EDITORIAL

Metadata as Data Intelligence

Jane Greenberg[1], **Mingfang Wu**[2], **Wei Liu**[3], **Fenghong Liu**[4,5†]

[1]Metadata Research Center, College of Computing and Informatics, Drexel University, 3675 Market St 10th floor,Philadelphia, PA 19104, USA

[2]Australian Research Data Commons, Melbourne, Victoria 3145, Australia

[3]Shanghai Library (Institute of Scientific and Technical Information of Shanghai), No. 1555 Huaihai Middle Road Xuhui District,
Shanghai 200031, China

[4]Department of Library, Information, and Archive, School of Economic Management, University of Chinese Academy of Sciences,
33 Beisihuanxilu, Haidian District, Beijing 100190, China

[5]Data Intelligence, National Science Library, Chinese Academy of Sciences, 33 Beisihuanxilu, Haidian District, Beijing 100190, China

Citation: Greenberg, J., Wu, M.F., Liu, W., et al.: Metadata as Data Intelligence. Data Intelligence 5 (2023). doi: 10.1162/dint_e_00212

1. INTRODUCTION

Metadata, as a type of data, describes content, provides context, documents transactions, and situates data. Interest in metadata has steadily grown over the last several decades, motivated initially by the increase in digital information, open access, early data sharing policies, and interoperability goals. This foundation has accelerated in more recent times, due to the increase in research data management policies and advances in AI. Specific to research data management, one of the larger factors has been the global adoption of the FAIR (findable, accessible, interoperable, and reusable) data principles [1, 2], which are highly metadata-driven. Additionally, researchers across nearly every domain are interested in leveraging metadata for machine learning and other AI applications. The accelerated interest in metadata expands across other communities as well. For example, industry seeks metadata to meet company goals; and users of information systems and social computing applications wish to know how their metadata is being used and demand greater control of who has access to their data and metadata. All of these developments underscore the fact that metadata is intelligent data, or what Riley has called value added data [3]. Overall, this intense and growing interest in metadata helps to frame the contributions included in this special issue of *Data Intelligence*.

This special issue of *Data Intelligence* includes a collection of 14 original articles covering metadata related research, practice, and theory. The contributions as a whole, include work complied by over 50 authors from eleven countries, including Australia, Canada, China, France, Germany, Italy, the Netherlands,

[†] Corresponding author: Fenghong Liu (E-mail: liufh@mail.las.ac.cn; ORCID: 0000-0002-3633-1464).

Portugal, Spain, the United Kingdoms, and the United States. Contributing authors are from range of organizations, national research data programs, such as the Australian Research Data Commons, academic and research institutions, such as the Chinese Academy of Sciences; government agencies and industry such as the United States Bureau of Labor Statistics among other national agencies. Among some of the overriding themes, contributions address FAIR principles in relation to metadata; metadata tools, practice or polices, innovative ideas, and theoretical approaches. The majority of contributions are arranged under "research" and "practice and implementation", and with three final contributions covering "vision and theory".

2. RESEARCH ARTICLES

Research in the study of any phenomena confirms a level of maturity. The rich collection of research articles in this special issue confirm that metadata work has expanded well beyond scheme development and implementation, and that metadata is a significant research topic [4, 5].

The series of research articles is initiated with a piece entitled, Improving Domain Repository Connectivity [6]. This contribution explores the notion of a connectivity metric, which is applied it to datasets collected and papers published by members of the UNAVCO community. The author explains that "as community members contribute to multiple datasets and articles, identifiers for them, once found, can be used multiple times". Identifiers found in DataCite and Crossref metadata that are shared through UNAVCO DataCite metadata can increase connectivity from less than 10% to nearly 50% for people and organizations.

Two of the research contributions interconnect with FAIR. First, FAIRification of Scanning Tunneling Microscopy [7] focuses on data management practices and services for making FAIR compliant a scientific archive of Scanning Tunneling Microscopy (STM) images. The authors report on a metadata database that includes metadata extracted from instruments and each image, which have been enriched via human annotation, machine learning techniques, and instrument metadata filtering. Additionally, the W3C PROV standard was explored for STM image.

A second work, FAIR Data and Metadata: GNSS Precise Positioning User Perspective [8], presents an analysis of current GNSS users' requirements in various application sectors on the way data, metadata and services are provided. We engaged with GNSS stakeholders to validate our findings and to gain understanding on their perception of the FAIR principles. Authors indicate that results confirm FAIR GNSS data and services are important for this community and have have had an impact standard compliant GNSS community metadata enabling FAIR GNSS data and service delivery for both humans and the machines.

