighting papers covering both national and international perspectives on linking E&E Data Ecosystems and Research Infrastructures (40 Minutes): Creating a National E&E distributed data ecosystem: catering for multiple users (Rebecca Farrington, AuScope); Global Ecosystem Research Infrastructures (Beryl Morris); Earth Science Research Infrastructures (Tim Rawling, AuScope); Oceans Data Information System (Speaker TBC); GBIF (Peggy Newman, Atlas of Living Australia); Connecting national E&E datasets to the WorldFAIR cross-domain project (Speaker TBC).
3. Community forum and determining the next steps (40 Minutes).

138

Increasing Resilience of Global Earth and Environmental Science Data Supply Chains

Authors: Joseph Gum[1]; Lesley Wyborn[2]; Megan Orlando[3]; Ruth Duerr[4]; Reyna Jenkyns[5]; Adrian Burton[6]

[1] *National Center for Atmospheric Research*

[2] *Australian National University*

[3] *ESIP*

[4] *Ronin Insititue*

[5] *World Data System*

[6] *ARDC/ ANU*

Corresponding Authors: rebecca@auscope.org.au, reyna@oceannetworks.ca, ruth.duerr3@gmail.com, jgum@ucar.edu, lesley.wyborn@anu.edu.au, meganorlando@esipfed.org, tim@auscope.org.au, adrian.burton@ardc.edu.au

Significance of the issues to be tackled:
Earth and environmental (E&E) datasets have a critical role to play in the Sustainable Development of our Planet. They contribute to the prediction of natural hazards, effective development of our natural resources, long term monitoring of vegetation changes, etc., and underpin many UN Sustainable Development Goals. Since the 1990s there have been concerted efforts within some E&E domains to develop the standards, protocols and best practices to enable global sharing of digital data (e.g., OneGeology, Federated Digital Seismology Network (FDSN), Global Biodiversity Information Facility (GBIF), Earth System Grid Federation (ESGF)). These networks can federate multiple repositories internationally and as they mature and stabilise, they form global data supply chains that can be fundamental inputs into modelling and research ranging from faster-than-real-time emergency warning systems (e.g., tsunami, flood, volcanic ash, wildfire) to longitudinal monitoring over many decades (e.g., desertification, sea level rises, anthropogenic contamination).

As these global networks become accepted as critical inputs into E&E data supply chains, their vulnerability is becoming of concern. Loss of digital data and physical assets within a repository has long been assessed by repository managers and archivists, and there are many published mitigating strategies. However, recent events have highlighted firstly, the need to formally assess the resilience of physical infrastructures underpinning these networks (e.g., individual repositories, research infrastructures, data networks, etc.). Concrete examples show how they can be severely damaged, if not annihilated due to natural disasters, political decisions, wars, funding cuts, cyberattacks and the inability to obtain/retain skilled staff to ensure continuity.

Secondly, contemporary happenings are showing the power of collaborations, at both a national and international level, to help each other overcome barriers and create resilient data professionals - individuals equipped with the knowledge and tools, most importantly, to create human-networks to keep pushing through obstacles. More resilient data professionals form the backbone of more resilient data facilities.

Recently members of the Sustainable Data Management Cluster of the Earth Science Information Partners (ESIP) focused on understanding the many factors that contribute to increasing resilience of repositories. A repository scorecard was published (https://zenodo.org/records/15122046) to enable any repository to measure how resilient it might be both in its normal state and during certain crises. This includes a measure of how well a repository might weather an example crisis, how easy it might be to restore metadata, and how much societal impact a missing repository would have. The scorecard was based around four scenarios:

1. Incoming Natural Disaster in 48 Hours. A natural disaster (e.g., hurricane, wildfire, tsunami or earthquake), is forecasted to hit the primary facility of the repository. The focus of the scenario is on local destruction of the facility, including all physical devices and the deposits on them, with an uncertain timeline of facility restoration.
2. Loss of Organizational Funding: The repository is being shut down in one month, with that one month to implement any plans. The focus of the scenario is on eventual total loss of the repository including staff, hardware, and software, but there is time for mitigating actions to be taken.
3. Cyberattack/Organizational Infiltration: The facility has been infiltrated and hostile agents have control of cyberinfrastructure. The focus of the scenario is on sudden denial of access to any and all deposits, but not necessarily deletion of deposits and systems.
4. Loss of Technical Expertise: Technical expertise critical to running the repository (e.g., knowing how to operate, maintain, and extend the software systems), are no longer available. The focus of this scenario is on the loss of knowledge to keep critical repository systems and processes running.

The ESIP-led work resulted in three major Earth Science Research Infrastructures from Australia (AuScope), United States (EarthScope), and Europe (European Plate Observing System (EPOS) starting discussions on how combined, they could provide international support in times of crises affecting one national system and the potential impacts on the global supply chain of a particular dataset.

It is highly probable not all datasets can be protected. To help focus on the more critical and impactful, the American Geophysical Union (AGU) is coordinating a community effort to determine the most impactful E&E datasets, based on three perspectives:

1. People: education, training, disaster response and prediction;
2. Planet: geophysical phenomena, conservation, climate, environmental indicators;
3. Prosperity: economic good, social equity, humanitarian relief, community resilience.

Connecting these three themes, there are suggestions that we require internationally agreed policies, agreements and infrastructure for sustaining critical global data supply chains, and ensuring that essential data is always available. These acknowledge the inherent universal and multilateral nature of science requiring corresponding data arrangements. This applies all the more so in times of crisis where there is a critical need for evidence-based approaches to preparedness, response, and recovery.

Approach, structure, format, and suggested agenda:
The session will focus on increasing resilience of global E&E data supply chains. It will comprise Ignition/Lightning Talks and Structured Discussion

Agenda

1. 0-10 Minutes: Introduction and background to the drivers for the session
2. 0-50 minutes: Short papers to set the scene including a) Measuring Resilience of Individual Repositories (Joseph Gum, ESIP Sustainable Data Management Cluster); b) Global networking of Earth Science Research Infrastructures for supporting times of crisis (Tim Rawling, AuScope); c) Identifying Impactful E&E Science Datasets (Speaker TBC); d) Developing policies for sustaining critical E&E data supply chains during times of crisis (Adrian Burton,ARDC).
3. 50-85 Minutes: Structured Discussion
4. 85-90 Minutes: Closing circle and next steps.

Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global / 139

Building a data infrastructure for Social Science and Humanities: A double perspective on quality and community from Italy and France

Authors: Alessia Spadi[1]; Edward Gray[2]; Nicolas Larrousse[3]; Emiliano Degl'Innocenti[4]

[1] *Consiglio Nazionale delle Ricerche (CNR), Italy*

[2] *IR* Huma-Num / DARIAH-EU*

[3] *Huma-Num CNRS*

[4] *CNR, DARIAH-IT*

Corresponding Authors: nicolas.larrousse@gmail.com, emiliano.deglinnocenti@cnr.it, edward.gray523@gmail.com, alessia.spadi@cnr.it

The Social Sciences and Humanities (SSH) disciplinary sector encompasses research studies that share an epistemological commitment to the critical investigation of human experience, cultural expression, and social organization. SSH research contributes to the development of theoretical frameworks and methodological approaches that are essential for understanding complex societal transformations.

European institutions, including the European Commission, recognized the importance of SSH researchers and research projects by integrating it into broader research frameworks, such as Horizon Europe and NextGeneration EU, and supporting international cooperation through initiatives like the European Research Infrastructure Consortium (ERIC). Yet, despite this support, SSH research faces distinct challenges when compared to STEM disciplines:

1. sparsity and fragmentation of its data;
2. heterogeneity of sources - textual, audiovisual, or contextual - generated under diverse, often non-standard conditions;
3. isolated datasets within institutional silos, limiting their discoverability, accessibility and reuse.

Such variability underscores a systemic challenge in ensuring data quality across SSH research, which involves evaluating the accuracy, completeness, consistency, and documentation of data to enable its meaningful interpretation, reuse, and long-term preservation.

Moving towards data-intensive and AI empowered research, RIs will see a further shift of their role from data suppliers to primary users of machine enabled workflows. To ensure sustainability

of workflows, given by their reproducibility and replicability, data quality becomes a fundamental aspect of good research and reliable results, in particular in the case of AI mediated processes and workflows.

DARIAH.it, as part of the H2IOSC project, and DARIAH-FR, particularly through the IR* Huma-Num, are two national research infrastructures that exist to serve SSH researchers and research projects in their respective countries. Both initiatives are inscribed in the ERIC DARIAH-EU, which seeks to structure the development and interaction of national research infrastructures in Europe. Though both Italy and France have different approaches, the common problem remains: how to improve quality of data, metadata, paradata and related tools based on an approach involving users?

Italy

In Italy, DARIAH.it is working to strengthen the national infrastructure, focusing on promoting standards for data, services and workflows, as well as developing platforms to support users in producing and managing FAIR data and resources, access a wide range of services and combine them into meaningful scientific workflows requiring the interaction of different & independent services, also leveraging on AI modules (i.e. AI-mediated DH).

Recently, due to the increasing number of attacks brought to different institutions (British Libraries) and resources (e.g.: Archive.org) dealing with cultural content - Italy started a process of transition towards Critical Infrastructures for SSH, bringing cybersecurity and resilience into DARIAH.it national infrastructure development plans.

Being a socio-technological environment, and aiming to bring measurable advancements for the research community, the upgrade of the technical infrastructures requires a strong investment on the human component: RIs can support research by providing quality data, tools and processes but researchers are encouraged to be the owner of their own algorithms and to know how they work to ensure they get the right answers to the right questions, hence transparence, explainability, standardization, alignment and training are the drivers to successfully complete this transition.

France

Huma-Num, the French national infrastructure for SSH has built a robust technical research infrastructure fed by community feedback. After its first decade, certain aspects need to be updated, most notably, the approach to data & metadata quality in its research data repository, NAKALA. Considering the large amount of existing data in NAKALA (over 1.4 million files), it was deemed unfeasible to manually curate all existing and future deposits..

To implement this quality plan, and in addition to purely technical controls, Huma-Num relies on communities to develop a network of curators, in articulation with the national ecosystem Recherche Data Gouv„ with their harmony assured by the development of a curation guide. Alongside this, Huma-Num has extensively developed its documentation, with a focus on data preparation, and has organized a series of webinar sessions for users. Finally, Huma-Num launched a global content analysis of the repository to gain a better understanding of current practices and develop quality indicators.
By combining technical developments with community feedback to improve quality, Huma-Num gradually evolves from a purely technical infrastructure to a knowledge infrastructure.

Conclusion

This paper will develop the two Italian and French approaches to building an infrastructure for SSH centered on quality and sustainability. The problems to be solved are quite similar despite differences in national organizations, but the crucial common point is the necessary involvement of users, without which actions are doomed to failure. These dual approaches are fruitful examples for others that are confronted with the thorny problem ensuring research data and metadata quality in SSH.

References

- Bellini, Emanuele and Emiliano Degl'Innocenti. 2024. Transitioning SSH European Research Infrastructures to Critical Infrastructure Through Resilience. IEEE International Conference on Cyber Security and Resilience (CSR): pp. 801-806, doi: 10.1109/CSR61664.2024.10679383.

- Edmond, Jennifer, ed. 2020. "Digital Technology and the Practices of Humanities Research."Open Book Publishers. https://doi.org/10.11647/obp.0192.
- Gray, Edward J., Nicolas Larrousse. Huma-Num IR*, 10 Years of Building a Research Infrastructure at the European level. Huma-Num, 10 Years of Building a Research Infrastructure at the European level, 2024. ⟨halshs-04573643⟩
- Lacagnina, Carlo, et al. «TOWARDS A DATA QUALITY FRAMEWORK FOR EOSC». Zenodo, 9 January 2023. https://doi.org/10.5281/zenodo.7515816.
- Spadi, Alessia, Emiliano Degl'Innocenti, and Carmen Di Meo. 2024. «DARIAH.It: Data Integration Strategies and Solutions for Digital Resources Management and Research in the Arts and Humanities». Mimesis Journal 13 (2):119-34. https://doi.org/10.13135/2389-6086/9920.

140

Data for Cognitive Health Equity: Shaping Global Data Ecosystems for Healthy Aging

Author: Mihoko Otake-Matsuura[1]

Co-authors: Alexandra Wolf [1]; Mike Martin [2]

[1] *RIKEN Center for Advanced Intelligence Project (AIP)*

[2] *University of Zurich*

Corresponding Authors: mihoko.otake@riken.jp, alexandra.wolf@riken.jp

Cognitive decline represents one of the most critical public health and societal challenges of the 21st century, with approximately 50 million people affected by dementia worldwide, and nearly 10 million new cases annually (World Health Organization, 2020). As populations age the incidence of age-related cognitive impairments is expected to rise dramatically. However, the global response is hampered by significant disparities in how cognitive health data is collected, integrated, and used.

Most large-scale datasets and longitudinal studies are rooted in high-income contexts, often overlooking socio-environmental and behavioral variability that characterizes aging across different cultural and geographical settings. The underrepresentation of low- and middle-income countries (LMICs), ethnically diverse populations, and marginalized communities creates a skewed evidence base that can limit the generalizability of interventions and hinder global progress toward equitable cognitive healthcare. Furthermore, much of the existing data infrastructure focuses on clinical endpoints and lacks integration with behavioral and contextual data critical for early detection, prevention, and personalized care strategies.

In response, this session explores how cross-domain data stewardship, including behavioral, physiological, sociocultural, and environmental data, can address these gaps and support cognitively healthy aging as a global public good. Organized by the CODATA Task Group (TG) on *Data-Driven Social Change Towards a Society Promoting Cognitively Healthy Aging*, in collaboration with the *Cognitive Behavioral Assistive Technology (CBAT) Team* at RIKEN AIP, the session aims to foreground the role of ethically governed, culturally adaptive, and community-centered data in enabling more inclusive and actionable insights into cognitive health.

While the Task Group's original scope emphasized social data and environmental design, this session introduces perspectives from behavioral data science, including non-invasive methodologies, as promising tools for cross-cultural cognitive assessment. Japan's experience as a front-runner in addressing the challenges of a super-aged society provides a compelling case study for the integration of local community-based research, policy implementation, and interdisciplinary technological innovation. Drawing on lessons from the Japanese context, the session will highlight how ethical data stewardship, guided by principles of FAIR (Findable, Accessible, Interoperable, and Reusable) and CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics) can foster more resilient

and cognitively inclusive societies. Finally, this session aligns with SciDataCon 2025's overarching theme, *Empowering the global data community for impact, equity, and inclusion*, by placing cognitive health equity at the intersection of data science, aging research, and global public policy.

Session Structure Highlights

Opening Presentation
Advancing Cognitive Health Equity through Cross-Domain Data Stewardship Speaker: Dr. Mihoko Otake (Team Director)
This keynote will explore how ethically governed, inclusive data infrastructures are critical to addressing cognitive health disparities.

Presentation
Visualizing Cognition: Data Insights from Eye-Tracking Research in Aging
Speaker: Dr. Alexandra Wolf
Most cognitive health datasets rely on clinical or survey-based data, often overlooking behavioral signals that can offer context-sensitive insights. The talk will emphasize the adaptability of eye-tracking technologies across cultures and contexts.

Interactive Group Discussions
Participants will join **small breakout groups to collaboratively examine the barriers and enablers to building inclusive cognitive health data ecosystems**. Themes will include ethical data governance, digital infrastructure disparities, culturally sensitive tool design, and interdisciplinary collaboration.

Real-Time Polling and Summary
The session will conclude with a **live poll capturing participant perspectives and proposed actions**, followed by a synthesis of discussion outcomes. These inputs will directly inform the CODATA TG's post-conference activities.

The proposed session will generate a structured and interdisciplinary dialogue around advancing cognitive health equity through data, leading to tangible contributions to international policy and practice. Expected outcomes include evidence-based policy guidelines for governments and public institutions on the ethical, inclusive collection and use of cognitive health data. The session will also offer practical frameworks for technology developers to design accessible, culturally adaptable, and cost-effective cognitive assistive technologies that meet the needs of diverse aging populations. In addition, the session will propose research roadmaps for integrating cross-domain datasets, including behavioral, clinical, and environmental data, to reduce bias in cognitive health research and enhance the responsible use of technology in interventions. These recommendations will support the development of inclusive digital and health infrastructures, particularly in low-resource settings. Ultimately, the session aims to support global efforts toward building cognitively inclusive, data-driven societies where people of all backgrounds can live and age with dignity, autonomy, and well-being.

141

The Sample Management Lifecycle in Action: Stages, Stakeholders, Identifiers, and Opportunities

Authors: Jens Klump[1]; Kirsten Elger[2]; Kerstin Lehnert[3]; Lesley Wyborn[4]; Fabian Kohlmann[5]; Rorie Edmunds[6]

[1] *CSIRO*

[2] *GFZ Helmholtz-Centre for Geosciences*

[3] *Lamont-Doherty Earth Observatory of Columbia University*

[4] *Australian National University*

[5] *Lithodat Pty Ltd*

[6] *DataCite*

Corresponding Authors: kelger@gfz-potsdam.de, athomer@arizona.edu, nraia@arizona.edu, fabian.kohlmann@lithodat.com, rorie.edmunds@datacite.org, lehnert@ldeo.columbia.edu, lesley.wyborn@anu.edu.au, jens.klump@csiro.au, anusuriya.devaraju@csiro.au

Significance of the issues to be tackled in the session

Effective sample management is essential to ensuring the integrity, reproducibility, and openness of research across diverse disciplines. From the Physical and Life Sciences to the Social Sciences and the Arts, material samples serve as the foundation for countless research projects. As the scale, complexity, and diversity of sample collections grow, the need for robust, interoperable management strategies becomes increasingly urgent.

Persistent identifiers (PIDs), such as International Generic Sample Numbers, play a central role in addressing this challenge. By providing globally unique and resolvable identifiers, PIDs enhance the traceability, discoverability, and reusability of material samples. They digitally connect samples to related datasets, publications, instruments, and contributors, strengthening research transparency and ensuring alignment with the FAIR Principles.

However, realizing the full potential of PIDs in sample management and the challenges due to the high granularity of objects to be managed, requires a coordinated effort among multiple stakeholders. Researchers, collection curators, data and informatics specialists, infrastructure providers, and publishers each contribute to different stages of the sample lifecycle—from collection and documentation to curation, dissemination, and reuse. Gaps in standardization, interoperability, and best practices still hinder effective sample management, limiting the ability to link samples across disciplines and repositories.

This session will explore the full sample lifecycle, examining current practices, identifying pain points, and highlighting opportunities for enhancement. By fostering dialogue across the community, the session aims to promote greater consensus around sample management standards, advocate for wider adoption of PIDs, and inspire collaborative solutions that drive scientific advancement.

Approach, structure, format, and suggested agenda for the session

To maximize engagement and impact, this session will feature a panel of presenters combined with a structured, interactive dialogue. The format is designed to capture the different perspectives of key stakeholders and encourage audience participation.

The session will begin with short presentations by the panellists (~45 min) with the objective of providing insights into the current challenges and opportunities in sample lifecycle management. The panel will be composed of representatives from the main stakeholder groups:

- Researchers: Needs and expectations for sample documentation and discovery.
- Collection Curators/Repository Managers: Sample preservation, access, and metadata standards, and tools for metadata collection.
- Data/Information Specialists: Technical considerations, such as PID assignment and integration with other research outputs.
- Publishers and Infrastructure Providers: Sample metadata in linking samples to publications and enhancing research integrity.

Following the presentations, there will be an interactive dialogue (~35 min), encouraging everyone in the room to share their experiences and explore solutions collaboratively. We will reflect on lessons learned, promoting cross-disciplinary understanding and identifying areas for improvement. The session will then end with a wrap-up (~10 min), summarizing key points and actionable insights, listing opportunities for future collaboration and any next steps to advance sample management practices.

Proposed speakers and the subject of their papers

The following stakeholder representatives have been invited and have confirmed their participation. Each speaker will present a paper that supplies their perspectives on the session themes of the importance of PIDs in sample lifecycle management, challenges in standardizing sample metadata and

protocols across disciplines, real-world examples of successful and problematic sample management practices, and strategies for improving sample discoverability and interoperability.

- Researchers: Andrea Thomer/Natalie Raia (ESIP Physical Sample Curation Cluster)
- Collection Curators/Repository Managers: Kerstin Lehnert (Interdisciplinary Earth Data Alliance)
- Data/Information Specialists: Anusuriya Devaraju (Commonwealth Scientific and Industrial Research Organisation)
- Publishers and Infrastructure Providers: Kirsten Elger (Earth System Science Data, Copernicus Publications)

Presentations Session 4: Data Stewardship / 142

Towards understanding identification, selection and appraisal in contemporary digital preservation practice

Authors: Laura Molloy[1]; Micky Lindlar[2]

[1] *CODATA*

[2] *Leibniz Information Centre for Science and Technology*

Corresponding Authors: laura@codata.org, micky.lindlar@tib.eu

Identification, selection and appraisal are key digital preservation activities when ingesting data objects for long-term preservation. This paper describes the approach taken to a major new global survey by the EOSC EDEN project, designed to improve understanding of current practices in this area, and the frameworks, standards and guidelines that support digital preservation professionals when tackling these challenges. We will situate the survey work in its wider context, outline our approach to question-building and analysis, and share what we hope our findings will tell us, including an overview of analysis and findings to date.

Regardless of which lifecycle model, conceptual framework or workflow is used by a digital preservation organisation, there is always a point at which data objects officially enter the digital preservation environment. However, these frameworks and models often don't clearly outline how, when and what criteria the decision to preserve that data is based upon. In addition, the term used to describe this decision process may be dependent on the domain in which the archive is embedded. Three terms commonly used for this process are: identification, selection, and appraisal.

These practices are key elements of digital preservation practice for any size of organisation, particularly when the data objects are intended for long-term data preservation, with the funding commitment and responsibility that implies. But how do on-the-job digital preservation professionals approach these key activities? Which standards, guidelines and frameworks do they refer to? What levels of quality are measured and how? Can these quality metrics be used for re-appraisal processes along the lifecycle if data is not to be stored for the long-term? And what does "long-term" mean anyway?

The newly-initiated European Open Science Cloud (EOSC) project, 'Enhancing Digital preservation strategies at European and National level'(EDEN) is a three-year (2025-2027) research initiative funded by the EU Horizon Europe programme, that aims to tackle these (among other) questions. The project targets research data archives and addresses the questions of how data quality for digital preservation can be defined, and how this definition can be used in decision-making processes for ongoing preservation. A first step within this work is a global survey to understand how the community currently conducts these decision-making processes for digital preservation.

The questioning approach is as follows. We will ask the participants about:
- Descriptive information from each participant.
- Familiarity with frameworks and guidelines for appraisal, identification and selection of data for

long-term preservation: a list of well-established frameworks and guidelines are presented. The respondent indicates how familiar they are with each and can list any further reference or guidance resources that they use.
- Definition of "long-term" preservation: the respondent is asked whether their organisation has a working or agreed definition of this and if so, the source of this definition. If they are working without an agreed definition, the respondent is invited to report why this is the case.
- Pathways of data into the archive: the respondent is asked how digital objects enter the organisation for long-term preservation. Choices include self-deposit on a voluntary basis, mandated deposit such as legal or funder mandated deposit, proactive collection building such as harvesting or classical collection building in libraries, and/or third-party data preservation on a contractual level.
- Practices in assessment of quality of data objects at ingest: the respondent is offered the choice of various aspects of data objects (technical quality, content/information quality, quality of technical metadata, quality of descriptive metadata, quality of administrative metadata). For each aspect, the respondent can specify the types of quality assessment undertaken. There is a further question on the frequency of reassessment.
- How long digital objects are initially preserved.
- What happens to digital objects after the agreed initial preservation period.
- Discipline-specific questions submitted by EDEN WP3 on requirements including metadata standards used, file formats preferred, handling of sensitive data, risks, and community needs, all with a discipline-specific lens.

Messaging about and within the survey is intended to avoid discipline-specific language or unnecessary technical terms. Also, our survey is designed to be used across the global research, cultural heritage, and industry sectors, so it was also necessary to ensure that our language is as sector-neutral and globally appropriate as possible.

In addition, the survey is designed for all those working within the digital preservation organisational context, whether at senior management, middle management, or practitioner levels. Our choice of language aims to accommodate these various staff levels.

By this work, we hope to better understand:

- Which quality aspects are currently checked (and if there's a stronger leaning towards content quality or towards technical quality depending on institution type);
- If there is a shared definition of 'long-term';
- How the organisations who define 'long-term'as a shorter period (e.g. less than 10 years) approach quality checks, and succession planning for these data objects.

The EOSC EDEN project will support the digital preservation community and contribute to the European Open Science Cloud (EOSC) through efforts to better understand current digital preservation practice and to provide appropriate guidance and resources to the community. This includes the survey described here, designed to better understand current identification, appraisal, and selection practices when dealing with the ingest of data objects, and the reference materials that digital preservation professionals currently use to guide them.

Presentations Session 8: Policy and Practice of Data in Research; Data, Society, Ethics and Politics / 144

Democratizing Data Management: Academia's Responsibility to Community Partners

Author: Kelsey Badger[1]

Co-author: Lariza Fenner-Lux [2]

[1] *The Ohio State University*

[2] *Erikson Institute*

Corresponding Authors: lariza.fenner@illinois.gov, badger.60@osu.edu

Community-based, not-for-profit organizations are critical partners in public interest research across disciplines. This presentation explores how community engaged researchers, data stewards, and librarians can embed data management training for community partners into existing workflows by leveraging data curation as a source of knowledge about training needs.

While it is well-known that the re-use value of data is greatly enhanced by management practices that begin as early as possible in the lifecycle of a project, many public service organizations have limited resources available for data management. The ad hoc processes they develop over time can create barriers to current and future use of their data. This disenfranchises the workforce who depend on timely and reliable recordkeeping for the delivery of community services and slows the progress of research partnerships. Rather than accepting this as an inherent limitation of community engaged scholarship, academia has a responsibility to make data management education more accessible to the organizations they partner with. Equitable and inclusive access to data management education benefits all stakeholders by promoting data stewardship practices that lessen the need for post-hoc curation.

Data Curation in the Public Interest is a project that addresses this gap in the modern workforce through a partnership between The Ohio State University (Ohio State) and the Erikson Institute Early Childhood Project (EC Project). The EC Project has operated for over 25 years under a government contract with the Illinois Department of Children and Family Services, a state-level social services agency in the United States. Every year, their team of 45 clinical and administrative staff collect developmental assessment data for thousands of young children who have experienced abuse or neglect, providing a unique source of historical information on a priority population for researchers.

A librarian-led team of data professionals at Ohio State are curating the EC Project data, delivering datasets that are well-structured, documented, and analysis-ready. Curating this data makes usable a valuable and at-risk data source, but it does not address the root challenges to data management that are likely to continue without additional training and resources. To better meet the ongoing needs of the EC Project, Ohio State is bringing together two sources of information necessary for customized data management education. First, the curation process itself has been re-envisioned as an opportunity to understand the data management practices that should be addressed based on the actual state of the data. Second, focus groups with EC Project staff have been used to understand the impact of administrative record-keeping on different job roles and how this leads to specific data management practices. These activities provide two different types of "ground truth"information: what is happening with the data and why. This work will culminate in a capacity-building workshop customized to the needs of the EC Project staff, focused on developing a bottom-up data culture in the organization and facilitating dialogue about the interaction between clinical expertise and recordkeeping.

Many research studies with community partners involve curation of administrative data sources as preparatory work before analysis. This presentation will provide recommendations for how this work can be repurposed to provide direct benefit to partner organizations. Attendees will learn about the successes and challenges of the EC Project case study that can be used as a foundation for customizing data management training with their own community partners.

Presentations Session 8: Policy and Practice of Data in Research; Data, Society, Ethics and Politics / 147

Rethinking Data Governance: A Three-Pillar Approach for Public Universities

Authors: Gabriela Pino Chacon[1]; Ricardo Castro Blanco[1]

[1] *Universidad Nacional, Costa Rica*

Corresponding Authors: gabriela.pino.chacon@una.ac.cr, ricardo.castro.blanco@una.cr

Strategic and responsible management of research data is essential for Higher Education Institutions (HEIs) such as Universidad Nacional in Costa Rica, particularly within the context of Low and Middle-Income Countries (LMICs) given the existing challenges such as budgetary constraints, demands for transparency, and increasing expectations concerning social, economic, and environmental impacts. The Universidad Nacional, Costa Rica (UNA) has just completed 52 years of education, during which it has generated knowledge through theoretical and methodological deepening, experience in teaching, research, and extension activities, as well as dialogue with the people and sectors with which it is linked.

With the goal of contributing to the transformation of that knowledge in a sustainable resource the Institution has developed a strategy for Open Science (OS), a process in which the Vice-Rectory for Research (VI) has been a leader finding ways to implement a systematic approach based on international standards that optimize the quality, accessibility, and usability of research data.

Three axes have energized the VI management of the strategy:

1. Institutional Data Management Capacity:
   In this axis the VI has encouraged and supported the generation and actualization of institutional regulations on open science including the protection for products and authors, as well as the appropriate technological platforms for open data management. A lot of intra-institutional negotiations had and must be done in this axis because there are a lot of stakeholders involved here, from the departments for technology to the Institutional Assembly and from secretarial personnel to high level researchers.
2. Community Capacity Building on Open Science:
   After assessing the needs and strengths on OS of UNA, the VI team developpe a capacity building plan that has strengthened the practice of OS in the research community, as well as that of publishers and librarians. The work with authorities and researchers from the faculties, centers and sites has been emphasized in the design of data management plans, the use of data sets, the data curation for the institutional repositories, the strengthening of diamond route in our journal and the implementation of the FAIR and CARE principles in the research.
3. Data Availability for Informed Decision-Making:
   UNA has made considerable progress in promoting OS over the past decade; however, greater VI involvement is required to standardize and streamline data availability, facilitating access to FAIR data for institutional decision-making. Clear adherence to FAIR principles will be central to the success of the Open Research Data Portal (Portal Abierto de Datos de Investigación - PADI), ensuring research data is efficiently discoverable, accessible, and shareable, both nationally and internationally.

The establishment of PADI marks a critical step in advancing UNA's scientific research and enhancing data-driven governance. It will enable authorities and researchers to effectively identify capacity-building needs, promote research communication, improve international visibility, foster strategic partnerships, and strengthen data literacy within the academic and broader community.

Currently, much of UNA's data is already findable and accessible through institutional repositories and various other platforms. However, data distribution across multiple locations hinders effective reuse. PADI will integrate and structure this data comprehensively, promoting a robust culture of data reusability. Encouraging this reuse among the research community remains challenging, necessitating continued advocacy and education.

Interoperability is particularly vital within LMIC academic contexts, facilitating multidisciplinary and cross-institutional collaboration while maximizing resource utilization. Establishing clear interoperability standards ensures seamless data integration across diverse platforms, maximizing efficiency and research impact.

Employing advanced tools for interactive data visualization offers considerable advantages. Technologies such as Tableau, Power BI, R (Shiny, ggplot2), Python (Dash, Plotly), and open platforms like CKAN or Dataverse allow intuitive and impactful interactions with open data. These tools provide swift, clear interpretation of complex datasets, supporting informed decision-making at academic, community, and governmental levels (https://rpubs.com/ricardoc_07/1269700).

In conclusion, PADI aims to deliver diversified content architecture combining interoperable databases, dynamic visualization capabilities, and open-access platforms. Such infrastructure promises efficient

and strategic data management, positioning Universidad Nacional as a leader in applying scientific knowledge to enhance social welfare, thereby reinforcing its institutional commitment as "the necessary university."

Session Format: Workshop

Objectives:
- Validate a prioritized pillar approach for strengthening data governance in HEIs.
- Gather inputs to develop an open data governance model applicable to LMIC higher education.

Desired Outcomes:
- Systematic compilation of participant recommendations for creating an open data governance framework in LMIC HEIs.

Agenda:
- Opening (10 min): Proposal Presentation.
- Dialogue and Experience Sharing (10 min): Interactive discussion.
- Exploration (30 min): Ideation board session exploring data management, governance, and the proposed three pillars.
- Closing (30 min): Kahoot session focused on essential elements for a governance model in LMIC contexts.
- Wrap-up (10 min): Session summary and closing remarks.

Additional Material:
https://www.canva.com/design/DAGg_WH8wro/6ZOIDAIp1huchQaqT5OcuQ/edit?utm_content=DAGg_WH8wro&u

148

An evolving role for Data Scientists in the Age of Intelligent Automation

Authors: Matthew Mayernik[1]; Gita Yadav[2]

Co-authors: Debasis Mohanty [3]; Mark Parsons ; Dimitris Symeonidis [4]; LILI ZHANG [5]

[1] *NSF National Center for Atmospheric Research*

[2] *National Institute of Plant Genome Research (NIPGR)*

[3] *National Institute of Immunology, New Delhi*

[4] *University of Tartu*

[5] *COMPUTER NETWORK INFORMATION CENTER, CAS*

Corresponding Authors: gy@nipgr.ac.in, mayernik@ucar.edu, zhll@cnic.cn, parsonsm.work@icloud.com, dimitrios.symeonidis@ut.ee, deb@nii.res.in

Data science has evolved significantly over the past two decades, becoming a force within academic research, public and private sector workplaces, and in government policies and practices. Exponentially increasing volumes of publicly available datasets throughout the social and scientific realms have contributed to an explosion of data science applications, including AI tools and Large Language Models (LLMs) that are being connected to digital technologies of all types, such as personal computers, cell phones, social media, smart devices, and sensor networks. Data science has become a broad term with very different meanings and instantiations, growing far beyond the traditional panoply of techniques applied to derive value (economical, intellectual, or cultural) out of data. A former editor of the CODATA Data Science Journal argued in a recent retrospective essay that, "The scientific data community requires more from data science than other communities"(Rumble, 2023, p. 2), arguing that topics like data preservation, provenance, and traceability cannot be ignored when developing data science applications for scientific research. In this session, we will present broad perspectives on this key question: What are the distinctive aspects of data science for the scholarly and scientific data community, and how should the present day data scientists adapt to address issues like FAIRification, AI-readiness and Ethics?

A diverse panel will engage the CODATA community in a discussion of the role of data science in relation to scientific data initiatives, and how this community is contributing to the growth and evolution of data science. The speakers in this session will address several important questions about the current status and future evolution of data science, including:

- What is data science, and how does it relate to the "science of data"?
- Is data science a scientific domain in its own right?
- What are the main gaps / opportunities that must be addressed in data science going forward?
- What are the key data science trends in relation to the scientific data community?
- Science has always been data driven; what is different now? How have changes in the way that data are or should be shared influenced data science?
- How do AI technologies impact data science and contribute to FAIR research data ecosystems?
- How should the Data Science Journal and SciDataCon respond to changes in the nature of data science?

The talks in this session will inform broader discussions of the implications of these questions for CODATA, the World Data System (WDS), the International Science Council, and the many international scientific associations and unions. How can these organizations both contribute and respond to the changing nature of data science via concrete actions, programmes, and initiatives?

References:
Rumble, J. (2023). Thoughts on starting the CODATA Data Science Journal. Data Science Journal, 22, 13. https://doi.org/10.5334/dsj-2023-013

Speakers:

- Matt Mayernik (Session co-Moderator) - NSF National Center for Atmospheric Research, USA, Editor-in-Chief, Data Science Journal
- Gita Yadav (Session co-Moderator) - Scientist, National Institute of Plant Genome Research (NIPGR), New Delhi India; Member, CODATA IDPC; Professor of Data Science, IISER Bhopal, India Founder, #SemanticClimate; Member Editorial Team, Data Science Journal
- Deb Mohanty - Director, National Institute of Immunology, DBT, Govt of India; Member, CODATA India National Committee; Chairperson, Indian Biologica Data Center
- Mark Parsons - Former Editor-in-Chief, Data Science Journal; Member, Arctic Data Committee for the International Arctic Science Committee (IASC) and the Sustaining Arctic Observing Networks (SAON) CODATA representative to the International Polar Year 2032-3 Planning Committee
- Dimitris Symeonidis - University of Tartu, Institute of Computer Science Member, Digital Government Society
- Dr. Lili Zhang - Executive Director, GOSC IPO, member of CODATA IDPC Senior Research Scientist, CSTCloud, CNIC,CAS; Member Editorial Team, Data Science Journal

Presentations Session 4: Data Stewardship / 150

Bridging Metadata Standards: Implementing the CDIF Framework for Enhanced Interoperability in Data Observatory catalog.

Authors: Alejandro Antilao[1]; Álvaro Paredes Lizama[1]

[1] *Data Observatory Foundation, Chile*

Corresponding Authors: alvaro.paredes@dataobservatory.net, alejandro.antilao@dataobservatory.net

The exponential growth of data has intensified challenges in achieving cross-disciplinary and cross-repository interoperability. Metadata standards—such as DataCite, Dublin Core, and OpenAIRE—play a pivotal role in data discovery and reuse, yet their heterogeneity creates fragmentation. This proposal presents the implementation of the CODATA-CDIF Conceptual Domain Interoperability Framework (CDIF) within the Chilean National Data Observatory catalog, a platform aligned with the Chilean National Research and Development Agency (ANID). Our work addresses critical interoperability challenges by harmonizing metadata practices across repositories, enabling federated search, and enhancing compliance with national and international funding mandates.

Key Challenges

1. **Compatibility Between Standards**: Mapping metadata elements across standards (e.g., OpenAIRE's "Embargo Period" vs. DataCite's "Date" fields) often involves complex many-to-one relationships, complicating automated workflows.
2. **Semantic Ambiguity**: Disparate naming conventions (e.g., "language" vs. "linguisticCoverage") hinder federated searches. Locating datasets in Spanish, for instance, requires querying variants like "es-CL", "Spanish", or "es".
3. **Value Standardization**: Free-text fields introduce ambiguity (e.g., "spatialCoverage" entries like "Chile" vs. "CL" vs. geographic coordinates). While controlled vocabularies (e.g., GeoNames, Wikidata) resolve this, their integration demands meticulous curation.
4. **Automation Barriers**: Schema mismatches impede scalable integration, necessitating flexible infrastructure to reconcile differences.
5. **Metadata Loss**: Crosswalks between standards risk losing properties. For example, OpenAIRE's "Citation properties" lack equivalents in DataCite.

Implementation Approach
Our CDIF-based framework harmonizes metadata across ANID-aligned repositories, DataCite, Dublin Core and OpenAIRE through:

- **Crosswalk Development**: Semantic mappings resolve ambiguities, such as distinguishing DataCite's "Date:dateType='Issued'" from Dublin Core's "dateIssued". These mappings are validated against ANID's grant reporting requirements, ensuring coverage of critical fields like funding attribution.
- **Controlled Vocabularies**: FAIR-aligned authorities (e.g., ISO 639-3 for languages, UNESCO Thesaurus for disciplines) standardize values, reducing ambiguity in searches.
- **Middleware Infrastructure**: A RESTful API dynamically translates metadata queries across standards, leveraging AWS services (OpenSearch for indexing, S3 \+ Athena for querying) to deliver fast, scalable federated search.
- **Modular Ontology Extensions**: Gaps in discipline-specific metadata (e.g., geospatial granularity in OpenAIRE) are addressed by extending CDIF's ontology, ensuring compatibility with domain-specific needs.

Case Study: Searching for "Cactaceae Species" Models
To illustrate interoperability challenges, consider searching for "Cactaceae species" models across repositories:

- **Zenodo**: Filters allow limiting results to "resource type \= model" and "access \= open", yielding relevant datasets.
- **DataCite Commons**: No "model" filter exists; available filters (e.g., "Work type", "License") fail to narrow results effectively.
- **Harvard Dataverse**: Similar limitations—filters like "Dataverse Category" or "File type" do not align with the query.

Outcome: Only Zenodo returned targeted results. The lack of standardized filters in other repositories forces users into manual, time-consuming searches with no guarantee of success.

CDIF-Driven Solution:
Our implementation standardizes metadata properties and values across repositories. For example, the "resource type" field is mapped to equivalent terms (e.g., "Model", "Simulation") in DataCite and OpenAIRE, while controlled vocabularies enforce consistency (e.g., "language:es-CL" instead of free-text variants). An example spreadsheet example demonstrates mappings for 20+ properties, including:

- **Resource Type**: Mapped to "Model" (DataCite's "Software"), "Dataset" (OpenAIRE's "Research Product").
- **Temporal Coverage**: Aligned with ISO 8601 across standards.
- **Discipline**: Normalized using UNESCO and OECD taxonomies.

While still in progress, we believe this approach holds the potential to enable efficient searches across repositories and platforms in the future.

Conclusion and Relevance to SciDataCon 2025
Our CDIF implementation underscores that interoperability requires both technical solutions (crosswalks, APIs) and governance frameworks to align stakeholders. This work aligns with SciDataCon's focus on actionable strategies for data integration, offering insights for:

- **Policymakers**: Balancing flexibility with standardization in national metadata mandates.
- **Repository Managers**: Adopting modular frameworks to avoid schema lock-in.
- **Global Communities**: Addressing multilingual and multidisciplinary challenges through shared vocabularies.

We invite collaboration to scale this approach, particularly for repositories in underrepresented regions. By bridging the metadata gap, we empower researchers to focus on discovery—not data hunting.

Audience: data librarians, repository managers, metadata specialists, policymakers involved in research data management, and researchers interested in data discovery, interoperability, and the application of metadata standards.

Keywords: Metadata interoperability, CDIF, crosswalks, controlled vocabularies, FAIR data, federated search.

Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data / 152

A General-Purpose Framework for Structured, Reproducible, and Transparent Data Harmonization

Author: Jimmy Yu[1]

Co-authors: Marcos Martínez-Romero [1]; Mark Musen [1]; Matthew Horridge [1]; Mete Akdogan [1]

[1] *Stanford University*

Corresponding Authors: marcosmr@stanford.edu, horridge@stanford.edu, jkyu@stanford.edu, mete@stanford.edu, musen@stanford.edu

Despite efforts to improve the availability and accessibility of research datasets, interoperability remains a serious barrier to reuse. Data harmonization, the process of aligning data from disparate sources to a standardized schema, plays a key role in addressing data heterogeneity and enabling

integration and reuse. Consider a data scientist curating patient blood-glucose measurements as a precursor to secondary data analysis. Secondary data analysis often requires the integration of multiple datasets into an aggregate dataset large enough to provide sufficient statistical power to investigate a hypothesis or to provide adequate training data for a computational model. If the datasets are heterogeneous, the data scientist must first align them by implementing a data harmonization strategy.

Data harmonization entails mapping the schema of an original dataset to a standardized target schema, often chosen by the consumer of the data or established by a data repository. For example, the data scientist may need to compile a dataset that represents blood-glucose measurements using a categorical variable for clinically-relevant blood-glucose levels (e.g., hypoglycemic, normal, prediabetic, and diabetic). If the data scientist encounters a dataset that instead records numerical blood-glucose levels (e.g., in units of mg/dL), those data must be mapped from the numerical to the categorical schema to enable integration.

Recent developments in the literature of data harmonization have focused on establishing best practices for implementing data harmonization [1,2]. However, current harmonization practices suffer from limitations in reproducibility and flexibility, and existing tools to support harmonization are either specialized to domain-specific datasets or part of large integrated data management systems. Modern harmonization methods rely on hard-coded, manually maintained scripts, which limit adaptability when data standards and research objectives evolve. Consequently, the revision of outdated data representations or harmonization protocols, especially those developed early in a project, can be prohibitively labor-intensive. Moreover, harmonization protocols are conventionally documented in text, which leaves room for interpretation when implemented by a reader, resulting in harmonized datasets with opaque provenance and limited reproducibility.

To address these limitations, we have developed a general-purpose data harmonization framework that emphasizes reproducibility and transparency 3. Our framework achieves reproducibility using a novel strategy of building data transformations from standardized building blocks called "primitive" operations. For example, to transform numerical blood-glucose measurements to labeled categories, the data scientist would leverage the "Bin"primitive, which assigns numerical values to categorical labels using histogram ranges, e.g., hypoglycemic (<70 mg/dL), normal (70–99 mg/dL), prediabetic (100–124 mg/dL), and diabetic (≥125 mg/dL). Our framework defines a standard vocabulary of named primitive operations, including, among others, the "Bin"primitive above, the "ConvertUnits"primitive for performing unit conversions (e.g., from mg/dL to mmol/L), and the "Round"primitive for rounding a decimal value to a specific precision (e.g., from 1.15 to 1.2 at one significant digit of precision). Each primitive operation can be parameterized to fine-tune its behavior, such as by providing labels and their corresponding numerical ranges for the "Bin"primitive as demonstrated by the harmonization of blood-glucose levels. Moreover, primitives can be composed to achieve complex transformations, such as by performing unit conversion followed by decimal rounding to conform to a standardized precision level. Our declarative language of primitive operations provides standardization for representing previously ad hoc data transformations, while the abilities to parameterize and compose primitives provide the flexibility to build complex harmonization procedures.

