Research Ideas and Outcomes 6: e53916
doi: 10.3897/rio.6.e53916



Method

# Methods & Proposal for Metadata Guiding Principles for Scholarly Communications

Kathryn Kaiser[‡], Jennifer Kemp[§], Laura Paglione[|], Howard Ratner[¶], David Schott[#], Helen Williams[¤]

‡ University of Alabama at Birmingham, Birmingham, United States of America
§ Crossref, Lynnfield, MA, United States of America
| Metadata 2020, New York, NY, United States of America
¶ CHORUS, Staten Island, New York, United States of America
# Copyright Clearance Center, Danvers, MA, United States of America
¤ The London School of Economics and Political Science, Library, London, United Kingdom

Corresponding author: Laura Paglione (lpaglione@metadata2020.org)

Reviewable v1

Received: 04 May 2020 | Published: 05 May 2020

Citation: Kaiser K, Kemp J, Paglione L, Ratner H, Schott D, Williams H (2020) Methods & Proposal for Metadata Guiding Principles for Scholarly Communications. Research Ideas and Outcomes 6: e53916.
https://doi.org/10.3897/rio.6.e53916

## Abstract

This article describes an international community-based effort to create metadata guiding principles for adopting and using richer metadata and advancing its application in scholarly communications. These principles can facilitate the dissemination, discoverability and use/reuse of many types of research and scholarly outputs. While much work remains to be done, these principles serve as a starting point for the evolution of processes that span communities including publishers, researchers, scholars, authors and other creators, librarians, curators, custodians, and consumers of scholarly works.

These aspirational Metadata 2020 Principles are designed to encompass the needs of our entire community while ensuring thoughtful, purposeful, and reusable metadata resources. They provide a framework for all of us to be good metadata citizens. They also provide a foundation for considering related work from Metadata 2020 and must be interpreted within the legal and practical context in which we operate. They are intended to guide the broadest possible cross-section of our community in improving research communications, publishing, and discoverability.

## Keywords

Metadata, Best Practices, Principles, Complete, Open, Interoperable, Provenance, Community

## Introduction

"*Metadata is a love note to the future." -- Jason Scott*

Metadata--i.e., data or descriptors about data or another object--is only useful to applications or future users if it is thoughtfully and consistently applied. Electronic generation and management of the scholarly record and scholarship globalization have driven and increased the importance of metadata to the scholarly ecosystem. While initial metadata efforts have created value in discovering and connecting research outputs, historical and current models/processes are not meeting present expectations as scholarly communication needs evolve.

Despite increases in scholarly cross-community collaboration and data sharing, issues in system interoperability and metadata compatibility persist. As a result, processes designed to facilitate, exchange, and reuse metadata can be costly, difficult to apply, and inefficient. In addition, metadata improvements can extend beyond discoverability. When thoughtfully applied, rich metadata should enable people also to discover the ecosystem and context of a work, in addition to the context of a specific final published output.

The purpose of this paper is to report the methodology used to create a set of metadata principles for research outputs and scholarly communications objects. Included are a discussion of the connection of these principles to existing work, assertions of how we hope the principles will be used, and descriptions of how they can serve as a foundation for extending to more concrete activities. To provide some context, it is important to acknowledge prior and ongoing work in this area by other groups; the work presented here is intended to complement and highlight these other efforts. This work considers the social and cultural challenges in changing how we think about and use metadata related to education, research, and scholarship.

## Origins of the work

Initiated by Crossref in 2017, Metadata 2020 expanded quickly to develop into an international community of stakeholders from across scholarly communications. It functions as a collaboration advocating for "richer, connected and reusable, open metadata for all research outputs in order to advance scholarly pursuits for the benefit of society."Metadata 2020 (2020) While many efforts have been made to address challenges in single communities, few have extended solutions that are targeted to be applied across them. Recognizing this situation as a strength, the Metadata 2020 conveners agreed on an agenda driven by the interests and needs of this broad community. This approach allowed

for unrestricted interactions among the subgroups participating in Metadata 2020 as determined by its volunteer community members and needs for additional expertise and review.

Initial community groups were organized by roles and interests as follows: researchers; publishers; librarians; data publishers and repositories; services, platforms and tools; and funders. Through insights gained by exploring each group's use cases, challenges, and opportunities, six projects were developed to explore and address core needs with representatives from all stakeholder groups:

1. Researcher communications
2. Metadata recommendations and element mappings
3. Terms and definitions
4. Incentives for quality improvements
5. **Shared best practices and principles** *(the focus of this report)*
6. Metadata evaluation and guidance

Participants in the Metadata 2020 Principles projects include the authors as well as other contributors from a variety of stakeholder communities who provided valuable contributions throughout. A full list of participants is included in the acknowledgements. Through a grassroots structure and broad stakeholder involvement, the Metadata 2020 projects aim to tackle global issues that need to be addressed in process and quality improvement, including lack of central core metadata principles, best practice and consistent guidance, and lack of interoperability.

