Contribution ID: **95** Type: **Session**

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

Thursday 16 October 2025 11:00 (1h 30m)

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 ELIXIR[2]—a European research infrastructure, bringing together life science resources across over 20 member countries. Part of ELIXIR's activities[3] 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 platform[4] 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/19c7CrCE2sILHnbI7i5DiaZ7fD1iYIOBnS
D-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

Primary authors: GOBEILL, Julien (SIB Text Mining group, Swiss Institute of Bioinformatics); Dr HARRISON, Melissa (EMBL's European Bioinformatics Institute (EMBL-EBI)); Dr RUCH, Patrick (SIB Text Mining group, Swiss Institute of Bioinformatics); NYBERG ÅKERSTRÖM, Wolmar

Presenters: GOBEILL, Julien (SIB Text Mining group, Swiss Institute of Bioinformatics); NYBERG ÅKERSTRÖM, Wolmar

Track Classification: SciDataCon Persistent Themes: Open Data, FAIR Data, Innovation, Industry and Development