The standardization established by the vocabulary of primitives affords an additional advantage: the ability to serialize harmonization protocols in a machine-readable format. A data transformation implemented using our framework is completely specified by naming the primitives that compose the transformation and by including the parameters used by each primitive. A harmonization protocol implemented as part of a computational harmonization workflow can be stored externally (e.g., in JSON format) and reconstructed from its serialization, thereby enabling the exact reproduction of a previously harmonized dataset and, in turn, guaranteeing reproducibility for downstream secondary data applications.

Additionally, our framework records all executed harmonization transformations in order to provide traceability for the resulting harmonized and integrated dataset. A different researcher can obtain the integrated blood-glucose dataset compiled by the data scientist in the earlier example and inspect the provenance of an individual element within the integrated dataset to trace its origin and identify the transformations applied to it as part of harmonization. This traceability allows the researcher to determine whether the harmonized dataset suits their research goals or whether to consider the original data under an alternative harmonization strategy.

In this presentation, we will detail the conceptual models and technical components of our harmonization framework and showcase real-world results from applying the framework within the RADx

Data Hub, a repository established by the National Institutes of Health (NIH) as part of the Rapid Acceleration of Diagnostics (RADx) program for hosting research data collected during the pandemic 4. The diversity of research programs, methods, and data types represented in the RADx repository demonstrates the framework's effectiveness in achieving principled, reproducible, and transparent data harmonization for complex heterogeneous data.

References
1. Fortier, I. et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int. J. Epidemiol. 46, 103–105 (2017).
2. Cheng, C. et al. A general primer for data harmonization. Sci. Data 11, 152 (2024).
3. Yu, J. K., et al. A general-purpose data harmonization framework: supporting reproducible and scalable data integration in the RADx Data Hub. Preprint at https://doi.org/10.48550/arXiv.2503.02115 (2025).
4. Martinez-Romero, M. et al. RADx Data Hub: a cloud platform for FAIR, harmonized COVID-19 data. Preprint at https://doi.org/10.48550/arXiv.2502.00265 (2025).

Poster Session / 155

Astronomy Data and Computing Services: Changing the way research software is developed and maintained

Author: Gregory Poole[1]

Co-author: Jarrod Hurley [1]

[1] *Swinburne University of Technology*

Corresponding Authors: jhurley@swin.edu.au, gpoole@swin.edu.au

Addressing the requirements for transparent, cost effective and high-impact research in this era of big data and cross-disciplinary research will require significant community changes to how research software is created, managed and maintained. In this talk I will introduce Astronomy Data And Computing Services (ADACS): a highly successful initiative of the Australian astronomy community established to address the need for coordinated national investment in software systems and training. Established by Astronomy Australia Limited (AAL) via funding from the federal National Collaborative Research Infrastructure Strategy (NCRIS), ADACS has operated since 2017 under the mandate of optimising the return from Australia's investment in astronomy computing infrastructure. In practice however, it is working to shift the culture and practices around the creation and maintenance of research software; towards a more modern, professional and sustainable model built upon maximising the expertise of cross-functional teams. I will describe how ADACS harnesses economies of both scope and scale to solve problems for researchers that otherwise could not have been solved; accelerating old science and enabling new science. I will also show how transferrable this model is to cross-disciplinary engagement - drawing from examples with commercial partners in satellite communications, mining, and medicine as well as other research domains including marine sciences and social network analysis - pointing the way to a model by which Australia could begin to address the challenges of effective research software management in the modern age.

Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research / 156

FAIR Challenges when using AI to Tailor Data for Climate Change Risks Applications

Author: Christian Pagé[1]

Co-author: Alessandro Spinuso [2]

[1] *CERFACS*

[2] *KNMI*

Corresponding Authors: alessandro.spinuso@knmi.nl, christian.page@cerfacs.fr

Conducting high-quality research increasingly involves complex workflows and the generation of numerous intermediate datasets. Achieving reproducibility requires the availability of extensive and well-structured information. To enable this, researchers need interfaces and tools that are not only user-friendly but also FAIR-aware.

Data analysis workflows typically span multiple tools and platforms, making interoperability of metadata a critical requirement for reproducibility. Ensuring that datasets and software comply with FAIR principles, particularly over time, remains a significant challenge. Such resources must be preserved in permanent, citable repositories to support long-term reuse and citation in scientific publications. This is especially relevant in the context of "long-tail research data," which encompasses not only datasets but also software and workflows.

The emergence of AI-powered tools has further amplified the need for transparency and reproducibility, as AI models are highly sensitive to training data, configuration parameters, and implementation details. Supporting the full lifecycle of research data—especially in domains like climate science—requires sustained and coordinated efforts.

In the climate research infrastructure community, various initiatives have emerged to meet these challenges. The Research Data Alliance (RDA) has been instrumental in providing guidelines and best practices to enhance FAIR compliance across data and software. Key RDA groups contributing to this effort include the FAIR Digital Object Fabric Interest Group (IG), the FAIR Data Maturity Model Working Group (WG), FAIR for Machine Learning (FAIR4ML) IG, and FAIR for Research Software (FAIR4RS) WG.

This presentation will explore the specific challenges of achieving FAIR compliance in the context of tailoring future climate simulation data for climate change risk applications. These efforts are part of the Horizon Europe IRISCC project, which involves multiple Demonstrators and Service Design Labs that rely on such data to support their objectives.

This project (IRISCC) is funded by the European Union under @HorizonEU research and innovation programm under grant agreement N°101131261.

Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global / 157

RACE: RMIT's Cloud Supercomputing Facility to Accelerate Data-Intensive Research

Authors: Robert Shen[1]; Matt Duckham[1]; Pier Marzocca[1]; Mark Easton[1]

[1] *RMIT University*

Corresponding Authors: pier.marzocca@rmit.edu.au, mark.easton@rmit.edu.au, robert.shen@rmit.edu.au, matt.duckham@rmit.edu

RMIT researchers are increasingly challenged by the size, dimension, and complexity of their data, the need to develop and run sophisticated data processing and analysis pipelines, and the need for computing-intensive simulations to compare with and interpret physical experiments. To partially address these challenges, the RMIT Advanced Computing Ecosystem (RACE) model provides scalable, high-performance resources through commercial cloud services. This pioneering facility, supported by AWS, Microsoft and AARNet partnerships and the Victorian Government's Higher Education Investment Fund, leverages the power of cloud computing to advance data-intensive research from local to global scales.

As Australia's first dedicated commercial cloud supercomputing facility at a university, RACE facilitates research excellence and strengthens partnerships across industry, government, and academia.

Through partnerships with AWS and Microsoft Azure, RACE offers cutting-edge cloud services, including advanced data storage, cloud computing, and supercomputing capabilities. Positioned at the forefront of technological innovation, RACE also facilitates access to emerging technologies like AI and quantum computing. This strategic integration accelerates the transition from data to knowledge, enabling swift, groundbreaking discoveries and advancements in research. By utilising state-of-the-art commercial cloud infrastructure, RACE enhances computational capacity and empowers researchers to explore new frontiers in data-intensive research.

Since its launch in October 2022, RACE has onboarded more than 800 researchers and PhD students, enabling rapid testing of ideas at speeds over 100 times faster than traditional on-site servers. In recognition of its outstanding contributions, RACE was recently awarded the CAUDIT Excellence in Research Support 2024, and two recent examples highlighted on the AWS website:
• Geospatial Data Management: Professor Matt Duckham partnered with RACE to manage, update, and learn from increasing volumes of geospatial data. They built a statewide knowledge graph of Victoria, with over 2 million interlinked geospatial records processed in under an hour, a task that would typically take days on a regular computer. These partnerships are pushing the boundaries of research and technology, contributing to Australian sovereign capability, and preparing RMIT for the next wave of geospatial data.
• Aerospace Simulations: The Sir Lawrence Wackett Defence & Aerospace Centre, led by Prof. Pier Marzocca, has partnered with RACE to accelerate aerospace simulations. This collaboration has significantly increased the number of analyses that can be run, reducing the time-to-solve by several orders of magnitude. The team is now able to handle 40 million simultaneous equations for more than 400,000 iterations per run, optimising designs and reducing the number of costly physical tests. Leveraging RACE's data and computing ecosystem, the centre has started storing input and output data securely, enabling seamless sharing with other collaborators. This enhances collaborative opportunities and allows for comprehensive data analysis and reuse, reducing the team's solving time for analyses from nearly 3 months to just 3.5 days using RACE.
In addition to its cloud services, RACE further enhances the research ecosystem through:
1. Comprehensive Training: Offers varied training formats to lower technical barriers, empowering users with essential cloud computing skills.
2. Tailored Consultation: Provides customised cloud solutions to optimise costs and maximise efficiency for innovative research projects.
3. Expert Support: Embeds specialists within project groups, providing direct assistance to overcome data and computing challenges, fostering a collaborative environment that drives excellence.

RACE leverages advanced computing capabilities, including AI and quantum computing, to transform complex data into actionable insights and groundbreaking discoveries through cross-disciplinary collaboration. By enhancing data management aligned with FAIR and CARE principles, RACE ensures data is efficiently accessible, interoperable, and ethically handled, empowering researchers to produce impactful, socially responsible insights. This approach fosters world-class outcomes in research and reinforces Australia's global standing in innovation.

Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI / 158

AI-Ready Data Workflows for Social Science and Humanities

Authors: Joan Giner Miguelez[1]; Raül Sirvent[1]; Eudald Lerga Felip[1]; Rosa M. Badia[1]; Mercè Crosas[1]

[1] *Barcelona Supercomputing Center*

Corresponding Authors: eudald.lerga@bsc.es, raul.sirvent@bsc.es, giner.joan@gmail.com, rosa.m.badia@bsc.es, merce.crosas@bsc.es

Recent advancements in artificial intelligence (AI) and access to new types of data have led to increased applications of AI in computational social science and humanities (SSH). A wide range of cutting-edge examples shows the results of bringing AI and SSH together, from the latest computer vision AI models used to detect archaeological traces in satellite imagery or to identify mounds on

historical maps 1, to recent language models used to analyze social network behaviors1 or to perform longitudinal studies on the entire scientific corpus between others 2. Furthermore, multimodal AI solutions combine different data types and are rapidly gaining popularity in social science and humanities. For instance, they are being used to generate language models capable of understanding ancient regional languages, thereby helping to enhance our understanding of history 3.

Despite AI's undeniable benefits in these fields, there are plenty of challenges to solve. Due to its inherent complexity and the need for high computational power, social and humanities scientists usually face a huge entry barrier to integrating these methods into their research 4. In addition, using AI methods entails a set of dangers that must be carefully considered through proper guardrails and validation methodologies 5 over the AI models and the data used to train them. Still, they are usually tied to specific use cases and continuously evolve in parallel with the evolution of AI technologies, which may be difficult for scientists to follow.

To address this challenge, we have been developing a set of computational workflows for social science and humanities at the Barcelona Supercomputing Center. Since every computational experiment can be described as a workflow, creation, execution, tracing, and validation techniques for workflows become essential to address the problems mentioned above. The proposed presentation aims to review the opportunities and challenges in the workflow design and gather insights from the data expert community.

A main goal of the workflows is to increase accessibility to the use of AI and computational power. First, the workflows aim to lower technical barriers by abstracting complexity as much as possible, letting scientists focus on the iterations between research questions, results, and refinements. Second, the workflows aim to optimize costs and allow for scaling-up experiments by optimally using shared public computational infrastructures (e.g., Exascale Supercomputers belonging to the EuroHPC network), making cutting-edge research more accessible and affordable to a broader community. The workflows community 6 tackles many of these challenges to lower the burden for end users while enabling the efficient and scalable execution of computational experiments, such as the convergence of AI and HPC workflows, the support of multi-facility workflows, exploiting heterogeneous HPC environments (GPUs, NVMs), and the achievement of FAIR computational workflows, among others.

However, workflows are not only built of technical components that abstract code complexity or optimize costs and performance. Workflows should also aim to provide state-of-the-art validation and guardrails techniques for integrating AI responsibly into scientific research. To this end, the goal is to facilitate the adoption of best practices throughout the validation of the AI models'outputs and the data curation and documentation. For instance, leveraging current AI-ready and machine-actionable metadata initiatives, such as Croissant 7 and DDI-CDI (project homepage https://zenodo.org/records/11236871), we can make the research data more interoperable and discoverable by design. Another key technique is what in the literature is known as eXplainable AI (XAI) 8, which encompasses the capture of relevant metadata during the execution of AI experiments that later helps to understand in detail both training and inference processes of AI algorithms. XAI techniques provide a way for users to learn not only how AI models have been trained, but also to better trust and understand the decisions taken by AI systems.

One of the pillars of science is to enable research reproducibility. Workflows, and in particular, tracking the provenance within a workflow, are key to enabling automatic computational reproducibility, as shown in 9. By establishing repeatable methods through workflows, we can define detailed logs and provenance records of the experiments, making them computationally reproducible by design and facilitating posterior verification processes. By integrating community-driven specifications such as RO-Crate 10 approaches into workflows and high-performance computing environments [11, 12], we can make computational AI workflows reproducible on demand. This work can allow computational social and humanities scientists to verify prior AI-based studies with less complexity.

In conclusion, our work aims to address the following issues with AI-Ready Data workflows: (i) Lowering the entry costs and reducing technical barriers for social and humanities scientists; (ii) Aiding in the adoption of best practices during data curation, preparation, and the validation of AI outputs; and (iii) Improving research reproducibility and posterior audit processes by integrating the AI workflows with RO-Crate-like solutions. While our proposal represents an initial effort to accelerate research and innovation in computational social science and humanities, maturing these AI workflows requires domain-specific expertise and a large variety of use cases. Consequently,

we advocate for creating open workflows as a community-driven effort to build better validation methodologies and practices that can be shared and utilized by a wide research community.

References:

1 I.Berganzo-Besga, et.al, Scientific Reports,2023.
2 A.Castro Torres et.al., ICCSI,2025.
3 M.Coll-Ardanuy et.al., Digital Humanities Conference,2025.
4 J.Calder, et.al., IEEE BITS,2022.
5 C.A.Bail, PNAS,2024.
6 R.F.Da Silva et.al., arXiv:2410.14943,2024.
7 M.Akhtar et.al., NeurIPS,2024.
8 R.Dwivedi et.al., ACM Computer Surveys,2023.
9 R.Sirvent et.al., IEE/ACM WORKS,2022.
10 S.Soiland-Reyes et.al., Data Science,2022.
11 S.Leo et.al., PLoS One,2024.
[12] R.F.Da Silva et.al., Computer,2024.

159

Research data stewardship in the Asia Pacific –What is happening now and how to move forward?

Authors: Hilary Shiue[None]; Meng-nan Lee[None]; Pei-shan Liao[1]; Shoichiro Hara[2]; Su Nee Goh[3]; Tyng-Ruey Chuang[4]; Willie Koh[5]; Yasuyuki Minamiyama[6]

[1] *Research Center for Humanities and Social Sciences, Academia Sinica*

[2] *Center for Southeast Asian Area Studies, Kyoto University*

[3] *Nanyang Technological University*

[4] *Academia Sinica, Taiwan*

[5] *Nanyang Technological University, Singapore*

[6] *Center for Social Research and Data archives, Institute of Social Science, The University of Tokyo, Japan*

Corresponding Authors: trc@iis.sinica.edu.tw, psliao@gate.sinica.edu.tw, hara.shoichiro.42w@st.kyoto-u.ac.jp, minamiyama@iss.u-tokyo.ac.jp, williekoh@ntu.edu.sg, mnl168@gate.sinica.edu.tw, hshiue@iis.sinica.edu.tw, sunee@ntu.edu.sg

Research funding organizations in the Asia Pacific are moving forward to develop support for research data management (RDM), especially for data-intensive research. At the same time, researchers and students are becoming more proactive in sharing datasets to enhance research visibility and impact. The convergence of top-down data policies and bottom-up initiatives is shaping a culture of data management and sharing that supports research integrity and trustworthy science.

This session offers a platform for institutions to share experiences with research data management support in the Asia Pacific. It will address what programs are in place, how they are implemented, how data communities engage with policymakers, and the challenges encountered. Four presenter groups from different research data contexts will share their ongoing stewardship efforts across the region.

1. Shoichiro Hara and Yasuyuki Minamiyama Research data management in Japanese academic research institutions has already been practiced in the natural sciences and engineering. However, it is difficult to say that there is sufficient understanding or implementation of research data in the humanities and social sciences. Academic institutions in Southeast Asian countries are also facing similar problems, and there is a delay in establishing appropriate systems and frameworks to manage, preserve, and reuse research data. Then, Kyoto University, which has built up networks with universities and various communities in Southeast Asia, started an international exchange project to support the promotion of RDM in this area. Through joint surveys of the current state of RDM in the Southeast Asian area, this exchange project aims to create a framework that promotes the development of human resources related to RDM activities adapted to each area's situation. To this end, we will provide training on GakuNin-RDM, a research data management platform, and

introduce examples of RDM implementation at higher research and educational institutions such as Kyoto University. At the same time, through this exchange, we will explore ways to promote RDM in the humanities and social sciences in Japan. In this session, we will provide an overview of our project and report on the current situation of RDM in Japan and Southeast Asia.

2. Su Nee GOH and Willie KOH At the Nanyang Technological University (NTU), Singapore, Data Management Plans (DMPs) became mandatory in 2016 and was integrated into the grant system. Principal investigators must submit a DMP as a prerequisite to access funding. NTU's DMP guides PIs to prepare for the sharing of non-sensitive research data on open access repositories, in accordance with the FAIR principles, Findable, Accessible, Interoperable, and Reusable. This contributes to research integrity, reproducibility, and efficient reuse.

However, the research landscape has evolved significantly in recent years due to geo-political developments and heightened attention to data security. In response, NTU is currently revamping its DMP template to better align with its institutional data security framework requirements, and to guide researchers also focus on protecting sensitive, personal, and confidential data, especially in collaborative projects with commercial and industrial companies.

The revised DMP will not only continue to support the planning for FAIR data but will also embrace FAIRER data principles, where 'E' stands for 'Ethical' and 'R' stand for 'Responsible'.

In this presentation, we will share our journey towards implementing a FAIRER DMP that reflects both transparency and the evolving responsibilities of data stewardship.

3. Pei-shan Liao and Meng-nan Lee Research data management has been an important issue in respond to open science and open data initiatives. Despite persistent challenges in sharing social science research data, Survey Research Data Archive (SRDA) in Taiwan has played an important role in the systematic acquisition, organization, and preservation of academic survey data, and its dissemination to scholars and researchers mainly for academic purposes. Standard data management procedures are applied to ensure data are compatible with the FAIR Data Principles.

In this presentation, we will share SRDA's experience with RDM practices. These include formatting and metadata requirements at the time of data submission, the standardization of questionnaires and codebooks during the curation stage, and strategies for format conversion and long-term preservation during storage and archiving. Throughout each stage, comprehensive regulations and systems are in place. As an important data infrastructure in Taiwan, SRDA has established clear principles for data categorization, de-identification, and access control based on the sensitivity level of the data. We will conclude by discussing the future challenges of RDM in managing social science data.

4. Tyng-Ruey Chuang and Hilary Szu Yin Shiue Despite global advances in open science, many regions still lack strong national policies or sustainable funding for research data management. The depositar lab in Taiwan has operated in this environment since 2013. Additionally, from 2017 and on, the lab has been developing a FAIR-aligned research data repository – the *depositar* – and supporting researchers' data sharing efforts largely through bottom-up initiatives in Academia Sinica, with some funding from Taiwan's National Science and Technology Council (NSTC). This presentation will share our experiences building and maintaining an RDM platform, advocating for institutional policy changes, and navigating some barriers to improved research data practices in different disciplines. We highlight our success in influencing the NSTC's DMP requirements for grants awarded for research database development, and reflect on the ongoing challenges of sustainability as the *depositar* transitions from a lab-hosted project to proposed institutional infrastructure. Our story shows the challenge for emerging open science cultures to move from grassroots actions to more stable and sustainable systems.

Poster Session / 161

Leveraging AI to Automatically Link Controlled Vocabulary Terms in Metadata

Author: Vyacheslav Tykhonov[1]

[1] *DANS-KNAW*

Corresponding Author: 4tykhonov@gmail.com

Automatically linking controlled vocabulary terms in metadata enhances semantic consistency and improves data interoperability across systems—particularly by connecting terms from frameworks such as OntoPortal, Skosmos, Wikidata, and others. This work presents an AI-driven approach that leverages Large Language Models (LLMs) in combination with knowledge graph techniques to identify and establish meaningful connections between controlled vocabulary terms. By harnessing the contextual understanding of LLMs and the structural capabilities of knowledge graphs, this method enables the automated enrichment and alignment of metadata vocabularies. The approach reduces manual curation efforts, supports scalable metadata harmonization, and opens new possibilities for intelligent data integration across domains.

165

FAIR Implementation Profiles, FAIRsharing, and FAIR²: Promoting the Informed and AI-Ready Reuse of Standards When Making Data FAIR

Authors: Erik Schultes[1]; Sean Hill[2]; Susanna Sansone[3]

Co-authors: Allyson Lister [3]; Cristina Gonzalez [2]; Tobias Kuhn [4]

[1] *GO FAIR Foundation*

[2] *Senscience*

[3] *FAIRsharing*

[4] *Knwledge Pixels*

Corresponding Authors: cristina.gonzalez@frontiersin.org, sean.hill@frontiersin.org, tobias@knowledgepixels.com, sa.sansone@gmail.com, allyson.lister@oerc.ox.ac.uk, erik@gofair.foundation

Overview:

This session aims to showcase the roles and relationships among FAIR Implementation Profiles (FIPs), FAIRsharing, and FAIR²—three complementary efforts that help researchers and data stewards to optimally reuse standards and make research data truly FAIR. The session will provide an overview of key challenges, introduce key technologies, and offer perspectives on how these tools are evolving to support responsible, reproducible, and increasingly AI-ready data reuse.

The Challenges:

At the core of putting FAIR into practice are the many and often complicated choices that must be made when selecting appropriate standards—terminologies, models, formats, minimal information requirements, identifier schemas—and suitable repositories and knowledge bases. These resources are essential to describe, report, and share research objects such as datasets, code, and workflows. Yet each project, group, or organisation typically follows its own norms. Without visibility into community preferences, it becomes difficult to find and reuse existing solutions. This increases the risk of needless reinvention and divergence in how standards are applied.

Even when datasets are declared FAIR, reuse in practice is often hindered by incomplete documentation, unclear standard usage, and metadata that is not structured for use in computational workflows. These issues are especially pronounced in interdisciplinary research and in AI-driven settings, where data needs to be both machine-actionable and richly contextualized to support automated discovery, integration, and analysis.

Practical Solutions: FAIR Implementation Profiles and FAIRsharing

To map this landscape and encourage convergence, the FAIR Implementation Profile (FIP) was introduced in 2019 by GO FAIR and developed in cooperation with ENVRI-FAIR. A FIP systematically represents a collection of declarations a community makes about its usage of FAIR Enabling Resources

(FERs). The FIP Wizard is a tool that supports the creation and publication of community-specific FIPs (now more than 450 intances of FIPs representing over 1200 accumulated FERs). Once published, FIPs from different communities can be openly searched with semantic precision and compared, providing critical insight into community norms and decision-making about standards.

FAIRsharing complements FIPs by offering an informative and educational service that describes and interrelates standards, databases, and data policies across all disciplines [https://blog.fairsharing.org/?p=971]. FAIRsharing records are curated, tagged by maturity, and continuously updated to reflect the dynamic evolution of the standards ecosystem. Communities can create FAIRsharing Collections to represent the resources they use in their FIPs or recommend to others. FAIRsharing also supports integration with tools such as the Data Stewardship Wizard (DSW), enabling data producers and stewards to generate FAIR assessments and data management plans based on trusted metadata.

FAIRsharing content is machine-actionable and accessible via API, enabling third-party tools to answer key questions about standards and repositories: "Which repositories support controlled access?", "Which identification schemas are used?", or "Which standards are suitable for describing software?"Where FIPs describe resources already registered in FAIRsharing, curated metadata is retrieved automatically and incorporated into FER nanopublications, including citations and provenance. If new resources are introduced, users are prompted to create corresponding records. This collaborative ecosystem supports the development of machine-learning approaches that assist communities in optimizing FAIR implementation strategies, improving alignment and interoperability across domains.

Expanding the Ecosystem: FAIR²

As a key use case, this session will also introduce FAIR², a new framework focused on enabling structured, reproducible, and AI-ready reuse of data. FAIR² responds to real-world challenges that remain even when data is technically FAIR—particularly those related to machine usability, provenance clarity, and contextual documentation. It introduces three new publication outputs: the FAIR² Data Article, FAIR² Data Package, and FAIR² Data Portal. These formats support deeply structured metadata (e.g., schema.org, Croissant, PROV-O), transparent provenance, and integration into modern data workflows.

As artificial intelligence becomes a common tool for knowledge discovery, synthesis, and prediction, FAIR² is designed to ensure datasets are not only discoverable but usable in AI systems. Its structured outputs provide the metadata and contextual scaffolding that intelligent agents and machine learning models require for interpreting, filtering, and applying data responsibly.

The presentation will explore how FAIR² can benefit from integration with FAIRsharing and FIPs—such as by referencing curated standards or reflecting community practices in structured metadata. These opportunities will be discussed as a pathway toward more coherent, machine-actionable, and ethically grounded data publication and reuse.

Session Format:

This 90-minute session will include three short presentations followed by open discussion and audience Q&A. The proposed agenda is:

Presenter 1 –GO FAIR Foundation (Schultes): An overview of FAIR Implementation Profiles (20 minutes, including 15-minute presentation)

Presenter 2 –FAIRsharing (Susanna Sansone): FAIRsharing and its role in FAIR assessment and assistance (20 minutes, including 15-minute presentation)

Presenter 3 –Senscience (Sean Hill): FAIR²: Structured publication for reproducible and AI-ready data reuse (20 minutes, including 15-minute presentation)

Discussion and Q&A –Integration, use cases, and future directions for FAIR data sharing (30 minutes)

Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI / 168

Evaluating the Effectiveness of an Open-Source Large Language Model in Drafting NIH Data Management Plans

Authors: Nahid Zeinali[1]; Bhavesh Patel[2]

Co-authors: Becky Hofstein Grady [3]; Maria Praetzellis [3]; Brian Riley [3]

[1] *FAIR Data Innovations Hub, California Medical Innovations Institute, San Diego, California, United States of America, 9212*

[2] *FAIR Data Innovations Hub, California Medical Innovations Institute, San Diego, California, United States of America, 92121*

[3] *California Digital Library, University of California Office of the President, Oakland, CA, United States of America, 94607*

Corresponding Authors: maria.praetzellis@ucop.edu, bpatel@calmi2.org, becky.grady@ucop.edu, nahidzeinali2021@gmail.com, brian.riley@ucop.edu

As funding agencies increasingly emphasize responsible data stewardship in alignment with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, Data Management Plans (DMPs) have become a core requirement in research proposals. This emphasis reflects a growing recognition that data serves as the foundation for scientific discovery and progress. Since 2023, the National Institutes of Health (NIH) requires the inclusion of a DMP in all grant applications, encouraging investigators to proactively think about how scientific data will be managed, preserved, and shared throughout the course of the research. Creating a high-quality, policy-compliant DMP is essential but remains a complex and time-consuming endeavor, particularly because researchers are often not trained in data management and/or lack support. Tools like DMP Tool, DMPonline, Data Steward Wizard are available to help but are limited to providing guidance and samples. Recent advancements in large language models (LLMs) present promising opportunities to streamline and automate aspects of data management planning. We envision a workflow where a DMP is drafted with the help of an LLM, given basic information about a grant proposal and data to be collected. The draft could then be reviewed and revised by researchers, effectively reducing time and complexity in the process. In this work, we evaluated the performance of an open source LLM in creating drafts of DMPs that are compliant with the NIH guidelines. The goal was to investigate if an LLM could be used off-the-shelf for DMP drafting without undergoing fine-tuning or other domain-specific performance improvements.

We assessed the capabilities of Meta's Llama 3.3 70B, which is one of the most advanced open-source LLMs with demonstrated high-performance for writing tasks. We tested its performance in reproducing 26 DMP examples provided by the NIH from previously funded proposals that cover different study and data types. As per the NIH guidelines, each DMP follows a standard structure consisting of 12 sections, each requiring information about different aspects of data management and sharing, including what data types will be collected, the standards that will be followed for formatting data and metadata, where and how data will be shared, and more. Our prompting strategy consisted of a single prompt that includes the name of the NIH Institute/Center where the proposal was submitted, details about the data collection (from human or non-human participants, number of subjects, data types to be collected) and the full NIH-provided DMP template. Since the research strategy was not included in the NIH-provided DMP examples, we included Section 1A of each DMP as an input into the prompt since this section is expected to provide details about data collection. We then compared side-by-side the 12 sections of the NIH DMP generated by the LLM with the related NIH-provided example to assess content accuracy and completeness using SBERT-semantic Similarity score.

Usually, a score of 0.7 or higher is considered a strong indication of semantic similarity, while scores above 0.8 suggest a very high degree of similarity. Our results show that the SBERT-semantic similarity score is the highest on average for Section 1A (0.86). This is expected since content from Section 1A was included as part of the prompt. The similarity score is low for the other 11 sections of the DMP, ranging between 0.46 and 0.64 across all 26 DMPs. Section 4A, which requires information about the repository where data will be archived, had the lowest similarity score on average (0.46). The second lowest was Section 1B, which requires details about the data that will be preserved and shared along with the rationale for the decision. The highest score on average (excluding section 1A) was for sections 4B, 4C, and 6 (all around 0.63-0.64), which require details about how data will be made findable, when data will be shared, and who will manage compliance with the DMP, respectively. Looking at the DMP-specific score (average score across all 12 sections), we observed the highest scores for a DMP about clinical and genomic data (0.67) and a DMP about survey and interview data (0.65). The lowest scores were observed for a DMP about non-human genomic data (0.55) and Hela cell whole genome sequence data (0.56).

These preliminary findings suggest that even a powerful open-source LLM like Llama 3.3 may not

be ready off-the-shelf to use for DMP drafting. Performance improvements are likely needed in certain areas of data management such as best practices for sharing data, and on specific data domains such as non-human genomic data. Such improvements can be achieved through several approaches that will be tested in future studies, such as segmented prompting, retrieval-augmented generation (RAG), and domain-specific fine-tuning. We may also find that additional details from the researcher are needed in the prompt that may help the LLM draft the DMP correctly for that particular study we will be exploring. Future work will also assess the performance of additional LLMs, including commercial ones, to achieve a thorough benchmarking of LLMs performance off-the-shelf prior to any improvements. To ensure more robust evaluation, we also plan to incorporate expert human reviews alongside automated metrics, offering deeper insight into the quality, completeness, and usability of LLM-drafted DMPs, as it is possible that the generated DMPs are perfectly valid options for a study in that domain, even if it didn't happen to match the examples in this case. This work has practical implications for research data managers, librarians, grant administrators, tool developers, and funders interested in leveraging LLMs to support compliance with evolving data-sharing policies.

Presentations Session 6: The Transformative Role of Data in SDGs and Disaster Resilience / 169

Adapting to Climate change with Open Science : Experiences from the CLIMATE-ADAPT4EOSC project

Authors: Alberto Azevedo[1]; Athanasios Sfetsos[2]; Catherine Freissinet[3]; Dimitra Panou[2]; Feroz Farazi[4]; Marc Pattinson[5]; Matti Heikkurinen[6]; Mohammad Azizur Rahman[7]; Simon Hodson[6]; Stian Soiland-Reyes[8]

[1] *Laboratório Nacional de Engenharia Civil (LNEC), Lisbon, Portugal*

[2] *National Centre for Scientific Research "Demokritos" Institute of Nuclear & Radiological Sciences & Technology, Energy & Safety Environmental Research Laboratory*

[3] *Artelia*

[4] *University of Cambridge*

[5] *GAC Group Sophia Antipolis*

[6] *CODATA*

[7] *Technovative Solutions LTD*

[8] *University of Manchester*

Corresponding Authors: catherine.freissinet@arteliagroup.com, mpattinson@group-gac.com, simon@codata.org, soiland-reyes@manchester.ac.uk, aazevedo@lnec.pt, panou@fleming.gr, msff2@cam.ac.uk, aziz@technovativesolutions.co.uk, ts@ipta.demokritos.gr, matti@codata.org

Core objectives of the EU Mission on Climate Adaptation include ensuring that all Europeans have access to information on climate risks by 2030, supporting local authorities in developing risk management plans, and designing transformative strategies for 150 communities and regions to lead healthier and more prosperous lives. A central point of this work is addressing the challenges of cross-domain research by integrating data, knowledge, and solutions from various scientific disciplines. It therefore calls for collaboration among civil society organisations, authorities, researchers, and other stakeholders to develop comprehensive and innovative strategies for climate resilience.

The integration of data spaces such as COPERNICUS, GEOSS, DRMKC, and CLIMATE-ADAPT into the European Open Science Cloud (EOSC) represents a strategic initiative to coordinate climate adaptation efforts across Europe. These climate data spaces are essential for capturing diverse, large-scale observational data and enabling interdisciplinary research critical to understanding and mitigating climate change impacts. While compliance with European Directives supports data harmonization across Europe, it remains limited to specific domains. By aligning with the EOSC vision, this integration fosters a holistic approach to enhance resilience, ensuring that insights and methodologies developed in the EU member states can be effectively applied and adapted to diverse contexts.

A significant challenge in climate change adaptation research is gaining access to high-quality datasets that comply with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. By prioritising FAIR and open data practices, the mission aims to overcome legal and technical barriers to

data sharing, improving data availability, quality, and interoperability. Implementing these principles is essential to unlocking the full potential of existing datasets, enabling easier discovery, access, use and re-use of data crucial for advancing climate adaptation research both in the EU and globally.

CLIMATE-ADAPT4EOSC aims to eliminate existing barriers to climate data access and interaction with operational climate data spaces, fostering a collaborative research environment where data can flow seamlessly between researchers and other stakeholders contributing to the climate adaptation mission. To achieve this, we will establish seamless interaction between the EOSC e-infrastructure and various EU and national climate data spaces by aligning precisely with EOSC's e-infrastructure plan.

CLIMATE-ADAPT4EOSC project places strong emphasis on ensuring that data management practices are compliant with legal standards, semantically aligned for cross-disciplinary research, technically robust for integration, and organisationally structured to support the diverse needs of the research community. Broadly, we aim to deploy the following two major services in EOSC-core (i) CLIMATE-ADAPTdata4EOSC: a service for EOSC to generate and share FAIR data, metadata, and digital research objects, and (ii) CLIMATE-ADAPTservice4EOSC: a suite of services for EOSC to analysing, processing, and modelling data to generate new insights.

We will embed a novel Ontology-Based EOSC Climate-Adapt Knowledge Graph within this service to enhance the FAIRness of data in climate change adaptation research. As part of this effort, we will recommend five interoperability frameworks to the EOSC community: (1) Technical Interoperability Framework, (2) Semantic Interoperability Framework, (3) Cross-domain Interoperability Framework, building on the work of the WorldFAIR project, (4) Organisational Interoperability Framework and (5) Legal Interoperability Framework. Taken together, these frameworks will provide a comprehensive and standards-based approach to enable data repeatable (and reproducible) data combination, integration and reuse for climate adaptation. We will use and further develop the SIMPL1, an open source middleware, to enable data FAIRness across EU data spaces and to ensure that EOSC stakeholders can efficiently and securely collaborate and promoting data interoperability.

To demonstrate the effectiveness of our CLIMATE-ADAPTdata4EOSC value of sharing and reusing research data, our CLIMATE-ADAPTservice4EOSC will include four novel services: (a) OPENHIDRA –a service designed to empower users and stakeholders to adapt to climate change in coastal regions (b) Shrink-Swell from Space2Earth Service (3SES)- a service for static and dynamic mapping of shrink-swell risks affecting both old and new infrastructures prone to climate change (c) Digital Twins for Just Climate Urban Resilience Service (Just-CURS) –a solution tailored for enhancing climate resilience in socially vulnerable communities (d) Big Data Analytics (BDAnalytics) –a framework-as-a-service for comprehensive climate risk assessment.

To showcase the robustness, effectiveness and impact of our innovations –and the value of data sharing –we will implement all CLIMATE-ADAPT4EOSC novel methods, tools, and services in eleven real-world case scenarios: three use cases (UCs) and eight replication use cases (R-UCs). These will be demonstrated in three rounds across five EU member states: France, Greece, Portugal, Cyprus, and Poland.

Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global / 172

Helmholtz Metadata Collaboration - Lessons Learned on the Path to a FAIR data space for Helmholtz

Author: Constanze Curdt[None]

Co-authors: Witold Arndt [1]; Oonagh Brendike-Mannix [2]; Volker Hofmann [3]; Thomas Jejkal [4]; Christine Lemster [5]; Sören Lorenz [6]; Marco Nolden [7]; Emanuel Söding [6]; Wolfgang Süß [4]

[1] *DLR*

[2] *HZB*

[3] *FZJ*

[4] *KIT*

[5] *GEOMAR Helmholtz Centre for Ocean Research Kiel*

[6] *GEOMAR*

[7] *DKFZ*

Corresponding Authors: ccurdt@geomar.de, clemster@geomar.de

In 2019, the Helmholtz Association of German Research Centres launched the Helmholtz Metadata Collaboration (HMC) platform as part of its Information and Data Science Framework to translate global metadata concepts into application and to harmonize scientific practice within the association. HMC's mission is to facilitate the visibility and reusability of data within the Helmholtz Association and beyond, and to promote the FAIRness of Helmholtz data. HMC aims to create a sustainable, distributed, and semantically enriched Helmholtz FAIR data space that spans across the 18 autonomous Helmholtz centers and its six research fields. (Aeronautics, Space & Transport, Earth & Environment, Energy, Health, Information, and Matter).

HMC focuses on three primary areas of activity: (1) assessing and monitoring the state of FAIR data, (2) facilitating connectivity of Helmholtz research data, and (3) transforming metadata recommendations into implementations. These activities address multiple stakeholders within Helmholtz, including the scientific community, data professionals, research data infrastructures, technicians, and administration. By increasing the coherence and connectivity of metadata, HMC aims to promote a more harmonized and efficient research environment.

Over the past years, we analyzed the state of research data management and data FAIRness within Helmholtz through surveys [1,2] and FAIR assessments 3. Awareness was raised about the importance of metadata through outreach events 4, training 5 and consulting. We developed a technical backbone for connecting FAIR data in Helmholtz, including a Helmholtz Knowledge Graph 6, to establish the Helmholtz FAIR data space. Together with our communities, we worked towards the implementation of aligned FAIR metadata practices and recommendations. Since 2020, HMC has also funded 36 community projects across Helmholtz 7 to address practical metadata challenges. Through our contribution to various working groups, projects and panels we are closely intertwined with national (e.g. NFDI) and international (e.g. EOSC, RDA, CODATA) initiatives in research data management.

This contribution will provide details on the HMC's approach, highlight key results, and share lessons learned on the path to creating a Helmholtz FAIR data space. By sharing our experiences, we want to engage in active discussions on metadata, relevant stakeholders, and the advancement of a FAIR data ecosystem. We hope that our progress so far will provide valuable insights for others who are on a similar path, and we look forward to exchanging ideas and best practices with the community.

References:
1 Arndt, W., Gerlich, S. C., Hofmann, V., Kubin, M., Kulla, L., Lemster, C., Mannix, O., Rink, K., Nolden, M., Schweikert, J., Shankar, S., Söding, E., Steinmeier, L., & Süß, W. (2022). A survey on research data management practices among researchers in the Helmholtz Association (HMC Report). https://doi.org/10.3289/HMC_publ_05.
2 Gerlich, S. C., Kubin, M., Kulla, L., Lemster, C., Schmidt, A., Schweikert, J., Shankar, S. & Stucky, K.-U. (2025). A survey on the status quo, gaps and needs among research data professionals in the Helmholtz Association (HMC Report). https://doi.org/10.3289/HMC_publ_08.
3 HMC FAIR Dashboard. https://fairdashboard.helmholtz-metadaten.de/ (accessed: 15 April 2025).
4 HMC FAIR Friday seminar series: https://helmholtz-metadaten.de/en/fair-friday (accessed: 15 April 2025).
5 Fundamentals of Scientific Metadata training course: https://carpentries-incubator.github.io/scientific-metadata/ (accessed: 15 April 2025).
6 Helmholtz Knowledge Graph: https://helmholtz-metadaten.de/en/unhide_helmholtz-kg (accessed: 15 April 2025).
7 HMC Projects: https://helmholtz-metadaten.de/projects (accessed: 15 April 2025).

Presentations Session 1: CAREful Indigenous Data Governance / 174

The implications of CODATA's priorities for countries such as South Africa

Author: Antony Cooper[1]

Co-author: Anwar Vahed [2]

[1] *CSIR*

[2] *Retired*

Corresponding Authors: avahed@live.com, acooper@csir.co.za

Founded in 1966, CODATA is the Committee on Data of the International Science Council (ISC). CODATA's vision "is of a world in which science is empowered to address universal challenges through the transparent, trustworthy and equitable use of data and information"[CODATA 2025]. CODATA's mission "is to connect data and people to advance science and improve our world"[CODATA 2025]. CODATA has three long-standing priorities:
• Data Policy, aimed at meeting current and urgent challenges at both the international and national levels.
• Data Science and Stewardship for the professions, through practical initiatives such as developing terminology and fundamental constants.
• Data skills capacity building activities, particularly for early career researchers, repository professionals and those researchers interested in the FAIR principles for interoperability: data that are findable, accessible, interoperable and reusable [CODATA 2025].

CODATA's Strategic Plan for 2023-2027 has four thematic priorities [CODATA 2023]:
1. Making Data Work for the Cross-Domain Grand Challenges, to support the plans of the ISC. The ISC identifies these grand challenges when necessary, such as climate change, sustainable development and reducing the risks of disasters. CODATA is doing this through the WorldFAIR+ initiative.
2. Improving Data Policy: this focuses on the principles that data should be FAIR, that is, findable, accessible, interoperable and reusable (see below).
3. Advancing the Science of Data and Data Stewardship: this is to promote evidence-based research and policies and the required systems, standards and infrastructure.
4. Enhancing Data Skills: these are needed to ensure that the data stewardship and science are trustworthy, equitable and transparent [CODATA 2023].

The FAIR principles have become influential and are widely cited. They were developed to facilitate reusing existing data holdings easily and correctly, not just by humans but also automatically by computers [Wilkinson et al 2016]. They are:
• Findable: such as using unambiguous, persistent identifiers and providing metadata that allow the data to be discovered.
• Accessible: such as explicit access conditions and well-described technical access protocols.
• Interoperable: such as standard machine-encoded definitions of the key concepts, variables, etc.
• Reusable: such as clear licensing and details of fair use, and suitable metadata, including provenance and quality [Wilkinson et al 2016].

The FAIR principles have led to related initiatives, such as:
• Machine-actionable FAIR Implementation Profiles (FIPs) to help different disciples implement the FAIR principles.
• The Cross-Domain Interoperability Framework (CDIF), which provides a framework of standards, particularly for the interoperable and reusable FAIR principles. See Gregory et al [2024].
• The Leiden Declaration on FAIR Digital Objects (FDOs), for individuals and organisations to commit to FAIR data, open standards and increased reliability and trustworthiness [FDO Forum 2022].

WorldFAIR was a successful project aimed at collaboration to implement the FAIR principles, done through 11 case studies. CODATA then launched the WorldFAIR+ initiative to focus on practical guidance and technical recommendations to increase the availability of FAIR data. WordlFAIR+ includes projects around the world, including Data Science Without Borders project, with several African countries participating.

