DCPAPERS

Future-proofing Metadata at Los Alamos National Laboratory

Mark Offtermatt[1,*,†], Megan Rose Kilidjian[1,†] and Laura McGuiness[1,†]

[1] *Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM 87545, United States*

Abstract

The Los Alamos National Laboratory's National Security Research Center is home to a distinct collection of technical documentation spanning the history of the laboratory. Recent current events have created a need for expedited scientific research. Access to information has been stymied by a lack of coherent metadata, generated by heavily siloed working groups managing information across disparate systems. The laboratory's Metadata Initiative, made up of a group of library science and metadata professionals, assembled in 2023 to begin standardizing metadata across the complex. The initiative is engaged in several projects that will promote consistency across systems, driven by the adoption of the Dublin Core metadata standard, and additional in-house development of tools designed to assist both staff and user alike. Our approach, as it is presented here, may be used as a guide for other institutions working to overcome similar challenges.

Keywords

Dublin Core, metadata, schemas, archives

1. History

Los Alamos National Laboratory's National Security Research Center (NSRC) serves as the laboratory's classified library. It maintains an extensive collection of unique, technical documentation spanning the history and evolution of the site. Tracing its lineage back to the technical library formed by J. Robert Oppenheimer in 1943, the NSRC contains one-of-a-kind research materials that support stockpile stewardship and help safeguard the United States's deterrence capabilities. Due to a history of compartmentalization, the laboratory exists as a place of data siloization. With over 15,000 employees, the lab generates an abundance of information every day. Groups working towards the same ends struggle to share the information they generate effectively due to a lack of communication, standardization, and tools.

For much of the lab's history, this problem has been exacerbated by a lack of metadata capture. Often, teams of scientists and technical personnel have been responsible for organizing

* Corresponding author.
† These authors contributed equally.
mofftermatt@lanl.gov (M. Offtermatt); mkilidjian@lanl.gov (M. R. Kilidjian); lauraleemcg@lanl.gov (L. McGuiness)
0009-0000-1323-0903 (M. Offtermatt); 0009-0006-2860-3390 (M. R. Kilidjian); 0009-0002-5232-2557 (L. McGuiness)

https://doi.org/10.23106/dcmi.952460454

the data they create. Standards and practices utilized by professionals in the field of information science have not usually been followed. This inconsistency has led to numerous problems with both finding and retaining information within the laboratory. The lack of useable metadata is made more difficult by the size of the NSRC's collection, which numbers in the tens of millions.

2. Metadata Initiative

In spring 2023, it was determined that improvements to standardization of metadata were necessary. The Metadata Initiative was formed with the objective to identify current gaps in data structure and content standardization, and improve the findability, accessibility, interoperability, and reusability (FAIR) of NSRC collections.[1] Previous research in information organization, such as Yarmey and Baker's paper highlighting the importance of standardization for coordination and cooperation, the Metadata Initiative received support to implement a metadata schema and develop best practices that would be applied across the NSRC.[2]

Since the NSRC consists of multiple teams focused on paper digitization, multimedia digitization, and collections management, the first goal of the Metadata Initiative was to determine metadata capture disparities between the teams. This type of assessment is a necessary first step with expansive special and archival collections, as it has been shown that digitized materials in large-scale collections often lack consistent, high-quality descriptive metadata, making them poorly suited for long-term use.[3,4] The NSRC's teams capture descriptive metadata on multiple spreadsheets. These have been created over time by different individuals working with those collections and formats. This has led to many inconsistencies.

The Metadata Initiative began correcting this practice by creating a series of tables that captured the language utilized on each spreadsheet, as well as the fields used in the corresponding software that the metadata was entered in to. This effectively captured the totality of the discrepancies within the NSRC, setting the groundwork for decisions to be made regarding best practices and standardization. Due to its flexibility, Dublin Core was selected as the framework for standardization and the basis for organizational policies and guidelines.

3. Next Steps

The implementation of Dublin Core has made a critical contribution to the standardization of metadata within Los Alamos National Laboratory's Weapons Program. The Metadata Initiative has created a data dictionary that crosswalks Dublin Core into other disparate databases and metadata schemas. The Controlled Vocabulary Committee is working on revising and updating unique, technical glossaries and thesauri to facilitate better search and retrieval. We are also exploring opportunities for collaboration with an ontology team with the hope of eventually developing knowledge graphs. A Metadata Committee is also being created to govern standardization and use. Finally, the NSRC is beginning to collaborate with colleagues at other Department of Energy laboratories to create custom qualifiers in Dublin Core. This will allow Los Alamos National Lab to meet enterprise-specific metadata needs when sharing resources between laboratories. However, for this to function properly at such a high level, there is also a need to establish custom namespaces on the back-end to facilitate validation for any structured metadata generated in extended markup or resource description framework.[5] In theory, this consistency will benefit the use of large language models, resulting in more meaningful search results across the range of information systems at its disposal.

https://doi.org/10.23106/dcmi.952460454

Acknowledgements

The authors would like to thank Charlie Nakhleh, Scott Doebling, Jason Kritter, Nanette Mayfield, Julie Maze, Elaine Rodriguez, Caroline Blackburn, Patty Templeton, Heath Robinson, and everyone involved with Los Alamos National Lab's Metadata Initiative for their support, without which this endeavor would not be possible.

References

[1] M. Wilkinson, M. Dumontier, I. Aalbersberg, et al., The FAIR guiding principles for scientific data management and stewardship, Sci Data 3 (2016). doi:10.1038/sdata.2016.18.

[2] L. Yarmey, K. S. Baker, Towards standardization: A participatory framework for scientific standard-making, Journal of Digital Curation 8 (2013) 157-72. doi: 10.2218/ijdc.v8i1.252.

[3] K. E. Martin, Marrying local metadata needs with accepted standards: The creation of a data dictionary at the University of Illinois at Chicago, Journal of Library Metadata 11 (2011) 33-50. doi: 10.1080/19386389.2011.545006.

[4] D. Salo, Retooling libraries for the data challenge, Ariadne 64 (2010). URL: http://www.ariadne.ac.uk/issue64/salo/.

[5] C. Mathieu, Practical application of the Dublin Core standard for enterprise metadata management, Bulletin of the Association for Information Science & Technology 43 (2017) 29-34.

https://doi.org/10.23106/dcmi.952460454