Another research piece includes, Research on Intelligent Organization and Application of Multi-source Heterogeneous Knowledge Resources for Energy Internet [9] focuses on improving the informaionization and intelligence of the energy Internet industry for enhancing the capability of knowledge services. The authors propose methods to synthesis and transform the original multiple, heterogeneous knowledge resources of the State Grid into a unified and well-organized knowledge system. The effectiveness of the proposed methods are demonstrated with knowledge resources in the field of human resources of the State Grid.

The collection of research articles also include, An analysis of crosswalks from research data schemas to Schema.org [10]. This work presents an extensive analysis of crosswalks from fourteen research data schemas to Schema.org. The analysis indicates that most descriptive metadata are interoperable among the schemas, the most inconsistent mapping is the rights metadata, and a large gap exists in the structural metadata and controlled vocabularies to specify various property values. The analysis and collated crosswalks can serve as a reference for data repositories when they develop crosswalks from their own schemas to Schema.org, and provide the research data community a benchmark of structured metadata implementation.

The research cluster also includes a contribution that underscores metadata as data intelligence with attention to AI/ML methods. The work entitled Automated metadata annotation: What is and is not possible with machine learning [11] presents three use cases that demonstrate the possibility of utilizing AI/ML models in improving subject indexing of culture or data catalogs, and it requires bringing process, technology and interdisciplinary team together to achieve quality of automated subject.

The last contribution in the research cluster: Provenance documentation to enable explainable and trustworthy AI: A literature review [12] discusses the importance of capturing and providing provenance information within the context of running AI/ML models, for making AI/ML results explainable, trustworthy and reproducible by capturing provenance metadata about each step of the AI process (e.g. data, AI models, software source code for data preparation and executing models).

3. PRACTICE AND IMPLEMENTATION

The practice and implementation articles included in this special issue report on innovative approaches and developing technologies. Three pieces contributions that target practices focus on FAIR principles. The FAIR (data)meta principles provide guidance to the documentation of metadata so that data can be Findable, Accessible, Interoperable and Reusable, by both human and machine. Making metadata fair is a progressive process, it requires to gauge gaps and set up goals for improvement, and methods and tools to assist the process. Two of the FAIR contributions are about tools relating to data repositories "FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication" [13] and "The FAIR Data Point: Interfaces and Toolings" [14] introduce the FAIR Data Point (FDP)—a general-purpose metadata repository that follows the DCAT schema and has been implemented by following the FAIR principles. The first paper introduces the software architecture, core components and features of the FAIR Data Point (FDP), and the second paper describes interfaces and tools for implementing FDP and facilitating the uptakes and utilities of FDP. The two papers will benefit those who want to adopt a metadata repository solution or existing metadata repositories for extending their repository functionalities for publishing semantically-rich and machine-actionable metadata by following the FAIR principles.

The third practice paper addressing FAIR, FAIREST: A Framework for Assessing Research Repositories [15], introduces a set of metrics for assessing and selecting solutions for creating digital repositories for research artifacts. The metrics are built on the FAIR principles, Engagement, Social Connections and Trust (this FAIREST). The paper also applied metrics for an assessment of 11 widespread solutions, with the goal to provide an overview of the current landscape of research data repository solutions, identifying gaps and research challenges to be addressed.

The last contribution in the practice and implementation contribution, Building Community Consensus for Scientific Metadata with YAMZ [16], introduces YAMZ (Yet Another Metadata Zoo). YAMZ was developed to help address the challenges with the formal metadata standardization process. The paper presents an exploratory demonstration of YAMZ within an academic research lab, where there is a need to standardize metadata to help with data management activities, but researchers lack of time and metadata expertise to proceed through the formal standardization process.

4. VISIONARY AND THEORY

Three contributions cover vision and theory. The work, Achieving Transparency: A Metadata Perspective [17], discusses what information should be captured in metadata (schema) and in consistent way (technical specification) to ensure metadata quality and transparency of data; in order to communicate better what data mean and why they should be trusted, within the context of providing datasets from the government and government agencies.

The work Continuous Metadata in Continuous Integration, Stream Processing and Enterprise Data Ops [18] argues that metadata is continuous in many real data context, thus one-off metadata collection may be inadequate for future analysis. Based on the review of some current tools in specifying, capturing and consuming metadata; the author suggests features and design patterns for future cloud native software, which could enable streamed metadata to power real time data fusion or fine turn automated reasoning through real time ontology updates.

Finally, a contribution entitled, Metadata as a Methodological Commons: From Aboutness Description to Cognitive Modeling [19], discusses the requirement and feasibility for semantic coding and cognitive metadata modeling, as the rise of huge volume of labeled data and ChatGPT, as well as the availability of emerging technologies (e.g Web 3.0, AI/ML, knowledge graph).