For centuries, the San peoples have been studied by academics, but with concern over being objectified, doubt over usefulness and even perceptions of actual harm, the San leaders initiated the San

Code of Research Ethics [South African San Institute 2017]. With similar initiatives by other Indigenous Peoples around the world, this led to the CARE Principles for Indigenous Data Governance, to balance protecting Indigenous rights and interests with open data, etc [Carroll et al 2020]. The CARE principles are:
• Collective Benefit: Indigenous data must help Indigenous Peoples achieve inclusive development and innovation and realise equitable outcomes.
• Authority to Control: in line with the United Nations Declaration on the Rights of Indigenous Peoples [UNGA 2007], Indigenous Peoples need to be able to govern their own data and have sovereignty to facilitate greater Indigenous self-determination [Hudson et al 2023].
• Responsibility: researchers need to nurture respectful relationships with Indigenous Peoples, including developing their capacities, and embedding the data within the languages and cultures of the Indigenous Peoples.
• Ethics: the research must protect the rights and wellbeing of the Indigenous Peoples throughout the data lifecycles to minimise harm, maximise benefits, promote justice and allow for future use – but the Indigenous Peoples must determine this [Carroll et al 2020].

The CARE principles work together with the FAIR principles, rather than contradicting or competing with them. For South Africa, the CARE principles are perhaps more important than the FAIR principles.

This paper will explore the implications of the priorities listed above for countries such as South Africa, southern Africa or Africa as a whole. Before a data set can become FAIR, it needs to exist, and there are concerns that much data have been lost in South Africa. The Promotion of Access to Information Act [South Africa 2000] is similar to freedom of information legislation in other countries. The Act requires public bodies to compile and make readily available a manual on what records can be accessed and how. However, compliance is poor, even though there are officially penalties such as fines or imprisonment for non-compliance [Fourie 2023]. Unsurprisingly, this does not bode well for the availability of data sets, never mind those that comply with the FAIR and CARE principles.

CODATA is looking for more partners to provide case studies for WorldFAIR+, but the obvious limitation for South African organisations is funding. Besides the global problems, the South African economy has been doing poorly for some years.

Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data / 175

FAIRifying at Scale: Lessons from NIAID's Ecosystem-Wide Approach to Repository Interoperability

Authors: Christine Kirkpatrick[1]; John Graybeal[2]

[1] *San Diego Supercomputer Center / CODATA*

[2] *San Diego Supercomputer Center / GO FAIR US*

Corresponding Authors: nhoebel@kmotifs.com, jbgraybeal@sdsc.edu, christine@sdsc.edu

As global communities continue to adopt the FAIR Principles, many organizations face the challenge of not just FAIRifying individual datasets, but entire ecosystems of data repositories and services. The many tools and assessments developed for FAIRtend to be customized to a particular data architecture, especially data organized in a file with a DOI and listed on a website. The National Institute of Allergy and Infectious Diseases (NIAID), an arm of the U.S. National Institutes of Health, manages a data landscape that includes infectious, allergic, and immunologic data. NIAID data has been at the forefront of research and therapeutics for today's biggest public health challenges, and this project targets increasing the discovery and reuse of these data. With this diversity of topic, use, and audience, assessing NIAID data resources for FAIRness brings a host of challenges: diverse data architectures, varying levels of entity management, declaration, and resolution, and data that has been collected over many time scales—decades in some cases. This session will share insights and

actionable strategies from the NIAID Data Landscaping and FAIRification project undertaken by GO FAIR US and partners including:

-Framing a data resource landscape for an organization
-Customized FAIR assessment instruments
-Approaching record-based repositories (as opposed to file-based ones)
-Constructing a FAIR common strategy for your organization
-Assessing individual data resources' progress towards a common approach
-Using Impact assessment to drive adoption of community- and standards-based metadata and PIDs
-Collecting and extending assessment data in a FAIR, AI-Ready way for further exploration
-Increasing the value and use of citations for all types of data products

The session will delve into a novel, stakeholder-informed method for assessing how well data repositories meet the FAIR Principles—developed and applied by GO FAIR US with support from GO FAIR Foundation with speakers from GO FAIR US, NIAID, and National Center for Atmospheric Research. The method builds on and extends prior impactful work1,3 from and including FAIRsFAIR2, the RDA Data Maturity Model Working Group4, the National Science Foundation's EarthCube5, GO FAIR Foundation's FAIR Implementation Profiles, and others. We will also present a number of novel tools that have been developed to maximize our understanding, and discuss how these have supported our progress on the project.

For this project we designed tailored questionnaires and interviews for different repository roles (managers, technical staff, data depositors, data users) to provide a multi-dimensional view of FAIR practices. The assessment products combine desk-based research and structured templates to create a detailed FAIR baseline for each repository—including metadata practices, technical infrastructure, and governance. This addresses the limitations of automated tools when assessing complex or secure repositories, and offers a qualitative, human-centered alternative. The assessment is then used to produce targeted FAIRification strategies that align with a repository's goals, constraints, and domain-specific practices. This method enables comparative analysis across multiple repositories and informs broader strategic planning for FAIR implementation, while supporting repository ownership and buy-in. We will discuss several common challenges including inconsistent understanding of metadata, the need for trust and buy-in from repository staff, and difficulty in addressing interoperability goals while supporting unique repository approaches. In addition, our API strategy considers a landscape with hundreds of diverse data systems that implement web-based Application Programming Interfaces (APIs) in many different ways. The sociotechnical aspects of top-down vs. bottom up change and how these impacted the FAIRification process also will be discussed.

We will present a toolkit of the instruments, processes, and strategies developed by the GO FAIR US team. The results support collecting, analyzing, and communicating FAIR knowledge of the ecosystem, and motivating desirable changes. Specific advances include a detailed framework for FAIR data collection (on-line and in-person role-based questionnaires), an analysis of methods to evaluate record-based repositories (i.e., that don't contain data files), development of repository profiles and FAIR summaries, and the use of the FAIR Implementation Profile (FIP) and FIP Wizard to collect raw repository information that is then analyzed by a FAIR Evaluator tool.

Lastly, the session will illuminate the steps and challenges related to FAIRifying the information. That is, how does one collect FAIR assessment data that is AI Ready and can be converted into a knowledge graph for further exploration? Audience questions, feedback, and discussion will be encouraged. The session is designed for data producers, data repository or resource owners and technical managers, data stewards, and those interested in gaining further data stewardship skills.

References:

1. European Commission: Directorate-General for Research and Innovation, European Research Data Landscape –Final report, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2777/3648
2. FAIRsFAIR's "Fostering FAIR Data Practices in Europe"project documentation
3. Mathers, B.J., L'Hours, H., Increasing the Reuse of Data through FAIR-enabling the Certification of Trustworthy Digital Repositories, https://doi.org/10.5334/dsj-2020-041 which explores the

alignment of the FAIR Data Principles with the CoreTrustSeal Trustworthy Digital Requirements

4. Recommendations from a Research Data Alliance Data Maturity Model Working Group, which identifies FAIR data maturity model indicators (Bahim et al., 2020)
5. Questions that EarthCube Office staff used to interview EarthCube project leads about the organization's impact (Stocks and Evans, 2022)
6. "Data Science Dispatch."NIAID NIH. 2025. https://www.niaid.nih.gov/research/data-science-dispatch
7. "GO FAIR US awarded a NIAID FAIR data and ecosystem contract by Frederick National Laboratory for Cancer Research."GO FAIR US, Dec.
   (a) Press release. https://www.gofair.us/post/go-fair-us-awarded-a-niaid-fair-data-and-ecosystem-contract-by-frederick-national-laboratory-for-can

Presentations Session 9: Empowering the global data community for impact, equity, and inclusion / Education / 176

The CODATA RDM Terminology: a community-focused approach to semantic interoperability

Author: L Molloy[1]

[1] *CODATA*

Corresponding Author: laura@codata.org

Introduction

As part of efforts to make research data more FAIR, semantic interoperability is important to consider. Standards, controlled vocabularies, and terminologies are well established types of FAIR-enabling resources that help us create interoperable systems and metadata. The CODATA Research Data Management Terminology (RDMT) is one such semantic resource that emerged from the former CASRAI Glossary to become a useful and usable, up-to-date reference tool for research data managers and other professionals involved in creating, managing and preserving research data. The CODATA RDMT is now published as a FAIR terminology through the Australian Research Data Commons (ARDC)'s Research Vocabularies Australia service.

This paper will discuss the purpose and value of the Terminology, and how we have taken a community-focused approach to its development to maximise the ability of the RDM community to contribute their expertise. We hope this will be of interest to those looking for a terminology for the RDM field for their own use, and/or those who are exploring approaches to terminology management and development.

Context

The CASRAI Glossary was intended as a practical reference for individuals and groups concerned with the improvement of research data management (RDM). In 2020, CASRAI asked CODATA to assume responsibility for the curation of this valued resource - a natural fit given CODATA's previous participation in the stewardship and development of the glossary, and our links with a heterogenous range of task groups, working groups, national committees, and projects, giving us a rich network of expertise on which to draw.

The goal of the refreshed Terminology is to gather the key terms needed for a common understanding of the research data management domain. In this context, RDM refers to research data management practices covering the entire lifecycle of the data, from planning research to conducting it, and from backing up data as it is created and used to long-term preservation of data objects after the research investigation has concluded.

We realised that the real power of the resource was its ability to support meaning across different contexts, making it a terminology rather than a glossary, which resulted in the change of name. Then

we recruited a Working Group to review the Terminology and contribute expertise from different relevant sectors, and established a rolling cycle of reviews with a fresh group of experts recruited for each review, to ensure diversity of input.

Method and approach

The RDMT is biennially reviewed and refreshed by an expert Working Group, which is responsible for creating a stable and sustainably governed standard terminology of community-accepted terms and definitions for concepts relevant to research data management, and keeping this terminology relevant by maintaining it as a 'living document'that is updated regularly. To those ends, the RDM Terminology Working Group uses a lightweight and pragmatic process to review the current Terminology and suggest any edits, additions and removals that are required to develop and improve the set of terms.

The Terminology is not an attempt to list every concept, tool and standard relevant to RDM; rather, it focuses on terms without easily found authoritative definitions elsewhere and offers an accessible definition in the context of contemporary RDM. Definitions are intended to be clear and unambiguous, and where possible, fit with common usage. We aim to produce definitions that are apposite across RDM activities of key stakeholders, including those working on research data management within the context of research, data management, digital curation and preservation, research management, research policy, open data advocacy, computer science, information management, research administration, library, scholarly publishing, digital archiving and research funding roles. Some terms may have more than one definition, in which case the relevant context is specified.

Reflections and future directions

Our aim for the RDMT is to create and maintain the highest quality terminology possible within the bounds of our resources. To that end, we are working towards bringing definitions closer to the format specified in the ISO standard 704, for interoperability as well as for quality reasons. The approach to publication is also influenced by the principles laid out in the influential paper, "Ten simple rules for making a vocabulary FAIR", and is supported by our collaboration with ARDC to publish the Terminology in machine-actionable form.

We are keenly aware that the Terminology should serve as much of the global RDM community as possible. The 2025 review cycle involved participants based across thirteen countries. This is a welcome marked increase in geographical spread compared to the previous cycle which had representation from six countries. We are also encouraged by the interest shown in creating translations of the reference version. To date we have received enquiries about translations into other variants of English, variants of Chinese, French, Spanish and Portuguese.

We are also delighted that, in contrast to the broader trends within science5, we attract a high level of female expert participation: in our current Working Group, thirteen participants from nineteen are female. In the previous round, thirteen of seventeen participants were female. We are keen to celebrate and continue this success and do what we can to further improve diverse, cross-community participation.

Our paper will provide an overview of these various aspects of the development and management of the RDM Terminology, our approach to community review, and our plans for future development of this key resource to help improve semantic interoperability within the RDM and FAIR data communities.

177

Bridging the Data Science and Research Data Communities Through Education and Shared Practices

Authors: Bonnie Carroll[1]; Christine Kirkpatrick[2]; Kelsey Druken[3]; Leo Lahti[4]; Padmanabhan Seshaiyer[5]; Rania Kosti[6]

[1] *Co-Chair, US National Committee for CODATA*

[2] *San Diego Supercomputer Center / CODATA*

[3] *ACCESS-NRI*

[4] *University of Turku*

[5] *George Mason University, US National Committee for CODATA*

[6] *National Academies of Sciences, Engineering, Medicine*

Corresponding Authors: leo.lahti@utu.fi, kelsey.drucken@anu.edu.au, christine@sdsc.edu, draban@univ.haifa.ac.il

The data science and research data communities share many common goals and challenges. Despite this, the two communities tend to have separate venues for convening, membership, and educational tracks. This session of short presentations and panel discussion will explore some of the ways that these two worlds can come together in the areas of education, training, and data stewardship practices.

This session explores the evolving intersection of the research data and data science communities through diverse lenses ranging from foundational stewardship to citizen engagement with speakers from across the world. Leo Lahti emphasizes the critical journey from observation to interpretation, underscoring the importance of data throughout the research lifecycle. Daphne Raban highlights how data stewardship serves as a vital bridge between research and data science, ensuring that data is managed, documented, and reused effectively. Phil Bourne further examines this bridge, focusing on practical integration between data science techniques and robust research data infrastructures and suggesting ways the organizations that support both communities can come together. Padmanabhan Seshaiyer brings an educational perspective, advocating for embedding research data principles into K-12 and community college bridge programs to foster inclusive data science literacy. Carolynne Hultquist (invited) offers a citizen science and earth sciences vantage point, illustrating how environmental hazard monitoring can blend data stewardship and data science in participatory ways. Kelsey Druken will discuss how ACCESS-NRI is embedding FAIR through software and data workflows for Australia's climate modelling infrastructure. Together, these contributions reveal key synergies and a shared commitment to building interoperable, ethical, and impactful data ecosystems.
The session is meant to elicit ideas and suggestions for action assembled from the audience as well. An outcome will be input for a roadmap and a similar session proposal at the next Academic Data Science Alliance (ADSA) meeting. This proposed session, as well as the one at ADSA, are unique and have not been held before. The closest approximation were the prior sessions that sought to bridge the gap between the research data community and high performance computing (HPC). Sessions were held at IDW 2022, 2024, ISC (IDW for HPC people in Europe) and Supercomputing 22 & 23. Establishing baselines of understanding and presenting shared priorities and goals was key for driving progress and creating awareness of the gap. Aside from companion presentations, a desired outcome might be a future CODATA Task Group, an RDA interest group, or an ADSA activity. Such a group would look at a roadmap for shared education - including professional development and training opportunities, as well as ecosystem tools and services, and shared research priorities.

Issues to Be Addressed by the Session

1) Fragmentation Across Communities

Despite overlapping goals, research data and data science communities often operate in siloed venues with different infrastructures, memberships, and training programs. This session will explore how to foster cross-community dialogue and collaboration.

2) Disconnect Between Practice and Infrastructure

Research data practices (e.g., data curation, stewardship, provenance) are not always integrated into the day-to-day workflows of data science, limiting reproducibility, transparency, and reuse. How can we better embed stewardship into data science infrastructures?

3) Educational Gaps and Opportunities

There is a lack of integrated education and training pathways that bridge research data management and data science skills. The session will address how to co-develop curricula, professional development, and early education programs that reflect both domains.

4) Institutional and Organizational Coordination

Many data-related and data science-related consortia (e.g., CODATA, ADSA, RDA, WDS, GO FAIR) operate in overlapping and unique domains. The session will discuss how these entities might coordinate activities, policies, and collaborate on funding priorities to support shared goals.

5) Roadmapping Shared Goals

There is currently no unified roadmap for aligning the research data and data science ecosystems. The session aims to collect input from the audience to help define shared priorities, pain points, and actions for a future roadmap and collaboration agenda.

6) Sustainability of Shared Ecosystems

Tools, standards, and services that support FAIR, open, and ethical data practices often struggle with sustainability. The session will explore how shared infrastructures and cross-community collaboration can improve the resilience and utility of data ecosystems.

7) Governance, Ethics, and Impact

As both fields intersect with sensitive data (health, environment, education), there's a need for shared approaches to governance, equity, and ethical AI/data practices. How can we co-create responsible frameworks for stewardship across domains?

Speaker Name, Affiliation, Topic
1. Leo Lahti , University of Turku, Finland, CODATA Executive Committee, From Observation to Interpretation
2. Daphne Raban, University of Haifa, Chair Israel CODATA NC, The Role of Data Stewardship in Research and Data Science
3. Phil Bourne, University of Virginia, US National Committee for CODATA, ADSA Board member, Bridging Data Science and Research Data
4. Padmanabhan Seshaiyer, George Mason University, US National Committee for CODATA, Weaving research data lessons into K-12 and community college data science bridge programs
5. Carolynne Hultquist (invited), University of Canterbury, New Zealand, Embedding FAIR through software and data workflows for Australia's climate modelling infrastructure
6. Kelsey Druken, Australian National University, Embedding FAIR through software and data workflows for Australia's climate modelling infrastructure

Panel moderated by Christine Kirkpatrick, San Diego Supercomputer Center, US National Committee for CODATA

Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data / 180

Methodological Approaches and Best Practices for Integrating Arctic Data and Research Infrastructure

Author: Marcin Wichorowski[1]

[1] *Institute of Oceanology, PAS*

Corresponding Author: wichor@iopan.pl

Arctic research faces persistent challenges, including limited accessibility, geopolitical complexities, and a general scarcity of high-quality data. In response, several international consortia—such as INTAROS and Arctic PASSION—have initiated collaborative efforts to address these issues through the development of integrated Earth observing systems (e.g., SIOS and GIOS). The Svalbard Integrated Arctic Earth Observing System (SIOS) builds upon an established and diverse portfolio of world-class observational and research infrastructure concentrated in Svalbard. This foundation supports SIOS' s aim to advance systematic methodologies for observational design and data integration, thereby enhancing the capacity for coordinated Arctic monitoring.

This session will explore methodological frameworks and operational practices that support collaboration among data providers, researchers, and infrastructure operators in the Arctic and Northern Polar regions. Emphasis will be placed on strategies for harmonizing distributed data sources and enhancing interoperability through coordinated data management systems. A keynote presentation will provide illustrative examples of the added value generated by the SIOS data management system, particularly in terms of its capacity to integrate heterogeneous observational data and infrastructure across institutional boundaries.
Through presentations and discussions, the session will highlight emerging tools, protocols, and governance models that enable more efficient sharing, discovery, and reuse of Arctic data. Special attention will be given to the role of FAIR (Findable, Accessible, Interoperable, Reusable) data principles in facilitating transnational collaboration and long-term data stewardship. Case studies from current and past projects will be presented to demonstrate how integrated observing systems can improve the scientific understanding of Arctic environmental processes and support decision-making in response to rapid climate and socio-ecological changes.

The session aims to bring together stakeholders from across the Arctic research community—including data scientists, infrastructure managers, environmental researchers, and policy advisors—to exchange knowledge, align efforts, and foster the development of interoperable, sustainable data ecosystems. By promoting best practices in data integration and infrastructure co-design, the session seeks to contribute to the resilience and responsiveness of Arctic research systems in the face of ongoing and future challenges.

Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI / 181

Co-Designing AI Readiness: CODATA's Call to the Global Data Community

Authors: Christine Kirkpatrick[1]; Mercè Crosas[2]; Simon Hodson[3]; Steven McEachern[4]; Tyng-Ruey Chuang[5]

[1] *San Diego Supercomputer Center / CODATA*

[2] *Barcelona Supercomputing Center*

[3] *CODATA*

[4] *UK Data Service*

[5] *Academia Sinica, Taiwan*

Corresponding Authors: trc@iis.sinica.edu.tw, simon@codata.org, s.mceachern@essex.ac.uk, merce.crosas@bsc.es, christine@sdsc.edu

Rapid advances in AI technology have the potential to ease or speed Research into research data management challenges. The CODATA and WDS communities are already coming up with ways to leverage AI for data stewardship. With much of the research data community dedicated to FAIR implementation and the colloquial second meaning of FAIR being 'Fully AI Ready', there is conflation and confusion about which of the FAIR Principles leads to data that can be consumed by AI or 'AI Ready Data'.

The data community had just begun to confront these questions as they relate to machine learning (ML) and then saw the emergence of deep learning technologies based on Large Language Models (LLMs) and Generative AI (Gen-AI). An even more recent development, the Model Context Protocol (MCP) brings a way to create agents and workflows on top of LLMs. This provides great opportunities, such as simpler ways to work with external data and LLMs. At the same time, the introduction of MCP brings open questions about how best to harness these capabilities in data stewardship practices. There are wide gaps of understanding and 'gut checks'between the FAIR/research data community and the computer science community.

For example, it is a closely held truth in computer science that more data is better for ML, and that lack of data quality can be overcome by having 'enough'training data. In the research data community, a common assumption is that FAIR data means 'AI Ready data'. This session and the

CODATA concept paper discuss these assumptions and examine community norms and tenets in the light of existing scholarship (publications) and state of the art (current best practices). This sets the stage to provide practical advice that can be used by data stewards, researchers, decision makers allocating resources, funders, and policymakers.

The term 'AI Ready Data'seems to suggest the scenarios where the datasets have been prepared and are ready-to-use for various AI systems and applications. There is less discussion, however, on whether and how to shape research data workflows to meet general AI needs. In addition, trustworthy AI can only be driven by trustworthy data. The issues of data provenance, integrity and measures of data quality, will only become more pressing in the face of rapid data turnover and changing workflow. Going further, we can view research datasets not just as end products to be consumed by AI, but as the carriers of information in collaborative research networks aided by AI.

This session introduces CODATA's new position paper that illuminates the opportunities and challenges as they relate to the CODATA community and current initiatives.

The session will:

1. Include a level-setting discussion introducing state of the art AI practices as they relate to:
2. AI for Data: Using AI for metadata enrichment and research data preparation. The session will present and discuss examples of this including current work at the OECD, in CGIAR, in GeoGPT, in ESIP, in FARR 1 and FAIR2. The importance of such techniques, of authoritative terminologies, ontologies and knowledge graphs will be explored.
3. Data for AI: Preparing data for the application of AI in science, including foundational AI model development, model training or fine-tuning, or using AI for inference. The session will also discuss key initiatives for standardising metadata for AI training data, including ML Commons'Croissant initiative. Approaches from various disciplines to communicating provenance and quality, including work in the Cross-Domain Interoperability Framework, will also be discussed as well as new developments such as MCP and its capabilities.
4. Explain the concept paper's key sections and recommendations
5. Solicit feedback, additional resources, and ask participants to help set priorities through interactive polls and collecting use cases of AI for data and data for AI.

The recommendations and future work should serve as a source for focusing new momentum at understudied and underdeveloped areas at the intersection of AI and data, while reorienting existing projects.

1 FAIR in Machine Learning AI Readiness, AI Reproducibility (FARR), NSF Award (2226453) led by Kirkpatrick

182

Integrating Ecosystem Observatories: Data Collaboration Across Continental-Scale Research Infrastructures

Authors: Siddeswara Guru[1]; Renée Brown[2]; Christine Laney[3]; Christoph Wohner[4]; Leo Chiloane[5]; Margareta Hellström[6]

[1] *University of Queensland*

[2] *The University of New Mexico*

[3] *NEON USA*

[4] *Environment agency Austria*

[5] *South African Environmental Observation Network*

[6] *Lund University*

Corresponding Authors: margareta.hellstrom@nateko.lu.se, claney@battelleecology.org, christoph.wohner@umweltbundesamt.at, leo@saeon.ac.za, rfbrown@unm.edu, s.guru@uq.edu.au

Research infrastructure initiatives play a critical role in enabling data-intensive science by providing the capabilities and services necessary for researchers to deliver innovative outcomes. In the environmental sciences, continental-scale research infrastructures facilitate consistent and standardised data collection across broad spatial and temporal scales. These datasets, collected through in-situ measurements, aerial and satellite-based remote sensing, long-term monitoring, citizen science, and model outputs, are diverse and complex. These data are crucial for detecting and quantifying environmental changes, validating remote sensing products, and calibrating models. Effective management of these vast and heterogeneous data collections requires a robust infrastructure to ensure they are Findable, Accessible, Interoperable, and Reusable (FAIR) at local, regional, and global scales.

This session will convene data infrastructure specialists, data scientists, data engineers, and researchers to explore how ecosystem research infrastructures, such as the Terrestrial Ecosystem Research Network (TERN, Australia), National Ecological Observatory Network (NEON, USA), South African Environmental Observation Network (SAEON), Integrated Carbon Observation System (ICOS), and European Long-Term Ecological Research (eLTER) are advancing data management and related infrastructure to build interoperable large-scale ecosystem observing networks to support global ecological understanding. For example, all of these research infrastructures, under the umbrella of the Global Ecosystem Research Infrastructure (GERI), are piloting data harmonisation across numerous data products for use in ecological drought research. This session will explore barriers, opportunities, and challenges to:

- Manage and process large-scale datasets from in-situ and remote sensorsList item
- Share indigenous data guided by CARE principles
- Apply FAIR principles, persistent identifiers, and semantic web technologies to enable data sharing across platforms and jurisdictions
- Develop data and metadata standards and controlled vocabularies for interoperability
- Leverage cloud-based infrastructure and high-performance computing platforms to support national and global collaborations
- Discuss strategies to foster data sharing and interoperability to enable data-intensive research at scale
- Address policy and governance challenges for the equitable use of infrastructure
- Develop and support next-generation infrastructure to provide curated training datasets and integrate Artificial Intelligence and Machine Learning (AI/ML) capabilities to advance global-scale research and innovation

The session will feature presentations, interactive and panel discussions. We invite contributions on the cyberinfrastructure capabilities of environmental research infrastructure that address barriers, opportunities, and challenges for cross-infrastructure collaboration. Panel discussions will explore the challenges of cross-continental data integration, highlight existing collaborative efforts, and identify opportunities and potential use cases that could benefit from global-scale research coordination. By the end of the session, participants will gain a shared understanding of current challenges, emerging opportunities, and actionable paths forward to advance collaboration across continental-scale research initiatives.

Presentations Session 6: The Transformative Role of Data in SDGs and Disaster Resilience / 185

From Data to Action: Supporting Coral Reef Conservation in the Pacific

Author: Julie Vercelloni[1]

Co-authors: Kerrie Mengersen [2]; Samuel Chan [1]

[1] *Australian Insitute of Marine Science*

[2] *Queensland University of Technology*

Corresponding Authors: sg.raine@qut.edu.au, jeremywicquart@gmail.com, a.gainsford@aims.gov.au, s.chan@aims.gov.au, j.vercelloni@aims.gov.au, k.mengersen@qut.edu.au

Session description:

The digital revolution is reshaping marine science by enabling unprecedented access to data efficiently and cost-effectively. These advancements are transformative for coral reef conservation, fostering a stronger connection between science and policy to safeguard biodiversity and support the communities that depend on these ecosystems. In this session, we highlight the role of data science in monitoring coral reefs across the Pacific, with the aim of empowering communities through actionable insights for effective management and conservation.

We introduce four digital platforms that leverage data-driven analytical solutions to enhance various forms of image recognition, including underwater imagery, remote sensing, and data fusion. The speakers will explore how these large-scale ocean programs responsibly and ethically manage data while advancing coral reef research, conservation, restoration and collaborations.

This agenda includes a combination of research and practical presentations, complemented by a short panel discussion and audience Q&A. The speakers (and later panellists) are a gender-balanced group of transdisciplinary, eminent and emerging researchers working alongside practitioners across the Pacific region.

Agenda:

Presentations (10min each, 2 min change):

- ReefCloud: Transforming Coral Reef Monitoring –Ashton Gainsford
- Detecting Coral Bleaching with AI –Nader Boutros (confirmed)
- MERMAID: An All-in-One Coral Reef Data Solution - Emily Darling (confirmed)
- Leveraging AI for a Real-Time Deployment of Baby Corals –Scarlett Raine
- The Global Coral Reef Monitoring Network –Jérémy Wicquart

Panel discussion (15min):

The discussion will explore the benefits and challenges of digital advancements in reef conservation.
Panellists: Ashton Gainsford, Nader Boutros, Emily Darling, Scarlett Raine, Jérémy Wicquart
Moderator: Julie Vercelloni

Q&A (15min):

We will address questions from the audience.
Moderator: Julie Vercelloni

Short bio of the speakers:

Dr Ashton Gainsford contributes to the leadership of the ReefCloud project, which is developed by the Australian Institute of Marine Science. Her background in coral reef ecology helps bridge and integrate the various components of ReefCloud, from fieldwork to data workflows, analyses and reporting.

Dr Nader Boutros is a machine learning engineer at the Australian Institute of Marine Science. He brings extensive expertise in automated image processing to develop digital tools that enhance coral reef monitoring and track environmental threats including coral bleaching.

Dr Emily Darling is an internationally recognized scientist and Director of Coral Reef Conservation at the Wildlife Conservation Society. Trained as a field biologist, her work investigates how tropical coral reefs are changing in the face of our climate crisis. She is a co-founder of MERMAID for coral reef monitoring.

Dr Scarlett Raine conducts research at the intersection of robotics, computer vision, artificial intelligence and coral reef conservation. Based at Queensland University of Technology, she develops cutting-edge AI methods for the Reef Restoration and Adaptation Program to support large-scale coral reef recovery.

Dr Jérémy Wicquart is a marine ecologist with extensive expertise in coral reef data analyses. He currently works as the Technical Coordinator of the Global Coral Reef Monitoring Network that provides information on the status and trends of coral reef ecosystems to support their conservation and management.

186

Leveraging Data Science and AI to Eradicate Modern Slavery

Author: Adriana-Eufrosina Bora[1]

Co-author: Kerrie Mengersen [2]

[1] *The Queensland University of Technology*

[2] *QUT Centre for Data Science, Queensland University of Technology*

Corresponding Authors: adrianaeufrosina.bora@hdr.qut.edu.au, k.mengersen@qut.edu.au

Abstract:

Modern slavery affects over 50 million individuals globally, millions being subjected to forced labour that infiltrates the supply chains of major corporations. This session explores the pivotal role of data science and artificial intelligence (AI) in combating modern slavery, bringing together perspectives from academia, non-profit organisations, and government agencies. By highlighting innovative methodologies and collaborative efforts, the session will demonstrate how data-driven approaches can enhance actions to fight against modern slavery. The session will conclude by celebrating the success of a pre-conference hackathon focused on this critical issue, showcasing creative data-driven solutions developed by participants and announcing the winning teams.

Significance of the Issue:

The pervasive nature of modern slavery presents significant challenges to human rights and economic development worldwide. Accurate data collection and analysis are essential for understanding the scope of the problem, identifying vulnerable populations, and formulating effective interventions. The integration of AI and data science offers unprecedented opportunities to process vast amounts of information, uncover hidden patterns, and predict risk factors associated with modern slavery. This session will address the critical need for interdisciplinary collaboration and technological innovation in tackling this global issue.

Session Structure and Agenda:

Introductory Remarks (5 min)

Understanding Modern Slavery Estimates and: (10 min)

A representative from the Walk Free Initiative, authors of the Global Slavery Index, will provide an in-depth analysis of current methodologies for estimating modern slavery prevalence and the challenges in data collection and interpretation. The Global Slavery Index offers national estimates of modern slavery for 160 countries, drawing on data from household surveys and assessments of national-level vulnerability.

Question from Moderator: "Could you please elaborate on the challenges faced in measuring the prevalence of modern slavery and how we can ensure the collection of robust and reliable data to accurately inform policy and action against modern slavery?"

Understanding the Legal Landscape Addressing Modern Slavery in Global Supply Chains: (5 min)

A representative from the Australian Attorney-General's Department will discuss the Modern Slavery Act, outlining expectations for businesses and situating the law within the global regulatory landscape. This presentation will provide an overview of compliance requirements and the role of legislation in combating modern slavery.

Question from Moderator: "What are the challenges your organisation faces when manually reviewing thousands of modern slavery reports annually, and how do you envision leveraging technology to enhance the efficiency and effectiveness of monitoring compliance with modern slavery legislation?"

Practical Applications of Data Science and AI in Business Compliance with the Modern Slavery Act:
Data Visualisation: Beyond Compliance Initiative: (10 min)

Insights from WikiRate and Walk Free on their Beyond Compliance project will be presented. This initiative includes a comprehensive visualisation dashboard assessing over 2,000 modern slavery statements, highlighting the role of businesses in addressing this issue and informing policy.

Question from Moderator: Given the importance of corporate transparency and accountability in combating modern slavery, how do you see this application of data science and AI creating more transparency in supply chains in response to the Modern Slavery Act and beyond?

Introduction to Project AIMS (Artificial Intelligence against Modern Slavery): (25 min)

Researchers from Queensland University of Technology (QUT) and Mila will present Project AIMS, which leverages ethical development of AI to analyse corporate reporting data and promote compliance with modern slavery laws. The project has developed the largest dataset of annotated modern slavery statements used to assess compliance with the Australian Modern Slavery Act and has fine-tuned and benchmarked AI models on this dataset. All the resources are shared *open source*, and the peer-reviewed research presented at conferences such as ICLR and ACL will be discussed.

Question from Moderator: What are some of the key findings from peer-reviewed academic papers?

Question from Moderator: What were some of the key developmental challenges and successes of the project?

Question from Moderator: How can this work scale to other jurisdictions?

Hackathon Outcomes: (15 min)

Presentation of finalist solutions'pitches from a recent hackathon focused on Project AIMS, culminating in the announcement of the winners. This segment will highlight the potential of collaborative, community-driven approaches in developing technological solutions to combat modern slavery.

Q&A (15 min)

Proposed Speakers:

Moderator: Distinguished Professor Kerrie Mengersen, Founding Director, QUT Centre for Data Science

Speakers:

Katharine Bryant: Director, Walk Free. (to be confirmed)

Representative from the Australian Attorney-General's Department (TBD) (to be confirmed)

Auréliane Froehlich –Program Manager, WikiRate (to be confirmed)

Adriana Eufrosina Bora –PhD Candidate, QUT and Project Lead, Mila.

Jerome Solis: Director, AI for Humanity, Mila (to be confirmed)

Intended Outcomes:

Enhance awareness of modern slavery issues and the importance of data-driven approaches in identifying and combating such practices.

Demonstrate the practical applications of data science and AI in assessing and ensuring compliance with modern slavery legislation.

Foster collaboration among academia, non-profits, and government entities to develop scalable solutions for eradicating modern slavery.

Emphasise the critical role of open source tools and data in enabling transparency, collaboration, and innovation in the fight against modern slavery.

Inspire the global data community to leverage technological innovations for social good.

Celebrate and showcase the hackathon's success, highlighting innovative solutions developed by participants and recognising outstanding contributions.

Contribute to SDG8, Target 8.7: "Take immediate and effective measures to eradicate forced labour, end modern slavery and human trafficking and secure the prohibition and elimination of the worst forms of child labour, including recruitment and use of child soldiers, and by 2025 end child labour in all its forms."

188

Early Career Researcher perspectives on data repositories across disciplines, geographies and cultures

Authors: Claire Rye[1]; Cyrus Walther[2]; Adrianna Eufrosina Bora[3]; Ntsundeni Louis Mapatagane[4]; Pragya Chaube[5]

[1] *University of Auckland/WDS ECR co-chair*

[2] *TU Dortmund University/CODATA Executive Committee*

[3] *Queensland University of Technology*

[4] *Walter Sisulu University/CODATA Connect*

[5] *UPES/CODATA Connect*

Corresponding Authors: cyrus.walther@tu-dortmund.de, lmapatagane@wsu.ac.za, adrianaeufrosina.bora@hdr.qut.edu.au, pragya.chaube@ddn.upes.ac.in, claire.rye@auckland.ac.nz

Data is of ever increasing value to the global research ecosystem, as underlined by recent emphasis on the FAIR principles and Open Science. Research infrastructures and data repositories are key to enabling Open Science and implementing the FAIR principles within the research ecosystem. Early Career Researchers (ECR) play an essential role in shaping and evolving new data practices and bringing fresh perspectives to well established data domains and research infrastructures.

Reflecting the importance of the ECR community, the two hosting organisations of SciDataCon have their own ECR networks, CODATA Connect and WDS ECR Network. They will be joined by local ECRs to host a session on research infrastructures and repositories across disciplines and geographies, highlighting the uneven landscape in their use and access.

The aim of the session is to showcase excellent work carried out by ECRs and to explore perspectives on use, accessibility and value of research infrastructures and repositories across a range of disciplines, geographies and cultural contexts. Extending to infrastructural, financial and policy related challenges, particularly in the Global South which often contrast sharply with the Global North. Aligning strongly with the conference themes, with international representation of speakers, the session will focus on open research, equity, global collaboration and in one example, CAREful Indigenous Data Governance.

We propose research presentations, sharing best practices and lessons learnt while underscoring where advancement is needed. Highlighting resources and thinking needed to improve the data repository structures, to enable more complete adoption of the FAIR and CARE principles. Followed by discussion with the audience.

Ntsundeni Louis Mapatagane, Uneven Data, Unequal Futures: Climate Change Data Disparities Between the Global North and South

Presenting the disparities in climate change data infrastructure on a global scale, highlighting how the predominance of data repositories situated in the Global North constrains the integration of localised data, particularly from entities in the Global South. Drawing on research conducted within South African universities, a microcosm for fostering sustainable innovation, this discourse will underscore the critical necessity for establishing regional data repositories. Such repositories are essential for enhancing climate resilience, facilitating education, and tracking progress toward the Sustainable Development Goals (SDGs). With the UN's 2024 SDG Report and COP29's forthcoming emphasis on education, the presentation will advocate for the pivotal role of universities in producing reliable, localised climate data that both enriches global datasets but also promotes equitable climate action.

Adrianna Eufrosina Bora, AI Against Modern Slavery: The Role of Open Data and Infrastructure
Sharing insights from the speaker's experience leading Project AIMS (Artificial Intelligence against Modern Slavery), an initiative that leverages AI to analyze corporate modern slavery statements for compliance with legislation in the UK and Australia. The project addresses the challenge of processing thousands of corporate disclosures by developing machine learning tools capable of reading and benchmarking these reports.
Drawing from this work, the talk will explore the critical importance of high-quality, accessible, and well-structured data in developing trustworthy AI systems for high-risk applications. Highlighting how inconsistencies in data formats, limited machine readability, and gaps in data availability can impede the effectiveness of AI tools, particularly in sensitive areas. The session will also discuss the infrastructural and policy-related challenges encountered in building and maintaining open-source AI initiatives for social good.

Pragya Chaube, Who Gets to Share? Understanding the Challenges of Open Research Data in India

Sharing perspectives on the author's experience and drawing from a qualitative study conducted at a leading research institution in India. This study examined attitudes toward open research data across academic disciplines and faculty career stages—from ECRs to senior professors. The study mapped the perceived value of open data, as well as the challenges faced in its adoption and implementation. These include infrastructural and institutional limitations, disciplinary norms, and varying levels of awareness and motivation to engage with open data practices. The findings offer insights into how openness is negotiated within the Indian research environment context, and how faculty perspectives differ based on career stage and discipline.

Claire Rye, Learnings from helping to build data repositories

Comparing and contrasting two experiences working to build data infrastructures, this talk will reflect on the journey of Ingestion service of the Human Cell Atlas Data Coordination Platform. The software infrastructure and metadata standards that support data sharing across the Human Cell Atlas project, while based at the European Bioinformatics Institute. Next on the development of Aotearoa Genomic Data Repository, an Aotearoa-based resource, that enables researchers and Māori communities to fulfil their obligations relating to the guardianship, management, sharing and use of genomic data from biological samples that are taonga (treasured).

Cyrus Walther, What do we do with all the data? Insights on Data Repositories in Large-Scale Multinational Collaborations in Physics

Addressing the challenges and respective approaches in high-energy experimental physics, a data-driven field of research, relying on large data volumes while requiring collaboration across borders and continents.
The talk will introduce several of these international collaborations, MAGIC, CERN, or SKAO, providing context and showcasing their necessity in advancing research. Furthermore, critical challenges in data storage, stewardship, and usage that originate from the nature of data acquisition are discussed, utilizing these research use cases. While respecting their individualities, the talk will highlight examples of such approaches for data repositories, demonstrating the opportunity for interdisciplinary transfer of these methods. Overarching best practices in high-energy experimental physics collaborations will be presented and perspectives towards implementation in other areas of the research will be discussed with the audience.

Presentations Session 8: Policy and Practice of Data in Research; Data, Society, Ethics and Politics / 189

Open data science and responsible research

Author: Leo Lahti[1]

[1] *University of Turku*

Corresponding Author: leo.lahti@utu.fi

From open data to open methods
Observation, interpretation, and communication are key elements of research. Whereas open science has traditionally emphasized open data and publications, the openness of research methods has received less attention. Methodology –the derivation of results and conclusions from the data –is as critical to the understanding and trust on scientific outcomes. The higher education and research community has, for a long time, recognised the transparent communication of methods as an essential part of research and dissemination. While digitalization has revolutionized open access to research, the rapidly expanding volume and complexity of digital resources has emphasized the need to reassess the overall requirements on good research practice. Whereas researchers have traditionally reported their methods as part of academic publications, the diversification of research, changes in technology and society, and the need to increase the impact of research through, for example, the adoption and reuse of methods, have set new challenges and opportunities towards responsible research practice. Methods are increasingly recognized as independent research outputs and disseminated through various channels, such as methods sections and supplementary materials, distinct data or methods publications, public protocols, code and material availability statements, open repositories, or in micropublications. Ensuring the early and long-term availability and preservation of methods may require new solutions to complement the more traditional forms of research dissemination. While the boundaries between research data and methods can be fuzzy, assessing the openness of research methods on its own right forms a necessary element of responsible data science.

Data science and responsible research
While openness has been recognized as a key element of research quality and impact, it has to be balanced by other aspects of responsible research. Researchers constantly face the tension between demands to support openness of research on the one hand, and respecting the necessary ethical and legal boundaries on sensitive information on the other hand. Even when data itself is sensitive and cannot be shared, making the methodology - or the interpretation of data - more transparent is often possible. The openness of research methods therefore becomes an essential element of maintaining scientific integrity. Open sharing of methods support the transparency and standardization of research also more broadly, facilitating scalable collaboration beyond national and institutional borders. Research organisations, funders, publishers, and infrastructure providers have an important role in supporting the early and broad dissemination of the research process, tools, and intermediate research outputs. Remarkable differences exist between fields, however, and the adoption of new practices and change in research culture is a gradual process that can greatly benefit from the translation of best practices between traditionally distinct research fields.

National policy work on open science
Whereas the lack of commonly accepted standards has slowed down progress, a more open research culture can be actively promoted by developing national and international guidelines. National policy work on open research data and methods in Finland has called for an active dialogue among the international research community towards defining the global standards and norms of openness of research methods. This talk provides an overview on the recent national policy work on open access to research data, methods and infrastructures as developed by the Finnish research community. The policies cover key themes of research quality and impact, support and infrastructures, and regulatory considerations. These overarching themes include specific recommendations on implementation for researchers, research organizations, research funders, publishers and other stakeholders. The talk will conclude by highlighting some of the practical challenges in advancing the transformation towards more open and reproducible research, and discuss the broader societal implications on the norms of evidence-based decision making.

190

The Planet Research Data Commons - delivering trusted environmental data and information supply chains

Authors: Hamish Holewa[1]; Kerry Levett[2]

[1] *Australian Research Data Commons (ARDC)*

[2] *Australian Research Data Commons*

Corresponding Authors: hamish.holewa@ardc.edu.au, kerry.levett@ardc.edu.au

A Thematic Research Data Commons is a vehicle for the ARDC and our national partners to collaboratively develop and deliver sustainable digital research infrastructure on a national scale. It is enabling us to best meet the needs of our diverse national research communities in a strategic and comprehensive way.

The Planet Research Data Commons (Planet RDC) is delivering enduring digital research infrastructure in the earth and environmental sciences. The initiative is establishing strong research translation pathways between research, government and industry.

The release of the 2021 State of the Environment report documents an unprecedented rate of deterioration in the state of Australia's environment, and in 2020 an independent review of the Environment Protection and Biodiversity Conservation Act asserted that "better data and information are needed to set clear outcomes, effectively plan and invest in a way that delivers them, and to efficiently regulate development."

The national data landscape for earth and environmental sciences is rich, diverse and complex – spanning multiple sectors, jurisdictions and data modalities. There is a critical need for digital research infrastructure that can support integrated and seamless national-scale research. Environmental managers and policy makers need trusted data supply chains and tools that enable them to make data-driven decisions.

With the help of accessible data and digital research tools, researchers can tackle the big challenges for our planet, which include adapting to climate change, saving threatened species, and reversing ecosystem deterioration.

The Planet RDC delivers infrastructure in 4 focus areas:

1. Trusted Environmental Data and Information Supply Chains
2. Integrated FAIR Datasets and Services
3. Modelling, Analytics and Decision Support Infrastructure
4. Governance of Indigenous Data and Skills.

