

CI Compass Fellows Develop Knowledge Graph to Enable New Uses for MagLab Metadata

David S. Butcher¹, Diya Adhikari², Veronika Saiadian³
1. National High Magnetic Field Laboratory, 2. Caldwell University, 3. University of Southern California



Funding Grants: K. M. Amm (NSF DMR-2128556); E. Deelman (NSF OAC-2127548)

In recent years, the value of metadata for research artifacts such as datasets and publications has become increasingly clear. Findable, Accessible, Interoperable, and Reusable (FAIR) datasets that are annotated with rich metadata can facilitate tracking of their provenance and impact, ensure the reproducibility of research, and enable novel artificial intelligence and machine learning (AI/ML) workflows to generate more scientific value.¹ To accomplish this, metadata must be conceptually linked to form knowledge graphs – networks of nodes representing specific pieces of knowledge linked by various semantic relationships. However, establishing these links can be challenging, as the necessary information is siloed in separate locations and is not standardized.

To develop new methods for extraction, standardization, and visualization of metadata in the form of knowledge graphs, and to explore their applications, the MagLab’s Center for FAIR and Open Science worked with the NSF Cyberinfrastructure Center of Excellence, CI Compass, to bring two summer fellows to the lab. They created PRISM (Provenance-Rich Index of Scientific Metadata), an extraction and formatting pipeline (Figure 1) that synthesizes instrumental parameters extracted from raw data files from the Ion Cyclotron Resonance facility, metadata extracted from publications using a large language model, and information from open databases such as Crossref (for citations) and ORCID (for researchers) into a graph database managed and visualized using Neo4j (Figure 2). Utilizing this graph, it is possible to trace the connections between specific instruments, their data acquisition parameters, the files they acquired, the public datasets the files belong to, the researchers who published them, and the journal articles that reference them. In this way, the full provenance of each data file can be known and the experimental context can be reproduced.

In the future, detailed knowledge graphs will be used for a broad variety of applications, including tracking the usage of MagLab instruments and datasets across the scientific ecosystem. Great potential lies in the fusion of knowledge graphs with AI agents, enabling the use of semantic reasoning capabilities, retrieval augmented generation to improve results and suppress “hallucinations”, and enable new methods of scientific discovery.²

Citation: (1) Greenberg, J.; Pepper, J.; Zhao, X.; Marciano, R.; Breen, D.; An, Y. The Metadata Ecosystem and AI: Enabling FAIR and AI-Ready Data. *AI Magazine* 2026, 47 (2), e70060. <https://doi.org/10.1002/aaai.70060>. (2) Presutti, V.; Motta, E.; Sabou, M. Opportunities for Knowledge Graphs in the AI Landscape — An Application-Centric Perspective. *Journal of Web Semantics* 2025, 100867. <https://doi.org/10.1016/j.websem.2025.100867>.

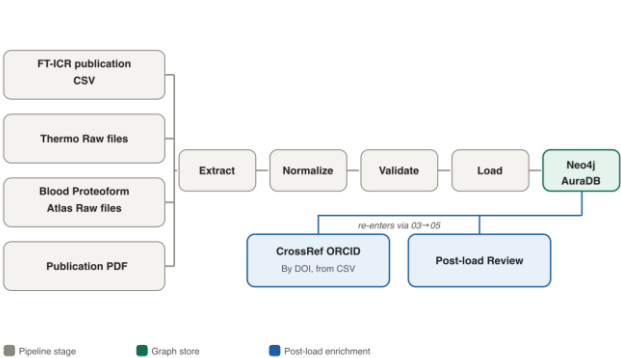


Figure 1. Data extraction and formatting pipeline for graph visualization.



Figure 2. The 13 most-published researchers and the 32 papers linking them (≥4 shared coauthors each).

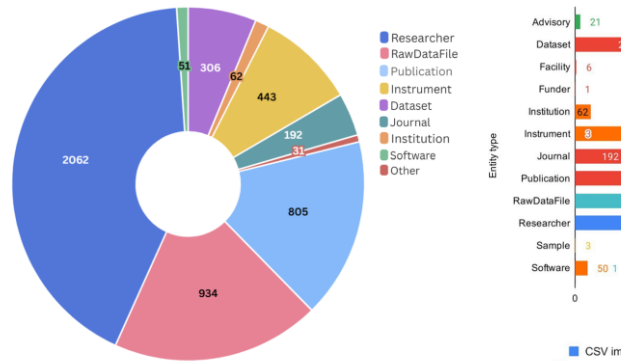


Figure 3. Composition of the knowledge graph.

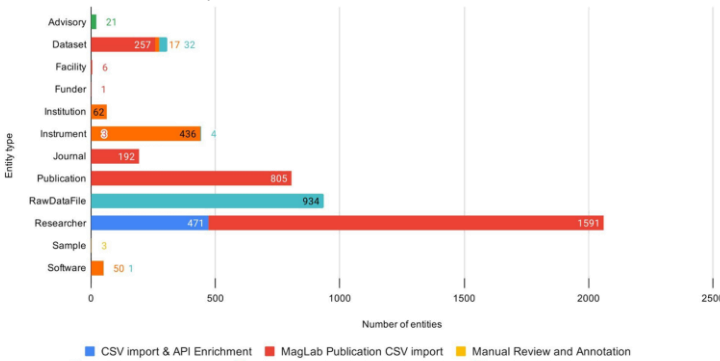


Figure 4. Knowledge graph node provenance by data acquisition workflow.