5. CONCLUSION

Metadata is often seen as a practical topic, not worthy of research and, yet, it is a topic that is central to data driven research [18]. This issue of *Data Intelligence* underscores the significance of metadata as a research worthy and critical topic in advancing our data infrastructure and achieving greater ‘data intelligence’. In closing, this issue presents a unique collection of articles and that confirms the importance of metadata research, practice and implementation, and visionary and theory as we aim to collectively advance our data infrastructure across all information sectors.

REFERENCES

[1] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(1), 1–9 (2016). http://www.nature.com/articles/sdata201618

[2] Jacobsen, A., de Miranda Azevedo, R., Juty, N. et al.: FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2(1–2), 10–29 (2020). doi: https://doi.org/10.1162/dint_r_00024

[3] Riley, J.: Understanding Metadata: What is Metadata, and What is it For?: A Primer. Front Matter. NISO (2017)

[4] Leipzig, J., Nüst, D., Tapley, C., et al.: The role of metadata in reproducible computational research. Patterns 2(9), 100322 (2021). https://doi.org/10.1016/j.patter.2021.100322

[5] Greenberg, J.: Big metadata, smart metadata, and metadata capital: Toward greater synergy between data science and metadata. Journal of Data and Information Science 2(3), 19 (2017)

[6] Habermann, T.: Improving Domain Repository Connectivity. Data Intelligence 5(1), 6–26 (2023). doi: 10.1162/dint_a_00120

[7] Rodani, T., Osmenaj, E., Cazzaniga, A., Panighel, M., Africh, C., Cozzini, S.: Towards the FAIRification of Scanning Tunneling Microscopy images. Data Intelligence 5(1), 27–42 (2023). doi: 10.1162/dint_a_00164

[8] Ivánová, I., Keenan, R., Marshall, C., et al.: FAIR data and metadata: GNSS precise positioning user perspective. Data Intelligence 5(1), 43–74 (2023). doi: 10.1162/dint_a_00185

[9] Wang, Y.X., Luo, L.Q., Li, G.J.: Research on Intelligent Organization and Application of Multi-source Heterogeneous Knowledge Resources for Energy Internet. Data Intelligence 5(1), 75–99 (2023). doi: dint_a_00158

[10] Wu, M.F., Richard, S.M., Verhey, C., et al.: An analysis of crosswalks from research data schemas to Schema.org. Data Intelligence 5(1), 100–121 (2023). doi: 10.1162/dint_a_00186

[11] Wu, M.F., Brandhorst, H., Marinescu, M.-C., et al.: Automated metadata annotation: What is and is not possible with machine learning. Data Intelligence 5(1), 122–138 (2023). doi: 10.1162/dint_a_00162

[12] Kale, A., Nguyen, T., Harris, Jr., et al.: Provenance documentation to enable explainable and trustworthy AI: A literature review. Data Intelligence 5(1), 139–162 (2023). doi: 10.1162/dint_a_00119

[13] da Silva Santos, L.O.B., Burger, K., Kaliyaperumal, R., et al.: FAIR Data Point: A FAIR-Oriented approach for metadata publication. Data Intelligence 5(1), 163–183 (2023). doi: 10.1162/dint_a_00160

[14] Mohammed Benhamed, O., Burger, K., Kaliyaperumal, R., et al.: The FAIR Data Point: Interfaces and Tooling. Data Intelligence 5(1), 184–201 (2023). doi: 10.1162/dint_a_00161

[15] Mathieu, A., Fabian, K., Daniela, O., et al.: FAIREST: A Framework for Assessing Research Repositories. Data Intelligence 5(1), 202–241 (2023). doi: 10.1162/dint_a_00159

[16] Greenberg, J., McClellan, S., Rauch C., et al.: Building community consensus for scientific metadata with YAMZ. Data Intelligence 5(1), 242–260 (2023). doi: 10.1162/dint_e_00211

[17] Gillman, D.: Achieving transparency: a metadata perspective. Data Intelligence 5(1), 261–274 (2023). doi: 10.1162/dint_a_00188

[18] Underwood, M.: Continuous metadata in continuous integration, stream processing and enterprise DataOps. Data Intelligence 5(1), 275–288 (2023). doi: 10.1162/dint_a_00193

[19] Liu, W., Fu, Y.M., Liu, Q.Q.: Metadata as a Methodological Commons: From Aboutness Description to Cognitive Modeling. Data Intelligence 5(1), 289–302 (2023). doi: doi.org/10.1162/dint_a_00189