This session will focus on the Trusted Environmental Data and Information Supply Chains focus area, which is working closely with Australian Government Department of Climate Change, Energy, the Environment and Water (DCCEEW), national research infrastructures, state agencies, industry partners, Traditional Owners, universities and NGOs to establish 'trusted data and information supply chains'for priority regional use cases.

The independent review of the Environment Protection and Biodiversity Conservation Act highlighted the need for an effective 'supply chain'of environmental information. As with more traditional supply chains, data collection, management and analysis activities can be carried out by different parties, and coordination is needed for an efficient chain that delivers the right products at the right time to the right customers.

A trusted environmental data and information supply chain requires an effective system of data-sharing agreements and information systems that can talk with each other, and reliable data and analytics products.

A series of **4 lightning talks** will include an overview of the program by Hamish Holewa, Director of the Planet Research Data Commons, and an explanation of the 3 exemplar projects by the project leads.

Two projects focus on shared data and analytics for environmental impact assessments and sustainable resource management in the Pilbara region and Cockburn Sound in Western Australia; and offshore renewable energy developments in the Bass Strait between Victoria and Tasmania. A third project aims to meet the needs of Traditional Owners, researchers and government agencies in managing and restoring the diverse wetland ecosystems of Gayini, NSW.

Project leads will explain how the cross-sector partnerships have been formed between Industry, Government and Research partners in each of the three projects, with the aim of delivering enduring data infrastructure that meets regional needs. The projects are developing technical architectures and data policies to enable FAIR, trusted data and analytics, that will be reusable in other regional partnerships, accelerating the development of other FAIR data infrastructures. The talks will highlight how collaborations with industry are supported with secure data systems that allow sensitive data to be shared with researchers, and ultimately made FAIR. They will also discuss the dataspaces model that is being piloted to streamline data sharing policies, agreements and data delivery.

After the lightning talks, attendees will have the opportunity to **ask questions of the panel of speakers**, and gain insights into lessons learned, and challenges overcome in the development and operation of the infrastructures.

191

Panel: How is data empowering Indigenous communities?

Author: Becki Cook[1]

Co-authors: Bernadette Hyland-Wood [1]; Kerrie Mengersen [1]; Raymond Brunker [2]; Robert McLellan [3]; Shani Gwin [4]

[1] *QUT Centre for Data Science*

[2] *Aboriginal and Torres Strait Islander Community Health Service Brisbane*

[3] *UQ, Language Data Commons of Australia (LDaCA)*

[4] *pipikwan pêhtâkwan*

Corresponding Author: r24.cook@qut.edu.au

This session will be conducted in a panel format and explore the central question "How is data empowering Indigenous communities?"It will bring together 5 speakers from diverse backgrounds across Australia and Canada, to present for 10-15 minutes each, followed by a facilitated discussion and Q&A with the audience.

The panel members will offer their own perspectives on the use of data in Indigenous contexts, including Indigenous-led research, community-controlled organisations, non-Indigenous allyship, and leveraging international standards. Each speaker will share insights from their experiences and areas of expertise, reflecting on the challenges and opportunities of working with data in ways that affirm Indigenous self-determination, governance, and knowledge systems.

This session aligns with the conference theme of *CAREful Indigenous Data Governance*, and will touch on:

- Indigenous-led research to improve data literacy within the Indigenous community;
- Data priorities, practices and current research initiatives at the Aboriginal and Torres Strait Islander Community Health Service Brisbane (ATSICHS), a community-controlled organisation;

- How researchers and practitioners outside Indigenous communities can work as allies in ways that are culturally sensitive and respect principles of Indigenous data governance and Indigenous Data Sovereignty; and
- The development of Indigenous-led AI tools.

Panel members:

Becki Cook: Becki is a proud Nunukul Aboriginal woman, educator, and researcher currently undertaking a PhD at the QUT Centre for Data Science. Her research explores Indigenous data literacy through Indigenous Research Methodologies, with a focus on elevating Aboriginal and Torres Strait Islander perspectives and priorities in the field of data science. Alongside her doctoral studies, Becki works as a Research Assistant in Indigenous Data Science and serves as the Early Career Researcher Co-Leader for the Data Science and AI in Society theme within the Centre for Data Science.

Robert McLellan: UQ, Language Data Commons of Australia (LDaCA).
Robert, a Gureng Gureng descendant from the Wide Bay region, is a community researcher, director and governance practitioner. He is an Industry Fellow at the University of QLD and Senior Program Manager for Language Data Commons of Australia (LDaCA), and a strong advocate for truth telling and speaking up for Aboriginal rights, justice, and economic advancement. Dedicated to authentic inclusion of First Nations voices, Robert is passionate about revitalising Indigenous languages, cultures and building culturally inclusive, honourable and cohesive communities.

Raymond Brunker: General Manager, Community Services, ATSICHS Brisbane.
Raymond is a proud Maramanindji man from Daly River, Northern Territory. He has a special interest in finding solutions that empower communities and families. He has a deep history and connection to the Logan and Brisbane community, having attend the Murri School and Woodridge State High School. Raymond has also completed a Bachelor of Education (Primary) and a Bachelor of Human Services. Currently, Raymond is an Atlantic Fellow, completing a Masters of Social Change Leadership with the University of Melbourne.

Dr Bernadette Hyland-Wood is a Research Fellow at the Centre for Data Science, Queensland University of Technology advancing in responsible AI and technological humanism. She is a research investigator on Indigenous-led data governance research programs and has chaired international data standards to support community advocacy and evidence-informed policy making. Dr Hyland-Wood actively engages with academia, industry, and public policy makers to advance human-centric technologies.

Shani Gwin : Shani Gwin is the Founder and CEO of pipikwan pêhtâkwan and wâsikan kisewâtisiwin. Shani is a proud, sixth generation Métis and a descendant of Michel First Nation. She has a passion for elevating Indigenous voices, truths and successes.
In her role, Shani has built one of Turtle Island's (North America) leading Indigenous-owned, -led and majority-staffed communications and engagement agencies. She helps guide clients and organizations with strategic guidance on decolonization, reconciliation, trauma informed communications, issues management and public relations.
She has over 15 years of professional communications experience in all sectors and is now developing a matriarchal and Indigenous powered artificial intelligence tool. The tool provides protection for Indigenous People online and supports non-Indigenous people with limiting bias and harm in their writing about Indigenous Peoples.

Moderator: Distinguished Professor Kerrie Mengersen
Kerrie Mengersen is a Distinguished Professor of Statistics at Queensland University of Technology. She is a Co-Leader of the Fundamental Methods in Data Science and AI Theme in the QUT Centre for Data Science. Kerrie's research sits at the intersection of computational and applied statistics and machine learning, and focuses on developing ways to efficiently collect, analyse, share and trust diverse data sources. Her applied work focuses on health, environment and industry.

The session will provide both theoretical reflections and tangible examples, drawing from work in academia, community health and governance, and cross-cultural collaboration.

The panel will appeal to a broad audience, including those engaged with data in community, research, advocacy, and policy contexts. It is designed to encourage thoughtful, values-based dialogue on how data can support Indigenous self-determination and community priorities. The 90-minute format includes time for audience engagement, fostering a space for shared learning and diverse perspectives on Indigenous data practices and potential.

Poster Session / 192

Open Ecoacoustics: A Platform to Manage, Share and Analyse Ecoacoustic Data for Scalable Fauna Monitoring

Authors: Anthony Truskinger[1]; Lola Lange[1]; Nelli Holopainen[1]; Paul Roe[2]; Philip Eichinski[1]; Robert Clemens[3]; Susan Fuller[1]

[1] *QUT*

[2] *JCU*

[3] *ARDC*

Corresponding Authors: philip.eichinski@qut.edu.au, a.truskinger@qut.edu.au, s.fuller@qut.edu.au, paul.roe@jcu.edu.au, l2.lange@qut.edu.au, nelli.holopainen@qut.edu.au, rob.clemens@ardc.edu.au

There is an urgent need for continental-scale monitoring of threatened species and ecosystems. Acoustic monitoring of the environment, ecoacoustics, provides a scalable way to achieve this. The Open Ecoacoustics platform supports ecoacoustics monitoring of the environment and is open to everyone to aggregate and share data, analyses and tools. The project goal is to enable open science and conservation through the development and promotion of open access ecoacoustics technologies, methodologies and standards.

There are a number of challenges in supporting large scale ecoacoustics, including how to aggregate, manage and share data; how to analyse and validate analyses; and how to interoperate with downstream services. The Open Ecoacoustics platforms supports FAIR data by developing standardised metadata and third-party analyses by moving to flexible workflow technologies. It accelerates data analysis by publishing a shared repository of annotated datasets and recognisers. It also interfaces to other systems through services and shared tools, to provide end to end workflows using ecoacoustic data.

The Open Ecoacoustics platform underlies the Australian Acoustic Observatory (A2O) database (www.acousticobservatory.org/) and the Ecosounds database (www.ecosounds.org/), together comprising over one Petabyte of acoustic data. The A2O is a single project collecting data using a standard protocol from over 360 sensors around Australia. The Ecosounds database comprises over 50 ecoacoustic monitoring projects.

Presentations Session 8: Policy and Practice of Data in Research; Data, Society, Ethics and Politics / 193

From RAiDs to Riches: how a local project ID got big global ideas

Authors: Natasha Simons[1]; Christine Kirkpatrick[2]; Clifford Tatum[3]; Joy Owango[4]

Co-author: Chris Erdmann [5]

[1] *Australian Research Data Commons (ARDC)*

[2] *San Diego Supercomputer Center / CODATA*

[3] *SURF Netherlands*

[4] *Training Centre in Communication*

[5] *SciLifeLabs Sweden*

Corresponding Authors: christopher.erdmann@scilifelab.uu.se, clifford.tatum@surf.nl, natasha.simons@ardc.edu.au, christine@sdsc.edu, joy.owango@tcc-africa.org

In 2017, RAiD was a budding new concept and the early beginnings of a technical system for identifying and tracking research projects. The idea of RAiD as a project identifier itself came out of a project - an Australian project to better track the research data lifecycle which put research projects

at centre stage. Today, RAiD is an ISO standard for Project Identifiers (23527:2022) with a global Registration Authority (Australian Research Data Commons) operating the RAiD system and partners setting up Registration Agencies for RAiD capability at SURF in the Netherlands (for the whole of Europe; as a component of EOSC infrastructure) and in the USA (as an NSF funded grant led by the San Diego Supercomputer Center). DataCite, the global DOI Registration Agency, issues the identifier component of RAiD through a collaboration with ARDC. The need for RAiD is growing: it is the fourth priority identifier listed in National PID Strategies (after ORCID, DOI and ROR) according to the work of the RDA National PID Strategies Interest Group. Recommendations to use RAiD are emerging internationally (e.g. the Czech R&D Council recommends use of RAiDs) and it is being used in related emerging infrastructures (e.g., the Africa PID Alliance is seeking to interoperate RAiD with its DocID system).

In this session, we propose a panel of lighting talks from international speakers involved in RAiD development and adoption to tell the RAiDs to Riches story of a little project ID with big global ideas. We will explore why RAiDs are important infrastructure in tracking data intensive research; facilitating responsible research reassessment; contributing to open research information; how they fill a key gap in the PID landscape; and what the future might hold for RAiD adoption and use globally. We will give examples of using the new data from RAiD to model impact as well as research networks. Through Q&A and discussion, we will address questions from the community, and take in feedback to improve future outreach.

Poster Session / 195

Measuring Data Matters‼

Authors: Ai Lin Soo[1]; Claire Rye[2]; Isabel Ceron[None]; Luc Betbeder-Matibet[1]; Nick Jones[2]; Rhys Francis[None]

[1] *UNSW Sydney*

[2] *The University of Auckland*

Corresponding Authors: rhyssfrancis@gmail.com, ai_lin.soo@unsw.edu.au, luc@unsw.edu.au, n.jones@auckland.ac.nz, isabel.ceron@uq.edu.au, claire.rye@auckland.ac.nz

The Macro View, reported at IDW2023, set out to estimate the national scale of research data that is under management for the purpose of future access in Australia and New Zealand. Two key observations can be made:

1. The participating institutions lacked internal reports on data as an asset, from which a total could be easily aggregated. Instead one off measurement tasks were undertaken.
2. While data was definitely counted, non-data digital content was also being counted.

Further work with a small subset of the institutions, revealed that an expected rising and falling of data volumes as a research project proceeds was never detected in practice. The events that might cause reduction of data towards a small set of refined outputs, either didn't occur in practice, or did not result in the deletion of the intermediate data. Instead, the operative policy could be summarised as "if researchers don't delete it - keep it".

We therefore postulate the existence of a significant volume of content held in institutional archives that is not research or scientific data. However, no measure of its extent is available.
We propose to label this content as 'the digital debris of research', given it arises from the day to day practice of performing research. Some of the digital debris is inherent such as copies of downloaded material, intermediate error filled software versions and their test outputs, and redundant faulty data superseded by correctly gathered data. Some examples of debris are more difficult to evaluate. For example, older data can lose its ability to influence the advancement of knowledge as that knowledge does in fact improve. This might involve prior versions of data falling into disuse as instruments and analysis evolve, creating simply better data (eg. The human reference genome and its downstream by-products is at version 38).

Data creating and supplying entities such as terrestrial observing platforms or population scale genomic libraries, know the digital objects they hold are data, and can measure their data and its use and reuse patterns directly. It appears the research performing institutions, by direct observation during the effort to establish the Macro View, could not. They counted data and debris together because as data volumes have grown, extensive homogeneous file systems have been developed to underpin their research activity. This means that the way we understand data from the experience of our formal data collections, is an incomplete narrative when applied to the management of digital content in research performing institutions. For instance, the digital debris of research retained within the digital corpus under management in an institution, should, most likely, not be made FAIR.

Our poster will highlight results from initial investigations into institutional data practice that support the following two provocations:

1. The digital debris of research is real, accumulates endlessly and over time uselessly and by doing so, renders curating valuable data held with it, increasingly inefficient and impractical.
2. Research performing institutions need to develop a debris policy to enhance their data policy if the desire to maximise the reuse of valuable data is to be realised.

Data and debris intermingle in the day to day research process within institutions, and therefore in institutional archives. Because they should be treated very differently, this is a major unresolved complexity in our research data management practice.

A possible response would be to articulate the life cycle of research debris as distinct to the life cycle of research data, develop policies and guidelines that can separate the pathways for data and debris, and measure all aspects of the journey they each take.

Poster Session / 196

Analysing Defence Mechanisms Against Gradient Attacks in Contrastive Federated Learning

Author: Achmad Ginanjar[1]

Co-authors: Xue Li [1]; Priyanka Singh [1]; Wen Hua [2]

[1] *The University of Queensland*

[2] *The Hong Kong Polytechnic University*

Corresponding Authors: xueli@eecs.uq.edu.au, a.ginanjar@uq.edu.au, priyanka.singh@uq.edu.au

Vertical Federated Learning (VFL) has emerged as a transformative approach in collaborative machine learning, enabling multiple parties to jointly train models while maintaining data privacy through vertical partitioning of features. This paradigm has gained significant traction in privacy-sensitive domains such as healthcare and finance, where different organisations possess distinct feature sets of the same entities.

Despite its promise in preserving data privacy, VFL faces inherent vulnerabilities related to information leakage during the intermediate computation sharing process. Research has shown that even partial information exchange can potentially expose sensitive data characteristics, compromising the system's fundamental privacy guarantees. These limitations have prompted researchers to seek more robust privacy-preserving solutions.

Contrastive Federated Learning (CFL) was introduced as an innovative approach to address these privacy concerns. By incorporating contrastive learning principles, CFL reduces the need for direct feature sharing while maintaining model performance through representation learning. This method has demonstrated promising results in minimising information leakage during the training process.

However, while CFL enhances privacy preservation in feature sharing, it does not fully address the broader spectrum of security threats in federated learning, particularly internal attacks. Among these, gradient-based attacks have emerged as a significant concern, where malicious participants can exploit gradient information to reconstruct private training data or compromise model integrity. These attacks pose a substantial threat to the security of federated learning systems, potentially undermining their practical applications.

In this paper, we conduct a comprehensive experimental analysis of gradient-based attacks in CFL settings and evaluate three defensive strategies: random client selection, gradient clipping-based client selection, and distance-based client selection. Our research aims to quantify the effectiveness of these defence mechanisms and provide empirical evidence for their practical implementation in securing federated learning systems against internal attacks.

197

Decentralizing for Resilience: Beyond Data Rescue in Global Climate Networks

Author: Stephen Diggs[1]

Co-author: Rebecca Cowley [2]

[1] *University of California Office of the President*

[2] *C*

Corresponding Authors: rebecca.cowley@csiro.au, sdiggs@gmail.com

Background

For 30+ years, the scientific community has worked toward unified ocean and climate data networks. While initiatives like EarthCube, RDA, and WDS established critical foundations, centralized systems remain vulnerable to political shifts, technical failures, and disasters.

This session moves beyond theoretical discussions to **practical solutions**, building on the hard-won successes of previous initiatives while addressing their limitations. It proposes a bold yet practical vision for a **decentralized global data ecosystem** that strengthens global data resilience, enhances FAIR compliance, and leverages cutting-edge technologies like AI/ML to meet the demands of our rapidly changing environment.

Vision

We propose a network that is:

- **Distributed**: Regional hubs prevent single-point failures (e.g., Pacific/Arctic nodes)
- **AI-Optimized**: Expert-trained machine learning accelerates FAIR data curation
- **Equitable**: Actively reduces barriers for Global South participation
- **Secure Yet Open**: Federated repositories with standardized APIs

Objectives

1. **Diagnose Centralized Risks**
2. Analyze vulnerabilities in current systems using case studies
3. Highlight successful decentralized models from oceanographic and climate networks
4. **Define Decentralized Infrastructure**

5. Demo open-source tools for federated storage and AI quality control
6. Propose governance frameworks with UN/WMO oversight
7. **Launch Pilots**
8. Establish Pacific/Arctic monitoring hubs with distributed technical support
9. Prototype a Global Climate Observation Consortium (GCOC) with G20 funding mechanisms
10. **Align Policy**
11. Map outputs to SDGs and Paris Agreement targets
12. Draft agreements for open climate data access

Expected Outcomes

Participants will leave with:

- A nascent **roadmap** for transitioning to decentralized systems
- **Improved awareness** of AI/ML tools for faster FAIR data processing
- **Actionable policy ideas** for international data governance

Why This Fits IDW 2025

- **Local Action**: Partners with Australian/Pacific Island researchers on sea-level resilience
- **Global Need**: Addresses instability in centralized climate data systems
- **Technical Innovation**: Features AI curation tools in-development

---

Session Format (90 min)

1. **Keynote (10 min)**: “Lessons from Vulnerable Centralized Networks”
2. **Lightning Talks (20 min)**:
3. **Workshop (30 min)**: Build a decentralization checklist using real Arctic datasets
4. **Panel (20 min)**: Policymakers \+ technologists debate implementation hurdles

Target Audience

- **Data Engineers** needing resilient architectures
- **Policy Teams** drafting international data agreements
- **Tool Developers** working on federated/AI solutions
- **Equity Advocates** for Global South access

Presenters:
- Steve Diggs (University of California Office of the President: CDL/UC3)
- Rebecca Cowley (CSIRO)

Track Classification:
- INFRA (Data Infrastructures)
- EQUITY (Inclusive Systems)

“This isn’t about abandoning existing systems—it’s about making them robust enough to survive the next 30 years of climate challenges.”

Poster Session / 198

Towards a Sustainable and Resilient Future: the Transformative Role of Data in Crisis Management

Authors: Shreya Srinivas[1]; Gnana Bharathy[2]; Salvatore Flavio Pileggi[1]; Ghassan Beydoun[1]

[1] *University of Technology Sydney*

[2] *ARDC/ UTS*

Corresponding Authors: ghassan.beydoun@uts.edu.au, shreya.srinivas@student.uts.edu.au, salvatoreflavio.pileggi@uts.edu.au, gnana.bharathy@ardc.edu.au

Crisis management plays a role in achieving a sustainable and resilient future by preparing governments and communities to effectively respond to and recover from disruptions . Crisis management generates large volumes of heterogeneous data, including spanning structured databases, unstructured government reports, real-time news reports, and social media channels. Despite such an availability of data, the information remains siloed, inconsistently classified, and inaccessible across systems, leading to delays and inefficiencies in crisis recovery and response. This inability to rapidly synthesise diverse data streams impacts the efforts to achieve global disaster risk identification and reduction as outlined by the UNGA and in the Sendai Framework for Disaster Risk Reduction .

A decade since the Sendai Framework, there exist challenges in understanding disaster risk (Priority 1) and increasing information availability (Target 7) within the framework . The heterogeneous nature of crisis data, often owned by different institutions with varying formats, hinders open access and seamless data integration. This creates challenges for effective emergency response. The impact is twofold: first, it hinders governments' ability to accurately report on disaster impact, thus obstructing recovery planning; second, the lack of understanding of current impacts limits governments' capacity to predict and plan for future disasters –a key concern for the National Emergency Management Agency (NEMA) in Australia is improving its ability to assess and provide data on local economic damage impacts . Given the predicted increase in disaster frequency in Australia, coupled with the Defence Strategic Review emphasis on decentralised emergency plans, establishing robust data infrastructure is crucial to ensure effective reporting of disaster events and more accurate economic impact reporting.

To address challenges of establishing an effective data infrastructure and reporting, there needs, we need to achieve interoperability, harmonization and governance across the heterogenous data sources and ontologies. This paper introduces the Ontology Alignment, Structure, Integration, and Synthesis (OASIS) framework, an AI-driven framework designed to create and align ontologies across diverse systems and domains. A key contribution of this framework lies in its ability to enable the harmonization and governance of dynamic, real-time collaboration between humans and systems. This research focuses on the sourcing and transformation of crisis-related information through the integration of ontologies, knowledge graphs, and LLMs to empower effective decision-making. OASIS directly tackles the need for semantic interoperability for ontology engineers within the crisis management domain. The framework supports the complete ontology lifecycle, from alignment with established standards such as the Disaster Management Metamodel (DMM) and EM-DAT taxonomy, to ontology engineering using Protégé , and finally, to the synthesis of interconnected knowledge graphs utilizing Neo4j .

This research adopts an Action Research methodology , actively engaging domain practitioners in the iterative design, development, and evaluation of the OASIS framework and its outcomes. The framework's practical implementation was iteratively tested through two real-world case studies: Cyclone Jasper (2023) and Cyclone Alfred (2025), impacting Northern and South East Queensland, respectively. These events provided valuable opportunities to assess OASIS in distinct crisis contexts. In both cases, unstructured data—including post-disaster reports, emergency declarations, and news articles—was processed using advanced LLMs (GPT-4 and Gemini) to extract key entities, relationships, and structured attributes. These extracted elements were then automatically mapped to the developed ontology and integrated into a knowledge graph, enabling practitioners to visualize, explore, and query critical information such as disaster impacts and recovery funding allocation.

A quantitative and qualitative validation was performed for both case studies. We achieved a F1-score of 0.89 for structured data extraction following iterative prompt refinement and ontology alignment, thus showcasing the potential of LLMs to significantly accelerate information modelling when guided by structured ontology. Qualitative feedback was gathered from practitioners through interviews and collaborative workshops. Beyond its specific technical implementation, the OASIS framework offers a reusable, domain-agnostic model adaptable to other sectors, providing a repeatable methodology for rapid, AI-assisted knowledge base construction in dynamic and time-critical scenarios.

While not all crises can be anticipated or fully mitigated, for the events we can proactively plan for, this research highlights critical areas for improvement in disaster resilience. As the global community progresses towards the SENDAI 2030 agenda, this research offers a timely and practical contribution, demonstrating how data can be effectively leveraged through the proposed OASIS framework to enhance disaster preparedness and response.

Poster Session / 200

Leveraging Large Language Models (LLMs) for enhanced access to polar datasets through Natural Language Queries

Author: Alice Cavaliere[1]

[1] *ISP*

Corresponding Author: alice.cavaliere@cnr.it

This presentation aims to build a local generative search engine that demonstrates how generative AI can be effectively integrated with semantic search to enhance information retrieval and user interaction. The proposed interface, powered by large language models (LLMs), will simplify access to polar datasets stored in catalogs applications such as GeoNetwork and ERDDAP. By leveraging LLMs, the system will enable users to query polar data repositories using everyday language, significantly enhancing accessibility for researchers, policymakers, and the general public. The project will focus on three key aspects: (1) LLM-driven query translation, allowing users to input natural language requests; (2) interactive query refinement, where the system engages in dynamic dialogue with users to clarify and adjust search parameters for more accurate results; and (3) enhanced result summaries, enabling LLMs to condense complex metadata into concise, relevant descriptions for quick interpretation.

Presentations Session 4: Data Stewardship / 201

Linking Data & Publications in Social Science and Humanities: the role of infrastructures in the French national context

Author: Nicolas Larrousse[1]

Co-authors: Bénedicte Kuntziger [2]; Charles Bourdot [3]; Dominique Roux [3]; Hélène Jouguet [1]; Julie Verleyen [1]; Sandra Guigonis [4]; Yannick Barborini [2]

[1] *Huma-Num CNRS - France*

[2] *CCSD CNRS - France*

[3] *Métopes - Université de Caen Normandie - France*

[4] *OpenEdition Aix-Marseille University - France*

Corresponding Authors: benedicte.kuntziger@ccsd.cnrs.fr, julie.verleyen@huma-num.fr, dominique.roux@unicaen.fr, sandra.guigonis@openedition.org, charles.bourdot@unicaen.fr, helene.jouguet@huma-num.fr, yannick.barborini@ccsd.cnrs.fr, nicolas.larrousse@huma-num.fr

In SSH (Social Science and Humanities) the link between data and publication can be seen from different angles depending on its potential use. The first use that comes to mind is to cite a dataset in a publication for the purposes of scientific verification. It can be done in a number of ways, from a simple text citation, both in the publication or in the description of a dataset, to a PID (Persistent IDentifier) in a specific metadata. Another possible type of link would be to show multimedia material (e.g. illustration, table, soundtrack, etc.) in a publication. Finally, in another vein, data papers can be considered as a case of linking data and publications.

At a national level, an ecosystem of repositories and publication platforms are involved in the process of creating links between data and publications. More specifically, as part of two national projects (HALiance and COMMONS), four infrastructures are working together on SSH resources: Huma-Num (Data repository NAKALA), CCSD (Open Archive for research papers - HAL-SHS), OpenEdition (Publication Platforms for books and journals) and Métopes (Publishing Process). A study of the content of the different platforms carried out at the beginning of these projects showed that link-building practices already existed. Unsurprisingly, the study illustrated the diversity of the means used to create the above-mentioned links, and the fact that they are generally unidirectional.

What role can infrastructures play in this area? First of all, there is a need to simplify things for users by building bridges between platforms. For instance, when a research paper referring to a dataset in NAKALA is deposited in HAL-SHS, it would be convenient to have direct access to the dataset from the HAL environment. Similarly, to include multimedia material in a book from OpenEditionBook may require access to the files in NAKALA, which represents a different level of granularity. This means that it is necessary to work on seamless integration between the platforms; in particular by working on the consistency of information systems, developing specific APIs, and then finally by adapting interfaces.
Another crucial role of infrastructures is to guarantee the consistency of link information and to disseminate it in a standardised way. This requires automated communication between platforms to be able, for example, to create reciprocal links and maintain them over time: for instance, what happens to the link if the data disappears from the repository or if a new version is created?

This process which involves adaptations at various levels, both technical and organisational, did not start from scratch. A POC (Proof of Concept) of creating links between NAKALA and HAL-SHS was performed as part of the European project EOSC-Pillar. This enabled us to determine important issues to be resolved. It also helped us to identify standards that are emerging on this topic: in this regard, it was decided to use the COAR Notify protocol for communication between the platforms and the SCHOLIX standard to disseminate data-publication links which is supported by DataCite and Crossref and used in the European context by OpenAire.
This work is well advanced and the link between the HAL-SHS archive and the NAKALA repository is already operational on the development platforms and is about to be put into production. This work will serve as the basis for the link between the NAKALA repository and the OpenEdition platforms, which will also use a structured standard developed by Métopes (COMMONS TEI-Publishing) for integrating data directly into a publication.

At this stage of the projects HALiance and COMMONS, communication between the various platforms via the COAR-notify protocol has been standardised and stabilised. This makes it possible to consider communication with other platforms in the national ecosystem, such as the national repository RDG (Recherche Data Gouv), which uses the same protocol. Technically, there is still work to be done: especially a better version management for publications and datasets and a better implementation of the SCHOLIX standard. Finally, it will be necessary to improve the integration between platforms through a potential common authentication, for instance. However, the key to successfully implementing the link between data and publications lies in informing and training future users, which is an important part of our projects.

Bibliography

McGillivray B, Marongiu P, Pedrazzini N, Ribary M, Wigdorowitz M, Zordan. (2022). E. Deep Impact: A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences. Publications. 10(4):39. https://doi.org/10.3390/publications10040039

Burton, Adrian, & Koers, Hylke. (2016). ICSU-WDS & RDA Publishing Data Services WG Interoperability Framework Recommendations (1.0). https://doi.org/10.15497/RDA00002

Edmond, J. (Ed.). (2020). Digital Technology and the Practices of Humanities Research. Open Book Publishers. https://doi.org/10.11647/obp.0192

Gassama, M., Szabo, D., Tang, C., & Bravo, S. (2024). Analyse de l'enquête sur les pratiques des scientifiques en matière de publication de data paper [Report, INRAE]. https://doi.org/10.17180/vrh7-r606

Harper, L. M. (2023). Data reuse among digital humanities scholars : A qualitative study of practices, challenges and opportunities [Université d'Ottawa / University of Ottawa]. http://hdl.handle.net/10393/45445

Arnold, T., Scagliola, S., Tilton, L., & Gorp, J. V. (2021). Introduction : Special issue on audiovisual data in dh. Digital Humanities Quarterly, 15(1). https://www.digitalhumanities.org/dhq/vol/15/1/000541/000541.html

Kinnaman, A., & Guimont, C. (2023). Dh as data : Establishing greater access through sustainability. Digital Humanities Quarterly, 17(3). https://www.digitalhumanities.org/dhq/vol/17/3/000715/000715.html

202

Interoperable and Federated Vocabulary Services

Author: Clement Jonquet[1]

Co-authors: Brandon Whitehead [2]; Nicholas Car [3]; Alexandra Kokkinaki [4]; Rob Atkinson [5]; Christelle Pierkot [6]; Naouel Karam [7]; John Graybeal [8]; Megan Wong [9]

[1] *INRAE (MISTEA) and INRAE (MISTEA)*

[2] *Manaaki Whenua—Landcare Research*

[3] *KurrawongAI*

[4] *National Oceanographic Centre (British Oceanographic Data Centre)*

[5] *Open Geospatial Consortium (OGC)*

[6] *CNRS - Centre national de la recherche scientifique*

[7] *Institute for Applied Informatics (InfAI), University of Leipzig*

[8] *San Diego Supercomputer Center / GO FAIR US*

[9] *Australian Research Data Commons and Federation University*

Corresponding Authors: nick@kurrawong.ai, ratkinson@ogc.org, whiteheadb@landcareresearch.co.nz, karam@infai.org, megan.wong@ardc.edu.au, jbgraybeal@ucsd.edu, christelle.pierkot@cnrs.fr, alexk@noc.ac.uk, clement.jonquet@inrae.fr

Vocabulary services are a critical component of data sharing infrastructures. If these can be shared across infrastructures, then a more global ecosystem for data sharing and reuse can be supported. They are a foundational enabler for the very concept of FAIR (Findable, Accessible, Interoperable, Reusable), for without common references to semantic concepts, no data can be interpreted safely. A vocabulary service—ranging from basic terminology catalogues to more sophisticated Semantic Artefact Catalogues (SACs) or ontology repositories—enables users to share, describe, discover, browse and download controlled vocabularies or other types of semantic artefact (such as terminologies, ontologies, thesauri). Those artefacts may be available in various representation languages (e.g. OWL, SKOS, RDF-S) and encoding formats (e.g., TTL, XML, JSON-LD, Notation3).

Interoperability between vocabulary services is essential to enable the discovery, access, and reuse of vocabularies across diverse domains. However, most vocabulary services are developed in isolation, and even when they adhere to relevant standards (such as SKOS or OWL), there are no agreed interconnections between them. A growing number of vocabulary services have emerged across different domains and communities, often developed independently with domain-specific priorities, different technology stacks and design principles, leading to fragmentation and interoperability challenges.

Federation of vocabulary services is achieved through API-level and UI-level integration, supporting seamless data sharing and a unified user experience. However, there are many challenges in federating vocabulary services and the semantic resources they deliver. A high-priority task is surfacing the operational services currently available and understanding how these differ, and why. Dedicated programs of work are required to progress the vision of vocabulary services federation. Solutions that work toward achieving this vision are likely to be multifaceted, addressing challenges such as:

-Finding and choosing terms and vocabularies across a wide range of vocabulary services, with different scopes, interfaces (user interfaces and application programming interfaces) and underlying technologies;

-Standardisation (with governance and maintenance) of APIs and data exchange formats for federation;
-Transparency in who is using which vocabularies (the inbound link problem) and how;
-Identification and replication of subsets; Standardisation of metadata profiles, including provenance and change management;
-FAIR crosswalks; and
-Good PIDs, efficiencies and trust in where and how they resolve.

This session will explore the required scope and evaluation criteria for a general solution and examine how several initiatives meet these requirements. One such initiative is FAIR-IMPACT, a European project within the EOSC program, which has identified and studied three complementary technical approaches to SAC interoperability. Each approach offers distinct benefits and trade-offs.

1. MOD-API: A Standardised API for Interoperable SACs
   This approach defines a common API specification, the MOD-API, based on the MOD ontology, which standardises metadata descriptions for semantic artefacts and catalogues. Widely adopted through an open call by FAIR-IMPACT, it supports uniform querying across catalogues and enhances FAIRness. Eleven SACs, including those built on SKOSMOS, OLS, OntoPortal, ShowVoc, and Prez, are currently adopting MOD-API.
2. OntoPortal Federation: Leveraging a Shared Technology Stack
   SACs based on the OntoPortal stack can easily federate by virtue of shared backend infrastructure. This enables both API-level compatibility and UI-level integration for browsing and searching across federated SACS that use OntoPortal. Currently operational across AgroPortal, EcoPortal, EarthPortal, and BiodivPortal, this marks a significant milestone in SAC interoperability and vocabulary services federation.
3. API Gateway: A Centralised Aggregation and Proxy Model
   This model introduces an external API gateway, such as the one currently developed by TS4NFDI, that connects to various SACs via tailored connectors. It fetches and transforms data from multiple sources into a unified model without requiring changes to the original SACs. Though convenient and inclusive, it depends on a central proxy and lacks incentives for SACs to adopt shared standards. It builds on earlier experiments, such as FAIRCat, and currently supports platforms using SKOSMOS, OLS, and OntoPortal.

Session structure:

This session is a structured panel discussion on advancing federated vocabulary services. It seeks to progress discourse regarding the scope of evaluation criteria toward broader adoption of general solutions.

Firstly, the problem space will be introduced for a federated ecosystem of vocabulary services, including Semantic Artifact Catalogues (SACS), which are essential for enabling FAIR data. Then, panel members will share a current practice or initiative, followed by their perspective on: ‘moving forward: how do we continue toward an ecosystem of federated vocabulary services?’(10 minutes each). This may include reflections on strategy for articulating vision, identifying key roadblocks, minimal viable product, priorities, and pathways forward (the next steps). The community audience will also be encouraged to contribute their reflections on challenges and next steps through a shared document. Panellists include:

-Clement Jonquet - Approaches to vocabulary services or semantic artefact catalogues interoperability studied in the FAIR-IMPACT project.
-Christelle Pierkot (or Clement Jonquet) - OntoPortal Federation illustrated with EarthPortal.
-Alexandra Kokkinaki - Adopting the MOD-API in the NVS.
-Naouel Karam (TBC) - Implementing the API Gateway as a unique endpoint to vocabulary services federation.
-Nicholas Car - A review of long-term, operational, vocabulary production and (re)use
-John Graybeal - OntoChoice: A collaborative project documenting evaluation strategies for choosing terms and ontologies, ontology selection and recommendation
-Rob Atkinson - Requirements for a general solution to implementing FAIR principles for vocabularies in a scalable global data ecosystem.

203

Legal and organisational aspects of data interoperability: climate adaptation case studies

Authors: Adrian Burton[1]; Hamish Holeva[1]; Hilde Orten[2]; Kelsey Drucken[3]; Lesley Wyborn[4]; Mark Rehbein[5]; Matti Heikkurinen[6]; Michelle Heupel[7]; Pascal Perez[8]; Rebecca Farrington[9]; Shaily Gandhi[10]; Simon Hodson[6]; Thanasis Sfetsos[11]; Tim Rawling[12]

[1] *ARDC*

[2] *Sikt*

[3] *ACCESS-NRI*

[4] *ANU*

[5] *AODN*

[6] *CODATA*

[7] *Coast RI*

[8] *AURIN*

[9] *AusScope*

[10] *ITU Linz*

[11] *Demokritos*

[12] *AuScope*

Corresponding Authors: simon@codata.org, rebecca@auscope.org.au, hilde.orten@sikt.no, pascal.perez@unimelb.edu.au, shaily.gandhi@it-u.at, mark.rehbein@utas.edu.au, michelle.heupel@utas.edu.au, kelsey.druken@anu.edu.au, ts@ipta.demokritos.gr, matti@codata.org, lesley.wyborn@anu.edu.au, hamish.holewa@ardc.edu.au, tim@auscope.org.au, adrian.burton@ardc.edu.au

Building data-driven solutions to support climate change adaptation is inherently a cross-disciplinary and cross-organisational challenge. Legal compliance and organisational practices and assumptions will provide requirements and constraints for the technical solutions that can be deployed in the operational environments. This session will investigate these challenges and present an initial roadmap towards standardised legal and organisational agreement frameworks that can facilitate solving privacy, IPR and other compliance issues in a transparent and traceable manner.

The methodology and analysis is based on the world of the EU-funded Climate-Adapt4EOSC project with three specific case studies: urban climate vulnerability, principally heat; coastal/estuarine/port hazard resilience, including overtopping; shrink-swell of clay soils, with attendant issues of building damage and insurance. The work builds on the EOSC Interoperability Framework and other work on Legal and Organisational Interoperability, to develop 1) a diagnostic list of commonly encountered issues, and 2) an easy-to-use approach to mapping and understanding data flows and exchanges, so as to highlight and address legal and organisational interoperability obstacles. Both of these approaches are being tested with the Climate-Adapt4EOSC case studies. The session will present, and seek feedback on, these methodologies as well as the organisational and legal challenges encountered as part of the requirement analysis and service design.

In addition to presenting the initial results of the agreement framework, the session aims at identifying opportunities to reuse and enhance the solutions based on an interactive workshop. The session will test this approach and to understand how issues of organisational and legal interoperability are being addressed in other geographies and settings. There will be an interactive activity to review the diagnostic list (and identify any gaps) and to explore the mapping approach, as well as the findings and recommendations. Research infrastructures encountering legal and organisational interoperability issues in analogous circumstances will present and discuss the approach they are taking to overcome them and the solutions identified.

Programme (short presentations highlighting key issues, followed by discussion and interactive exercises with the diagnostic list and mapping approach.)

Simon Hodson and Matti Heikkurinen (CODATA), Hilde Orten (Sikt), The Climate-Adapt4EOSC approach of legal and organisational interoperability.

Hamish Holewa, Rebecca Farrington, Tim Rawling, Legal and organisational interoperability in the ARDC Planet Commons

Adrian Burton, Legal and organisational interoperability in the ARDC People Commons

Michelle Heupel, Legal and organisational interoperability in Coast RI

Pascal Perez, Legal and organisational interoperability in Aurin

Shaily Gandhi (or MHT contact?), Legal and organizational interoperability for urban heat mitigation in India.

Thanasis Sfetsos, Legal and Organizational interoperability for socially just climate adaptation, the case of Egaleo Greece

Discussion.

Exercises around the diagnostic list and mapping of data flows and exchanges.

Feedback.

Presentations Session 9: Empowering the global data community for impact, equity, and inclusion / Education / 205

Mitigating Equity Challenges to Foster Open Science Practices in Emerging Countries

Author: JIBAN KRISHNA PAL[1]

Co-author: Mohamad Mostafa [2]

[1] *Indian Statistical Institute*

[2] *DataCite*

Corresponding Authors: jiban@isical.ac.in, mohamad.mostafa@datacite.org

Open Science practices are essential for promoting transparency, collaboration, and accessibility in research. However, developing countries often face significant equity challenges that hinder their participation in the global research ecosystem. These challenges include capacity gaps, infrastructure disparities, lack of awareness, and digital divides. This session aims to address these barriers by fostering open science practices in underserved communities, ultimately creating a more equitable and interconnected research environment.

The goal of this session is to enhance understanding and cultivate a culture of open science through capacity-building training and strategic outreach activities, focusing on the perspective of developing countries. It has planned to build scholarly communities that empower equitable access to information resources, particularly in South Asian countries.

This session will develop a comprehensive strategy for building scholarly infrastructure through open science utilities (such as FAIR, CAIR, PID, CC licensing, and mandate policies), emphasizing the need for capacity-building, awareness programs, and strategic outreach.

The current landscape of scholarly communication in South Asian countries faces significant challenges, including fragmentation of data, limited accessibility, and lack of standardization. The implementation of open science practices based on FAIR Guiding Principles can help bridge these gaps, enabling researchers to share and access scientific knowledge more efficiently.

This initiative seeks to mobilize open science practices in South Asian countries by focusing on collaboration and networking among the Open Science players (including academic institutions, government bodies, NGSs, and other stakeholders) to support open science initiatives.

Ultimately, it will foster a more interconnected research ecosystem that aligns with global trends.

207

Australian Health Data Evidence Network (AHDEN): Building a National Data Infrastructure for Standardised, Federated Health Data Research

Author: Nicole Pratt[1]

Co-authors: Clair Sullivan [2]; Graeme Hart [3]; Roger Ward [4]

[1] *University of South Australia*

[2] *Queensland Health*

[3] *University of Melbourne*

[4] *ARDC*

Corresponding Authors: gkhart@unimelb.edu.au, nicole.pratt@unisa.edu.au, roger.ward@ardc.edu.au, clair.sullivan@health.qld.gov

Significance of the Issues
Australia's healthcare system generates a vast amount of data, however, data systems are highly fragmented, with information captured across diverse and often incompatible systems. This lack of interoperability creates major barriers to the integration and analysis of health data at scale, limiting the nation's ability to conduct efficient, multi-centre research and generate timely, actionable evidence for health policy and clinical care.

Internationally, federated data networks such as the European Health Data and Evidence Network (EHDEN) and the Observational Health Data Sciences and Informatics (OHDSI) community have demonstrated the value of a harmonised infrastructure using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). These models promote standardisation, preserve privacy, and enable research that is scalable, reproducible, and globally collaborative.

To address this critical need in the Australian context, The University of South Australia (UniSA), with co-investment from the Australian Research Data Commons (ARDC), has established the Australian Health Data Evidence Network (AHDEN). AHDEN's mission is to build a nationally coordinated infrastructure that supports the transformation of hospital-based Electronic Medical Record (EMR) data into the OMOP CDM format across Australia. The OMOP CDM is and open-source, internationally recognised data standard that enables consistent structuring of observational health data across multiple different sources such as hospitals, primary care systems, and disease registries. By transforming diverse local healthcare data into the OMOP CDM format, AHDEN will enable researchers to apply a common suite of analytic tools and methods to gain insights more efficiently, enhancing reproducibility and supporting collaborative research without compromising data security or privacy. Ultimately, AHDEN will strengthen national research capacity, foster data-driven health policy, and unlock the full potential of real-world data for improving population health.

Approach, Structure, and Format
This Session will showcase the AHDEN initiative and the power of the OMOP CDM to enable a scalable, federated infrastructure for health data research. Presentations will demonstrate the technical foundations of the OMOP CDM format, including syntactic and semantic harmonisation, as well as approaches to governance, security, and distributed analytics. The session will highlight the strengths of the OMOP CDM in enabling privacy-preserving data analytics at local, national, and international scales. Participants will gain practical insights into the challenges and successes of mapping Australian EMR data to the OMOP CDM and see how this infrastructure supports the generation of evidence for regulatory, clinical, and policy applications.

