DCPAPERS

Metadata provenance and explainability

Inkyung Choi[1,*], Jessica Yi-Yun Cheng[2], Josh Falconer[3] and Anne Washington[1]

[1]*OCLC Global Headquarters, 6565 Kilgour Pl, Dublin, OH 43017*

[2]*Rutgers, the State University of New Jersey, 4 Huntington St., New Brunswick, NJ, USA 08901*

[3]*Bloomberg, 100 Business Park Drive Skillman, NJ 08558*

Abstract

From its role in record integrity to its significance in digital trust, provenance remains indispensable when navigating in these rapidly evolving information landscapes. This recognition prompts a reevaluation of provenance's role, particularly in the context of metadata preservation.

Keywords

metadata provenance, explainability, entity management, KOS, bibliographic classification, ontology,

1. Introduction

In recent years, standards such as Preservation Metadata: Implementation Strategies (PREMIS) and the W3C PROV standard have emerged as key frameworks, indicating a growing awareness of the need to represent provenance in digital preservation and web resources. Li and Sugimoto [1] advocate for extending the practice of provenance description to encompass metadata instances, highlighting the inherent digital nature of metadata and emphasizing the distinction between preserving metadata and preserving digital objects.

According to the W3C Provenance Working Group, provenance serves as a comprehensive record, characterizing the involvement of agents, entities, and activities in the creation, influence, or delivery of data or objects. Given its usefulness in decision-making, provenance plays a key role in assessing information trustworthiness and reconstructing data generation processes [2]. Toward these ends, metadata must be updated continually for consistent interpretability. A reliable evidential record of the provenance of such metadata enables its ongoing maintainability. This may be achieved to some extent by managing metadata vocabularies using essential services such as the Open Metadata Registry (OMR). However, even when a record of provenance is kept, challenges persist in maintaining interpretability in machine processing over time [3]. This in turn can affect explainability for humans to the extent that machine interpretability is required for reconstructing the data generation process.

Despite the recognition of metadata provenance's importance, research in this area remains underdeveloped. This is evident even as diverse standards proliferate: implementations become

DCMI-2024 International Conference on Dublin Core and Metadata Applications

*Corresponding author.

choii@oclc.org (I. Choi); yiyun.cheng@rutgers.edu (J. Y. Cheng); jfalconer16@bloomberg.net (J. Falconer); washinga@oclc.org (A. Washington)

0000-0001-5048-8516 (I. Choi); 0000-0001-6123-7595 (J. Y. Cheng); 0009-0006-7663-1177 (J. Falconer); 0000-0002-2955-4352 (A. Washington)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952454951

more complicated, and metadata management processes become increasingly dependent on human intervention. Moreover, the widespread adoption of machine-consumable metadata underscores the critical need for metadata that is both interpretable and explainable. For example, in domains such as finance, where the demand for explainable AI is high, discussions have focused on leveraging metadata as fundamental principles for achieving transparent data governance [4]. This highlights the importance of ensuring clear and understandable data lineage, which in turn enables robust metadata provenance practice.

In the context of these problems, our panel seeks to initiate a dialogue on metadata provenance, emphasizing its role in transparency and explainability. We aim to explore these concepts from various perspectives with interrelated use cases, including identifying and documenting the reasons and motivations that justify changes. This exploration will help us to clarify the diverse dimensions of metadata provenance and address pertinent issues such as the types and scopes of provenance. We hope to foster a deeper understanding about the implications of metadata provenance on the diverse information ecosystems.

2. Panel Structure

Opening up the dialogue, first, panelists will discuss the role of provenance within knowledge organization systems (KOSs) and metadata by showcasing relevant cases. Anne Washington will focus on the OCLC Meridian Entity Changes API's role in robust provenance tracking within entity management editors. Jessica Cheng and Inkyung Choi will present the modeling practices in their current research project, focusing on capturing the Dewey Decimal Classification (DDC) editing logs to enrich data context, transcending the current scope of DDC. Josh Falconer will speak about recognizing patterns of motivation behind ontology changes, underscoring the need for transparent documentation of provenance information to ensure comprehensive evaluation and understanding.

We will have an interactive Q&A session to address questions such as:

- How do you ensure the integrity and reliability of metadata provenance throughout the data lifecycle?
- What are some of the ways that explainability is impacted differently when metadata provenance is manually documented (e.g., editorial notes in KOSs) versus when it is automatically generated (e.g., machine-created logs)?
- How do advancements in generative AI technologies affect the explainability, transparency, and accountability of metadata provenance?

Join us as we navigate the evolving landscape of metadata for its expanding role in providing transparent, trustworthy, and effective representation of data, information, and knowledge. We hope to initiate collaborative discussions through this panel and reach a collective understanding and application of this vital concept.

3. Moderator/Speaker Bios

Inkyung Choi: an Associate Research Scientist at OCLC Research, specializing in data science, metadata research. Her work has been published in JASIST, Journal of Documentation,

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952454951

Knowledge Organization Journal. She focuses on embedding contextual references within classification systems, improving information retrieval and advances the field of knowledge organization and metadata management. In this panel, Dr. Choi will moderate the discussions and share her ongoing project about enhancing the provenance descriptions with editorial notes from OCLC in the Dewey Decimal Systems.

Jessica Yi-Yun Cheng: an Assistant Professor at the School of Communication and Information, Rutgers, the State University of New Jersey. Her research and teaching focuses on how systems of organization affect the ways data are represented. In her work, she resolves interoperability problems in taxonomies, metadata, ontologies, and other knowledge organization systems (KOSs) in biodiversity and geographic contexts. Her work has been published in JASIST, Journal of Documentation, Knowledge Organization Journal, and she has co-authored a book about provenance metadata.

Anne Washington: a Product Analyst on the Metadata Services Team at OCLC focusing on linked data applications and services. She has over a decade of experience in academic libraries, museums, and archives; with previous roles including Metadata Services Coordinator at the University of Houston, and Metadata Librarian at the University of Virginia. Her work has been published in the Journal of Library Metadata, Library Resources Technical Services, and she co-authored a book on library linked data. Anne received her MLIS at the University of Wisconsin-Milwaukee.

Josh Falconer: a Senior Ontologist at Bloomberg in Princeton, New Jersey. Previously, he was an ontologist at Indeed. For more than a decade, he served in various bibliographic metadata cataloging roles, with a focus on manuscript library collections from the Middle East and North Africa. He has cataloged thousands of Syriac, Arabic, Garshuni, and Ethiopic manuscripts from diverse collections including the Hill Museum Manuscript Library and the Library of Congress. His recent research interests converge on semantics, knowledge representation, and living systems. He has earned several advanced degrees in a wide range of intersecting disciplines, including linguistics, philosophy, and information science.

References

[1] C. Li, S. Sugimoto, Provenance description of metadata application profiles for long-term maintenance of metadata schemas, Journal of Documentation 74 (2018) 36–61. doi:10.1108/JD-03-2017-0042.

[2] Y. Gil, S. Miles, PROV model primer, 2013. URL: https://www.w3.org/TR/prov-primer/.

[3] C. Li, S. Sugimoto, Provenance description of metadata vocabularies for the long-term maintenance of metadata, Journal of Data and Information Science 2 (2017) 41–55. doi:10.1515/jdis-2017-0007.

[4] J. Chen, Ontology drift is a challenge for explainable data governance, 2021. URL: http://arxiv.org/abs/2108.05401. arXiv:2108.05401 [cs].

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952454951