This group of diverse stakeholders delivers an initial set of aspirational and foundational metadata principles to guide the broadest possible cross-section of our communities in creating and promoting thoughtful, purposeful and reusable metadata content. In promoting good metadata citizenship, the goal is to elevate and share these principles with a wider audience so that more groups and efforts can help formalize their adoption, use, and evolution.

## How does this movement compare to others?

Many prior institutions and technology movements can inform this and related paradigm shifts. One example is HL7 Health Level Seven International (2007) in the medical informatics industry. This effort arose from the need to facilitate data sharing and clinical data transfer within and among organizations made up of often heterogeneous components that are highly customized with many moving parts and/or disparate formats. This work developed industry standards by creating a convergence of processes and technical schema from a once highly disjointed industry. Facing similar challenges within scholarly metadata, the Metadata 2020 Principles group focused on a set of ideals that represent a convergence of technology and process to aid data sharing, interoperability, and discoverability.

These metadata principles also address key aspects of the fair principles: Findability, Accessibility, Interoperability, and Reusability.GOFair (2016), Wilkinson et al. (2016) As groups begin to apply other similar principles, more global synergy and momentum can be leveraged. Overlap among principles is an indication of their universality. Indeed, because so much related work has already been done, our principles took a high-level, unifying approach with a lay audience in mind. Whether this approach is successful remains to be seen but the need to establish metadata as its own primary output alongside content seems clear as is the need to bring business and strategy people into the conversation. The group benefits from a mix of contributors ranging from those with deep technical metadata knowledge to metadata novices and those in between. Involving both practitioners and non-practitioners in these efforts is a calculated and important element of our methodology.

## How the principles were prepared

Work was led by the co-chairs and organized by a central coordinator who scheduled meetings, recorded minutes, generated documents to capture work and provide for additional collaborations, and organized two in-person workshops in 2018 in New York (September) and London (October). Additional sharing beyond the project group took the form of blog posts, symposia, and presentations at a range of international conferences.

The Metadata 2020 Principles project group discussed at length how to generate a resource (a set of core principles) distinct from other Metadata 2020 efforts that also fit within the overall Metadata 2020 mission in a logical and practical way. Two simple but essential efforts started the work. First, the project team collected and reviewed as many existing best practices as possible, publishing the list of them as the initiative's first resource. Next, the group crowdsourced suggestions for the most requested, most misunderstood metadata elements, independent of specific schema. This foundational work was instrumental in defining the project's boundaries, identifying a focus on broadly applicable principles inclusive of FAIR guidelines, and contextualizing the best practices to maximize their utility.

Our high-level aim was to focus on a resource relevant to business and decision makers as well as creators, with the intention of driving advocacy rather than compliance. The work done for this goal led to recognition of other groups responsible for standards creation and promotion and to emphasize the distinctive aims of this project. However, it created a challenge for the group. Collaborative metadata initiatives are typically focused around standards with practical day-to-day workplace application. In contrast, these high-level principles strive to clarify the characteristics of improved metadata content (the "what") without dictating the actions one should take to achieve them (the "how"). The needed actions are the subject of a separate output, the Metadata Practices. (See the "How do we live these principles?" section below.)

During the 2018 workshops, the proposal for each community to have a separate set of principles was rejected in favor of one, over-arching set. This decision was based on the

idea that these principles may apply across a variety of metadata communities to minimize current interoperability challenges that arise from working in silos. Once the initial principles were developed and shared within the Metadata 2020 community, we made minor revisions based on input from the other project teams and posted them on the Metadata 2020 blog for community comment in May 2019.

The context of these principles is critical, as they need to be interpreted within the practical context in which communities operate. This aspiration guided much of the discussion around how to limit the specificity of how they are phrased. Thus, the initial list and scope of the principles aim to provide attributes and rationales at the highest and most inclusive level.

## The Metadata 2020 Principles

For metadata to support the community, it should be

**COMPATIBLE**: provide a guide to content for machines and people

> *So, metadata must be as open, interoperable, parsable, machine actionable, and human readable as possible*.

**COMPLETE**: reflect the content, components and relationships as published

> *So, metadata must be as complete and comprehensive as possible*.

**CREDIBLE**: enable content discoverability and longevity

> *So, metadata must be of clear provenance, trustworthy and accurate*.

**CURATED**: reflect updates and new elements

> *So, metadata must be maintained over time*.

## Why the principles matter

These principles are a resource for moving the community toward a set of shared goals by providing a common understanding of the required metadata characteristics needed to meaningfully support scholarly communications. They may serve to stimulate further work and development of ideas and workflows in the scholarly ecosystem if framed as guidance on the "metadata supply chain."Gregg et al. (2019)

These principles address not only metadata creation, but also their curation and custodianship in order to keep it optimally useful for as long as possible. The principles are in accord with other Metadata 2020 project work by taking perspectives from and informing other aspects of metadata improvement. For example, outputs developed by Metadata

2020 project groups such as the metadata evaluation and guidance project team (Project 6) will work closely under these initial principles.