Proposed Speakers and Topics
1) Professor Nicole Pratt, AHDEN Project Lead, University of South Australia. (10 minutes)
Professor Pratt will introduce the vision and mission of AHDEN and outline the national and international significance of building a federated health data infrastructure.
2) Associate Professor Graeme Hart, University of Melbourne (15 minutes)
Associate Professor Hart will describe the technical processes involved in syntactic and semantic harmonisation of health data using the OMOP CDM, highlighting the challenges and solutions in applying this model to Australian EMR systems

3) Roger Ward, Solutions Architect, Australian Research Data Commons (15 minutes)
Mr Ward will focus on the data governance and privacy-preserving infrastructure that underpins AHDEN. He will explain how the federated model ensures compliance with privacy regulations while enabling scalable research.
4) Professor Clair Sullivan, University of Queensland (15 minutes)
Professor Sullivan will present progress on mapping Queensland's statewide EMR observation data to the OMOP CDM. She will highlight the clinical and operational benefits of standardisation at a jurisdictional level.
5) Professor Nicole Pratt, AHDEN Project Lead, University of South Australia) (15 minutes) Professor Pratt will present a series of impactful clinical applications that have been enabled through the use of the OMOP CDM, including evidence generation of the safety and effectiveness of all second-line diabetes medications, real-time response to emerging health threats during the COVID-19 pandemic, and investigation of rare adverse events supporting global pharmacovigilance efforts for medicine regulators.
6) Panel discussion (20 minutes)
The session will conclude with a 20-minute panel discussion where all speakers will respond to audience questions and reflect on the future of federated research in Australia

This session directly addresses the growing demand for national-scale health data infrastructure that is privacy-preserving, methodologically rigorous, and interoperable. By showcasing AHDEN' s approach to standardising EMR data through the OMOP CDM, the session will demonstrate the feasibility and value of federated health data research in Australia. Attendees will leave with a clear understanding of how this infrastructure supports timely, high-quality evidence generation that can inform clinical practice, regulatory decision-making, and health policy both nationally and globally.

Contribution Type: Session

208

The CARE Data Maturity model in practice

Authors: Cassandra Sedran-Price[1]; Riley taitingfong[1]; Rose Barrowcliffe[2]; Stephanie Russo-Carroll[1]

[1] *Global Indigenous Data Alliance*

[2] *Maiam Nayri Wingara*

Corresponding Authors: rtaitingfong@arizona.edu, cassandra.price@utas.edu.au, stephaniecarroll@arizona.edu, rose.barrowcliffe@mq.edu.au

The proposed CARE Data Maturity model and in practice Sessions at SciDataCon will explore research and practice. The interactive session will include presentations and original research.This session highlights the development of the CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, and Ethics) emerged from a workshop convened at the IDW 2018/RDA 12th Plenary in Botswana and were originally published by the RDA International Indigenous Data Sovereignty (IDSov) Interest Group (IG).

The session include three presentations:

Abstract 1: The CARE Data Maturity Model
Presenter/s: Cassandra Sedran-Price, Riley Taitingfong and Stephaine Russo Carroll
Format: Interactive session

Since 2019, the CARE Principles have become a leading resource guiding the development of policies and practices for the governance of Indigenous data. The CARE Principles have informed national and international policies around the world, such as the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) Code of Ethics, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) Recommendation on Open Science, and the Policy Partnership on Science, Technology, and Innovation (PPSTI) Statement on Open Science. As universities, repositories, governments, and other data-holding institutions increasingly acknowledge and endorse the

CARE Principles, and use them within their organizational policies, new tools are needed to guide, assess, monitor, and ensure that CARE implementation adheres to the rights, interests, and protocols of the Indigenous Peoples and communities.

The CARE Data Maturity Model (CARE DMM) will meet this need, serving to iteratively assess and improve the strength of CARE implementation in data ecosystems and infrastructures. After the release of CARE, RDA International IDSov IG members collaborated with FAIR Data Maturity Model WG members to plan the development of a CARE DMM. The CARE DMM core team will present progress on the CARE DMM, including criteria and indicators (i.e., measurable actions for the governance of Indigenous data) to evaluate an organization or project's policies, practices, relationships, and data infrastructure for alignment to the CARE Principles. Feedback will be sought on the applicability of the CARE DMM to data practitioners'including on criteria and indicators, and how to translate the indicators into a web-based tool. Finally, progress on the application and endorsement of the tool will also be shared. As data actors increasingly look to the CARE Principles to uphold their commitments to Indigenous Data Sovereignty, the CARE DMM will provide concrete guidance for evaluating and enhancing their data practices and policies to uphold Indigenous Peoples'data rights, priorities, and protocols.

The following two sessions will highlight application of CARE:
Title Abstract 2: Reconnecting Indigenous Data to Country
Presenter/s: Rose Barrowcliffe
Indigenous data sits in repositories around the world in the form of legacy records. These records have been created by non-Indigenous people to serve the functions of non-Indigenous governments, organisations or researchers. Due to the lack of Indigenous Data Sovereignty in the creation of this data, the records'metadata is largely absent of fields or keywords that indicate to which First Nations the data relates. This results in a findability gap for Indigenous people trying to find and access their Indigenous data. In this Aboriginal-led project partnership with New York Botanical Gardens (NYBG), we sought to develop processes for reconnecting Indigenous data back to Country even when the record lacks Indigenous metadata.
The NYBG Herbarium holds over 20,000 specimens collected in Australia from the beginning of colonisation. The specimen records'metadata mentions the collector, the species, and sometimes some contextual information in the field notes, but they don't say whose Country the specimens come from. This presentation discusses the considerations and attempts to reconnect the Indigenous data held at NYBG to the Country from which it was collected. The project was grounded in Indigenous Data Sovereignty and tested different geospatial analysis techniques to propose the First Nation/s that relate to that record with the hope that we can eventually connect the specimen records back to Country.

Title Abstract 3: Embedding Indigenous Data Sovereignty in Environmental Research
Presenter/s: Cassandra Sedran-Price, Rachel A. Ankeny, Riley Taitingfong, Jess Melbourne-Thomas, Rose Barrowcliffe, Lydia Jennings and Stephaine Russo Carroll

In ecology, climate science, and natural resource management, Indigenous data are increasingly used to develop models that guide decision-making for biodiversity conservation, sustainable agrifood systems, and emerging environmental technologies. Supporting Indigenous rights and integrating both Indigenous and Western Knowledges is essential for improving environmental outcomes in the management of Country This inclusive approach benefits Indigenous Peoples and can enhance the development and application of ecological models. For instance, incorporating Indigenous Knowledge into species distribution modelling for threatened and culturally significant species in Australia has led to more accurate models and better-informed management strategies.
Widely promoted data principles enhance data sharing, such as the FAIR Principles (Findable, Accessible, Interoperable, Reusable) while other principles that remind data users of the people and purpose orientation, such as CARE Principles (Collective Benefit, Authority to Control, Responsibility, Ethics), remain underutilised. Embedding the CARE Principles presents a critical opportunity for modellers and researchers to ensure that their work respects Indigenous knowledge systems, strengthens partnerships with Indigenous communities, supports fair and equitable benefits for those on whose Country they work, and benefits from the expertise of Indigenous Peoples intergenerational knowledge of their ecosystems and from where the data and associated modelling derive.
This presentation explores how operationalising the CARE Principles can support Indigenous rights, strengthen the ethical foundations of modelling practices, and ensure Indigenous communities maintain sovereignty over their data while receiving equitable benefits from research outcomes.

Presentations Session 1: CAREful Indigenous Data Governance / 209

Implementing Indigenous Data Sovereignty within a Government system

Authors: Bobby Maher[1]; Raymond Lovett[2]

[1] *Maiam nayri Wingara*

[2] *Global Indigenous Data Sovereignty Alliance*

Corresponding Authors: raymond.lovett@anu.edu.au, bobby.maher@anu.edu.au

Indigenous Data Sovereignty (IDSov) is a global Indigenous developed and led philosophy asserting Indigenous rights in data. In Australia, an Anglo-colonised state, IDSov has increasingly been identified and communicated as a high priority for the nation state. The nation state has committed to supporting 'data sharing'through a National Partnership Agreement for Closing the Gap by providing access to data and information at a regional level for Aboriginal and Torres Strait Islander communities. To ensure that the data agenda benefits Aboriginal and Torres Strait Islander communities requires commitment and systems change in those state agencies. Embedding Indigenous Data Governance (IDGov) structures and processes Indigenous data priorities, data sharing from the state to communities and state actors facilitating access to Indigenous data for Indigenous communities can be operationalised.

The Maiam nayri Wingara Indigenous Data Sovereignty Collective together with the Australian Capital Territory (ACT) Aboriginal and Torres Strait Islander community and the ACT Health Directorate have embarked on a process to embed IDSov through IDGov. This presentation will share the IDGov development and implementation processes of: (1) Socialising IDSov and IDGov concepts with community, the ACT Health Directorate, including the executive, (2) Undertaking priority setting and data mapping exercise, (3) Developing an IDGov structure including policy and procedures and (4) Mapping the system to evaluate the systems change.

The project may act as an example for other state agencies across the Australia to embed IDGov structures within their systems. The result will ensure that Aboriginal and Torres Strait Islander people can take control of their data for their self-determination and nation (re)building.

Poster Session / 210

Pedagogical Innovation and the Maiam nayri Wingara Indigenous Data Sovereignty Fundamentals Course

Authors: Bhiamie Williamson[1]; Cassandra Sedran-Price[1]; Jacob Prehn[1]; Maggie Walter[2]; Raymond Lovett[3]; Sam Provost[None]

[1] *Maiam nayri Wingara*

[2] *Maiam Nayri Wingara*

[3] *Global Indigenous Data Sovereignty Alliance*

Corresponding Authors: bhiamie.williamson@monash.edu, raymond.lovett@anu.edu.au, maggie.walter@utas.edu.au, cassandra.price@utas.edu.au, sam.provost@anu.edu.au, jacob.prehn@utas.edu.au

In Australia, the interest in Indigenous Data Sovereignty (IDSov) has increased over the past decade. There are now emerging competing interests between the state as holders of vast Indigenous data assets and how these data are governed and Indigenous communities to drive their agenda on Indigenous data priorities including data sharing from the state to communities and state actors facilitating access to Indigenous data for Indigenous communities.

This presentation details pedagogical developments by the Maiam nayri Wingara (MnW) Indigenous Data Sovereignty collective on their Foundations of IDSov and IDGov short course to guide the operationalisation of IDSov principles. We will discuss the Foundations course and how it socialises the concepts of IDSov and IDGov, including how the content aligns with the global philosophy of

IDSov as an Indigenous-led movement. Additionally, we will share lessons learned thus far relating to the delivery of the training modules that make up the course.

We conclude by sharing potential new additions to the Maiam nayri Wingara course offerings and seeking additional suggestions for future training.

211

Coalition building to support sustainable digital data standards

Authors: Ian Bruno[1]; Kerstin Lehnert[2]; Leah McEwen[3]; Lesley Wyborn[4]; Mark Rattenbury[5]; Oliver Koepler[6]; Richard Hartshorn[7]; Simon Hodson[8]

[1] *CCDC*

[2] *Lamont-Doherty Earth Observatory of Columbia University*

[3] *Cornell University*

[4] *Australian National University*

[5] *GNS Science*

[6] *TIB - Leibniz Information Centre for Science and Technology*

[7] *University of Canterbury*

[8] *CODATA*

Corresponding Authors: simon@codata.org, bruno@ccdc.cam.ac.uk, oliver.koepler@tib.eu, richard.hartshorn@canterbury.ac.nz, lrm1@cornell.edu, lehnert@ldeo.columbia.edu, lesley.wyborn@anu.edu.au, m.rattenbury@gns.cri.nz

This session will bring together members of the research data community with experience of and interest in developing consortia and coalitions that advance the development and application of practices, principles and standards relating to research data. It will aim to identify common considerations and challenges encountered when building coalitions that can inform those embarking on a similar journey, and identify areas for cross-community collaboration to help ensure the long-term sustainability and success of such activities.

We will consider how proposed ideas and specifications have become accepted by communities as agreed-upon 'standards', what sort of community support, governance, and organization 'ownership'is necessary to maintain and sustain digital resources, and the challenges of sustaining the ongoing development and maintenance of outputs in order to meet evolving circumstances. To do this, we will draw on examples of international initiatives at various stages of maturity from across the chemical, earth and life sciences.

In chemistry, stakeholders from across industry and academia, including publishers, data organizations, scientific unions and research institutes, have been convening in workshops focused on the sustainable development of digital standards for describing chemical information across disciplines. These workshops originated from the WorldFAIR Chemistry project, led by IUPAC, with an aim of establishing a coalition of organizations and individuals contributing to long term, pre-competitive services for discovery, adoption, and validation of standards.

At a national level, NFDI4Chem, the chemistry consortium within the NFDI initiative in Germany, brings together research institutions, information providers and learned societies with the aim of digitizing all key steps in chemistry research. In the UK, the Physical Sciences Data Infrastructure project (PSDI) is aiming to connect across various data systems in use by researchers to support process recording, facilitate data compilation and advance computational analysis. Services and infrastructure being developed in these projects require standards for ontologies, metadata and data, and international collaboration with researchers, RDM specialists, ontology engineers, and data scientists to develop guidelines, data models and strategies that support standardization.

The Earth Sciences are global in scope and many international coalitions have been formed to develop digital data standards that enable integration of data collected around the planet. Some of the oldest are the GeoSciML and EarthResourceML, logical data models developed by the Commission for the

Management and Application of Geoscience Information (CGI) of IUGS, which are now governed and maintained by the Open Geospatial Consortium and CGI. Sustaining and promoting these standards without ongoing project support is proving challenging.

A more recent earth sciences initiative, OneGeochemistry, brings together major geochemical data providers and the research community with the objective of establishing global standards and best practices for FAIR laboratory analytical data. Standardization and promulgation of best practices and protocols will be a long process as the global geochemistry community is very heterogeneous and fragmented due to a multitude of analytical techniques applied to a wide variety of materials. Developing sustainable long-term coalitions with other international initiatives to facilitate cross-domain interoperability will be critical.

The Genomic Standards Consortium (GSC) is an international organization focused on the development of genomic and environmental metadata standards to facilitate sharing and reuse of genomic data across databases and between studies. The GSC teams up with domain experts to develop and integrate novel minimum information standards (MIxS) through monthly working group sessions and annual meetings. Striving for true reproducibility of genomic data, the GSC teams up with outreach initiatives run by other national and international microbiome and multi-omics data alliances and consortia.

The Worldwide Protein Data Bank (wwPDB) represents a consortium of international organizations in America, Europe and Asia that collaborate to jointly manage a global repository of 3D biological macromolecular structures. As part of this, the wwPDB oversees the governance of mmCIF, a community standard for the exchange, annotation, validation and archiving of macromolecular structural data. The partnership of organizations in the wwPDB enables funds to be channeled from different regions in support of data standards and curation.

More broadly, the Global Biodata Coalition brings together a range of organizations who fund life science research to jointly address the need for sustainable financial support for global biodata resources and share approaches and strategies for efficient management and growth of infrastructure in this area. The challenge for databases and knowledgebases that archive and add value to research data is that they constitute infrastructure requiring stable long-term support, but are generally funded through short-term grants and contracts.

This session will be designed as a panel discussion with brief presentations from a broad sample of established and emerging coalitions such as those described above, several of whom have agreed to participate. We will share current progress and goals, hear about success stories, how barriers were overcome and ongoing challenges that might be addressed by wider cooperation and collaboration.

Specifically we will ask panelists to reflect on the challenges encountered, strategies adopted and outcomes achieved in the following areas:

-Core goals and target stakeholders
-Leadership and governance structures
-Engaging and coordinating across diverse stakeholders
-Communicating vision and activities to wider communities
-Attracting resources needed to advance aims
-Maintaining momentum of activities over time
-Timescales - initial aspirations vs reality
-Strategies for ensuring sustainability of activities and outputs

We anticipate that the result of this session will be an understanding of strategies that can successfully catalyze the development of initiatives aiming to organize across stakeholders and disciplines, how to avoid common pitfalls, and opportunities for broad community collaboration of benefit across coalition building activities.

Poster Session / 212

KeyPoint: Trusted Research Environment for sensitive data.

Author: Peter Marendy[1]

[1] *QCIF*

Corresponding Author: peter.marendy@qcif.edu.au

QCIF has developed a purpose-built trusted research environment named KeyPoint to address the increasing need for secure and trusted digital environments for sensitive data in various research fields, including population health, biosecurity, food security, environmental science, and social science. KeyPoint provides a remote analysis environment for sensitive data which enables robust governance, management, and sharing of sensitive research data with approved researchers in a scalable, highly secure platform.

KeyPoint ensures data governance at scale and expandability by employing self-contained research environments with strong role-based access controls and complete separation of research activities. KeyPoint uses a novel approach which associates roles and contexts to an Australian Access Federation identity through personas, achieving strict role and project separation. This is particularly relevant at the virtual desktop layer, eliminating inadvertent opportunities for data linkage or masquerading data ingress and egress across different projects. KeyPoint's capabilities support globally distributed collaborators on research projects.

KeyPoint has been developed with security controls, such as ISO/IEC 27001, in mind. Further, it aligns with the Five Safes Data Sharing Principles, providing governed and highly secure environments for collaborative research analysis.

This presentation will provide an overview of KeyPoint, focusing on its innovative data governance model and its scalability models. It will also describe the approach taken to ensure strict project separation at the virtual desktop. Additionally, the presentation will address future work and planned capabilities.

KeyPoint's advanced capabilities in data analysis, including AI and machine learning, have already been adopted by ground-breaking projects, empowering researchers to tackle complex research challenges across any research domain. This highlights the importance of robust infrastructures in supporting data-intensive research and fostering global collaboration for sensitive data.

214

AI for Metadata Enhancement, Metadata for AI Readiness: how do we ensure a virtuous rather than a vicious circle?

Authors: Arofan Gregory[1]; Cristina Gonzalez[2]; Deirdre Lungley[3]; Doug Fils[4]; Isabel Ceron[5]; Jieping Ye[6]; Kelsey Drucken[7]; Mercè Crosas[8]; Pascal Heus[9]; Rebecca Farrington[10]; Sean Hill[2]; Simon Hodson[1]; Stephen Richard[4]; Vyacheslav Tikhonov[11]; Yitian Xiao[6]

[1] *CODATA*

[2] *SenScienceAI*

[3] *UKDS*

[4] *Consultant / CODATA*

[5] *Australian Academy of Social Sciences*

[6] *GeoGPT / Zhejiang Lab*

[7] *ANU*

[8] *BSC / CODATA*

[9] *Postman / CODATA*

[10] *AuScope*

[11] *DANS*

Corresponding Authors: doug@fils.network, simon@codata.org, rebecca@auscope.org.au, dmlung@essex.ac.uk, cristina.gonzalez@senscience.ai, 4tikhonov@gmail.com, jieping@gmail.com, isabel.ceron@socialsciences.org.au, yitian.xiao@yahoo.com, pascal@codata.org, sean.hill@senscience.ai, kelsey.druken@anu.edu.au, smrtucson@gmail.com, arofan@codata.org, merce.crosas@bsc.es

This session will explore the intriguing and potentially urgent interaction (and even codependency) between high quality metadata and semantic richness on the one hand and Generative Artificial Intelligence (AI) and Large Language Models (LLMs) on the other. A lot of work is going on to improve the richness, quality and standardisation of metadata and semantics in order to make data sets 'AI ready'. At the same time, the potential of generative AI is being explored precisely to enrich metadata and semantics. Exercising caution in this endeavour is critical, however, as the quality of the outputs is directly tied to the quality of the underlying data and documentation. The topic of this session is to explore, through numerous examples and discussion, the latest work in this area. AI for Metadata Enhancement, Metadata for AI Readiness: how do we ensure a virtuous rather than a vicious circle?

On the side of metadata for AI readiness, we have:

- ML Commons' Croissant
- The Cross-Domain Interoperability Framework (CDIF) is aligning with Croissant and has important components on data description (a profile of DDI-CDI) and work underway on provenance and data quality, which is important in this context.
- Work on semantic mappings and knowledge graphs.

On the side of AI for metadata enhancement and inference we have:

- The remarkable work of the SenScience team with FAIR2 and a compelling example of metadata enhancement for a Frontiers data article and science article on biodiversity off the Basque coast.
- The work at Closer, UK on metadata inference, enrichment and 'uplift'.
- GeoGPT assisted classification and application of geological terminologies and semantics.

In parallel, there is a growing realisation that maintaining the quality of AI metadata enhancement and inference, requires the LLMs being able to access key knowledge, for example through a Model Context Protocol server, as expressed in authoritative terminologies or other sources of reference:

- The idea of a Model Context Protocol (MCP) server for the SI Reference Point to make the underlying knowledge accessible to LLMs and agent.
- The idea of implementing a MCP server for Croissant and for CDIF.
- Work to predict semantic mappings (including GeoGPT and the work of Vyacheslav Tykhonov).

The session will be composed of a number of quick-fire presentations covering various aspects of the issues raised here, thus below, in many instances, we give not titles but issues and examples to be introduced and discussed. We intend this to be a rapid exchange of ideas rather than a series of formal presentations. There will be significant time for discussion. One outcome will be a quick report surveying the landscape and covering the issues raised. Above all, however, we will seek to identify concrete steps that scientific communities and the Research Infrastructures that serve them can take, drawing on these examples and emerging practices, to address issues of AI readiness and metadata enhancement, while ensuring we achieve a virtuous circle.

Programme:

Simon Hodson, Arofan Gregory, CODATA:
- Introduction to the issues: AI for Metadata Enhancement, Metadata for AI Readiness: how do we ensure a virtuous rather than a vicious circle?

Doug Fils, Consultant / CODATA; Vyacheslav Tykhonov, DANS; Pascal Heus, CODATA / Postman:
- Croissant, Semantic Croissant and GeoCroissant.

Pascal Heus, CODATA / Postman:
- The critical role of FAIR Open Data APIs for AI
- Findings of the CDIF AI Readiness Working Group
- Related R&D and topics

Vyacheslav Tykhonov, DANS:
- AI-powered Semantic Mappings with RAG for ontology alignment
- Leveraging AI to Automatically Link Controlled Vocabulary Terms in Metadata
- Semantic Croissant: Enabling FAIR Data for AI Applications with the Model Context Protocol (MCP)
- MCP Server Library: A Foundation for AI Applications and FAIR Data Workflows
- Dataverse: Building a Distributed Data Network Ready for AI

Deirdre Lungley, UKDS:
- AI for Metadata Enhancement and Inference: the example of UKDS metadata uplift.
- Metacurate-ML: UK ESRC funded project to improve curation tooling, enabling semi-automated metadata uplift at scale. Workstreams:
- Questionnaire Extraction from PDFs using LLMs
- Subsequent Question alignment using AI
- LLM topic classification of these questions/variables
- Harnessing the knowledge graph produced in these preceding steps, together with further LLM identification of indirect identifier variables to power dataset ingest, including Statistical Disclosure Control (SDC)

Sean Hill, Cristina Gonzales, SenScience:
- FAIR2

Jieping Ye, Zhejiang Lab / GeoGPT:
- GeoGPT for classification.

Rebecca Farrington, AuScope; Kelsey Druken, ANU; Isabel Ceron, Australian Academy of Social Sciences:
- AI readiness and metadata inference in Australia and beyond

Discussion:
- Landscape
- Future collaborations
- Recommendations

217

The transformative impact of the CARE Principles and Māori Data Sovereignty: Lessons from Aotearoa New Zealand

Authors: Andrew Sporle[1]; Daniel Wilson[2]; Kiri West[2]; Lara Greaves[3]; Larissa Renfrew[2]; Phil Wilcox[4]; Tahu Kukutai[5]; Tori Diamond[6]

[1] *iNZight Analytics Ltd*

[2] *University of Auckland*

[3] *Victoria University of Wellington*

[4] *University of Otago*

[5] *University of Waikato*

[6] *University of Auckland, iNZight Analytics Ltd*

Corresponding Authors: larissa.renfrew@auckland.ac.nz, andrew@inzight.co.nz, tori.diamond@auckland.ac.nz, daniel.wilson@auckland.ac.nz, kiri.west@auckland.ac.nz, tahu.kukutai@waikato.ac.nz, phillip.wilcox@otago.ac.nz, lara.greaves@vuw.ac.nz

The development of Indigenous data sovereignty as a global movement and the creation of the CARE principles have resulted in a diversity of solutions to meet local issues and contexts. This movement

from principles to practice has involved both locally-specific and global principles that are informing change to both Indigenous and non-Indigenous data practice.

This session focuses on Aotearoa New Zealand, where Māori data sovereignty and data governance has been advanced over the past decade using many strategies, including Te Mana Raraunga's principles of Māori data sovereignty, and the newer Māori data governance model from Te Kāhui Raraunga. Some strategies have been more impactful than others in producing change. This session presents case studies from a decade of work in Māori data sovereignty and governance to share lessons with other Indigenous groups and allies.

The session comprises a series of short talks from Māori speakers across a range of disciplines and career stages. Examples of these speakers and topics include:

Considering the application of CARE principles and the Māori Data Governance Model, in the governance of Māori research data held by universities - Dr Kiri West, University of Auckland

Ethical considerations and data management in visual research with Māori youth - Larissa Renfrew, University of Auckland

Applying the CARE principles, Māori data sovereignty and governance to data used in the exercise of Māori political and citizenship rights - AProf Lara Greaves, Victoria University of Wellington

Implementing the CARE principles to enable Māori genomic data sovereignty in precision health. - AProf Phillip WIlcox, University of Otago

Applying Māori data sovereignty to promote equitable public health policy: Using insights from age standardisation and population denominators - Tori Diamond, University of Auckland

The session will end with both a Q&A and a panel discussion with tuākana (senior) and teina (junior, emerging) discussants wrapping up the past ten years of developments, key lessons for international contexts, and looking forward to the next decade.

Poster Session / 218

Facilitating Cross-Domain Interoperability of X-Ray Absorption Spectroscopy (XAS) Data: Developing a CDIF Profile for the Galaxy Platform.

Authors: Abraham Nieva de la Hidalga[1]; Arofan Gregory[2]; Heike Görzig[3]; Leandro Liborio[4]; Markus Kubin[3]; Patrick Austin[4]; Rolf Krahl[5]; Simon Hodson[2]

[1] *School of Computer Science and Informatics, Cardiff University, Wales, UK.*

[2] *CODATA, the Committee on Data of the International Science Council, France.*

[3] *Helmholtz Zentrum Berlin für Materialien und Energie (HZB), and Helmholtz Metadata Collaboration (HMC), Germany.*

[4] *Scientific Computing Department, STFC, UKRI, UK*

[5] *Helmholtz Zentrum Berlin für Materialien und Energie (HZB), Germany.*

Corresponding Authors: nievadelahidalgaa@cardiff.ac.uk, heike.goerzig@helmholtz-berlin.de, ilg21@yahoo.com, simon@codata.org, patrick.austin@stfc.ac.uk, leandro.liborio@stfc.ac.uk, rolf.krahl@helmholtz-berlin.de, markus.kubin@helmholtz-berlin.de

The Cross Domain Interoperability Framework (CDIF) provides a set of implementation guidelines designed to lower the barriers to cross-domain research data reuse. CDIF provides standards and methodologies for addressing interoperability issues preventing cross-domain research data utilization. CDIF's initial version comprises five core profiles: Discovery, Access, Controlled Vocabularies, Data Description for Integration, and Universals, which collectively support the cross-disciplinary implementation of the FAIR principles.

The challenge of embedding FAIR principles into research outputs applies not only to data and metadata, but also to the methods used to analyse them. The metadata associated with raw data must be Interoperable support reusability. However, if the parameters used in the analysis of the data -and the corresponding metadata- are not recorded properly, Reusability will also be compromised.

Two European projects are addressing these challenges of cross domain interoperability and data analysis reproducibility in X-ray Absorption Spectroscopy (XAS). The first one is the OSCARS-funded CDIF-4-XAS project that is applying CDIF to enhance the interoperability and reusability of XAS data. The objective is to streamline data exchange between applications, databases, and institutions, making XAS data interoperable across multiple research disciplines. The second project is the EuroScienceGateway, which aims to leverage European computing infrastructures for data-intensive research guided by FAIR principles.

Recently, the CDIF-4-XAS project published its first deliverable: a comprehensive landscape analysis of standards, vocabularies, ontologies, data formats, and practices in XAS. For the EuroScienceGateway project, we have contributed a set of custom tools that can be used in the Galaxy platform for managing workflows associated to XAS data processing and analysis. Galaxy offers a number of features that ensure that the workflows'outputs retain all the metadata needed for them to be reproduced: histories store the data and parameter inputs associated to all output data; software tools are strictly versioned and run in containers, and the execution of workflows can be exported as Research Object Crates.

Building on this, the CDIF-4-XAS is developing semantic descriptions of two XAS community standards (NXxas for multi-spectra raw and processed data and XDI for single spectra data) that will produce a CDIF profile (XAS-CDIF). This includes: exploration of the use of the CDIF Discovery profile and related standards such as schema.org, DCAT, and PROV-O and of DDI-CDI for data description of variables; characterisation of the HDF5 data structure using DDI-CDI and mappings of key NXxas and XDI concepts. The XAS-CDIF profile will be used to extend, and update, the existing XAS Galaxy tools and workflows, which will facilitate the seamless integration of data from various sources into processing and analysis workflows.

This paper will briefly present the OSCARS and EuroSciencegateway projects and explain how the XAS-CDIF may facilitate seamless integration of XAS datasets from different beamlines and laboratories, promoting data reuse across diverse research domains. The prototype implementation of XAS-CDIF into Galaxy tools and workflows will be presented as an example of how the XAS-CDIF profile facilitates building advanced tools that take advantage of data interoperability.

Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data / 219

FAIR mappings for data transformation and semantic alignment using Metadata Schema and Crosswalk Registry - Case Research Data Cloud (NII)

Authors: Joonas Kesäniemi[1]; Toshiyuki Hiraki[2]

[1] *CSC*

[2] *National Institute of Informatics, the Research Center for Open Science and Data Platform*

Corresponding Authors: hiraki@nii.ac.jp, joonas.kesaniemi@csc.fi

Mappings are an essential component in making research data interoperable across infrastructures, domains and disciplines. Correspondences between official and de facto standards related to conceptual models, structures and vocabularies are required to share meaning and transfer information between both humans and machines. Despite their importance, these correspondences, mappings or crosswalks (i.e. collections of mappings 1), can be hard to find or reuse as they are often scattered across different systems and available in a variety of formats and structures. Hence, making mappings FAIR (for Findable, Accessible, Interoperable, and Reusable) and machine-actionable has a great potential for reducing resources needed to create and maintain mappings. With FAIR principles

the creation of high quality mappings can be supported by tools that help with discovery, access and evaluation of existing mappings at scale. Making mappings FAIR comes with specific requirements for mapping models and formats as well as associated metadata and services 1. It has been proposed in 2 that these responsibilities should be handled by specialised mapping repositories to ensure the findability, availability, and reusability of these potentially highly valuable resources.

Metadata Schema and Crosswalk Registry 9 (MSCR) is a new and innovative service developed as part of the EU funded FAIRCORE4EOSC project and now operated by the IT Centre for Science, Finland. The MSCR offers multiple features that support and guide users through the different steps of the schema registration, mapping creation and development as well as publication. The MSCR supports versioning, lifecycle and content management and provides a persistent identifier for all published schemas, crosswalks and mappings for reliable referencing. By using MSCR, users have the possibility to map between any two registered schemas on the platform, which makes it possible to bridge ontologies and vocabularies as well as create crosswalks for data structure, format, and value transformation. The crosswalk editor supports mappings with different source and target element cardinalities (e.g. 1-to-1 and 1-to-many) and provides a set of processing functions that can be used to apply more complex mapping logic such as filtering or concatenation. Crosswalks created with the editor can be exported in different formats, such as SSSOM 7, XSLT and RML 8 to facilitate data transformation in other systems.

National Institute of Informatics (NII) in Japan manages the Research Data Cloud (RDC) 3, a research data infrastructure supporting research activities throughout their research data lifecycle. RDC consists of GakuNIn RDM (GRDM) for research data management, JAICO Cloud for papers and data publication, and CiNii Research for discovery. The NII RDC application profile (NII RDC-AP) was recently developed 4 to enhance the interoperability of information exchanged among the services in the NII RDC. The challenge for NII is now to make the NII RDC-AP resources interoperable with other research data related application profiles in order to make research data within the NII RDC reusable. We present three mapping scenarios that showcase features of the MSCR and work towards concrete solutions to the interoperability challenges at NII and RDC:

1. **Transformation of metadata** from the GakuNin RDM to the NII data governance function for monitoring and supporting data management plans. There already exists a working solution called NII-DG-manager that implements JSON-to-JSON data transformation based on a custom mapping configuration. This scenario evaluates MSCR's mapping capabilities against the original python implementation from technical and user experience perspectives.
2. **Semantic alignment;** mapping of NII RDC-AP with other research data management platforms through ontology mapping between the NII-RDM ontology 5 and the SKG-IF ontology 6.
3. **Operationalization;** by creating a crosswalk between the mappings models of the MSCR and NII-DG-manager, which allows MSCR crosswalks (with certain restrictions) to be executed using NII-DG-manager. This is an interesting meta-mapping type of case that can be extended to other target schemas as well.

The offering of the presentation is two-fold: First, we introduce an innovative and operational MSCR service, which supports a variety of mapping cases and fosters FAIR practices. Then, we present a practical example of the use and operationalisation of the MSCR in the context of NII national data infrastructure. The presentation targets a wide audience, including both newcomers and knowledge management practitioners (ontologists, metadata experts,...) who are interested in interoperable mapping practices.

References:

1 Le Franc, Y., Grau, N., Juty, N., Mejias, G., Reed, P., Ramezani, P., Poveda-Villalon, M., Garijo, D., Goble, C., & van Horik, R. (2025). D4.5 - Guidelines and methodology to create, document and share mappings and crosswalks (V1.1). Zenodo. https://doi.org/10.5281/zenodo.15111167

2 Broeder, D., Budroni, P., Degl'Innocenti, E., Le Franc, Y., Hugo, W., Jeffery, K., Weiland, C., Wittenburg, P., & Zwolf, C. M. (2021). SEMAF: A Proposal for a Flexible Semantic Mapping Framework (1.0). Zenodo. https://doi.org/10.5281/zenodo.4651421

3 Research Center for Open Science and Data Platform (RCOS), National Institute of Informatics. "Overview of the NII Research Data Cloud," https://rcos.nii.ac.jp/en/service/ (accessed 2025-04-25)

4 Minamiyama, Y., Hayashi, M., Fujiwara, I., Onami , J.- ichi, Yokoyama, S., Komiyama, Y., & Yamaji, K. (2023). Toward the Development of NII RDC Application Profile Using Ontology Technology. Proceedings of the Conference on Research Data Infrastructure, 1. https://doi.org/10.52825/cordi.v1i.260

5 RDM Ontology Usage Guidelines, Research Center for Open Science and data platform (RCOS) in National Institute of Informatics, https://rcosdp.github.io/RDM/ (accessed 2025-04-25)

6 The SKG-IF Ontology, Scientific Knowledge Graphs –Interoperability Framework (SKG-IF) WG https://skg-if.github.io/data-model/ontology/current/skg-o.html (accessed 2025-04-25)

|7] https://mapping-commons.github.io/sssom/

8 https://rml.io/

9 Kesäniemi, J., Suominen, T., Broeder, D., & Puuska, H. Implementing interoperability - Metadata Schema and Crosswalk Registry approach to FAIR metadata mappings. EPiC Series in Computing

Presentations Session 1: CAREful Indigenous Data Governance / 223

The Aotearoa Genomic Data Repository: A haven for digital sequence information enabling Māori Data Sovereignty

Authors: Carvin Rui Chen[1]; Claire Rye[2]; E. Owen Perkins[1]; Jun Huh[1]; Libby Liggins[3]; Mik Black[4]; Nathalie Giraudon[1]; Rudiger Brauning[5]; Tanis Godwin[4]; Tracey Godfery[4]

[1] *National eScience Infrastructure (NeSI)*

[2] *New Zealand eScience Infrastructure*

[3] *University of Auckland*

[4] *University of Otago*

[5] *AgResearch*

Corresponding Authors: carvin.chen@nesi.org.nz, tanis.godwin@otago.ac.nz, jun.huh@nesi.org.nz, rudiger.brauning@agresearch.c libby.liggins@auckland.ac.nz, tracey.godfery@otago.ac.nz, mik.black@otago.ac.nz, eirian.perkins@nesi.org.nz, nathalie.giraudon@nes claire.rye@nesi.org.nz

The Aotearoa Genomic Data Repository (AGDR, https://data.agdr.org.nz/) provides secure within-nation storage, management, and sharing of non-human genomic data generated from biological and environmental samples originating in Aotearoa New Zealand. Te ao Māori, the worldview of the Indigenous people of Aotearoa New Zealand, recognises all living entities as taonga (treasured or precious) that require protection through kaitiakitanga (guardianship). This responsibility extends to any data generated through the study of taonga species. For genomic data derived from taonga species, there is also the potential to obtain additional information about whakapapa (genealogy, social and ecological relationships, and ancestral inheritances), which is itself taonga, and must also be protected as part of the kaitiaki (guardian) role. The creation of a within-nation data storage facility provides Māori iwi, hapū and whānau (tribes, kinship groups and families) with the crucial ability to control access to these taonga.

The AGDR was jointly established by Genomics Aotearoa (https://www.genomics-aotearoa.org.nz/) and the New Zealand eScience Infrastructure (https://www.nesi.org.nz/), with funding from the Ministry of Business Innovation and Employment (https://www.mbie.govt.nz/). The repository has been designed with the FAIR Guiding Principles for scientific data management and stewardship in mind —making data findable and interoperable with data held in other genomic repositories. However, its development has been guided primarily by the principles of Māori Data Sovereignty. The decision-making process regarding who can access each data set is entirely in the hands of the iwi, hapū and whānau that are kaitiaki of the data, thus upholding the globally-relevant CARE Principles for Indigenous Data Governance.

In this presentation, we will provide an overview of the approach and development activities in support of the AGDR, and describe the benefits of the repository, along with proposed future developments. Our presentation will be relevant to researchers generating and managing genomic data in partnership with Indigenous communities, repository developers who work with Indigenous data, and data scientists who operate in contexts where Indigenous Data Sovereignty is relevant. Four years since its initial launch, we will describe the uptake and use of the AGDR, development features and adherence to established best-practices that ensure interoperability of the AGDR datasets with those of other genomic data repositories, and the mechanisms used to support Indigenous Data Governance, such as the use of Local Contexts (https://localcontexts.org/) Biocultural Notices and Labels and the formation of an AGDR Advisory Board to provide oversight and cultural guidance.

225

From Guidance to Practice: Implementing Open Science Data Policies in Crisis Situations

Authors: Ana Persic[1]; Areeq Chowdhury[2]; Ingvill Constanze Odegaard[3]; Jacqueline Stephens[4]; LILI ZHANG[5]; Nicole Mwananshiki[2]; Rania Sabo[1]; Simon Hodson[6]; Virginia Murray[7]; Francis P. Crawley[8]; Gnana Bharathy[9]; Allyson Lister[10]

[1] *UNESCO*

[2] *The Royal Society*

[3] *CBOW*

[4] *Flinders University*

[5] *COMPUTER NETWORK INFORMATION CENTER, CAS*

[6] *CODATA*

[7] *UKHSA*

[8] *CODATA IDPC*

[9] *ARDC/ UTS*

[10] *FAIRsharing*

Corresponding Authors: ingvillconstanze.odegaard@cbowproject.org, jacqueline.stephens@flinders.edu.au, a.persic@unesco.org, r.sabo@unesco.org, areeq.chowdhury@royalsociety.org, nicole.mwananshiku@royalsociety.org, simon@codata.org, fpc@gcpalliance.org, virginia.murray@ukhsa.gov.uk, allyson.lister@oerc.ox.ac.uk, zhll@cnic.cn, gnana.bharathy@ardc.edu.au

Introduction
This workshop addresses the urgent need to translate the high-level vision of the UNESCO Recommendation on Open Science into concrete, actionable data governance mechanisms tailored for crisis contexts. Whether triggered by natural disasters, health emergencies, climate change, or geopolitical conflicts, in times of crisis effective and ethical data management is essential for informed decision-making, resource coordination, and community resilience. Drawing on the practical outputs of the UNESCO-CODATA project Data Policies for Times of Crisis Facilitated by Open Science (DPTC) (including the Guidance, Checklist, Factsheet, and Technical Report), participants will explore how to operationalize these tools to design agile, transparent, and context-sensitive data policies. These tools are part of the UNESCO Open Science Toolkit and are aligned with global instruments such as the Sendai Framework for Disaster Risk Reduction, UNDRR's Hazard Information Profiles (HIPs), and the WHO's International Health Regulations (IHR).

Through structured, interactive exercises, the session will engage participants in applying the DPTC Toolkit to simulated crisis scenarios grounded in real-world complexities. Using open science-aligned policy frameworks, participants will collaboratively address barriers such as data sovereignty, fragmentation of data systems, legal and ethical constraints on real-time data sharing, protection of sensitive information, equity of access, and responsible use of artificial intelligence. The workshop will emphasize alignment with the FAIR Data Principles (Findable, Accessible, Interoperable, Reusable), CARE Data Principles (Collective Benefit, Authority to Control, Responsibility, and Ethics), and TRUST Digital Repository Principles (Transparency, Responsibility, User Focus, Sustainability, and

Technology). It highlights the importance of integrating traditional and Indigenous knowledge systems and multilingual accessibility.

The goal of the session is to empower multi-stakeholder communities, including researchers, policymakers, data scientists, data stewards, emergency responders, and representatives from affected regions, to co-create crisis-ready data policy strategies that are ethical, inclusive, and resilient. Drawing on international frameworks such as the SDGs, the UN Pact for the Future, WHO Health Emergency and Disaster Risk Management (Health-EDRM) Framework, and the Royal Society's guidance on trusted data systems, the workshop will produce a prototype implementation roadmap. This roadmap will support regional and disciplinary customization of open science data policies and serve as a living document for advancing capacity-building, governance, and preparedness in the face of complex and cascading crises. The session will centre equity and interoperability, ensuring that communities most at risk are not excluded from the benefits of scientific data and that data infrastructures are robust enough to support collaborative, cross-border crisis response.

Workshop (90 minutes) –Format: Interactive, Practical Implementation
All of the listed authors are invited to participate as either presenters, panelists, or breakout group leaders.
It includes a mix of brief presentations, group work, and structured discussion, designed to enable hands-on learning and policy co-creation. Here's the breakdown:

Opening Presentation (10 minutes)
Introduction to the UNESCO-CODATA DPTC tools (Checklist, Guidance, Factsheet) and their global policy context (UNESCO Open Science, Sendai, WHO IHR, etc.)

Lightning Case Presentations (20 minutes)
Short, real-world examples from diverse crisis contexts (e.g., WHO Health-EDRM, Royal Society data privacy tools, Ukraine conflict, Australia's federated systems)

Breakout Group Exercises (35 minutes)
Participants work in small groups to simulate applying the DPTC Checklist to a hypothetical crisis (e.g., multi-hazard urban flooding)
Focus: data governance, stakeholder roles, ethical AI, equity, infrastructure, privacy

Plenary Debrief and Synthesis (20 minutes)
Group insights shared; common challenges and local adaptations identified
Draft "Implementation Roadmap"template introduced for further use

Closing Remarks (5 minutes)
Summary, feedback, and invitation to join the DPTC community of practice

Key features
• Highly interactive and participatory
• Co-creative: builds practical outputs (e.g., a draft roadmap)
• Integrates real examples with hands-on application
• Designed to bridge guidance and action in crisis data policy

Expected outcomes
• Deeper understanding of how to apply UNESCO Open Science and DPTC tools to crises
• A prototype "Implementation Roadmap" for crisis data governance
• Contributions toward a community of practice around crisis-ready open data policy
• Cross-regional collaboration and capacity building

Presentations Session 1: CAREful Indigenous Data Governance / 226

Decolonizing Data Discovery: Metadata Syndication Model for FAIR and CAREful Health Data Governance in Africa

Author: David Amadi[1]

Co-authors: Agnes Kiragga [2]; Bylhah Mugotitsa [2]; Dorothy Mailosi ; Emma Slaymaker [1]; Jay Greenfield [3]; Michael Ochola [4]

[1] *LSHTM*

[2] *African Population And Health Research Center*

[3] *Committee on Data of the International Science Council- CODATA*

[4] *African Population and Health Research Center (APHRC)*

Corresponding Authors: emma.slaymaker@lshtm.ac.uk, dorothymailosi@gmail.com, david.amadi@lshtm.ac.uk, mochola@aphrc.org, akiragga@aphrc.org, jay@codata.org, bmugotitsa@aphrc.org

Fragmented health data systems across Africa perpetuate inequities in crisis response and research participation, echoing colonial legacies of extractive data practices. In a decisive move toward sovereignty, African Ministries of Health and national data custodians are advancing federated approaches that retain local governance while enabling cross-border collaboration through standardized metadata and interoperable models. This shift not only reclaims agency over sensitive health information but also redefines partnerships in global health research centering African leadership in balancing local priorities with transnational scientific goals.

