Information Services & Use 43 (2023) 323–326
DOI 10.3233/ISU-230219
IOS Press

Making data citations machine readable in article and other content metadata

Patricia Feeney*
Crossref, USA

Abstract. Research data is essential to the production of research, and is increasingly made available in repositories for reuse and reproducing research. Data citations play a crucial role in connecting this data to journal articles, and also to acknowledge and attribute research data correctly in scholarly publications. To ensure the interoperability and accessibility of data citations, it is essential to make data citations machine readable. This paper explores the common practices in making data citations machine readable within journal article citations, focusing on the markup of data citations in the Journal Article Tag Suite (JATS) and the use of Crossref as a data citation metadata endpoint.

Keywords: Data citations metadata, machine-readable citations, Crossref, Journal Article Tag Suite, JATS

1. Introduction

Research data is essential to the production of research, and is increasingly made available in repositories for reuse and reproducing research. Data citations play a crucial role in connecting this data to journal articles, and also to acknowledge and attribute research data correctly in scholarly publications. To ensure the interoperability and accessibility of data citations, it is essential to make data citations machine readable.

Structured metadata is used to make information about scholarly content machine readable – the metadata is the key to everything, and XML, JSON, and other formats are just packaging. Inconsistencies in the packaging can, however, lead to challenges.

Many journal publishers use the Journal Article Tag Suite (JATS) as the underlying markup language for their content [1]. JATS is a NISO [2] standard and is actively maintained and updated by a standing committee. Since JATS is a specification, not a schema, it is considered to be flexible and not overly proscriptive. There are three versions of JATS (authoring, publishing, and archiving).

Crossref [3] is a global nonprofit organization with a mission to "make scholarly communication better". This is done through the collection and distribution of scholarly metadata [4], including data and other citation metadata. Crossref has its own set of strict rules defining the metadata it collects.

JATS4R [5] is a NISO working group that is, in their own words, "devoted to optimizing the reusability of scholarly content by developing best-practice recommendations for tagging content in JATS XML".

*E-mail: pfeeney@crossref.org. Orcid: https://orcid.org/0000-0002-4011-3590.

Publishers may opt to follow the recommendations of JATS4R, or they may implement their own version of JATS. JATS4R has recommendations for citations and specifically for data citations, the advice can be extended to apply to software as well [6]. The recommendations are designed to make data citations machine readable and are designed to align as best they can with Crossref support for data and other citations.

Below is an analysis of how the JATS4R recommendations and Crossref support for data citation align, based on a talk given at the NISO Plus conference in February 2023.

2. Discrepancies

2.1. Citation types

The JATS4R data citation recommendations detail specific pieces of metadata that should be collected for data citations, and give recommendations on how to mark them up in JATS XML, as well as provide some example XML. An important piece of metadata in the recommendation is the citation type. If you flag a citation as 'data' using the publication type attribute, it will be read as a data citation.

Crossref does not currently accept citation types. Crossref's reference metadata has historically been used to match DOIs to the citations registered as part of an item record, and publication type has not been necessary to perform that task. This makes it a challenge to identify what is a data citation, and what needs to be matched with an identifier or item outside of the Crossref corpus. Data citation is new enough that citation practices are still being adopted and refined and the metadata collected by Crossref is not tailored to that. Crossref does intend to support citation types in their metadata in the future.

JATS4R also recommends providing metadata about how a data citation is used, the suggested values are "supporting," "generated," "analyzed," and "non-analyzed." Crossref similarly allows members to provide some information about the relationship between data and the article or other resource being registered, but the values do not align exactly. JATS4R maps the suggested JATS values and Crossref relationship types within their recommendations (Fig. 1)

2.2. Contributors

The people who generate data are 'curators,' not authors or editors as are traditionally included in citations. They can be identified as such within JATS markup, but the Crossref schema for citation metadata only supports an 'author' element. In practice, if a data curator or set of curators are provided to Crossref in an 'author' element, they will be identified as an overall contributor for DOI matching purposes, but they will be mislabeled.

Crossref has no plans to change this in their citation metadata support, but may reconsider as citation metadata sees more use and if Crossref membership expresses a need for more granular metadata.