We expect that the most immediate application of these principles will be in recommendations for best practices and their development in community contexts. These principles may also serve as a theoretical framework for businesses, non-technical stakeholders, and management audiences. These functions are part of a multi-layered and highly-dynamic ecosystem, and these principles are aimed at metadata generation, provision, and maintenance. As metadata becomes more robust and useful, end users will drive further refinement.

The principles are at an early stage of evolution. An iterative process of feedback and refinement will be necessary to increase their usefulness. We acknowledge that these contributions can evolve the principles. But, we trust that the small limited number of statements in and simplicity of the principles will support the breadth of needs and uses.

## Implementing these principles

### How do we live these principles?

In developing these principles, the group worked to generalize the concepts that would apply to the broad set of individual metadata standards or schemas. This approach enables us to harmonize present and future efforts by distilling activities to minimal necessary function and scope. The Metadata 2020 Principles will be lived through their application and specific relationship to how the community incorporates them into the practices that are adopted and the use cases that are addressed by them. The group recognizes the importance of both practices and use cases, and has adopted them as their next planned resource development activity.

### Key points

Reflecting on the process of creating the Metadata 2020 Principles, we offer the following key points about metadata to inform the use of these principles in adoption and best practice creation:

- Metadata must be as complete and comprehensive as possible, so the context in which it is used is important to define what metadata "quality" means for specific use cases.
- Cross-community awareness about metadata makes sense at a high level, but is difficult in practice because metadata value is found in domain-specific applications. Different communities use different languages for a reason. Best practices need to allow for flexibility in language and use across communities.
- Metadata custodians should use persistent identifiers (PIDs) to connect resources so that they can become interoperable. We encourage use of PIDs (that include

related metadata) rather than simply copying metadata from these resources, and for metadata creators to create PIDs where they are not yet established.
- One metadata schema will never fit all needs, though there will always be some overlap. Interoperability is strengthened through as much re-use as possible from existing standards and minimal redundancy.
- Metadata will always be evolving; updates will require versioning. Keeping metadata current and updated over time presents technical challenges that require a landscape change and clear roles and responsibility for these changes.

### Final thoughts

We expect that all Metadata 2020 project activities and outputs, including this article, will be fully described and publicly posted in the Metadata 2020 collection of articles in RIO Journal. Future work from this group will include demonstration of what creators, custodians, curators, and consumers will gain from better metadata through a set of Metadata Practices. We believe that this cumulative set of resources will provide the foundation for additional context and directions for next steps in driving further investment in compatible, complete, credible, curated metadata.

## Acknowledgements

This work wouldn't be possible without the contributions from the Metadata 2020 project team that focused on metadata Shared Best Practice and Principles. The project team's chairs, Jennifer Kemp and Howard Ratner, thank the project team members including Tony Alves, Aries Systems; Magaly Bascones, Bloomsbury; Fiona Bradley; Marleen Burger, Technische Informationsbibliothek (TIB); Laura Cagnazzo, Abertay University; Emilie David, AAAS; Paul Dlug, American Physical Society; Mark Donoghue, IEEE; Bethany Drehman, Independent; Lola Estelle, SPIE; Laurel Haak, ORCID; Kristi Holmes, Northwestern University; John Horodyski, Optimity Advisors; Maria Johnsson, Lund University; Melissa Jones, Silverchair; Kathryn Kaiser, UAB School of Public Health; Nettie Lagace, NISO; David Mellor, The Center for Open Science; Eva Mendez, Open Science Policy Platform (OSPP), Universidad Carlos III de Madrid (UC3M); Dom Mitchell, DOAJ; Ed Moore, SAGE Publishing; Nannette Naught, Innovative; Nancy Pontika, CORE; Kaci Resau, Washington & Lee University; Tyler Ruse, Digital Science; David Schott, Copyright Clearance Center; Julie Stoner, ASBMB; Peter Strickland, IUCr Journals; Michelle Urberg, ProQuest;Sarah Whalen, AAAS; Helen Williams, "LSE Library, The London School of Economics and Political Science"; Stephanie Williams, Ohio University Press; and Julie Zhu, IEEE.

## References

- GOFair (2016) FAIR Principles. https://www.go-fair.org/fair-principles/. Accessed on: 2020-4-13.

- Gregg W, Erdmann C, Paglione L, Schneider J, Dean C (2019) A literature review of scholarly communications metadata. Research Ideas and Outcomes 5 https://doi.org/10.3897/rio.5.e38698
- Health Level Seven International (2007) HL7 Website. http://www.hl7.org/. Accessed on: 2020-4-13.
- Metadata 2020 (2020) About Metadata 2020. http://www.metadata2020.org/about/. Accessed on: 2020-4-12.
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18