We introduce a federated metadata syndication framework developed with African partners through the INSPIRE Network to advance FAIR (Findable, Accessible, Interoperable, Reusable) and CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) aligned population health data governance. By enabling collaboration at the metadata level, the framework ensures that data remains under local control, where sensitive information stays at home and only FAIR metadata and analytical results are shared across borders. This approach bridges fragmented data systems, fosters equitable access to valuable research insights, and respects local sovereignty demonstrating a scalable model for ethical, inclusive, and decolonized health data systems. Importantly, the framework aligns with the wave of new national data protection laws across Africa, which assert citizens' rights over their personal data and mandate stronger governance in data sharing and research partnerships.

Key components include:

1. Data Documentation Initiative Lifecycle (DDI-Lifecycle) for standardized documentation of variables, study designs, and provenance, ensuring robust data governance aligned with indigenous data principles.
2. Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) to enable standardized, interoperable analytics and machine learning on locally stored data without moving sensitive information.
3. Schema.org annotations to enhance search engine discoverability, facilitating open research and collaboration across interdisciplinary health data ecosystems.
4. Low-code/no-code platforms to empower local institutions to actively participate in federated research workflows, such as secure multi-site machine learning, enhancing responsibility and reproducibility in data science.

The framework's practical application spans mental health research, population-based longitudinal surveys, and national data hubs supporting Sustainable Development Goals (SDGs) and pandemic preparedness.
By embedding ethical and technical design principles in real-world infrastructure, this approach:

- Demonstrates interoperability without centralization;
- Amplifies African agency in data-driven health innovation;
- Fosters inclusive global knowledge systems where African institutions are equal stakeholders, not merely data sources.

By aligning technical frameworks with ethical imperatives and legal developments, this work contributes to decolonizing global health data practices, ensuring African institutions are not just data providers but equal partners in setting research agendas and specifying governance models. Our proposal outlines a pathway for data-intensive research infrastructures that empower African countries to actively shape health research and drive positive change both locally and globally

227

AI without borders? Navigating data sovereignty and human rights in a fragmented world

Authors: Alexander Kriebitz[1]; Caitlin C. Corrigan[1]; Francis P. Crawley[2]; Perihan Elif Ekmekci[None]

[1] *Technical University of Munich (TUM)*

[2] *CODATA IDPC*

Corresponding Authors: fpc@gcpalliance.org, a.kriebitz@tum.de, c.corrigan@tum.de, drpelifek@gmail.com

Introduction

This workshop will examine how the evolving architectures of data policy and AI governance challenge traditional notions of political sovereignty, including state sovereignty, the sovereignty of the people, and the sovereignty of the individual human being, as recognized in the United Nations Charter and the Universal Declaration of Human Rights (UDHR).

The United Nations Charter upholds the "sovereign equality of all its Members"(Article 2.1), affirming the right of states to self-determination and governance without external domination. However, the dominance of transnational tech corporations, extraterritorial data flows, and private AI infrastructures now blur jurisdictional boundaries and pose direct challenges to state control over digital infrastructures and decision-making processes. This has led to a growing demand for digital and data sovereignty—where states assert the right to regulate, localize, and govern data produced within their borders.

The UDHR, in parallel, affirms the inherent dignity, freedom, and equality of all individuals (Articles 1 and 2) and protects individual sovereignty through rights such as privacy (Article 12), freedom of opinion and expression (Article 19), and participation in governance (Article 21). Yet, automated systems that profile, manipulate, or surveil citizens challenge these very freedoms.

The session will draw on practical and policy insights from global processes such as:
• The United Nations Charter
• The Universal Declaration of Human Rights (UDHR)
• The Munich Convention on AI, Data, and Human Rights (2024–25)
• The whitepaper: 'Promoting and Advancing Human Rights in Global AI Ecosystems: The Need for A Comprehensive Framework under International Law'
• The African Commission's Draft Study on Human and Peoples'Rights and AI (2025)
• The UN Global Digital Compact (2024)
• The UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021).

Format
Workshop Format and Detailed Agenda
Interactive Workshop Session –90 minutes
This session is designed to foster deep engagement, practical insight-sharing, and collaborative policy thinking across global, regional, and disciplinary divides. It combines expert framing with structured group discussions and plenary synthesis to address complex questions around sovereignty, AI, and human rights.
The session format follows four core principles:
1. Participatory: All attendees contribute to dialogue, not just listen.
2. Multi-perspective: Structured to reflect legal, ethical, cultural, and geopolitical diversity.
3. Action-oriented: Designed to generate recommendations and pathways for future work.
4. Grounded: Built on real frameworks like the UN Charter, UDHR, Munich Convention, and regional AI strategies.

Workshop Agenda (90 Minutes)
Additional speakers tbc:
• African Commission on Human and Peoples'Rights representative (TBC) –Regional perspectives on collective and peoples'rights in AI governance
• Global South civil society or technologist (TBC) –Challenges of asserting local digital sovereignty in the face of global AI platforms

1. Opening and Scene-Setting Presentations (20 minutes)
   Moderated by session chair
   Purpose: Establish foundational concepts of sovereignty (state, peoples, individual) and introduce global legal and policy frameworks relating to AI and human rights.
   • 5 min —Introduction and workshop goals
   • 5 min —Presentation: "Sovereignty in the Age of AI: From the UN Charter to the Digital Realm"
   • 5 min —Regional insight: African and Latin American perspectives on sovereignty and AI
   • 5 min —Legal perspective: The Munich Convention on AI, Data, and Human Rights
   Speakers: Global experts from law, ethics, and policy communities, including contributors to the Munich Convention.

2. Breakout Group Discussions (30 minutes)
   Participants will divide into four facilitated breakout groups, each tackling a key sovereignty dimension. Each group includes a facilitator and rapporteur and will address 2–3 guiding questions rooted in current international law and AI ethics frameworks.
   Group 1 –State Sovereignty
   Topic: National governance of cross-border AI and data regimes
   Focus: Data localization, extraterritorial digital infrastructure, compliance with human rights
   Group 2 –Sovereignty of Peoples
   Topic: Collective rights, digital democracy, and algorithmic governance
   Focus: Citizen participation, access to information, community data rights
   Group 3 –Individual Sovereignty
   Topic: Autonomy, data protection, and consent in AI systems
   Focus: Privacy (UDHR Article 12), meaningful consent, surveillance ethics
   Group 4 –Bridging Sovereignties
   Topic: Harmonizing local values with global digital governance
   Focus: Cross-cultural rights recognition, UN principles vs. local ethics, legal interoperability
   Participants will discuss:
   • Opportunities to use AI and sovereignty to advance human rights
   • Key risks and legal/policy gaps
   • Practical actions for future governance

3. Plenary Synthesis and Panel Response (30 minutes)
   Each group reports back (5 min per group), followed by discussion.
   • 20 min —Plenary synthesis: Group rapporteurs present key findings
   • 10 min —Panel response: Short reflections from a multi-region expert panel (Africa, Asia-Pacific, Latin America, Europe)
   This segment will explore cross-cutting tensions and possible reconciliations between sovereignty and universal rights in AI governance.

4. Closing and Outcome Framing (10 minutes)
   Moderator summarizes key outcomes and presents a draft outline for a post-workshop document: "Action Note on Sovereignty and Rights in AI Governance"
   • Invitation for ongoing collaboration via CODATA, CoARA-ERIP, or a dedicated follow-up task force
   • Next steps for aligning this work with SciDataCon and IDW open science values
   Participants will be invited to co-sign or contribute to the post-workshop output.

Expected Outcomes
• A deeper understanding of how international human rights frameworks interact with emerging digital sovereignties
• Strategic insights into how sovereignty claims can be balanced in AI governance across state, community, and individual levels
• Contribution to the formation of a global action framework or policy brief on AI, sovereignty, and rights-based digital governance

Presentations Session 6: The Transformative Role of Data in SDGs and Disaster Resilience / 228

FAIRer Hazard Information: principles, implementation and novel uses of the updated UNDRR/ISC Hazard Information Profiles

Author: Virginia Murray[1]

Co-authors: Helene JACOT DES COMBES ; Matti Heikkurinen [2]; Simon Hodson [2]

[1] *UKHSA*

[2] *CODATA*

Corresponding Authors: simon@codata.org, helene.jacotdescombes@council.science, virginia.murray@ukhsa.gov.uk, matti@codata.org

Standardized hazard definitions are a key element of the analysis of disasters. Without them, monitoring and reporting of the impacts of the hazards is difficult, and so is the development of effective early warning systems and response plans. Forecasting of future events and the generation of disaster risks reduction strategies are also hindered by a lack of standardized definition. To address this gap, in 2019 the UN Office for Disaster Risk Reduction (UNDRR) and the International Science Council (ISC) established a Technical Working Group to identify the full scope of hazards relevant to the Sendai Framework for Disaster Risk Reduction as a basis for countries and other actors to review and strengthen risk reduction policies and risk management practices. The resulting UNDRR/ISC Hazard Information Profiles (HIPs) were published in 2021 1.They provide to a broad range of users standardised definition and information on more than 302 hazards organized into 8 groups: meteorological and hydrological, extraterrestrial, environmental, geological, chemical, biological, technological and societal.

Following on from the recommendation in the UNDRR/ISC HIPs for regular review and update, experts from different disciplines, types of organizations (United Nations agencies, academia, government agencies, intergovernmental organizations and the private sector) and geographical regions are again working together to review the UNDRR/ISC HIPs. This process is systematically reviewing all sections of the current HIPs to identify potential updates in alignment with new scientific information. and decide on the inclusion of additional evidence additionally addressing the multi-hazard context of each hazard.
One of the main additions to the updated version of the HIPs is a section on multi-hazard context. The experts are specifically reviewing the interrelations between the hazards in a multi-hazard approach. The HIPs aim to summarize direct interactions between hazards in a concise and visual way.

In the future, the HIPs will be coded to be machine actionable, to support a broader range of applications when machine readability is extremely useful, for example, for analysis of large databases and datasets. This is especially relevant in the context of disaster risk management and of loss and damage associated to climate change.

This second review concludes in 2025, with the release of the enhanced UNDRR/ISC Hazard Information Profiles at the Global Platform for Disaster Risk Reduction. The updated document will continue to inform a broad community and support data analysis resulting in better early warning and event forecast and disaster risk management and planning.
In addition to presenting the HIPs and their background, the session intends to collect ideas for innovative uses of the HIPs and ways to encourage the reuse of the information

REFERENCE
1. Murray, Virginia; Abrahams, Jonathan; Abdallah, Chadi; et al. (2021) Hazard Information Profiles: Supplement to UNDRR-ISC Hazard Definition & Classification Review: Technical Report: Geneva, Switzerland, United Nations Office for Disaster Risk Reduction; Paris, France, International Science Council. DOI: 10.24948/2021.05

Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI / 230

Integrated Reference Architecture for AI-Enabled Healthcare Research: An Australian Harmonized Approach

Author: Gnana Bharathy[1]

Co-author: Adrian Burton [2]

[1] *ARDC/ UTS*

[2] *ARDC/ ANU*

Corresponding Authors: gnana.bharathy@ardc.edu.au, adrian.burton@ardc.edu.au

Integrated Reference Architecture for AI-Enabled Healthcare Research:
An Australian Harmonized Approach

Gnana K Bharathy and Adrian Burton, Australian Research Data Commons (ARDC)
gnana.bharathy@ardc.edu.au
(SciDATACon Abstract ID: 61 suitable for Healthcare Data, Analytics & AI Commons Session)

Introduction

This paper presents a co-design approach to developing a reference architecture for AI and advanced analytics infrastructure supporting healthcare research in Australia.
The Australian Research Data Commons (ARDC), as the national research data infrastructure provider, recognized various transformative opportunities and challenges in employing and providing a research infrastructure for AI and advanced analytics, and initiated a systematic co-design process to develop a framework. This systematic, inclusive co-design process yielded comprehensive insights into research community needs while identifying potential partners.

Approach

ARDC collaborated with the Australian Data Science Network and Australian Cancer Data Network through a comprehensive twin study involving a survey (n=110), multiple workshops (6), environmental scanning, and interviews (6). The ADSN-ARDC "Framework Project" provided broad coverage of key issues, while the ACDN-ARDC "Pathfinder Project" examined sensitive healthcare data challenges where movement is restricted by privacy regulations and organizational policies.

Indico rendering error

```
Could not include image: Cannot read image data. Maybe not an image file?
```

Armstrong et.al. (2024) DOI:. 10.5281/zenodo.13831386 | Holloway et.al. (2024). DOI: 10.5281/zenodo.13831454
Diagram (Figure 1) summarises the systematic and inclusive process ARDC has gone through in arriving at the co-investment projects.

Findings from Co-Design

Researchers identified needs across the advanced analytics lifecycle including secure, scalable infrastructure, interoperable tools with no-code capabilities, and AI-ready data. Many prioritized upskilling resources and guidelines matching their technical proficiency.

Indico rendering error

```
Could not include image: Cannot read image data. Maybe not an image file?
```

The Pathfinder project, relating to sensitive data and federated learning, revealed socio-technical barriers like computational requirements, manual inspection processes in Trusted Research Environments (TREs), standardization gaps, and complex cross-institutional governance.
These findings collectively demonstrated the necessity for an integrated approach addressing both technical infrastructure and socio-technical dimensions of advanced analytics.

Reference Architecture

Drawing on the outputs of the co-design, we developed a Reference Architecture for advanced analytics infrastructure.
This architecture is organized through the lens of the infrastructure, namely underpinning the cloud (Nectar), tools and platforms, data assets, socio-technical resources, and all being made available through virtual research environments (refer to Figure 3: AI & Advanced Analytics Reference Architecture).

Indico rendering error

```
Could not include image: Cannot read image data. Maybe not an image file?
```

The Reference Architecture created by ARDC provides a comprehensive foundation for integrating clouds, tools, platforms, data assets, and socio-technical resources into Virtual Research Environments (VREs).
This architecture explicitly embeds AI capabilities, supporting researchers with GPU-based computing resources, intelligent delivery mechanisms, and contextual training. For example, not only does the architecture provide resources to carry out AI research (compute GPU), access to tools, models, and data through a single platform, but it also provides contextual training and socio-technical resources through intelligent delivery vehicles, such as co-pilots and intelligent LMSs, to support Virtual Labs and direct learning.
In the architecture, we also outline key principles guiding the architecture development, including prioritization of FAIR data practices, modularity, interoperability, and sustainability.

Leveraging architectural thinking, we identified common patterns across virtual research environments to conceptualize a unified Virtual Research Environment (xVRE) framework (Table 1). This framework positions VREs along a complexity-security continuum from Open to Federated systems, accommodating diverse research workflows and governance requirements through modular, incremental development. This would potentially reduce duplication and cost over-runs.

Indico rendering error

```
Could not include image: Cannot read image data. Maybe not an image file?
```

Face Validation

The framework and reference architecture underwent validation through sector engagement via consultation drafts, workshops at national AI Month events, and presentations at forums including eResearch Australia and ADSN conference, demonstrating ARDC's alignment with national and sectoral needs.

Architecture to Program Design

Translating the reference architecture, we have four extensible and interconnected projects, the foundational cloud upgrade, the AI Resource Hub, Federated Machine Learning Network, and AI Virtual Research Environments. These provide a cohesive ecosystem that addresses both broad researcher needs and specialized requirements such as machine learning with sensitive healthcare data.

Discussion and Conclusions

The reference architecture addresses critical gaps identified during co-design by tackling both technical and socio-technical complexities. It integrates skill development through embedded tools and intelligent delivery mechanisms, while streamlining governance with standardized compliance frameworks for sensitive healthcare data.
The architecture resolves five key challenges: (1) reducing technical barriers through accessible interfaces, (2) democratizing specialized computing resources including GPUs, (3) enabling collaborative analysis of distributed datasets via federated learning, (4) simplifying regulatory compliance, and (5) supporting researcher upskilling through integrated training resources.
Our contributions encompass a co-design methodology for architecture development, a holistic reference architecture addressing socio-technical requirements, implementation strategies through interconnected projects, conceptualization and modularization of VREs on a continuum, and an eval-

uation framework for measuring impact.
The AI & Advanced Analytics Reference Architecture represents a strategic approach to healthcare research infrastructure in Australia. By demonstrating the value of co-design methodologies and federated approaches for sensitive data, this work strengthens Australia's capabilities in AI-driven healthcare research while ensuring responsible implementation. As a cornerstone of ARDC's strategic framework, this architecture employs principles of accessibility, interoperability, and sustainability to accelerate responsible AI adoption in healthcare research, ultimately advancing data-driven health outcomes.

Images

Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research / 231

The challenges of data sovereignty and AI in the European Health Data Space (EHDS)

Author: Francis P. Crawley[1]

[1] *CODATA IDPC*

Corresponding Author: fpc@gcpalliance.org

The challenges of data sovereignty and AI
in the European Health Data Space (EHDS)
Talk abstract contribution to
SciDATACon Abstract ID: 61
Advancing Healthcare Research with Data, Analytics & AI Commons
Francis P. Crawley
Chair, International Data Policy Committee, CODATA
francis@codata.org
Version 3.0, 24 April 2025

The European Health Data Space (EHDS) represents a groundbreaking initiative by the European Union to establish a harmonized framework for the sharing, access, and secondary use of health data across member states. This initiative is rooted in the vision of enabling high-quality research, informed policymaking, public health innovation, and improved clinical care by creating a trustworthy, interoperable digital infrastructure for health data. At its core, the EHDS aspires to transform healthcare systems by facilitating secure, ethical, and equitable access to data, fostering collaboration across borders and sectors.
However, realizing this vision presents complex challenges. Chief among them is the need to balance the dual imperatives of openness and innovation with national data sovereignty, individual privacy rights, and ethical governance, particularly in the context of increasing reliance on artificial intelligence (AI) in healthcare. These tensions are further complicated by the uneven implementation of data protection regulations, varied interpretations of ethical standards, and differing cultural attitudes toward data sharing among EU member states.
This session will examine these intersecting issues, focusing on the governance of AI and the operationalization of data sovereignty within the EHDS. Participants will explore best practices for managing patient data responsibly, while enabling cross-border collaboration and digital innovation. Topics include the role of federated data infrastructures, privacy-preserving technologies (e.g., differential privacy and federated learning), and frameworks for trustworthy AI that comply with legal standards and uphold ethical norms. In addition to EU-specific dynamics, the session will draw comparative insights from other national and regional systems engaging with similar challenges.
A key focus will be how the FAIR (Findable, Accessible, Interoperable, Reusable) and CARE (Collective benefit, Authority to control, Responsibility, and Ethics) data principles can be applied to guide the responsible design of data ecosystems. Strategies for building public trust, ensuring participatory governance, and enabling accountability in AI-driven healthcare decision-making will be emphasized.
Ultimately, this session seeks to advance practical and inclusive approaches to health data governance that respect national sovereignty and individual rights, while also supporting the broader

goals of open science, digital health innovation, and equitable healthcare. It aims to contribute to a resilient, future-ready EHDS and inspire similar efforts globally.

Poster Session / 232

Globalizing Space Weather Data Infrastructure: The IMCP Framework for Collaborative Data Sharing and Utilization

Authors: QI XU[1]; XiaoYan HU[2]; Ziming ZOU[2]

[1] 中国科学院国家空间科学中心

[2] *National Space Science Center, Chinese Academy of Sciences*

Corresponding Authors: xuqi@nssc.ac.cn, huxiaoyan@nssc.ac.cn, mzou@nssc.ac.cn

The International Meridian Circle Program (IMCP) represents a pivotal international initiative aimed at advancing coordinated space weather observations to address critical global scientific challenges and enhance operational applications. Effective data governance, sharing, and utilization form the cornerstone for transforming multi-national observational synergies into scientific breakthroughs and downstream services.
To realize this vision, the NSSDC team has developed a comprehensive architecture for IMCP Data Facilities, featuring:
1. Establishment of standardized space weather data protocols, specifications, and Data Management Plans (DMPs) tailored for global mega-projects;
2. Deployment of a dedicated data transmission and exchange network connecting IMCP instruments and worldwide ground-based observatories;
3. Implementation of a hybrid centralized-distributed mirrored database system for space weather data;
4. Development of community-driven open-access services with machine-actionable interfaces to support diverse research and operational needs;
5. Strategic integration of big data analytics and AI technologies to enable next-generation monitoring capabilities and research innovation.
Through this infrastructure, IMCP seeks to establish an open, trusted, and interoperable space weather data ecosystem that significantly enhances the FAIRness (Findable, Accessible, Interoperable, Reusable) and AI-readiness of observational data on a global scale. This concerted effort will provide the foundational data framework required to tackle key space weather science and application challenges worldwide.

233

Building Trust in Data Repositories: Lessons from Global Certification Efforts

Author: Daniela Santos Oliveira[1]

Co-author: Dale Peters [1]

[1] *World Data System*

Corresponding Author: doliveir@utk.edu

The World Data System (WDS) seeks to cultivate and support a global network of members committed to scientific data repositories and effective data stewardship. To keep pace with advancing technologies, growing data volumes and types, shifting user demands, and deeper integration into scientific workflows, WDS member repositories must continuously evolve. Ensuring that data remains accessible alongside scholarly publications and processes is crucial. A key function of WDS is

to collaborate on developing standards and certifications that uphold the reliability and trustworthiness of scientific data repositories while actively tackling the challenges repositories face in adapting to changes.
In November 2023, the WDS Scientific Committee established a subcommittee to address these challenges and evaluate and recommend standards and certifications relevant to WDS member repositories. The subcommittee's work included gathering information on existing certifications like CoreTrustSeal (CTS), nestor, TRAC, and PuRE, as well as conducting a survey to understand the certification landscape across different scientific fields and geographic boundaries.
During the proposed session, Daniela Santos Oliveira and Dale Peters will present the results from the WDS subcommittee's work and discuss recommendations regarding extended certification for lowering entry barriers and fostering a more inclusive expansion of WDS membership.
Moreover, this session will feature representatives from the global community who will share their experiences as members of cohorts of data repositories working together toward CTS certification. Key lessons from their experience will be shared. Among these is Dr. Chiware, Library Director at Cape Peninsula University of Technology, South Africa, who will discuss the importance of forming a Community of Practice (COP) for support through the process of applying for CTS certification, and how the cohort approach made the daunting task of certification more manageable.
Additionally, we are recruiting two or three representatives from other nations who have formed their own CTS support cohorts to give testimony of their experiences. We are currently exploring the possibility of having participation from the CoreTrustSeal Certification Support Cohorts of Canada and Australia, and from a digital data repository in South America. We aim to have testimonies that cover a broad geographical distribution and ample perspectives on the certification process.

Forming COPs, either regional or discipline-specific, is potentially one of the ways WDS can support its members by providing mentorship during the application process. The testimonies will highlight the benefits of community efforts in achieving certification and contribute to what we expect to be an enriching discussion with the data repository community regarding the importance of certification and how WDS can support its members in achieving that milestone.
This session aims to provide valuable insights into the certification process, promote community collaboration, and encourage inclusive participation in WDS membership.

Confirmed speakers:
- Dr. Daniela Santos Oliveira, WDS-IPO Program Manager
- Dr. Dale Peters, Vice-chair, WDS Scientific Committee; Chair, WDS SC Subcommittee on Standards and Certifications.
- Dr. Elisha Chiware, Library Director at Cape Peninsula University of Technology.
- Dr. Richard Ferrers, Data Consultant at Australian Research Data Commons (ARDC)
- Dr. Marcel Garcia de Souza, Coordinator for Processing, Analysis and Dissemination of Scientific Information, Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
- Dr. Lee Wilson, Director of Research Data Management at the Digital Research Alliance of Canada

Session Structure
1. Welcome and Interactive Polling Exercise (10 minutes)
Speaker: Daniela Santos Oliveira
Mentimeter will be used to identify the characteristics of the audience.
2. WDS Subcommittee on Standards and Certifications (15 minutes)
Speakers: Daniela Santos Oliveira & Dale Peters
- Overview of WDS Subcommittee work
- Importance of standards and certifications
- Recommendations
3. Presentations from International Representatives highlighting their experiences/testimonies with certification (40 minutes)
Speakers:
- Elisha Rufaro Chiware (10 minutes):
* South Africa: Experience with COP
- Lee Wilson (10 minutes)
* Collaborative efforts in certification, Digital Research Alliance of Canada's CoreTrustSeal Certification Support Cohort
- Marcel Garcia de Souza (10 minutes)
* Brazil: Regional perspectives on certification
- Richard Ferrers (10 minutes)
* Trusted Data Repositories Community of Practice

4. Q&A (15 minutes)
5. Closing (5 minutes)

Poster Session / 234

Data governance for development: An empirical assessment of open government data management quality and SDG performance

Author: Sang Hyeok Han[1]

Co-authors: Han Sol Lee [1]; Min Jae Park [1]

[1] *Ajou University*

Corresponding Authors: geglove@ajou.ac.kr, lhs15@ajou.ac.kr, hsh8086@ajou.ac.kr

In the current data-driven era, open government data serves as a catalyst for open innovation and the development of value-added services, while also promoting governmental transparency. These attributes collectively contribute to advancing the United Nations' Sustainable Development Goals (SDGs), a set of 17 objectives outlined in the 2030 Agenda for Sustainable Development, which encompass social, environmental, and economic dimensions aimed at achieving sustainable global transformation for future generations.

To ensure that open government data translates into tangible social value, it is essential to effectively manage key data quality dimensions—namely openness and coverage. The Open Data Inventory (ODIN) provides a systematic and quantitative evaluation of these dimensions by assessing the openness and coverage of official statistics and data across countries, covering sectors such as social, economic, and environmental data. ODIN evaluates specific attributes, including indicator coverage and disaggregation, data available last 5 years, administrative level of data, machine readability, metadata availability.

Despite the richness of ODIN data, limited research has quantitatively assessed the extent to which these data management dimensions contribute to the improvement of the SDGs to which the data are most closely aligned. To address this gap, the present study employs regression analysis to investigate the impact of aspects such as openness, coverage of open government data on the SDG outcomes. Specifically, we examine how the quality of data management in ODIN's "Education Facilities" and "Education Outcomes" subsections correlates with progress in SDG 4 (Quality Education), and how indicators related to "Health Facilities," "Health Outcomes," "Reproductive Health," "Population & Vital Statistics," and "Pollution" relate to SDG 3 (Good Health and Well-being).

For each SDG, multiple performance indicators are standardized and averaged to construct composite indices, which serve as the dependent variables. Additionally, we incorporate a two-year lag in ODIN scores to examine the causal relationship between open data quality and subsequent SDG performance.

By conducting this empirical analysis, the study aims to identify which specific dimensions of open data management most significantly influence SDG progress. The findings also contribute theoretically by clarifying how sector-specific data governance practices influence SDG outcomes, and practically by identifying which aspects of open data management should be prioritized to maximize measurable progress on SDG indicators.

236

Pitch Your Research: 3-Minute Scientific Research Pitch Competition

Authors: Claire Rye[1]; Cyrus Walther[2]; Louis Mapatagne[None]; Pragya Chaube[3]

[1] *New Zealand eScience Infrastructure*

[2] *TU Dortmund University*

[3] *UPES/CODATA Connect*

Corresponding Authors: louis.mapatagane@gmail.com, cyrus.walther@tu-dortmund.de, pragya.chaube@ddn.upes.ac.in, claire.rye@nesi.org.nz

As part of International Data Week, we (the 3 Early Career Researcher networks of the hosting organisations, CODATA Connect, WDS ECR Network and RDA Early Career and Engagement IG) are organising a dynamic 3-minute scientific research pitch competition targeted at early career researchers. This event will provide a platform for emerging scientists to showcase their innovative research, enhance their scientific communication skills, and connect with peers and mentors within the global research community.
Participants will have exactly three minutes to present their research idea, methodology, key findings (if applicable), and the potential impact of their work. This time-constrained format will promote clarity, precision, and persuasive communication while offering researchers the opportunity to gain visibility and recognition for their efforts.

| Agenda Item | Time | Details
Welcome Address	10 minutes	Welcome from CODATA, WDS & RDA
Pitch Block 1	15 minutes	3-minute pitches (5 pitches)
Community Voting 1	5 minutes	Voting with mobile devices in small groups
Introduction of CODATA Connect, WDS ECR Network and RDA Early Career and Engagement IG	15 minutes	3 x 5 minute pitches
Pitch Block 2	15 minutes	3-minute pitches (5 pitches)
Community Voting 2	5 minutes	Voting with mobile devices in small groups
Networking and Discussion	30 minutes	Open Networking with Drinks and Snacks while results are developed from the voting
Awarding of Prizes	10 minutes	Awarding of Prizes through representatives of CODATA, WDS & RDA

Detailed Overview:
1. Introduction (10 minutes):
- Brief overview of the event's objectives and format.
- Explanation of the judging criteria: clarity, innovation, societal impact, and communication skills.
- Introduction of the community judging procedure.
2. 3-Minute Pitch Sessions (Two of 15 minutes each):
- Each participant will deliver a 3-minute pitch about their scientific research.
- The pitch should cover the following:
o Research Question: What is the central question or problem being addressed?
o Methodology: How is the research being conducted, and what makes it innovative?
o Findings/Impact: What are the expected or discovered results, and how could they impact the scientific community or society?
During the Networking and Discussion there will be time to approach the pitching scientists for questions and discussions.
3. Judging & Prize Announcement (10 minutes):
- The organizers will deliberate given the voting results and announce the winners based on the selection of the participants.
- Prizes: Certificates and small prizes will be awarded to the top three participants in recognition of their outstanding research presentations.
o 1st Place: Certificate of Excellence for Best Research Pitch.
o 2nd Place: Certificate of Recognition for Innovative Research.
o 3rd Place: Certificate of Achievement for Outstanding Potential.

Objectives:
- To provide early career researchers with an opportunity to present their research ideas concisely and effectively.
- To create a platform for networking, collaboration, and the exchange of ideas among emerging researchers from various disciplines.
- To help researchers enhance their communication skills, particularly in presenting complex scientific concepts to a broad audience.
- To encourage the growth of early career researchers through constructive feedback and recognition.
Target Audience:

- Early career researchers, including PhD students, postdoctoral researchers, and young professionals in the fields of data science, technology, and other related disciplines.
Expected Outcomes:
- Increased visibility and recognition of early career researchers in the global research community and improved CODATA Connect, WDS ECR network and RDA Early Career and Engagement IG participation.
- Improved research communication skills among participants, enabling them to effectively engage with diverse audiences.
- New opportunities for collaboration and networking within the international data science community.
- A valuable and interactive addition to the International Data Week program that fosters professional growth and development.

Presentations Session 8: Policy and Practice of Data in Research; Data, Society, Ethics and Politics / 237

Toward a FAIR Data Policy for Chile: Building a National Ecosystem for Open and Responsible Data

Author: Soledad Quiroz Valenzuela[1]

Co-authors: Rodrigo Roa [1]; Álvaro Paredes Lizama [1]

[1] *Data Observatory*

Corresponding Authors: alvaro.paredes@dataobservatory.net, soledad.quiroz@dataobservatory.net, rodrigo.roa@dataobservatory.n

In an era where data is central to scientific discovery, innovation, and public policy, Chile is taking significant steps toward developing a comprehensive national strategy for FAIR data management. This presentation introduces the Estrategia FAIR, a multi-institutional initiative aimed at embedding the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) into the country's research ecosystem. Developed collaboratively by universities, government agencies, and private sector partners under the leadership of the Data Observatory, the strategy provides a roadmap for data governance that aligns with global standards and regional priorities.

The initiative is rooted in the recognition of Chile's unique position as an open-air laboratory with a diverse landscape of scientific data. However, challenges such as data fragmentation, limited infrastructure, and a lack of common standards hinder the full potential of data reuse and collaboration. The strategy addresses these gaps by proposing the establishment of a national FAIR office, the creation of interoperable repositories, and the implementation of policies that promote responsible data sharing.

Key components of the strategy include: (1) governance and coordination through a National FAIR Network; (2) investment in infrastructure and the adoption of persistent identifiers and metadata standards; (3) capacity building through targeted training for data stewards, researchers, and decision-makers; and (4) regulatory and financial frameworks that support sustainable data practices. Moreover, the strategy emphasizes the importance of cultural change, academic recognition of open science practices, and the integration of FAIR principles into institutional evaluation systems.

Drawing from successful pilot projects and institutional case studies—such as those developed under the InES Ciencia Abierta program—the strategy outlines an actionable plan to enhance data quality, foster interdisciplinary collaboration, and elevate Chile's leadership in regional data governance. This presentation will share insights from the strategy development process, highlight its alignment with international initiatives such as GO FAIR, and propose pathways for broader Latin American collaboration.

Ultimately, the Estrategia FAIR positions Chile as a pioneer in building a national ecosystem for open, ethical, and high-quality data management, offering a replicable model for other countries in the Global South seeking to advance data-driven science and innovation.

Presentations Session 1: CAREful Indigenous Data Governance / 245

The Language Data Commons of Australia: Supporting research for diverse communities

Author: Michael Haugh[1]

Co-authors: Ben Foley [2]; Sam Hames [2]; Simon Musgrave

[1] *The University of Queensland*

[2] *Language Data Commons of Australia*

Corresponding Authors: s.musgrave@uq.edu.au, sam.hames@uq.edu.au, ben.foley@uq.edu.au, michael.haugh@uq.edu.au

Australia is a massively multilingual country, in one of the world's most linguistically diverse regions. Significant collections of this intangible cultural heritage have been amassed, including collections of Aboriginal and Torres Strait Islander languages, Australian Englishes, and regional languages of the Pacific, as well as collections important for cyber-security and for emergency communication. The Language Data Commons of Australia (LDaCA) is integrating this existing work into a national research infrastructure while also securing at-risk collections and improving access to under-utilised collections. LDaCA is thus ensuring that these invaluable resources will be available for analysis and reuse in the future, and that they will be managed in a culturally, ethically and legally appropriate manner guided by FAIR (Wilkinson et al. 2016) and CARE (Carroll et al. 2020) principles. The project aims to make nationally significant language data available for academic and non-academic use while providing a model for ensuring continued access with appropriate community control.

To deliver on the above-mentioned aims, the LDaCA is being developed through five key activity streams:

1. Developing the social and technical foundations for a national, distributed archival repositories ecosystem.
2. Securing vulnerable and nationally significant collections of Aboriginal and Torres Strait Islander languages, Indigenous languages in Australia's Pacific region, (varieties of) Australian English and migrant languages, and sign languages of Australia and its region.
3. Developing a national portal for accessing and repurposing language data of significance to researchers and communities.
4. Establishing an integrated analytics environment for researchers to create fully described, reproducible research on written, spoken, multimodal and signed text in accordance with Open Science principles, and aligned with community expectations for research of practical benefit.
5. Providing training and resources for researchers and communities to support best practice in accessing, analysing and archiving language data in line with FAIR and CARE principles.

LDaCA is based at The University of Queensland (Brisbane, Australia). The project is part of the Australian Research Data Commons (ARDC) HASS and Indigenous Research Data Commons (HASS&I RDC) currently with co-investment from nine institutions and organisations. Given its aims, the establishment of LDaCA over the past four years has been dependent on developing collaborative partnerships across institutional boundaries and closely engaging with a range of different communities. This has involved, in turn, the need to foster new connections and raise awareness and capacity amongst a diverse range of stakeholders. The success of LDaCA depends not only on leveraging existing and previous language infrastructures in Australia in ways that support their respective aims (Musgrave & Haugh 2020), but also in ways that ensure researchers and communities have a real voice in the development of Australia's languages infrastructure. It has become increasingly evident that the success of LDaCA not only involves meeting a diverse range of needs, but also ensuring that researchers and communities can see themselves in that infrastructure.

We illustrate these points with two case studies from opposite ends of the continuum between local and global. Locating, securing and improving access for materials from Indigenous languages is a key activity for LDaCA. In many cases, the group for whom such material is relevant (even crucial) is a small community with close internal connections. Access control for the material may be important to such a community and LDaCA works to implement such control under community guidance, even if this may restrict the possibilities for academic research (Foley et al. 2024). At the opposite end of the scale, LDaCA has begun working with public interest documents, such as Federal Hansard, the record of the Commonwealth parliament. This material is openly accessible, but there is nevertheless a role for LDaCA in making it easier to work with for researchers from a wide range of disciplines, and also in collaborating with the ParlaMint project associated with the CLARIN network in Europe (Erjavec et al. 2024) to make Australian data available and useable for an international research community.

References:

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal 19. 43. https://doi.org/10.5334/dsj-2020-043.
Erjavec, Tomaž, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, et al. 2024. ParlaMint II: advancing comparable parliamentary corpora across Europe. Language Resources and Evaluation. https://doi.org/10.1007/s10579-024-09798-w.
Foley, Ben, Peter Sefton, Simon Musgrave & Moises Sacal Bonequi. 2024. Access control framework for language collections. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024), 113–121. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.10.
Musgrave, Simon & Michael Haugh. 2020. The Australian National Corpus (and beyond). In Louisa Willoughby & Howard Manns (eds.), Australian English Reimagined. Abingdon: Routledge.
Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1). 160018. https://doi.org/10.1038/sdata.2016.18.

Poster Session / 246

Baseline protocols for archiving

Author: Moises Sacal Bonequi[1]

Co-author: Peter Sefton [1]

[1] *Language Data Commons of Australia*

Many current solutions for data management are expensive or require considerable technical underpinnings (or both). The global data community needs to consider simpler approaches in order to include more participants and to improve equity, but this requires guidance about minimal requirements. The Protocols for the Implementation of Archival Repository Services are an attempt to start the process of establishing a baseline for reliable archiving which can guide the development of tools and services which will then be available to a broad range of organisations.
Working with researchers and community groups who have data of interest to our project, the Language Data Commons of Australia (LDaCA), we encounter a variety of scenarios. Some data has made its way into Archival Repositories, places of safe keeping, such as the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), which takes a principled approach to managing resources for the long term, including having simple access controls in place to make sure that access to some materials can be restricted.
However there are two categories of data which are very common where there is no effective governance or technology in place to ensure data is well cared for:

Firstly, a large amount of data is sitting in researchers'offices, garages, or community centres on analogue media and/or at-risk digital media such as shelved hard drives or tapes. A large amount of contemporary research currently being conducted has ad-hoc data storage on individual computers, institutional or individually provided cloud services (Dropbox, One Drive etc). If metadata is present at all it is ad hoc and not linked to data assets.

Secondly, a great deal of data is 'locked in'to applications such as websites or proprietary content management systems, many of which use the term 'archive'in their sales materials but which do not follow archival practices.

These current 'solutions'for data management are sub-optimal in terms of sustainability as they can be expensive financially and require complex technical underpinnings (or both).
The LDaCA team collaborated with PARADISEC staff and our network of partners and stakeholders to produce a set of protocols which are aimed to ensure that data can be managed for the long term, Protocols for Implementing Long-term Archival Repositories Services (PILARS).
The high-level aims of these PILARS protocols are to:
• Maximise autonomy for data custodians/stewards
• Maximise return on investment in data and data infrastructure
• Maximise long-term sustainability for data and for data systems and management

(Source: https://w3id.org/ldac/pilars)
We will present the protocols with examples of how they have been implemented and show the extensive Open Source toolkit that has been created to implement the protocols, not just for language data but for any data that needs to be stored for the long term.

Poster Session / 247

Enabling transparent, open research processes using (not-always-open) RO-Crate data packages

Author: Peter Sefton[1]

[1] *Language Data Commons of Australia*

The RO-Crate (Research Object Crate) specification (Sefton et al. 2023) is a method for describing data sets with rich, interoperable Linked Data metadata. This presentation will show how we, the Language Data Commons of Australia project (LDaCA), use well described RO-Crate data packages (Soiland-Reyes et al. 2022) to enable CARE (Carroll et al. 2020) and FAIR (Wilkinson et al. 2016) compliant research with language data and also touch on some examples from other disciplines.
RO-Crates in the LDaCA environment are self contained packages that describe data resources, collections which aggregate the data at a variety of scales of granularity from a whole collection in one package to individual files in a set of packages, with linked-data metadata fields hasMember and memberOf that establish relationships between packages. The main reasons for the differences in granularity of packages are firstly, practical limitations on size; and secondly, licensing of data, where it is desirable to have a single licence apply to a package. We are dealing with human-created data which may be subject to a variety of participant rights, including copyright, university policies, privacy legislation and Indigenous Cultural and Intellectual Property Rights (ICIP) and these may apply in different ways to different parts of a collection.
The LDaCA team have developed data access systems, which are all available under open source licences. All mediate access to data packages held in data portals backed by Archival Repository systems which enforce licensing requirements; agents requesting data must hold an appropriate licence to use it. Sometimes licensing is automatic, as in the use of a Creative Commons license, but in other cases access may require permission from a researcher or a community.
With the licensing in place, the linked-data nature of RO-Crate allows seamless processing of collections of data, and the creation of virtual collections of data which aggregate distributed packages or reference metadata and data from multiple packages.
Researchers can openly publish code which accesses resources showing a transparent research workflow, while those wishing to re-run the code have to obtain licenses to the data, which may involve applying to a chief investigator on an academic study with an approved Ethics plan from their institution, or being vetted by a community authority.
We will illustrate the benefits of RO-Crate linked data packages with examples that show the possibility of text analytics such as topic modelling (Blei 2012) using a single tool on multiple datasets with vastly different structure and provenance. This interoperability is possible because the RO-Crate Linked Data metadata allows for declarative configuration files to map between different data sets.
One of the main advantages of RO-Crate is that it is discipline-agnostic and is now widely used in a variety of research contexts, mostly science based (e.g. Weiland et al. 2024). We will conclude with a 'tour'of our tools that show how they can be used to create an RO-Crate environment introducing RO-Crate Metadata profiles that describe the method of storing data, data packaging and validation services, show how consistent RO-Crate metadata allows for data discovery by humans and enables machine access via an API and talk about how these are being applied in other disciples than the study of language.
References:
Blei, David M. 2012. Probabilistic topic models. Communications of the ACM 55(4). 77–84. https://doi.org/10.1145/2133806.2133
Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal 19. 43. https://doi.org/10.5334/dsj-2020-043.
Sefton, Peter, Ó Carragáin, Eoghan, Soiland-Reyes, Stian, Corcho, Oscar, Garijo, Daniel, Palma, Raul, Coppens, Frederik, et al. 2023. RO-Crate Metadata Specification 1.1.3. Zenodo. https://doi.org/10.5281/ZENODO.3406497.
Soiland-Reyes, Stian, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, et al. 2022. Packaging research artefacts with RO-Crate. Data Science. IOS Press 5(2). 97–138. https://doi.org/10.3233/DS-210053.

Weiland, Claus, Jonas Grieb, Daniel Bauer, Desalegn Chala, Erik Kusch, Carrie Andrew & Dag Endresen. 2024. Dataspace Integration for Agrobiodiversity Digital Twins with RO-Crate. Biodiversity Information Science and Standards 8. e134479. https://doi.org/10.3897/biss.8.134479.
Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1). 160018. https://doi.org/10.1038/sdata.2016.18.

Poster Session / 250

CAREful Indigenous Data and the National Statement: Early reflections on co-designing an Indigenous community-created, client-centric digital platform

Author: Bernadette Hyland-Wood[1]

Co-author: Jarryd Aleckson [2]

[1] *Queensland University of Technology*

[2] *Aboriginal and Torres Strait Island Community Health Service Brisbane*

Corresponding Authors: b.hylandwood@qut.edu.au, jarryd.aleckson@atsichsbrisbane.org.au

As demand grows to improve health outcomes for Aboriginal and Torres Strait Islander clients, health practitioners and researchers are increasingly embedding Indigenous perspectives through co-design and applying Indigenous data governance frameworks. This paper shares the preliminary reflections from a research program co-led by a Queensland-based Indigenous community-controlled organisation and an interdisciplinary team of public health experts, data scientists and social scientists based in Australia. The program is grounded in a collaborative governance model involving frontline staff–many of whom identify as Aboriginal and Torres Strait Islander—alongside community members, study investigators, and key staff from the Aboriginal and Torres Strait Islander Community Health Service (ATSICHS-Brisbane) and Queensland University of Technology responsible for the project's implementation.

