Contribution ID: **181** Type: **Presentation**

Co-Designing AI Readiness: CODATA's Call to the Global Data Community

Tuesday 14 October 2025 12:36 (11 minutes)

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

Primary authors: KIRKPATRICK, Christine (San Diego Supercomputer Center / CODATA); CROSAS, Mercè (Barcelona Supercomputing Center); HODSON, Simon (CODATA); MCEACHERN, Steven (UK Data Service); CHUANG, Tyng-Ruey (Academia Sinica, Taiwan)

Presenters: KIRKPATRICK, Christine (San Diego Supercomputer Center / CODATA); CROSAS, Mercè (Barcelona Supercomputing Center); HODSON, Simon (CODATA); CHUANG, Tyng-Ruey (Academia Sinica, Taiwan)

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

Track Classification: SciDataCon2025 Specific Themes: Rigorous, responsible and reproducible science in the era of FAIR data and AI