Contribution ID: **214** Type: **Session**

AI for Metadata Enhancement, Metadata for AI Readiness: how do we ensure a virtuous rather than a vicious circle?

Tuesday 14 October 2025 16:00 (1h 30m)

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

Primary authors: Mr GREGORY, Arofan (CODATA); Dr GONZALEZ, Cristina (SenScienceAI); Dr LUNGLEY, Deirdre (UKDS); Mr FILS, Doug (Consultant / CODATA); Dr CERON, Isabel (Australian Academy of Social Sciences); Dr YE, Jieping (GeoGPT / Zhejiang Lab); Dr DRUCKEN, Kelsey (ANU); Prof. CROSAS, Mercè (BSC / CODATA); Mr HEUS, Pascal (Postman / CODATA); Dr FARRINGTON, Rebecca (AuScope); Dr HILL, Sean (SenScienceAI); HODSON, Simon (CODATA); Dr RICHARD, Stephen (Consultant / CODATA); Mr TIKHONOV, Vyacheslav (DANS); Dr XIAO, Yitian (GeoGPT / Zhejiang Lab)

Presenters: Dr GONZALEZ, Cristina (SenScienceAI); Dr LUNGLEY, Deirdre (UKDS); Mr FILS, Doug (Consultant / CODATA); Dr YE, Jieping (GeoGPT / Zhejiang Lab); Mr HEUS, Pascal (Postman / CODATA); Dr FARRINGTON, Rebecca (AuScope); Dr HILL, Sean (SenScienceAI); HODSON, Simon (CODATA); Mr TIKHONOV, Vyacheslav (DANS)

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