Co-Design Approach

We discuss the early progress of a public-private funded collaborative program of work to co-produce culturally informed tools and approaches to physical health, social and emotional well-being, guided by principles of CAREful Indigenous Data Governance, Indigenous Data Sovereignty and Australia's National Statement. CAREful Indigenous Data Governance is a principle-based framework developed in response to the need for culturally appropriate, respectful, and reciprocal data practices for Indigenous peoples. Research programs aligning with CAREful Indigenous Data Governance affirm the rights of Aboriginal and Torres Strait Islander communities'rights to collect, access, and govern data about themselves, in alignment with principles of Indigenous autonomy and self-determination. This commitment is reflected in the research program's management practices, including establishing community stewardship structures and integrating client voices from the outset. We examine how these principles co-exist alongside Australia's National Statement on Ethical Conduct in Human Research 2023, published by the National Health and Medical Research Council (NHMRC), Australian Research Council (ARC), and Universities Australia, commonly called the National Statement. The National Statement guides researchers on conducting research with high ethical standards that protect research participants' rights, dignity, and welfare.

Reflections and Learnings

We contend that articulating community-defined needs and goals, respecting cultural protocols, and embedding co-creation principles—particularly those that prioritise cultural safety and First Nations values such as storytelling, relationality, and respect—are critical from the inception of a research initiative and must be sustained throughout its development and implementation. By integrating CAREful Indigenous Data Governance principles, we argue that the research program is better equipped to manage the risks and benefits of community-led research. This integration strengthens

the project team's capacity to ensure the program upholds respect for human dignity and adheres to the principles of justice, beneficence, and respect.

Conclusion and Future Directions

The digital tools and protocols resulting from this research program will guide service delivery and support Aboriginal and Torres Strait Islander Peoples so that they can play a direct role in their own health and well-being. When the client-centric digital platform is completed, the efficacy of community-led implementation and reflective practice can be assessed and evaluated.

The contribution of this research program includes a discussion of novel approaches and methods for building an Indigenous community-driven, client-centric digital platform. Our presentation aimed at data science researchers, research software engineers, and research infrastructure specialists, data professionals and early career researchers seeking to expand their knowledge on data governance using culturally appropriate, strengths-based approaches for improving the health and wellbeing of First Nations peoples.

Poster Session / 252

Uncovering the AMR Data Landscape across the Horticulture, Water, and Wine sectors in Australia

Authors: Green Cherry[1]; Noorul Amin[1]; Ricardo Soares Magalhaes[1]; Sahil Arora[1]; Tatiana Proboste Ibertti[1]

[1] *The University of Queensland*

Corresponding Authors: cherry.green@uq.edu.au, r.magalhaes@uq.edu.au, sahil.arora@uq.edu.au, noorul.amin@uq.edu.au, t.probosteibertti@uq.edu.au

Background: Antimicrobial resistance (AMR) is a growing concern in agribusiness sectors with serious consequences to productivity and public health. A data-centric approach is needed to support Australian agribusinesses and water sectors to understand the impact of antimicrobial usage on the emergence of resistance for diseases that farmers are faced on a daily basis. The SAAFE CRC Analytics Program has partnered with the Australian Research Data Commons (ARDC) and has undertaken a comprehensive study to assess the current AMR data landscape and needs in the Australian horticulture, water and wine sectors.

Methods: To map out the existing data generating processes, sources, data flow and its parameters, a structured interview instrument was developed and validated before deployment. Data types considered in the interviews included microorganism data, antimicrobial usage information and residues of antimicrobials, which were broadly categorised into three distinct categories: a) Compliance data, b) Operational data, and c) Research data. Furthermore, the interviews and workshops covered key challenges faced in managing and integrating the AMR data in these sectors. The interview instrument was deployed in 45-minute interviews of key domain experts from each sector (Horticulture N=4; Wine Sector N=2; Water Sector N=4). The information retrieved from interviews was then validated during 90-minute sector-specific workshops involving a broader group of key domain experts of each sector. The final data landscape information was summarised into directed acyclic graphs (DAG) depicting the interconnection of sector intervenient across all data types.

Results: The information that has been elicited during the interviews indicated that compliance data is tightly regulated, with a small set of parameters as part of the data flow between the data generator and regulatory authorities. Our results indicated that operational data collection is extensive, privacy-sensitive, and used for process monitoring and optimization. Research data is more ad-hoc and not always shared back with sector utilities. The data landscape DAG for each sector shows important data interactions between the horticulture, water and wine sectors that include the type of data, its flow, and the challenges and opportunities in these sectors.

Conclusion: This study will lay the foundation for digital adaptation in the aforementioned sectors. This process has resulted in invaluable insights and recommendations that will lay the foundation of future AMR data infrastructure in the Australian horticulture, water, and wine sectors.

Poster Session / 253

Metavaluation: A Participatory Framework for Valuing and Incentivising Diverse Research Contributions

Author: Cooper Smout[1]

[1] *Institute for Globally Distributed Open Research and Education; Open Heart Mind*

Corresponding Author: cooper.smout@gmail.com

Systemic reform in science continues to face a collective action problem: researchers agree that contributions such as data sharing, peer review, software development, and community engagement are essential, yet these remain structurally undervalued in current incentive systems. Although the Open Science movement has promoted greater transparency and expanded recognition, uptake of alternative evaluation tools remains low. Most systems rely on unpaid labor and lack meaningful incentives, creating a paradox where improved evaluation tools fail to attract enough evaluators to function effectively.

This talk introduces Metavaluation—an open-source, participatory evaluation protocol designed to overcome this barrier by embedding incentives directly into the evaluation process. Rather than treating peer review and contribution recognition as separate, unlinked processes, Metavaluation treats evaluations themselves as contributions, subject to the same collective valuation process. By feeding pairwise evaluations into a recursive feedback loop—where evaluations are used to value other evaluations—the system generates standardized, reproducible value metrics across multiple dimensions (e.g. Gratitude, Value to the Community and Mission). These metrics support rigorous, transparent, and inclusive science by informing recognition, funding, governance, and strategic planning.

Unlike conventional reputation systems, Metavaluation is designed from first principles to be open, inclusive, interoperable, and resistant to gaming. It avoids absolute scoring and popularity biases by relying on pairwise comparisons—a simple, accessible format proven to generate high-quality judgments with minimal cognitive load. It also controls for exposure and attention biases through random sampling. The key innovation is treating these comparisons as a "base unit"of community value, then embedding them within the same evaluative process. This allows all other contributions to be scaled in relative terms, making every value score comparable to the value of a single pairwise comparison. The result is a participatory, decentralized incentive mechanism that mitigates gatekeeping while encouraging broad engagement.

Since its inception, the framework has been prototyped and validated across multiple domains—academic conferences, community arts festivals, and decentralized projects—demonstrating its flexibility in valuing diverse contributions. These real-world experiments have shown how communities can use Metavaluation to make the invisible visible, recognize essential work, and begin aligning incentives with shared values. Its use of community-defined metrics and reward structures offers promising applications for open infrastructure, responsible AI, and sustainable funding models.

This presentation will:
• Share the conceptual foundations and system design of Metavaluation.
• Present results from recent prototypes in academic and grassroots settings.
• Reflect on learnings related to participation, data quality, and interoperability.
• Explore implications for open research infrastructures, value-aligned funding, and future data governance systems.

The session is especially relevant to attendees interested in reproducible science, decentralized governance, data stewardship, equity and inclusion, and cross-community infrastructure. It will complement a related workshop that introduces the hands-on use of the Metavaluation app for valuing contributions to the SciDataCon conference itself.

By centering reward and recognition on what communities truly value—rather than what is easiest to count—Metavaluation offers a scalable, interoperable, and fair approach to evaluation. It enables research communities to coordinate through shared metrics, reward engagement, and align

practices with values—laying the groundwork for a more inclusive, resilient, and participatory data ecosystem.

Presentations Session 9: Empowering the global data community for impact, equity, and inclusion / Education / 254

Metavaluation in Practice: A Workshop on Valuing Diverse Research Contributions

Author: Cooper Smout[1]

[1] *Institute for Globally Distributed Open Research and Education; Open Heart Mind*

Corresponding Author: cooper.smout@gmail.com

Introduction

The scientific community faces a critical challenge: essential contributions to research progress—including data sharing, code development, peer review, mentorship, and community engagement—remain systematically undervalued. Despite the Open Science movement's efforts to broaden recognition beyond traditional publications, engagement with alternative evaluation systems remains persistently low. Researchers continue to prioritize activities that advance their careers within traditional reward structures, creating a paradox in which systems designed to improve evaluation cannot attract enough evaluators to function effectively.

This workshop introduces **Metavaluation**, a novel participatory framework that directly addresses this engagement challenge by treating **evaluations themselves** as valuable contributions. Through an innovative self-referential system—where evaluations are also evaluated—participants are incentivized to provide high-quality assessments, creating a virtuous cycle of participation. Unlike existing systems that struggle with low uptake, Metavaluation includes **direct rewards for evaluators**, making the process engaging, equitable, and sustainable.

Significance and Relevance to SciDataCon Themes

Metavaluation's six-stage protocol aligns directly with several of SciDataCon 2025's key themes:

- **Record** –*Open Research Through Interoperable Data.* Contributions and evaluation dimensions are transparently recorded, forming an open registry that enables interoperable value metrics across research communities.
- **Review** –*Empowering Global Participation.* Evaluations are conducted through simple pairwise comparisons, allowing anyone—regardless of expertise level—to meaningfully contribute. This reduces gatekeeping and fosters inclusive global participation.
- **Recognize** –*Supporting Rigorous, Responsible Science.* A meta-evaluation process generates standardized, reproducible metrics for diverse contribution types, improving transparency and scientific integrity.
- **Reward** –*Aligning Incentives Toward Openness.* Tokenized or symbolic rewards for both contributions and evaluations incentivize sustained participation, aligning individual and collective goals.
- **Respect** –*Building Responsible AI and Governance.* The system tracks reliability over time, enabling merit-based governance models and providing value-aligned training data for future AI tools.
- **Research** –*Advancing Reproducible Meta-Research.* Metavaluation supports continuous learning by documenting and analyzing the valuation process itself, contributing to open, reproducible meta-science.

Workshop Format and Structure

This 90-minute workshop will combine a concise overview of the Metavaluation model with a live, hands-on session where participants apply it in real time to value contributions to SciDataCon 2025. The session will be structured as follows:

1. Introduction and Framing (15 min)

Overview of the collective action problem in research evaluation and core blockers to cultural change. Introduction to Metavaluation's six-stage protocol and its role as a participatory incentive layer for research communities. Overview of prototype data collected from diverse communities spanning the arts, sciences, and technology sectors, along with key learnings from three years of prototyping radical participatory governance models.

2. Live Demonstration (10 min)

Walkthrough of the newly released, open-source Metavaluation app. Demonstration of how to nominate contributions, conduct evaluations, and view output metrics.

3. Interactive Collective Valuation (40 min)

Participants will:
- Nominate real contributions from the conference (e.g. talks, organizing, registration fees).
- Evaluate them via pairwise comparisons across multiple dimensions (e.g. Gratitude, Value to the Community and Mission).
- Review system-generated metrics showing relative value rankings.

4. Discussion and Synthesis (20 min)

Group discussion on:
- What values emerged from the experiment
- Participant experience of the evaluation process
- Use cases in other communities (e.g. data stewardship, governance)
- Feedback and ideas for broader adoption

5. Wrap-Up and Invitation (5 min)

Invitation to continue using the system beyond the workshop. Encouragement to co-develop value taxonomies, nominate more contributions, and contribute to a growing commons of interoperable, community-generated data.

No technical background is required to attend this workshop —just your perspective, values, and voice.

256

Bridging Data Gaps with Citizen Science for People and Policy

Authors: Amanda Vilchez[1]; Carolynne Hultquist[2]; Dilek Fraisl[3]; Elaine Faustman[4]; Karen Soacha[5]; Kehinde Baruwa[6]; Maryam Rabbie[7]; Oluwatimilehin Shonowo[8]; Peter Elias[6]; Yaqian Wu[9]

[1] *Cornell University*

[2] *University of Canterbury*

[3] *Citizen Science Global Partnership*

[4] *University of Washington*

[5] *EMBIMOS Research Group (ICM-CSIC) CitSci - Environmental and Sustainability Participatory Information Systems*

[6] *University of Lagos*

[7] *Sustainable Development Solutions Network*

[8] *University of Glasgow*

[9] *University College London*

Corresponding Authors: kbaruwa@unilag.edu.ng, soacha@icm.csic.es, o.shonowo.1@research.gla.ac.uk, av442@cornell.edu, maryam.rabiee@unsdsn.org, yaqian.wu.18@ucl.ac.uk, pelias@unilag.edu.ng, carolynne.hultquist@canterbury.ac.nz, fraisl@iiasa.ac.at, faustman@uw.edu

**

Bridging Data Gaps with Citizen Science for People and Policy

**

There abound evidences to demonstrate how citizen science is making efforts to bridge data gaps for people and policy. In the last few years, the CODATA-WDS Task Group (TG) on Citizen Science for the Sustainable Development Goals has prioritized global report of citizen-generated data use and connection to SDG indicators. This entails documenting the characteristics, quality, ethics, and sustainment of citizen-generated data in Africa, Asia, Oceania and Latin America in the official monitoring of the SDGs. The relevance of this panel session is the recognition of the importance of representation in reporting, especially on marginalized and vulnerable populations. Some countries are leading efforts to prioritize inclusive community participation in monitoring through intentional engagement and subsequent civic outcomes with action to support progress. This session shall demonstrate the use of citizen-generated data through case studies of underrepresented groups (e.g., slum, refugee, extreme poverty, Pacific island communities) in relation to global challenges (e.g. health, flooding, biodiversity monitoring, etc.) for different cases and regional contexts.

Our proposed session fits into Theme 4: Empowering the global data community for impact, equity, and inclusion.

The panel session shall consist of high quality invited papers and case studies in line with the objectives of the CODATA_TG as follows:

Session Introduction: Peter Elias (Co-Chair, University of Lagos & CODATA Task Group on Citizen Science for the SDGs)

Part One: Citizen Generated Data, SDG Indicators and Progress on SDGs

Lead Paper: Challenges and Opportunities of incorporating Citizen Science in SDGs by Carolynne Hultquist Co-Chair CODATA Task Group on Citizen Science for the SDGs & University of New Zealand

Copenhagen Framework and Citizen Science for Policy –Emerging Opportunities. Presenter - Haoyi Chen, United Nations Statistical Division

Sustainable Development Solutions Network –Citizen science/citizen-generated data towards inclusive impact at local and global level. Presenter - Maryam Rabbie, SDG Today, New York, USA

Part Two: Regional landscapes of citizen-generated data priorities

How Citizen Science is Shaping Progress for SDGs - Example of SDG 14.1.1 (Marine Litter in Ghana). Presenter - Dilek Fraisl, Executive Director, Citizen Science Global Partnership

Ensuring Citizen Science and Political Voice in Issues of Marine Equity: Sharing Lessons from the Nippon Foundation Fellowships –Presenter - Yoshi Ota (Designated Fellow) & E.M. Faustman

Report from Africa - Kehinde Baruwa & Peter Elias (Lagos Urban Studies Group, University of Lagos

Report from Oceania –Carolynne Hultquist

Report from Latin America –Amanda Mayte Vilchez (Cornell University, USA) & Karen Soacha (EMBIMOS Research Group (ICM-CSIC)
CitSci - Environmental and Sustainability Participatory Information System)

Report from Asia - Yaqian Wu (University College, London, UK)

Discussion and Next Steps

Wrap up

Poster Session / 258

An Evolving Approach to Supporting Indigenous Data Sovereignty in an Institutional Data Repository

Authors: Alicia Zuniga[1]; Shanda Hunt[2]; Wanda Marsolek[2]; Kent Gerber[2]; Shannon Farrell[3]

[1] *California State University, Sacramento*

[2] *University of Minnesota*

[3] *University of Minnesota Libraries*

Corresponding Authors: sfarrell@umn.edu, hunt0081@umn.edu, mars0215@umn.edu, gerbe240@umn.edu, alicia.zuniga@csus.edu

Our university is built on the traditional and contemporary homelands of the Dakota people, a federally recognized Tribal Nation made up of four communities and their sovereign governments. We recognize the importance of acknowledging the People on whose land we live, learn, and work, but understand that words are not enough. Within our institutional data repository, we seek to improve and strengthen our relations with Minnesota Native Tribal Nations (Dakota and Anishinaabe) by supporting and advocating for Indigenous data sovereignty (IDSov), which affirms the rights of Indigenous Peoples, communities, and Nations to govern the collection, ownership, and application of data pertaining to them, their lands, and their non-human relations.

We began thinking about Indigenous data in earnest in 2021, when we received a dataset about animal locations and recognized shortly thereafter that the data collection occurred on Tribal lands. This marked the first time we requested documentation of Tribal approval for data sharing.

In the years since, we have been actively engaged in learning more about steps we could take to better manage Indigenous datasets and preserve Indigenous data governance. We implemented new language in our data repository's automated email, encouraging submitters to clarify Tribal data ownership and consent if applicable.

Next, we engaged in conversations with the US Indigenous Data Sovereignty Network where we received generous guidance from 16 collaborative members, and in April 2024 we attended the US Indigenous Data Sovereignty and Governance Summit. These experiences inspired a systematic assessment of Indigenous data within our repository as a first step. We utilized Data Services Continuing Professional Education to find a partner who could help to launch this effort. Her formidable efforts set us up to take immediate, effective action - she developed a list of Indigenous data search terms, created a spreadsheet to document datasets existing in our repository, curated relevant policies and resources, and drafted recommendations for us as we proceed.

We are currently focused on relationship development within our university which has resulted in the Native American Affairs Office offering to review and provide feedback on the list of Indigenous search terms. The list - now finalized and tailored to our repository - will be utilized to audit our data repository. We will document data that is potentially governed by Indigenous Peoples and collect information about that data, such as whether or not there is evidence of Indigenous consent to share the data in a public repository. This is just the beginning as we advocate for the ethical and responsible stewardship of Indigenous data, aligned with IDSov and the CARE Principles.

This poster will describe the background work that we did to better understand Indigenous data governance, repository audit details, early outcomes, and possible next steps.

Poster Session / 260

Federated Mental Health Data Analysis Using Standard Tools in OMOP CDM-Based Ecosystems

Author: Michael Ochola[1]

Co-authors: Agnes Kiragga [1]; Steve Cygu [1]

[1] *African Population and Health Research Center (APHRC)*

Corresponding Authors: akiragga@aphrc.org, mochola@aphrc.org, scygu@aphrc.org

Background
This study investigated mental health research data analysis across institutions where privacy of data is key, regulatory restrictions, and variation in how data is structured and stored. These limitations are especially pronounced in resource-constrained settings and federated data analysis offers a promising solution. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) served as a robust standard for harmonizing diverse mental health datasets, while DataSHIELD enables federated analytics across distributed environments. We aimed to demonstrate the use of standardized mental health tools mapped to OMOP CDM for federated analysis of mental health indicators.
Methods
Mental health datasets from distributed sources were mapped to the OMOP CDM, The study utilized standardized mental health tools such as Patient Health Questionnaire (PHQ-9), a screening tool for depression and Generalized Anxiety Disorder (GAD-7), a screening tool for anxiety. These were mapped into LOINC and SNOMED codes within the OMOP vocabulary. The harmonization process followed a typical extract, transform, and load (ETL) pipeline, with PostgreSQL serving as the database backend. Core tables mapped included person, observation, condition occurrence, and measurement standard OMOP tables. DataSHIELD was used to deploy federated analysis routines across harmonized OMOP CDM databases without sharing individual-level records
Results
Standardized tables were successfully generated across participating datasets, enabling federated queries on variables such as anxiety and depression severity, and demographic trends. DataSHIELD enabled seamless execution of statistical routines across all sites, returning only aggregate, non-disclosive outputs. This setup demonstrated that meaningful mental health analytics —such as trend comparisons and stratified summaries —can be conducted across institutions without compromising privacy. The solution adhered to FAIR principles and promoted collaborative mental health research across geographically distributed institutions.
Conclusion
The application of OMOP CDM and DataSHIELD has enabled privacy-preserving, federated analysis of mental health indicators using standardized tools. This approach demonstrates the feasibility of expanding traditionally siloed mental health data into collaborative analytic environments.

Poster Session / 261

Improving Australia's Food Security: Lessons learned from the ARDC's Food Security Data Challenges program

Authors: Sheida Hadavi[1]; Stefanie Kethers[1]

[1] *ARDC*

Corresponding Authors: stefanie.kethers@ardc.edu.au, sheida.hadavi@ardc.edu.au

Australia is one of the most food secure countries in the world. However, long-term strategies are needed to ensure Australia has a resilient and sustainable food industry that maintains its ability and reputation for delivering high-quality food nationally and internationally.

The Australian Research Data Commons (ARDC) established its Food Security Data Challenges program to support the Sustainable Development Goals (SDGs) identified by the United Nations by creating innovative digital infrastructure solutions to enable, support and improve research into Australia's production, consumption and distribution of safe and high-quality food. As part of Australia' s National Collaborative Research Infrastructure Strategy (NCRIS), the program has been developing national scale data and digital infrastructure capability aligned with national priority research areas and the UN Sustainable Development Goals through its portfolio of 10 mission driven projects (https://ardc.edu.au/multiproject/food-security-data-challenges-projects/). The projects aimed at providing data solutions in areas including agriculture, aquaculture, antimicrobial resistance, traceability and food provenance, biosecurity, nutrition, food equity, and food relief. Nine out of the 10 projects ran from March 2023 until June 2025, while the tenth project is running from May 2024 until June 2026.

Following on from our presentation at SciDataCon 2022, where we provided an overview and initial lessons learned at the early stages of the ARDC's Food Security Data Challenges program, this poster will report on the program's key outcomes and lessons learned during the program. Examples of our learnings included the definition and evaluation of our co-design process, opportunities and challenges we encountered when establishing and working on projects with different types of partner organisations (e.g. universities, NGOs, government agencies, etc), and how specific activities that we set up to create synergies between the projects were received by the projects.

We expect that our experiences will be useful to a variety of key parties in the data and research infrastructure ecosystem, including project managers, program designers and organisations investing in data projects.

Poster Session / **262**

RADAR - a flexible FAIR research data repository

Author: Felix Bach[1]

Co-authors: Christian Bonatto Minella [1]; Kerstin Soltau [1]; Sandra Göller [1]

[1] *FIZ Karlsruhe*

Corresponding Authors: felix.bach@fiz-karlsruhe.de, sandra.goeller@fiz-karlsruhe.de, christian.bonatto-minella@fiz-karlsruhe.de, kerstin.soltau@fiz-karlsruhe.de

RADAR, developed and operated by FIZ Karlsruhe –Leibniz Institute for Information Infrastructure, provides a robust and versatile research data repository designed to facilitate adherence to FAIR principles—ensuring data is Findable, Accessible, Interoperable, and Reusable. Since its inception, RADAR has evolved significantly, offering enhanced services tailored to diverse research communities.

The FAIR principles are deeply integrated into RADAR's infrastructure through comprehensive metadata schemas, persistent identifiers (DOIs), and standardised licensing, ensuring that research data remain discoverable and accessible. RADAR supports interoperability by adopting widely accepted metadata standards such as DataCite and Dublin Core, alongside integration with discipline-specific terminologies facilitated by the recent implementation of TS4NFDI (Terminology Service for NFDI). This inclusion significantly enriches metadata quality and consistency, facilitating semantic interoperability across various research domains.

A central strength of RADAR is its flexibility, offering tailored solutions through deployment models: RADAR Cloud as a centrally hosted, turnkey solution, and RADAR Local for institutions preferring to maintain full control over their infrastructure. These deployment models accommodate a broad spectrum of institutional requirements and technical preferences.

Recently introduced functionalities further enhance RADAR's usability and integration capabilities. Notably, researchers can now seamlessly import data via GitHub and WebDAV, streamlining data

publication workflows directly from commonly used platforms. RADAR also provides a comprehensive RESTful API that facilitates automated interactions, enabling efficient data submission, metadata management, and integration into existing research workflows.

In addition, RADAR has implemented FAIR Signposting, a machine-actionable mechanism for signaling FAIR-aligned metadata, thereby improving automated discovery and reuse of research data. This advancement underscores RADAR's commitment to supporting emerging best practices in FAIR data management.

Specialized RADAR variants have been developed under the umbrella of Germany's National Research Data Infrastructure (NFDI) initiative to meet community-specific needs:

RADAR4Chem, in collaboration with NFDI4Chem, supports chemical research data, including specialized features like DOI assignments, embargo periods, and peer-review integration.

RADAR4Culture addresses data publication and preservation needs within cultural heritage.

RADAR4Memory supports the historical humanities, offering tailored metadata standards and community-specific functionalities for managing and preserving historical research data.

Through continuous development informed by community feedback, RADAR remains at the forefront of data management infrastructures. Its commitment to FAIR principles, combined with innovative integrations such as TS4NFDI terminologies, GitHub and WebDAV import options, robust API access, and FAIR Signposting, positions RADAR as a leading repository, significantly advancing open science and interdisciplinary data reuse.

Poster Session / **267**

Title: Enabling Trustworthy and FAIR AI for Transboundary Aquifer Resilience: Challenges and Opportunities for Reproducible, Responsible, and Open Science

Authors: Ilya Zaslavsky[1]; Ashley Atkins[1]; Christine Kirkpatrick[2]

[1] *San Diego Supercomputer Center, UCSD*

[2] *San Diego Supercomputer Center / CODATA*

Corresponding Authors: asatkins@ucsd.edu, izaslavsky@ucsd.edu, christine@sdsc.edu

Artificial intelligence (AI) offers powerful potential to address pressing challenges in transboundary water management, especially in regions with insufficient infrastructure for in-situ water quantity and quality monitoring and modeling. However, the successful application of AI in this context depends on more than algorithmic accuracy and can be challenging to achieve even in a system with robust data that follows best practices to ensure AI readiness. In transboundary regions where groundwater resources are shared by more than one country, data collection and standards can vary dramatically across borders. Challenges associated with ensuring reproducibility and interoperability are exacerbated for transboundary systems due to this reality. These issues intersect directly with broader themes of responsible science in the era of FAIR (Findable, Accessible, Interoperable, and Reusable) data and AI.

The Groundwater Resilience Assessment through iNtegrated Data Exploration for Ukraine (GRANDE-U) project exemplifies both the promise and complexity of such transboundary collaborations. Focused on aquifer resilience in Ukraine and neighboring countries, the GRANDE-U project brings together researchers from the U.S., Ukraine, Poland, Latvia, Lithuania, and Estonia, with support from the U.S. National Science Foundation and parallel research agencies in the European partner countries. It combines physics-based and machine learning models with satellite data to monitor and predict groundwater storage and flows across national borders. This integrated approach hinges on the willingness and ability of each country to share data openly, follow interoperable data standards and database conventions, and conduct joint data analysis and AI modeling, leveraging the complementary expertise of the country teams.

The FAIR data principles provide a crucial foundation for such collaborations. Data collection protocols and hydrogeologic descriptions vary across GRANDE-U's partner countries, as does the density of in-situ observations. Thus, the GRANDE-U project has prioritized the development of a consolidated spatio-temporal database of satellite and in-situ groundwater and surface water observations. The database is structured to support consistent and interoperable data standards across the transboundary regions to create an AI-ready foundation.

The database schema was iteratively refined with feedback from all partner countries, ensuring alignment with both scientific objectives and policy priorities. This process included ensuring clarity across the international research team on data provenance, modeling assumptions, and validation methods. Reproducibility was prioritized across every stage and component of the research. All machine learning workflows were shared as executable Jupyter notebooks, detailing training data exploration, feature engineering, hyperparameter tuning, and spatial transferability. These transparent modeling practices were reinforced through technical webinars.

This collaboratively developed database serves as the foundation for two critical components: downscaling satellite-based groundwater estimates using local hydrogeologic context, and extending AI models to broader transboundary regions. A major milestone was the development and validation of novel algorithms to downscale GRACE/GRACE-FO satellite data, with successful application along the Poland-Ukraine border. These methods leveraged high-resolution geologic, topographic, and land cover datasets, and in-situ groundwater monitoring wells, to achieve accurate modeling using a range of machine learning models, including random forest regressor and boosting techniques (Pearson R > 0.8 in porous aquifers). The results underscore the potential of machine learning to fill observational gaps and enable groundwater modeling in regions with sparse field data —capabilities particularly crucial for conflict-affected areas such as Ukraine.

Analyzing collaboration networks in transboundary groundwater research offers valuable insights for strengthening international partnerships and designing more effective workforce development strategies. By identifying key contributors, interdisciplinary linkages, and institutional gaps, such network analysis can guide the organization of complementary expertise across hydrogeology, remote sensing, and AI, and foster scientific ecosystems that leverage unique expertise to move innovation forward in ways that are not possible otherwise. In GRANDE-U, this approach has been used to study the evolving structure of global groundwater research networks, leveraging bibliographic data and interactive visual analytics. These findings have directly informed the project's training and capacity-building efforts, including webinars and workshops for students and early-career researchers focused on reproducible AI workflows, FAIR data practices, and spatial modeling. Participants from Ukraine, the U.S., and multiple European countries have engaged in these sessions, helping build an internationally connected, interdisciplinary community. By making training materials and data pipelines openly available, GRANDE-U advances a culture of transparency, collaboration, and responsible data stewardship across borders.
While AI offers transformative opportunities for sustainable groundwater management across borders, its benefits can only be realized within a purposeful framework that emphasizes rigorous, responsible, and reproducible science. The GRANDE-U experience demonstrates how such a framework can succeed by prioritizing FAIR data and open research practices, and by actively engaging diverse transdisciplinary teams in collaborative modeling and knowledge-sharing across political and institutional boundaries. GRANDE-U leverages state-of-the-art practices catalogued by the FAIR in ML, AI Readiness, AI Reproducibility (FARR) Research Coordination Network, especially for maintaining AI readiness in its data products.

GRANDE-U support under the NSF IMPRESS-U program (awards 2409395 and 2409396) is gratefully acknowledged.

Poster Session / 269

Leveraging Corpus Linguistics for Linguistic Research in Kazakh: A Data-Driven Approach

Author: assel ormanova[1]

[1] *Astana IT University*

Corresponding Author: asel_86@mail.ru

The field of corpus linguistics has revolutionised linguistic research by providing data-driven insights into the structure, usage, and evolution of languages. By leveraging large-scale text corpora, researchers can uncover patterns in grammar, vocabulary, syntax, and language use that are not easily observable through traditional methods (Omarova et al., 2025). This data-driven approach offers a powerful tool for both theoretical and applied linguistic studies, particularly for languages such as Kazakh, which face challenges in terms of linguistic resources and computational tools. In our study, we explore the application of corpus linguistics to the Kazakh language, focusing on the creation, analysis, and implications of Kazakh language corpora for linguistic research. More specifically, our research lies in investigating the language contact called interference (Ormanova & Anafinova, 2022).

We present the methodologies employed in building Kazakh language corpora, including the selection of texts, the annotation process, and the use of computational tools for text analysis. We discuss the specific challenges encountered in working with Kazakh, including issues related to its agglutinative nature, complex morphology, and the lack of comprehensive, digitised language resources. A significant portion of the presentation focuses on how corpus linguistics has been utilized to investigate the borrowings from English and Russian due to the policy of trilingualism in Kazakhstan.

Furthermore, we will discuss the potential applications of Kazakh language corpora beyond academic research. These corpora hold significant promise for practical uses, such as in language education, machine translation, and speech recognition technologies. Our study outlines how corpus-based insights can be used to inform language teaching materials, contribute to the development of language resources for artificial intelligence, and support language preservation efforts for Kazakh, particularly in light of the ongoing sociolinguistic shifts within Kazakhstan.

Materials and methods. We generated a corpus of media texts in the Kazakh language (700 texts, 374087 words); we carried out a comparative analysis of statistical linguistic data (word occurrences) by using the computer program #LancsBox 6.0.

Results: A corpus analysis showed that borrowings from English and Russian are actively used in Kazakh (e.g., guide, speaker, PR, draft, team building, etc.). Along with widespread use, most borrowings are not included either in dictionaries of the Kazakh language or the official terminological base Termincom.kz.

There is a tendency in the use of borrowings when both borrowings and equivalent national variants are used in texts at the same time. The difference is observed in the number of occurrences. On the one hand, Kazakh words are dominant. For example, the Kazakh мәселе [masele] / problem has 696 occurrences in the corpus. However, along with this, there is a Russian translation equivalent проблема [problema] / problem with 109 occurrences in Kazakh texts. At the same time, there are cases where the dominance of a borrowed word is observed along with the already existing and approved Kazakh version. So, in the generated corpus, the borrowed word музей [muzei] / museum has 20 occurrences in Kazakh texts, while the Kazakh мұражай [murazhay] / museum has only 14 occurrences, which indicates the prevailing norms of a foreign language. Even though the difference is not significant, the fact of using the word indicates interference from the Russian language to the Kazakh vocabulary.

Thus, the implementation of the tools of corpus linguistics can enhance the research in linguistic data. A data-driven approach could highlight how the corpus data has been used to study the Kazakh language, providing insights into syntax, semantics, language variation, or language change. Through the creation and analysis of extensive corpora, linguists and researchers are better equipped to understand and document the linguistic intricacies, contributing to the broader field of corpus linguistics and the preservation of linguistic diversity in the digital age.

References:

Ormanova A. B. & Anafinova M.L. (2022) Linguistic Interference in Information Space Terms: A Corpus-Based Study in Kazakh. Theory and Practice in Language Studies, 12 (12), 2497-2507. DOI: https://doi.org/10.17507/tpls.1212.04 https://tpls.academypublication.com/index.php/tpls/article/view/5095

Omarova, S., Ospanova, D., Aitova, N., Tokenkyzy, G., Ormanova, A., & Alshynbekova, M. (2025) A Corpus Approach in Language Discovery: A Word Frequency Analysis Based on the Corpus Outcomes in Kazakh. Forum for Linguistic Studies, 7(2), 869–881. https://doi.org/10.30564/fls.v7i2.8317

Poster Session / **272**

Implementing the CARE Principles for Datasets with Local Con-

texts Labels

Author: Chantel Ridsdale[1]

Co-author: Sarvenaz Ghafourian [1]

[1] *Ocean Networks Canada*

Corresponding Authors: cridsdale@uvic.ca, sarvenazghbm@oceannetworks.ca

Data repositories recognize that the CARE Principles (Collective Benefit, Authority to Control, Responsibility and Ethics) and data sovereignty are integral when working with indigenous communities, but it can be difficult to put words into action. Ocean Networks Canada (ONC) has been working with Local Contexts to address this gap.

ONC has been working on integrating Local Context Labels into our infrastructure, to ensure that our indigenous community partners have the ability to communicate the values important to them, traditional protocols, and specific conditions for reuse of the data they collected. In order to support this, ONC had to develop pathways between our infrastructure and the Local Contexts API, as well as further develop our ISO 19115 and DataCite metadata profiles, making metadata human- and machine-readable, so that it is properly discoverable to users.

Local Contexts Labels have not been widely implemented within data repositories, so we encountered the need to develop parts of our metadata profile from scratch. It was a challenge, and we believe that other data repositories would benefit from our experiences. If our work can be leveraged by others, it would be a stride towards implementation of the CARE Principles and indigenous data sovereignty, that would add value to the entire research community.

The selection of Local Contexts Labels, and the drafting of the text associated, is intended to be done by the community, not the data repository. When ONC approached our partners about this initiative, we overwhelmingly received agreement that this is a great opportunity, but none of them had capacity to undertake this effort at the time. In an effort to be prepared for when our community partners are able to be involved, we began a Proof of Concept using ONC-owned data within the Local Contexts Test Hub.

We worked within the Local Contexts Test Hub to create a test case, using ONC-owned data allowing us to plan for technical integration, as well as provide communities with support and documentation, based on our experiences. This Proof of Concept has been valuable, and has provided us with some great insights into the Local Contexts Infrastructure, as well as the challenges our community partners will likely encounter, allowing us to develop supports in advance.

Poster Session / 275

Proactive, Risk-Based Thresholds for Dengue Early-Warning

Author: Wala Areed[1]

[1] *Postdoctoral research fellow*

Corresponding Author: w.areed@uq.edu.au

Dengue surveillance in many countries still relies on a simple outbreak rule: declare an alert when reported cases exceed the historical mean + 2 standard deviations. Although easy to apply, this cut-off does not adapt to changing transmission patterns, generates frequent false positives, and ignores forecast uncertainty. We develop and evaluate risk-based outbreak thresholds that incorporate both the probability of future cases and their potential magnitude. Monthly dengue counts from 114 districts in the Mekong Delta, Viet Nam (2012–2022) were modelled with an ensemble of probabilistic forecasts covering 1- to 3-month horizons. From each forecast distribution, we derived three decision metrics:
Absolute Risk Score –probability of exceeding a fixed case threshold multiplied by that threshold;Relative Risk Score –probability-weighted difference between the forecast and the historical mean Predicted Mean –

forecast mean compared directly with the baseline threshold. We assessed the metrics, alone and in combination with the fixed rule, in a simulation study that triggered vector-control interventions and counted avoided cases. Integrating the three risk-based metrics with the conventional threshold reduced simulated annual dengue incidence by at least 16 % relative to reactive control, while lowering false-alert rates. Probabilistic, risk-weighted thresholds offer a transparent and reproducible alternative to fixed rules, enabling earlier and more targeted interventions. All code and processed data will be released to facilitate reuse in other settings and diseases.

Poster Session / 278

Research on the Trustworthiness Evaluation of Scientific Data Management Platforms in Chinese Universities

Authors: Shenqin Yin[1]; Jilong Zhang[1]; Anna Fu[1]; Dongwei Wang[1]; Yuxiao Du[1]; Song Xue[1]

[1] *Fudan University*

Corresponding Authors: duyuxiao@fudan.edu.cn, fuanna@fudan.edu.cn, ysq@fudan.edu.cn, jlzhfd@fudan.edu.cn, sxue@fudan.edu.cn, wdw@fudan.edu.cn

This study investigates the current research status of trustworthiness evaluation in China through literature review and web-based surveys, revealing a lack of tailored evaluation frameworks and practices specifically targeting scientific data management platforms in Chinese universities. Building upon the FAIR principles (Findability, Accessibility, Interoperability, and Reusability), this research aims to develop a localized trustworthiness evaluation system aligned with the developmental needs of university scientific data management platforms in China. The proposed system is applied to evaluate platforms, enabling them to assess their trustworthiness status and identify pathways for improvement. The detailed approach includes:

Conducting a multi-dimensional comparative analysis and LDA (Latent Dirichlet Allocation) topic clustering of three international digital repository certification standards—Nestor Seal, ISO 16363, and CoreTrustSeal—to extract trustworthiness-related indicators. Concurrently, core elements affecting trustworthiness in university scientific data platforms are analyzed, focusing on organizational governance, data management functions, and technical infrastructure/security. By integrating these trustworthiness indicators with platform-specific elements and incorporating FAIR principles, a comprehensive set of trustworthiness evaluation indicators for university scientific data platforms is derived.

Synthesizing the extracted trustworthiness indicators with China's contextual realities and existing research, a preliminary draft of trustworthiness evaluation criteria for Chinese university scientific data platforms is formulated. This draft undergoes iterative refinement through expert reviews, resulting in a finalized hierarchical framework comprising **3 primary indicators, 14 secondary indicators, and 41 tertiary indicators**. Leveraging surveys from experts at nine universities within the China University Research Data Management Working Group, the Analytic Hierarchy Process (AHP) is employed to assign indicator weights, while the Fuzzy Comprehensive Evaluation Method is adopted to construct a localized trustworthiness evaluation system. The system rigorously adheres to FAIR principles to ensure data findability, accessibility, interoperability, and reusability.

In empirical validation, four scientific data platforms from China's top 10 universities (participants in the China University Research Data Management Working Group) are selected for testing. The evaluation results preliminarily validate the system's effectiveness and practical guidance.

Based on the evaluation outcomes, recommendations are proposed across three dimensions—organizational infrastructure, digital object management, and technical infrastructure/security**—to enhance the trustworthiness of Chinese university scientific data platforms. The study concludes with insights and implications for advancing trustworthiness evaluation in this domain.

In summary, this research represents a pioneering exploration of trustworthiness evaluation for scientific data management platforms in Chinese universities, demonstrating innovation in research

focus, methodology, and practical application. By embedding FAIR principles, the study holds significant value for promoting localized trustworthiness evaluation and certification in China, as well as advancing the long-term reliable management of scientific data.

Poster Session / 281

Developing a Data-Driven ESG Framework Integrating Carbon Emissions, Financial Performance and Supply Chain Risk Analysis

Author: hwang eunhye[None]

Co-author: Juyoung Kang [1]

[1] *Ajou university*

Corresponding Authors: jykang@ajou.ac.kr, heh3800@ajou.ac.kr

As global awareness of climate change risks deepens, companies are facing increasing pressure from key stakeholders such as investors, regulators and consumers to adopt more transparent and structured approaches to environmental, social and governance (ESG) practices. ESG disclosures have emerged as critical tools for demonstrating corporate accountability and long-term value creation. Multinational corporations with complex global supply chains are under intensified regulatory scrutiny, particularly in response to new climate policy instruments such as the Carbon Border Adjustment Mechanism.

In addition to this mechanism, regulatory requirements related to supply chain due diligence, climate disclosure obligations and RE100 commitments are rapidly expanding. Recent European policies including the Corporate Sustainability Reporting Directive, the Corporate Sustainability Due Diligence Directive and the Digital Product Passport demand disclosure of emissions data and reduction targets across parent companies, subsidiaries and suppliers. In response, leading global firms such as Microsoft, Samsung Electronics and Hyundai Motor have committed to full supply chain net zero targets and RE100 membership, reflecting a growing recognition that carbon accounting capabilities are now central to global market competitiveness and access.

This study proposes a data driven ESG assessment framework that integrates key performance indicators, value chain emissions commonly referred to as Scope 3 and TS2000 based financial performance metrics. The framework is designed to diagnose corporate sustainability by linking operational efficiency with climate risk exposure. Unlike traditional ESG approaches that rely on qualitative reporting, the model leverages structured quantitative data to enhance objectivity and cross-company comparability.

At its core, the framework employs an artificial intelligence powered analytical engine that assesses upstream and downstream emissions, supplier level risks and financial resilience in the face of shifting climate regulations. By linking ESG metrics with financial data, the model provides an integrated view of how carbon intensity and sustainability strategies affect long-term competitiveness and regulatory vulnerability.

The framework also incorporates emissions trading system data and corporate carbon allowance allocations to quantify the financial implications of surplus or excess emissions. It evaluates the effects of these variables on profitability, operational efficiency and investment capacity, ultimately enabling a more precise valuation of carbon assets and exposure.

Additionally, the model includes a practical climate response portfolio composed of emissions compliance tools, supplier diagnostics and emissions mitigation planning. It is particularly relevant for firms operating in jurisdictions preparing for carbon adjustment enforcement and similar pricing regimes. The model supports transparency by identifying emission hotspots using Scope 3 data and enables collaborative decarbonization strategies across the supply chain.

The framework further introduces a materiality based supply chain risk management approach that focuses on early identification of ESG risks, incorporates them into supplier selection and establishes risk mitigation strategies through continuous monitoring. More than an evaluative tool, this system serves as a due diligence mechanism directly linked to strategic corporate decision making.

Ultimately, this framework offers actionable insights for companies aiming to align with evolving sustainability regulations. It equips organizations with robust, data centric tools to proactively manage ESG performance, carbon responsibility and long-term resilience in a changing global environment.