2.3. Version

JATs supports providing a display and machine-readable version number. Crossref does not support version information, but will add this in the future.

Data type (@specific-use)	Description	Map to this Crossref relationship type
"supporting"	Data that supports the study's findings. Use this generic value if you do not wish to further distinguish whether the supporting data were generated or analyzed	"references"
"generated"	Supporting data that were generated for the study	"isSupplementedBy"
"analyzed"	Supporting data that were analyzed (but not generated) for the study	"references"
"non-analyzed"	Referenced data that were neither generated nor analyzed for the study	"references"
–	Data type is not indicated (no @specific-use value is supplied)	"references"

Fig. 1. Use recommendations from JATS4R data citation recommendations (https://jats4r.org/data-citations/).

3. Alignment

3.1. Titles

JATS supports a 'data-title' element to identify dataset titles, but Crossref uses a more generic general 'title' label. Though inconsistent, they functionally are the same and align well.

3.2. Source

JATS4R recommends including the name of a data repository within the 'source' element. Crossref does not support an equivalent value.

3.3. Year

Both JATS and Crossref support year metadata in data citation markup. JATS does accept and recommend supplying an ISO-8601 machine readable date as well if the date supplied is not already machine readable.

3.4. Persistent Identifiers (PIDS) and data citations

Within their identifier markup JATS accepts the type of ID (*pub-id-type*) and the assigning authority of the PID (*assigning-authority*). This allows metadata users to distinguish between, for example, a Crossref and DataCite DOI. A specific URL may also be provided (*xlink-href*) to provide persistent linkage for identifiers such as accession numbers which may not have a universally applied URL prefix.

Crossref currently supports DOIs, but not other identifiers within their citation markup, but is considering expanding support for other identifiers within their citation markup to help with matching citations to DOIs (when they exist).

Structured metadata is useful when identifying data citations, particularly for citations that do not have a persistent identifier (PID) such as a DOI, but as scholarly data becomes more interconnected, PIDs are increasingly essential. An ideal citation for Crossref would contain a DOI that can be used to link and retrieve metadata, as well as an unstructured citation that can be reused for display. A PID both links persistently to the dataset or resource being cited, but also is a path to complete and hopefully consistent metadata that eliminates the need to provide granular structured metadata that is difficult to align across organizations.

Integrating metadata that is used for different purposes is challenging, but conversations [7] between the JATS community, Crossref, and other organizations are helping to identify pain points and make changes for the better. Gaps between what is collected and defined in JATS by journal publishers and what ends up being registered with Crossref may impact the identification and dissemination of data citations, but the use of PIDs can help immensely.

About the author

Patricia Feeney has been Head of Metadata at Crossref since March 2018. This role was created to bring together all aspects of metadata, such as Crossref's strategy and overall vision, review, and introduction of new content types, best practice around inputs (Content Registration) as well as outputs (representations through Crossref APIs), and consulting with the community about metadata. During Feeney's tenure at Crossref which she joined in 2017, she has helped thousands of publishers understand how to record and distribute metadata for millions of scholarly items. Prior to Crossref she had worked in various scholarly publishing roles and as a systems librarian and cataloger. She has a B.A. in English from the University of Maryland, an M.A. in publications design from the University of Baltimore (2000), and a master's in library information studies from the University of Rhode Island (2005). E-mail: pfeeney@crossref.org (https://orcid.org/0000-0002-4011-3590).

References

[1] https://jats.nlm.nih.gov/, accessed September 18, 2023.
[2] https://www.niso.org/, accessed September 18, 2023.
[3] https://www.crossref.org, accessed September 18,m 2023.
[4] G. Hendricks, D. Tkaczyk, J. Lin and P. Feeney, Crossref: The sustainable source of community-owned scholarly metadata, *Quantitative Science Studies* **1**(1) (2020), 414–427. doi:10.1162/qss_a_00022, accessed September 18, 2023.
[5] https://jats4r.org.
[6] https://jats4r.org/data-citations/.
[7] P. Feeney, How JATs and Crossref can evolve together. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2019 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2019. Available from: https://www.ncbi.nlm.nih.gov/books/NBK540951/, accessed September 18, 2023.