Poster Session / 282

Building a National Persistent Identifier Toolkit to Enhance Research Quality, Provenance, and Impact

Authors: Hélène Draux[1]; Linda O'Brien[None]; Lyle Winton[2]; Matthias Liffers[2]; Natasha Simons[3]; Simon Porter[1]; Wastl Juergen[1]

[1] *Digital Science*

[2] *ARDC*

[3] *Australian Research Data Commons (ARDC)*

Corresponding Authors: matthias.liffers@ardc.edu.au, lindasobrien@gmail.com, j.wastl@digital-science.com, s.porter@digital-science.com, h.draux@digital-science.com, lyle.winton@ardc.edu.au, natasha.simons@ardc.edu.au

The Australian National Persistent Identifier (PID) Strategy is a critical national initiative that aims to accelerate Australian research quality, efficiency and impact through universal use of connected persistent identifiers. It supports a vision where researchers, institutions, and infrastructures are connected through a universal, trusted, and interoperable system of PIDs. This strategy promotes better discovery and reuse of research inputs and outputs, improved reproducibility and attribution, and more effective national planning. To realise the vision requires shared action and accountability across the research ecosystem stakeholders.

To support implementation of this vision, Digital Science, in collaboration with the Australian Research Data Commons (ARDC), has been commissioned to create a benchmarking toolkit. This toolkit is designed to help the Australian research ecosystem assess their progress toward PID adoption using measurable, SMART benchmarks (Specific, Measurable, Achievable, Relevant, and Time-bound) aligned to the strategy's five objectives. It enables stakeholders to understand national maturity in using PIDs for entities such as researchers, projects, grants, facilities, and outputs. The toolkit contextualizes local efforts within a broader national and global infrastructure, fostering alignment and targeted improvement.

Each benchmark approach aligns with one of the five strategic objectives: (1) enhancing the FAIRness of research inputs; (2) increasing the discoverability and reuse of research outputs; (3) improving reproducibility and reducing administrative burden; (4) enabling robust impact assessment through linked metadata; and (5) supporting national capability mapping through PID integration. Together, these benchmarks provide a comprehensive structure for evaluation, planning and improvement..

Ultimately, this benchmarking toolkit provides a mechanism for tracking collective progress, informing stakeholder action plans, and ensuring that Australia's investment in digital research infrastructure delivers measurable and sustainable benefits.

Poster Session / 283

Privacy-Enhancing AI-based Whole Slide Image Analysis

Authors: Benjamin Hong Meng Tan[1]; Chee Long Cheng[2]; Jin Chao[1]; Khin Mi Mi Aung[1]; Rodel Miguel[1]

[1] *A*STAR - Institute for Infocomm Research*

[2] *SingHealth*

Corresponding Authors: jinc@i2r.a-star.edu.sg, cheng.chee.leong@singhealth.com.sg, tanhm@i2r.a-star.edu.sg, mmaung@i2r.a-star.edu.sg, miguelrf@i2r.a-star.edu.sg

Whole-slide images (WSIs) drive state-of-the-art computational pathology, but hospitals typically restrict their analysis to isolated, air-gapped workstations because these gigapixel slides contain highly sensitive patient data. On such systems the workflow for a single case is onerous: (i) technicians copy the multi-gigabyte WSI to a removable medium and walk it to the secure workstation; (ii) the slide is partitioned into patches (≈7 min); and (iii) deep-learning inference —runs for ≈20 min. With sequential processing and manual hand-offs, throughput stalls well below the 50 cases per day target for routine diagnostics.

We present a privacy-preserving, cloud-enabled pipeline that removes the physical-transfer bottleneck while maintaining strict confidentiality guarantees. The solution hinges on hardware-based trusted execution environments (TEEs):

- TEE Encryptor (on-premises). After verifying a remote enclave's attestation, it establishes an ephemeral session key, preprocesses each WSI locally, slices it into patches, encrypts the patches with the session key, and transmits them over TLS.
- TEE Analyzer (cloud). Hosted in an AMD SEV or Intel TDX enclave, it decrypts patches only inside protected memory, executes the two-stage deep-learning cascade, re-encrypts results with the same session key, and stores all artefacts in an object store accessible solely within the private analysis service. Encrypted results return to the hospital and are decrypted by the TEE Encryptor.

Because computation now runs on elastic cloud hardware, multiple TEE Analyzer instances can be launched in parallel. A deployment with ten enclaves cuts effective turnaround to minutes per case and comfortably exceeds the 50-case-per-day target, all without exposing WSIs or predictions in plaintext to the cloud operator. Regulatory mandates such as HIPAA and GDPR are satisfied because the data never leave trusted memory or the hospital in unencrypted form.

Our results show that confidential-computing clouds can deliver order-of-magnitude improvements in digital-pathology throughput while preserving patient privacy. The proposed architecture decouples sensitive data from untrusted infrastructure, offering clinicians a scalable, secure path to real-time, image-centric diagnostics.

Poster Session / 284

Data stewardship for PalMod - A FAIR-based strategy for data handling in large climate modeling projects

Author: Swati Gehlot[1]

Co-authors: Andrea Lammert [1]; Hannes Thiemann [1]; Martin Schupfner [1]

[1] *DKRZ*

Corresponding Author: gehlot@dkrz.de

German climate research initiative Paleo Modeling or PalMod¹ (currently in phase III) is presented here as an exclusive example where the project end-product is unique, scientific paleo-climate data. PalMod data products include simulated climate data from three state-of-the-art coupled climate models of varying complexity and spatial resolutions. Integrating this simulated or modeled climate of the past 130,000 years, with a comprehensive compilation of paleo-proxy reconstruction data facilitate model-model and model-proxy intercomparison/evaluation leading to a more credible climate projections for the future. Being a large multidisciplinary project, a dedicated RDM (Research Data Management) approach is applied within the cross-cutting working group for PalMod. The DMP (Data Management Plan), as a living document, is used for documenting the data-workflow framework that defines the details of paleo-climate data life-cycle. The workflow containing the organisation, storage, preservation, sharing and long-term curation of the data is defined and tested. In order to make the modeling data inter-comparable across the PalMod models and easily analyzable by the global paleo-climate community, model data standardization (CMORization) workflows are defined for individual PalMod models and their sub-models. The CMORization workflows contain setup, definition, and quality assurance testing of CMIP6² based standardization processes adapted to PalMod model simulation output requirements with a final aim of data publication via ESGF³. PalMod data publication via ESGF makes the paleo-climate data an asset which is (re-)usable beyond the project life-time. Along with ESGF publication, the standardized data is long term archived for the use of paleo-climate research community.

The presented RDM infrastructure enables common research data management according to the FAIR⁴ data principles across all the working groups of PalMod. Common workflows defined for the exchange of data and information along the process chain(s) are an important asset which could be applicable to other large scale climate modeling projects. Applying data management planning within PalMod made sure that all the data related workflows were defined, continuously updated if needed and made available to the project stakeholders. End products of PalMod which consist of unique long term scientific paleo-climate data (model as well as paleo-proxy data) are made available for re-use via the paleo-climate research community as well as other research disciplines (e.g., land-use, socio-economic studies etc.). The PalMod data stewardship ensures a FAIR based data dissemination for the climate model datasets as well as the various data standardization workflows developed within the project RDM.

1. www.palmod.de
2. Coupled Model Intercomparison Project phase 6 (https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6).
3. Earth System Grid Federation (https://esgf.llnl.gov).
4. Findable, Accessible, Interoperable, Reusable.

Poster Session / 285

An interoperable and secure model supporting data mobility across the research ecosystem

Author: Greg D'Arcy[1]

Co-author: Alex Ip [1]

[1] *AARNet Pty Ltd*

Corresponding Authors: greg.darcy@aarnet.edu.au, alex.ip@aarnet.edu.au

With the pace of research accelerating into the age of Quantum computing and AI, data mobility has become the lifeblood of modern scientific research. As unprecedented volumes of research data are being created at increasing speed, it is imperative that data can be easily moved and shared to be findable and accessible. Yet achieving data mobility in scientific research faces significant challenges that span technical, ethical, and institutional domains. If moving data from point A to

point B is difficult, it undermines all existing investments in the latest instruments, high-performance computing and virtual research environments.

Looking to the future, AARNet is advancing Australia's national digital infrastructure to manage network speeds of up to 400Gbps and meet the evolving needs of modern, data-driven research fields. This poster will present some of the latest case studies from Australia where Globus has been used to drive data transfers and automation across diverse research domains, including Genomics, Astronomy, Advanced Microscopy & Materials Characterisation, Health and Medicine.

AARNet is Australia's national research and education network. For more than 30 years AARNet has provided reliable telecommunications services to the Australian academic and research sectors, along with an expanding range of cyber security, data and collaboration services. More than just a network. AARNet is about providing cost-effective solutions that drive powerful outcomes for research and education. Globus is one such high-value service that AARNet supports in Australia.

Globus is world-leading research cyberinfrastructure, developed and operated as a not-for-profit service by the University of Chicago. Globus allows researchers to easily, reliably and securely move, share, & discover data, no matter where it resides –from a supercomputer, lab cluster, tape archive, public cloud or laptop.

Globus allows users to leverage AARNet's high-speed networks by managing the efficient transfer of data up to petabyte-scale, and orchestrating distributed workflows across multiple facilities. Jobs are scheduled using a simple web interface: a feat that would be impossible using traditional file transfer tools.

Poster Session / 287

The Chinese Experience of Using Data as a New Production Factor to Realize Economic Value

Authors: Jie Yang[1]; Lin Huang[1]

[1] *Beijing Academy of Science and Technology*

Corresponding Authors: yangjie@bjast.ac.cn, huanglin@bjast.ac.cn

In the context of the digital economy, data has become an important strategic foundational resource and economic growth engine for a country. China is the first country to elevate data to the level of production factors in its policy system, and officially listed data as another key production factor after land, labor, capital, and technology in 2019. When the transformation of production factors triggered by data begins to change the organizational operation of the entire economy and society, how to unleash the value of data factors becomes a key issue. A large amount of low-cost raw data itself does not have value, and data as a production factor need to be empowered for digital economic growth through market-oriented processes such as resource transformation, productization, assetization, and capitalization.

The starting point is data resource transformation, which is the process of transforming previously scattered and disordered data into reusable, organized, and valuable information resources through steps such as collection, consolidation, cleaning, sorting, and labeling. According to 2024 National Data Resource Survey Report, the total annual national data production is 41.06 zettabytes (ZB), which means China has abundant data resources. In terms of public data, more than 60% of provinces and cities specifically designated in the national plan have started the authorization and operation of public data.

The second step is data productization, which refers to process data resources into tradable products or services with clear application scenarios. For example, in order to further accelerate the development and utilization of public data, the National Data Bureau of China has officially released 70 demonstration scenarios in 2025 with clear data products, covering various fields such as highway emergency rescue, smart farmland construction, enterprise service "zero proof", data empowerment

of precision medicine, ecological governance of the Yellow River Basin, and innovation of scientific research paradigms.

The third step is data assetization, which refers to transform data resources into measurable and manageable intangible assets with clear property rights and transaction value through legal and market mechanisms. The Interim Provisions on Accounting Treatment of Enterprise Data Resources issued by the Ministry of Finance of China was officially implemented on January 1st, 2024, clarifying the accounting treatment methods for data resources. At least 100 Chinese A-share listed companies have included "data assets" in their balance sheets in their 2024 annual reports, involving a total amount of 2245 million RMB.

The fourth step is data capitalization, which refers to the process of financing, pledging, and other financial operations of confirmed data assets through the financial market to realize the financial value of data assets. For example, on March 2025, Changzhou (a city near Shanghai) implemented 240 million RMB of public data pledge financing. The current data pledge cases are mostly concentrated in the fields of transportation, energy, etc. In the future, they can be extended to medical, agricultural, cultural, tourism and other industries.

Through these four steps, raw data will gradually be transformed into resources, products, assets, and capital, and then governments and enterprises can better utilize data, improve decision-making efficiency, and generate new revenues. Although there is still some disagreement among policy makers regarding the ownership of data factors, and some contradiction between public data openness and assetization, it is necessary to further study the basic concepts, operating rules, and implementation paths of data factor in practice, and continuously improve the relevant policy system in strengthening data infrastructure construction, improving data governance, unleashing the synergistic effects of data factor, promoting data openness and sharing, and fully unleashing the data "multiplier effect" in various industries.

Poster Session / **289**

LETNER: Label-EfficienT Named Entity Recognition for Cyber Threat Intelligence

Author: Yue Wang[1]

Co-authors: Duoyi Zhang [2]; Md Abul Bashar [2]; Richi Nayak [2]

[1] *Centre for Data Science, Queensland University of Technology*

[2] *The Queensland University of Technology*

Corresponding Authors: y355.wang@hdr.qut.edu.au, r.nayak@qut.edu.au, duoyi.zhang@hdr.qut.edu.au, m1.bashar@qut.edu.au

With the rise of cyber threats, automating Named Entity Recognition (NER) in open-source documents is crucial for Cyber Threat Intelligence (CTI). However, cybersecurity NER models face challenges in maintaining large annotated datasets due to the ever-evolving threat landscape. To address this, we introduce LETNER, a label-efficient NER framework that balances performance and annotation demands. LETNER features Span-CNN-Gate, a convolutional gating module that enhances span-based entity representation, and integrates metric learning to effectively capture entity-span relationships in a shared metric space, improving adaptability in low-resource settings. We also propose a systematic evaluation framework for label efficiency in supervised NER models. Experimental results demonstrate that LETNER achieves state-of-the-art label efficiency, significantly reducing annotation costs while maintaining high performance. On a complex CTI dataset with 21 fine-grained entity classes, LETNER outperforms the widely adopted Flair NER framework by 11.8\% in F1 score while using only 10\% of the training data.

Poster Session / **290**

Label Propagation Assisted Soft-constrained Deep Non-negative Matrix Factorization for Semi-supervised Multi-view Clustering

Author: Sohan Gunawardena[1]

Co-authors: Khanh Luong [1]; Richi Nayak [2]; Thirunavukarasu Balasubramaniam [1]

[1] *Queensland University of Technology*

[2] *The Queensland University of Technology*

Corresponding Authors: thirunavukarasu.balas@qut.edu.au, sohan.gunawardena@hdr.qut.edu.au, khanh.luong@qut.edu.au, r.nayak@qut.edu.au

To address the pressing challenge of capturing complex non-linear structures in semi-supervised multi-view clustering, we introduce a fundamentally novel framework:Label Propagation Assisted Soft-constrained Deep Non-negative Matrix Factorization for Semi-supervised Multi-view Clustering (LapSDNMF). Unlike prior approaches,LapSDNMF innovatively integrates deep hierarchical modelling with label propagation to jointly exploit the power of non-linear representation learning and the guidance from limited labelled data. By embedding a predictive membership matrix as a soft constraint within a deep architecture, LapSDNMF enables seamless propagation of label information, guided by graph-based regularization that reflects local data geometry. LapSDNMF unifies deep learning and graph-theoretic techniques in a principled optimisation framework. We also develop a novel, efficient algorithm based on multiplicative update rules to solve the resulting optimisation problem. Extensive experiments on diverse real-world datasets demonstrate that LapSDNMF consistently and significantly outperforms existing state-of-the-art multi-view clustering methods.

Poster Session / 291

Supporting the Life Sciences: The Role of the German Network for Bioinformatics Infrastructure and ELIXIR Germany

Authors: Alexander Sczyrba[1]; Daniel Wibberg[1]; Helena Schnitzer[1]; Irena Maus[1]; Nils Hoffmann[1]; Nils-Christian Lübke[1]; Oliver Kohlbacher[2]; Sebastian Jünemann[1]; Tanja Dammann-Kalinowski[1]

[1] *IBG-5, Forschungszentrum Jülich*

[2] *Eberhard-Karls-Universität Tübingen*

Corresponding Authors: d.wibberg@fz-juelich.de, a.sczyrba@fz-juelich.de, s.juenemann@fz-juelich.de, h.schnitzer@fz-juelich.de, n.hoffmann@fz-juelich.de, i.maus@fz-juelich.de, t.dammann-kalinowski@fz-juelich.de, oliver.kohlbacher@uni-tuebingen.de, n.luebke@fz-juelich.de

Modern life science research increasingly relies on complex data analysis, demanding robust bioinformatics tools, substantial computational resources, and specialised expertise. The German Network for Bioinformatics Infrastructure (de.NBI) addresses these challenges as a national, academic research infrastructure funded by the German Federal Ministry of Education and Research (BMBF) and coordinated by Forschungszentrum Jülich, within the Helmholtz Society. de.NBI offers a wide range of comprehensive bioinformatics services spanning nearly the full spectrum of life sciences, along with training and cloud resources. While these services are primarily aimed at researchers in Germany and Europe, many are also accessible globally. Comprising eight interconnected service centres and a coordination unit with 24 institutional partners at public universities and research institutes, de.NBI fosters collaboration within the German bioinformatics community and facilitates knowledge transfer between academia and industry via its Industrial Forum.

Since 2016, de.NBI constitutes the German node of ELIXIR, the European life science infrastructure for biological information. Connecting life science resources from 22 countries and EMBL, ELIXIR Europe brings together bioinformatics expertise and tools to support data-driven discovery. By streamlining access to research data and best practices, it empowers scientists to collaborate, share

knowledge, and accelerate research across disciplines. ELIXIR Germany integrates national bioinformatics resources into the broader European landscape, actively contributing to ELIXIR's scientific program and working groups. Key achievements include the designation of de.NBI data resources – BRENDA, BacDive and SILVA –as ELIXIR Core Data Resources and as Global Core Biodata Resources, the implementation of the Life Science Authentication and Authorization Infrastructure (AAI) with the de.NBI Cloud and affiliated services, and leadership in ELIXIR Platforms (Compute, Tools and Data) and Communities (Proteomics, Metabolomics, Galaxy and Plants). Since 2019, ELIXIR Germany's engagement is demonstrated by contributions to over 68 technological and scientific implementation studies and related publications.

Recognising the critical need for advanced data management and processing capabilities, ELIXIR Germany also hosts the German Competence Center Cloud Technologies (de.KCD). de.KCD focuses on building expertise and providing cloud infrastructure for efficient data handling, standardised analysis, and skill development, fostering a collaborative data space for national and international research.

Overall, de.NBI & ELIXIR Germany forms a vital infrastructure supporting cutting-edge life science research, promoting innovation, and ensuring a strong foundation for bioinformatics excellence.

Poster Session / 293

Enhancing Capacity for Ethical Data Sharing in Clinical Research

Authors: Amany Gouda-Vossos[1]; Kristan Kang[1]; Lisa Eckstein[2]

[1] *Australian Research Data Commons*

[2] *Bellberry (Clinical Trials IQ - CT:IQ)*

Corresponding Authors: kristan.kang@gmail.com, amany.gouda-vossos@ardc.edu.au

With the onset of the Open Science movement, research sites and clinical research sponsors are becoming increasingly entrusted with the storage of large amounts of research data and samples. The prospect of sharing a wide array of health data is an exciting one, as the collaboration of ideas and the expansion of shared knowledge promises to lead to accelerated research outcomes. However, identifying the appropriate ethical and governance arrangements for sharing data, especially clinical trials data, can be difficult. Further, unduly restrictive governance may prevent valuable new research from proceeding. Finally, insufficiently informed governance may breach participant privacy and autonomy interests.

Responsible sharing of clinical research data requires consideration of complex ethical, regulatory, legal and institutional requirements. To support researchers navigating this space, the Australian Research Data Commons (ARDC) via the Health Studies Australian National Data Asset initiative (HeSANDA) in collaboration with Clinical Trial IQ (CT:IQ) is developing practical principles and guidance for researchers, HRECs, data custodians, research institutions, and consumers to support trustworthy sharing of clinical research data in Australia.

Here we showcase our multifaceted approach to address the legal, ethical, and practical challenges of data sharing for clinical research, which included a series of interlinked activities. This includes the creation of a *Governance Framework*, which outlines the regulatory and ethical responsibilities for data sharing. This was supplemented by a *Consultation Report*, which captured insights from workshops with key stakeholders and identified knowledge gaps around secondary data use. To evaluate how ethics review bodies apply existing standards, a *Benchmarking exercise* tested the consistency of decisions on a simulated complex data sharing application. Findings from these activities are directly informing the design of a *Resource Toolkit*, a suite of skills-based materials aimed at improving awareness, decision-making, and practice in ethical data sharing across the clinical research community.

The development of these practical principles and guidance resouces aims to foster more trustworthy and responsible data sharing practices, contributing to a more impactful and efficient data sharing ecosystem in Australia.

Poster Session / 294

CAREful Linking of FAIR Language Data to Reproducible Jupyter Notebooks

Author: Steele Cooke[1]

Co-authors: Adam Bell [1]; Alex Ip [2]; Rosanna Smith [3]; Simon Musgrave

[1] *AARNet*

[2] *AARNet Pty Ltd*

[3] *UQ*

Corresponding Authors: rosanna.smith@uq.edu.au, adam.bell@aarnet.edu.au, steele.cooke@aarnet.edu.au, alex.ip@aarnet.edu.au, s.musgrave@uq.edu.au

The reliable reuse of language data largely depends on both managing the data in ways that respect the rights, responsibilities and communities from whom it originates, and allowing any user with the appropriate skills and resources to inspect, rerun and extend the analyses that underlie the published findings.

In practice, these goals often collide where data may be preserved in one location at an institutional or disciplinary repository, while the code that generates the tables, figures and models is scattered across personal repositories.

Jupyter Notebooks are web-based shareable documents that combine code, visualisations, rich text and interactive controls, allowing users to execute code in steps directly within the notebook, making it ideal for exploratory data analysis and interactive experimentation. Over time, however, library upgrades, version changes, missing credentials and undocumented requirements can break once-working Jupyter Notebooks, which make it harder for readers to verify and reproduce results or build on them.

BinderHubs solve this problem and further enhances reproducibility by allowing users to launch pre-configured Jupyter Notebooks as interactive computing environments from code repositories with explicitly defined hardware and software requirements.

The Language Data Commons of Australia (LDaCA) in collaboration with Australia's Academic and Research Network (AARNet) has addressed this problem by developing CAREful and FAIR data infrastructure that focuses on the preservation and access of distributed, multi-modal language data collections.

The LDaCA data portal has been configured to display the Jupyter Notebooks associated with a language data collection allowing users to automatically launch them in one of several available BinderHubs. RO-Crates, a data packaging specification, are used to describe the hardware, resources and dependencies of each Jupyter Notebook for FAIR reproducibility.

Current work involves adding additional language collections to the data portal. Some material may be sensitive and require access control, which would require a user to request access to data where appropriate. Once approved by the data custodian the user can inspect, rerun and extend the research findings using the original dataset and published analyses.

Poster Session / 297

Building national research infrastructure to share health research data: Lessons from HeSANDA and Health Data Australia.

Authors: Amany Gouda-Vossos[1]; Jo Savill[1]; Rhys Williams[1]

Co-authors: Amanda Del Valle [1]; Omer Waraich [1]; Richard Ferrers [1]

[1] *Australian Research Data Commons*

Corresponding Authors: richard.ferrers@ardc.edu.au, amanda.delvalle@ardc.edu.au, omer.waraich@ardc.edu.au, jo.savill@ardc.edu.au, rhys.williams@ardc.edu.au, amany.gouda-vossos@ardc.edu.au

The Health Studies Australian National Data Asset (HeSANDA), led by the Australian Research Data Commons (ARDC), is building national research infrastructure to enhance the discoverability, access, and reuse of data from health research studies across Australia. HeSANDA was established as a response to the critical need for more accessible and interoperable health research data.

The HeSANDA initiative has brought together diverse technical and institutional stakeholders across Australia to facilitate a cohesive and collaborative endeavour, formally known as the HeSANDA Node Network. This network comprises nine nodes across Australia, each representing a consortium of health research organisations. Collectively, these nodes encompass over 70 health research organisations, including universities, medical research institutes, health services, and clinical trial networks.

In 2023, the HeSANDA Node Network, alongside the ARDC, launched Health Data Australia (HDA). HDA is a national catalogue of Australian health data for researchers to discover and request access to data for their research (researchdata.edu.au/health/). The development of the platform was grounded in extensive consultations with the research community and key stakeholders. The framework for sharing clinical trial data was co-designed with ARDC and the HeSANDA Node Network.

HDA operates on a federated infrastructure model, where data remains under the control of the original data providers. The platform hosts metadata descriptions of datasets, allowing researchers to discover and request access to data without the data itself being stored centrally. This approach respects data custodianship and governance while promoting data discoverability and reuse.

In our poster, we outline the multi-layered infrastructure model that supports the control of the data with the data provider, while enabling national discoverability. We also highlight the collaborative work undertaken with the HeSANDA Node Network to develop standards procedures, best practices, and shared governance approaches. Our experiences addressing challenges we have encountered in harmonising metadata standards, policies and procedures, effective strategies for community co-design, and how we address consent and privacy concerns will also be shared.

The lessons learned during the development of HDA through the HeSANDA initiative provide insights for others seeking to build interoperable, researcher-focused data infrastructure.

Poster Session / 301

Understanding injury-related bloodstream infections in Queensland: a data linkage study

Author: Binuri Perera[1]

Co-authors: Kevin Laupland [1]; Kirsten Vallmuur [1]; Susanna Cramb

[1] *Queensland University of Technology*

Corresponding Authors: b2.perera@hdr.qut.edu.au, kevin.laupland@qut.edu.au, k.vallmuur@qut.edu.au, susanna.cramb@qut.edu.au

Defined as the presence of any infectious microorganisms in the bloodstream, bloodstream infections (BSIs) pose a major threat to public health. BSI is an important complication that may affect the recovery time, treatments of injured patients. The studies on patients with injury-related BSIs

report data from single or selected hospitals. No population-based studies have been conducted on injury-related BSIs. Hence, the burden of injury-related BSI is not quantified across the world. In Australia, injuries are a leading cause of mortality, hospitalisations, and disability making injury prevention and control a national health priority area, while BSIs have a significant economic impact. Queensland is the second-largest state in Australia by area and has the second-highest injury hospitalisation rates in Australia. We conducted a retrospective population-based data linkage study to understand the patient demographics, incidence rates, isolates and antimicrobial resistance among Queenslanders with injury-related BSIs.

The study population consisted of all residents of Queensland, Australia, with incident BSI, admitted to a Queensland Health public hospital between 1 January 2000 and 31 December 2019. The data used for this study consists of three population-based data collections: the Queensland pathology data providing positive blood cultures, Queensland Hospital Admitted Patient Data Collection (QHAPDC) and Death Registrations. All blood cultures taken from public healthcare institutions were recorded in the Queensland pathology data and linkage to QHAPDC data identifies the hospitalisation, demographics, clinical and outcome data of these patients. The Registry of General Death confirmed mortality as of 31 December 2020.

During the study period, a total of 3428 injury encounters and 3586 BSI episodes were identified among 3402 individuals. The median age of this cohort was 63 years, and the majority of the individuals were males (65%). Among these patients, around 50% had a 30-day mortality. The age-standardised rates of injury-related BSIs showed an increasing trend over the years and males reported a higher rate than females. The most commonly identified isolate among these patients was Staphylococcus aureus.
Through innovatively linking administrative datasets, this study provides the first population-based disease burden of injury-related BSIs and a strong foundation for future research to improve patient outcomes.

Poster Session / **303**

Powering Ecological Research and Environmental Decision-Making: Inside TERN's Data Infrastructure

Author: Siddeswara Guru[1]

Co-authors: Arun Singh Ramesh [2]; Avinash Chandra [1]; Enzhen Luo [2]; Gerhard Weis [1]; Javier Sanchez Gonzalez [2]; Junrong Yu [2]; Keion Larsen [2]; Megan Edward [1]; Mosheh Eliyahu [3]; Tiancheng Lan [1]; Yong Liaw [1]; wilma Karsdorp [1]

[1] *University of Queensland*

[2] *Terrestrial Ecosystem Research Network*

[3] *University of Adelaide*

Corresponding Authors: avinash.chandra@uq.edu.au, g.weis@uq.edu.au, y.liaw@uq.edu.au, t.lan@uq.edu.au, w.karsdorp@uq.edu.au, m.edward@uq.edu.au, mosheh.eliyahu@adelaide.edu.au, junrong.yu@uq.edu.au, j.sanchezgonzalez@uq.edu.a k.larsen@uq.edu.au, e.luo@uq.edu.au, a.singhramesh@uq.edu.au, s.guru@uq.edu.au

The Terrestrial Ecosystem Research Network (TERN) is Australia's national collaborative research Infrastructure for long-term environmental monitoring, data-driven ecological research, and evidence-based decision-making. TERN provides an integrated, standardised, and openly accessible data infrastructure that facilitates collecting, curating, analysing, and distributing high-quality ecological and biogeophysical data across Australia's diverse landscapes.

TERN's data infrastructure is built on the FAIR data principles—Findable, Accessible, Interoperable, and Reusable—to ensure that its data sets are not only openly available but also are structured to be of value to researchers, data scientists, government agencies, and other stakeholders. The infrastructure integrates several data streams from ecosystem monitoring sites, automated in-situ sensors (flux towers, phenocams, acoustic), remote sensing platforms (satellite and drone), and model-derived data. These data products cover a wide range of ecological attributes like vegetation structure, soils, climate variables, biodiversity, biogeophysical metrics, and water and carbon fluxes.

TERN's data infrastructure provides tools and services for effective data lifecycle management. The TERN Data Discovery Portal is at the heart of TERN's data infrastructure, enabling discovery and access to all TERN published data collections with a free-text and map-based search. The portal facilitates robust filtering based on platforms and parameters, spatial and temporal extent, and visualisation through the data visualiser web application and download capabilities.
The infrastructure leverages a robust cyberinfrastructure stack that includes scalable cloud storage, geospatial data services, data pipelines in automation, persistent identifiers, and rich metadata frameworks based on international standards (e.g. ISO 19115, DCAT). All datasets are also available as RO-Crate, a lightweight approach to package research data with its metadata.
Furthermore, TERN enables access to systematic survey data through EcoPlots, a data integration platform. The platform allows for search and access to observation-level data from systematic surveys conducted from TERN observatories and states and territories. In addition, an image repository-EcoImages has been created to collect all ecological images collected by TERN observatories. EcoImages enables users to search, query and download pictures of the same or different image types.

TERN presents a unique opportunity for the data science community to develop and deploy analytical techniques to high-dimensional, multiscale, and longitudinal environmental data. TERN data can be used to train machine learning algorithms for ecological applications, validate satellite remote sensing products, support conservation and land management and contribute towards the state of the environment reports. TERN allows integration with multi-disciplinary external data platforms to support research and innovation. In collaboration with national and global research infrastructure initiatives (e.g. DataOne, NEON, FLUXNET, ARDC), TERN enables its data products to be discovered and utilised in broader environmental and data science communities.

The CoreTrustSeal-certified TERN data infrastructure bridges the gap between environmental monitoring and data science by providing high-quality, interoperable data and tools to enable scalable ecological analysis and modelling. With rising demands for data-driven solutions to address specific environmental challenges such as biodiversity loss, land degradation, and climate change, TERN enables research and innovation through open, appropriately governed, and future-proofed environmental data infrastructure.

The poster will highlight TERN data infrastructure structure, governance, capabilities, and impact, showcasing how an open environmental data infrastructure would drive collaboration, innovation, and effective ecosystem management.

Poster Session / 305

Advancing Federated Open Science Infrastructures for FAIR and Responsible Research

Authors: Elli Papadopoulou[1]; Tassos Stavropoulos[2]

[1] *Athena*

[2] *OpenAIRE*

Corresponding Authors: elli.p@athenarc.gr, tassos.stavropoulos@openaire.eu

The EU-funded OSTrails project is advancing a federated approach to Open Science by addressing a key challenge: the fragmentation of research data management (RDM) practices across disciplines, tools, and institutions. By building a network of interoperable services for planning, tracking, and assessing research activities, OSTrails promotes reproducible, FAIR-aligned, and responsible science.

With 41 partners and 25 pilots spanning national, thematic, and European infrastructures, OSTrails demonstrates how a modular and standards-based ecosystem can support flexible integration while preserving domain-specific autonomy. Central to this approach are the Interoperability Reference Architecture and the Plan-Track-Assess (PTA) framework, which enable coordination across diverse tools such as machine-actionable Data Management Plans (maDMPs), Scientific Knowledge Graphs, and FAIR assessment services.

OSTrails embeds FAIR principles into day-to-day research workflows by providing machine-assisted guidance, modular metrics, and contextual user support. This empowers researchers to incorporate and assess FAIR practices throughout the research lifecycle, promoting greater rigour and transparency.

Through co-designed pilots, the project validates interoperability at scale—from local institutional settings to federation with the European Open Science Cloud (EOSC). The emerging OSTrails Commons, a shared environment for services, methods, and training, will support long-term sustainability and community-driven adoption across the Open Science ecosystem.

This poster will present how the architecture scales across disciplines and infrastructures, and explore OSTrails'contribution to reproducibility, interoperable research ecosystems, and the development of a trusted Web of FAIR Data and Services.

Poster Session / **306**

Crisis Map: Revolutionizing Emergency Response with Predictive Analytics & AI

Author: Indrasena Manga[None]

Corresponding Author: indrasenamanga@gmail.com

Natural disasters occur more frequently and intensely and impact predominantly vulnerable and under-resourced communities. Crisis Map suggests an end-to-end real-time, privacy respecting platform based on federated data mesh architecture and predictive artificial intelligence models for better disaster response and resource deployment.

With the integration of satellite imageries, public communication channels and IoT sensor data, Crisis Map applies the use of Graph Neural Network and ConvLSTM(Convolution Long Short Term Memory) for forecasting disaster impact locations and relief requirements. With cross agency partnerships enabled under federated learning, ensured compliance with territorial sovereignty, HIPAA, and CCPA is guaranteed. Simulation results confirm up to 88% accuracy in prediction, 40% reduced response time, and 30% equitable distribution of supplies.

This poster showcases the system architecture, key results in simulation, and its potential for scaling disaster resilience worldwide, putting data governance, equity, and regional adaptability as priority.

Poster Session / **307**

Development of a data search navigation tool to support data linkage information for comparative effectiveness research –a service provided by Taiwan Gateway to Health Data

Authors: Hsiu-An Lee[1]; Wen-Chang Tseng[1]; Yi-Hsin Yang[1]

[1] *National Health Research Institutes*

Corresponding Authors: yhyang@nhri.edu.tw, billy72325@gmail.com, gdi89009@nhri.edu.tw

The Taiwan Gateway to Health Data (GHD TW) is a government-funded project that collaborates with all primary data custodians and data controllers in Taiwan. We establish a data portal for various data users, including industrial and academic researchers, to promote community health and clinical and biomedical research. Our primary responsibility is to provide services that enhance the findability, accessibility, interoperability, and reusability (FAIR) of our data partners. For comparative effectiveness research, the search for fit-for-purpose data often involves linking data from

multiple sources to obtain complete information on outcomes, exposures, and confounders. In Taiwan, data linkage could be a challenge to data users. Since different data controllers may govern various data sources, the availability for linkage is generally subject to various restrictions. It is usually time-consuming for data users to determine whether a fit-for-purpose collection of data is available. We decided to develop a data search navigation tool to support data linkage information, which is available in both traditional Chinese and English.
The framework is built upon a MeSH-based synonym. We first convert English MeSH keywords into traditional Chinese and utilize a bilingual (Chinese-English) synonym dictionary and a hierarchical tag classification system, specifically designed to enhance data discoverability and relevance in Taiwan's national health and biomedical research context. We then leverage over 130,000 curated synonym entries and 182 concept tags derived from and extended beyond MeSH. We designed a search engine tailored for domestic datasets. Each dataset is annotated with weighted relevance scores across standardized tags, enabling the system to recommend associated dataset bundles based on term input and relevance ranking. The framework enables users to identify suitable datasets quickly, comprehend their fit-for-purpose structure, and assess their interoperability based on the feasibility of data linkage and the characteristics of the data. In addition to building the search interface, we incorporated user feedback mechanisms to dynamically adjust relevance scores, thereby improving recommendation precision over time. This architecture not only supports rigorous and reproducible research but also aligns with FAIR principles. Our approach demonstrates how domain-specific, language-sensitive design can bridge the gap between global data standards and local research needs.

Poster Session / 308

Towards the FAIRRREST Principles in Health Data Sharing

Author: Susan Smith[1]

Co-author: Clair Sullivan [2]

[1] *The Prince Charles Hospital, Metro North*

[2] *The University of Queensland, Queensland Digital Health Centre, (Centre for Health Service Research); Metro North Hospital and Health Service, Clinical Informatics (Digital Metro North)*

Corresponding Authors: susan.smith@health.qld.gov.au, clair.sullivan@health.qld.gov.au

As digital health technologies proliferate, the potential to harness real-world data (RWD) for improving healthcare outcomes grows dramatically. However, the realization of a truly responsive Learning Health System remains hindered by the complexities surrounding health data sharing. These complexities span technical, legal, regulatory, financial, organizational, and ethical domains and are influenced by factors including consent, purpose of use, data type, and stakeholder incentives.
Bringing order to the complexity of health data sharing has been on the research and organisational agenda for some time, however a mature vocabulary and shared vision for health data sharing governance among stakeholders is yet to emerge. It is worth recognising that the ubiquitous term 'data sharing'in health refers to 4 distinguishable levels:
1. Patient Care –Primary use of data for direct clinical treatment, reliant on system interoperability.
2. Quality Improvement and Contextual Evidence Generation (SeConts) –Secondary use for audits, surveillance, patient safety, and non-interventional research.
3 (a) Original Research –Data use for generating new knowledge, with increasing overlap with quality improvement initiatives.
3 (b) Completed Research –Sharing datasets from published studies to support reproducibility and broader scientific use.
The ubiquitous use of the term 'data sharing'amongst these activities, leads to further complexity in efforts to define health data sharing governance frameworks. The most recognisable are likely the FAIR principles (Findable, Accessible, Interoperable, Reusable) which were developed to improve data utility and reproducibility, focussing on re-use and sharing of scientific and scholarly data. While broadly adopted, their application in health contexts often overlooks key socio-ethical considerations. FAIR was never intended to operate in isolation, and it notably does not address moral or ethical questions around openness and data misuse.
To address this gap, several complementary frameworks are emerging. The TRUST principles (Transparency, Responsibility, User focus, Sustainability, Technology) emphasize societal accountability. The CARE principles (Collective Benefit, Authority to Control, Responsibility, Ethics), developed by

the Global Indigenous Data Alliance, foreground Indigenous data sovereignty and ethics. Similarly, FAIR-Health and other adaptations suggest additional factors like data quality, privacy-respecting practices, and incentives for data stewardship.
Despite this proliferation of frameworks, key recurring themes emerge across governance documents, especially within North American and European contexts. These include the societal value of data, equitable distribution of risks and benefits, respect for data contributors, and the imperative to build public trust through engagement and reciprocity.
Drawing from these insights, this article proposes the FAIRRREST principles—a unified governance framework that builds upon the foundational FAIR principles while incorporating vital ethical, societal, and sustainability dimensions:
• Findability: Ensuring that data and metadata are easily discoverable by both humans and machines.
• Accessibility: Establishing clear, transparent procedures for data access.
• Interoperability: Facilitating seamless integration and analysis across diverse platforms and systems.
• Reusability: Enabling meaningful reuse through standardized metadata, documentation and practices across different settings.
• Responsibility: Emphasizing accountability and shared obligations among all stakeholders, from data donors to users, when giving, receiving, or using data.
• Reciprocity: Ensuring data sharing meets community expectations, enables community collective benefit and social justice and also allows beneficial social institutions to grow, collaborate and receive recognition.
• Ethicality: Upholding the objective application of ethical values including respect, merit and integrity, justice, beneficence and balance of risk and benefit in all activities involving data re-use.
• Sustainability: Supporting long-term preservation and access to data and related services.
• Transparency: Promoting decision making with respect to who gets access to data and who does not, and for what purposes and what counts as appropriate involvement in the data-sharing and data-access policy process
The FAIRRREST principles address the need for nuanced and inclusive governance mechanisms that can navigate the complexity of health data sharing. They are particularly suited for guiding data reuse across the full spectrum of health-related activities—from clinical care and service improvement to academic and public health research.
By fostering shared understanding and trust across all data-sharing stakeholders, FAIRRREST offers a platform for aligning the diverse objectives of patients, providers, researchers, data custodians, and communities. It also lays the foundation for further stakeholder-driven consensus building to refine and implement ethical, practical, and sustainable data sharing policies. In doing so, it supports the realization of digital health's full potential to deliver equitable and impactful care.

Poster Session / 311

Implementation of the OMOP ETL pipeline for the standardization and integration of data from inpatients with respiratory diseases in Douala General Hospital, Cameroon

Authors: Agnes Kiragga[1]; Bertrand Hugo Mbatchou Ngahane[2]; Brenda Mbouamba Yankam[2]; François Anicet Onana Akoa[2]; Jean Blaise Ebimbe[3]; Luc Baudoin Fankoua Tchaptchet[2]; Miranda Barasa[4]; Pauline Andeso[4]; Samuel Iddi[5]

[1] *Infectious Diseases Institute, College of Health Sciences, Makerere University, Kampala, Uganda*

[2] *Douala General Hospital, Data Science Without Borders project, Cameroon*

[3] *Douala Gynaeco-Obstetric and Pediatric Hospital (DGOPEH), Cameroun*

[4] *African Population and Health Research Center (APHRC), Nairobi, Kenya*

[5] *Department of Statistics, University of Ghana, Accra, Ghana*

Corresponding Authors: mbatchou.ngahane@yahoo.com, pandeso@aphrc.org, brenda.yankam@gmail.com, ebimbejb@gmail.com, onanaanicet@gmail.com, fankoualuc@gmail.com, mbarasa@aphrc.org, siddi@aphrc.org, akiragga@aphrc.org

Douala General Hospital is a first-class hospital in Cameroon where we meet a multidisciplinary medical team treating several thousand patients each year. This hospital hosts numerous patient records that may be useful for public health research. However, majority of these records are paper-based, hence limiting their exploitation. For some cases, particularly

the pulmonology department, the data of hospitalized patients are recorded in heterogeneous datasheets mainly collected for research purposes without standardization or uniform structure. This data system also limits the exploitation for clinical research, care management and decision-making. Furthermore, the lack of standardization limited the integration of data within broader health systems, thus hindering its secure and reusable sharing.
To address this challenge, we undertook the implementation of a complete ETL pipeline aligned with the OMOP Common Data Model (CDM) version 5.4, an internationally recognized framework for standardizing health data. Our objective was to transform, standardize, and integrate patient data from the pulmonology department of this hospital into a database compliant with the FAIR (Findable, Accessible, Interoperable, Reusable) principles to facilitate their reuse for research, clinical management, and patient monitoring. This approach aimed to improve the quality of hospital data, strengthen its interoperability with other systems, and lay the foundation for advanced use in data science for healthcare.
We made use of patient level dataset on Respiratory illness, which included more than 120 variables covering a wide range of clinical and administrative variables such as sociodemographic data, medical history, clinical signs and symptoms, laboratory results, final and secondary diagnoses, and medical observations, as well as information on hospital stays. This data was extracted from datasheets, medical records, and papers, presenting varied forms, heterogeneous levels of completeness, and missing observations.
To standardize this data, we made use of a number of OHDSI tools such as: WhiteRabbit for data profiling, USAGI for vocabulary mapping from our source vocabularies to OMOP standardized concepts/vocabularies and Rabbit in a Hat for data mapping of the source tables to the standard OMOP CDM tables, including: Person, drug_exposure, Measurement, visit_occurrence, condition_occurrence, and observation. The concepts used in the mappings were derived from SNOMED, LOINC, and RxNORM vocabularies, while integrating adaptations specific to the local context.
The ETL pipeline of this data was based on SQL skeleton files exported from Rabbit in Hat after the data mapping. These scripts were then customized using pgAdmin interface for PostgreSQL. After creating the OMOP tables using scripts from OHDSI's GitHub, we loaded our transformed data into the OMOP PostgreSQL databases, structuring them in accordance with the model.
To ensure the quality and compliance of our standardized database, we used the Achilles tool (developed in R), which automatically checked the completeness, conformance, and plausibility of the transformed data. This tool achieved an overall data quality score of 97%, thus attesting to the reliability of the ETL pipeline and the robustness of the database.
Standardizing Respiratory illness data to OMOP CDM in the African context is a novel and promising field that will set pace for collaboration, data sharing and interoperability across health systems. This work made it possible to establish a standardized database that meets all FAIR requirements and is ready to be used for analysis, clinical research, and data science, thus opening up a perspective for decision-makers who can benefit from data-driven decisions to improve pulmonology care practices in Cameroon.

315

Pres session 2

316

Presentations Session 2: Data and Research & Data Science and Data Analysis