Work Item Title:
Integrating Publisher and Repository Workflows to Improve Research Data-Article Links

Proposal for Consideration by the NISO Voting Membership
Approval Ballot Period: August 26 – September 27, 2021

Submitted by:
Ian Bruno, CCDC (The Cambridge Crystallographic Data Centre); ***Amanda Casari***, Google; ***Megan Force***, Clarivate; ***Vincent Lizzi***, Taylor & Francis; ***Johanssen Obanda***, AfricArXiv; ***Wendy C Robertson***, University of Iowa; ***Paul Stokes***, Jisc; ***Marta Teperek***, 4TU.ResearchData & TU Delft

Proposal Last Modified: August 16, 2021
Approved by the NISO Information Creation & Curation Topic Committee: August 25, 2021

Background and Problem Statement:

We are looking to enrich data-article linkages by defining simple, machine-readable terminologies which may be shared between stakeholders to improve workflows, analysis, discovery, and reuse.

The practice of research data deposition, publication, and citation has increased in recent years, as government entities, publishers, funders, and other stakeholders have built policies and requirements to support open data. Yet gaps remain. While some limitations to the practice of data citation could be attributed to cultural barriers, we wish to lower technical barriers to make it easier to link data and other research objects with literature in a consistent way, by specifying how publisher and repository systems should interoperate so that link creation happens as a matter of course.

As published articles remain a driving force in academia, publisher requirements for the inclusion of links to datasets in research papers encourage data citation and reuse. However, as research datasets are published in data repositories, and research articles are published in scholarly journals, the act of associating these research objects can be ineffective and inefficient.

There is some agreement (JATS4R Recommendation on Data Citations; JATS4R Recommendation on Data Availability Statements) on which metadata terms should be applied to describe associations and what recommendations need to be made to publishers and repositories so that machines can process them at scale, but wider adoption is certainly necessary; currently available metadata terms may lack sufficient richness to record the necessary information. For publishing workflows to be as effective as possible, publishers and repositories need to work in tandem. Organizations that fund research and create policy could also be benefited by a shared understanding of metadata terms.

Efforts to build Open Science Knowledge Graphs that enable traversal of networks of research objects will be hindered if links between objects are not reliably captured. While Crossref has capabilities to capture data citations where these are identified, such identifications are not being performed consistently by publishers. If links are not identified, institutions do not know if data has been shared in compliance with institutional or funder policies, researchers wishing to examine data underpinning an article or reuse it may not be able to access it for those purposes, and additional context for a dataset that may be in an article is hidden from someone looking at a dataset. The current absence of reliable and consistent linking between research data and articles makes it exceedingly difficult to create metrics and trace what articles have used what data.

Additionally, it is necessary to include precision in citation types for data-article links to support reproducibility, researcher workflows, and bibliometric analysis. Context must be included in a link between a dataset and a publication: in what way are the two linked? Is this the main dataset which supports the results described in the publication, or a dataset which may have inspired the publication? Did it inform the publication, or was used in the study as a benchmark? Once a link is identified, it may be difficult to determine why an article is referencing a dataset or vice versa. Discerning a data citation/link type is key to understanding reuse and citation analysis; this should be a fundamental part of referring to data. Crossref and JATS4R recommend that this information be included; where/why these recommendations may not be taken up needs to be better understood and addressed.

The current environment is looking for a machine-readable standard that tracks data during its lifespan from creation through validation, availability, and use. We seek a simple, bidirectional solution which links datasets, related versions and related works. This solution would establish which publications a dataset has been used in, in what way it has been used, when it has been used, and how and whether it has been changed. Other valuable information would include the dataset's origins, version information, and availability at various points in the publication process. Systematic machine capture of information from the point of deposition would reduce pain for researchers through standard and widely agreed-upon labels and fields, and data repositories would be informed of an event in the scholarly space relevant to the dataset. If faults are found in a dataset, it would be easy to determine any published works that may be affected.

The risk of not doing this project is a continuing lack of transparency between research data creation and use. Once a data set is made available, further information is needed to capture when and how it is used; projects such as Scholix capture such links when they are known, enabling the easy discovery of associated material. This information also supports the measurement of compliance for policies on data sharing, to better monitor progress toward goals. The project will increase a funder's ability to evaluate their return on research investments, increase the efficiency of data publication and discovery, and provide an infrastructure to track data validation and verification over its lifespan.

Statement of Work:

This project aims to define what information at each end of the bidirectional link between datasets and other research works matters to resolve the issues described, including what notifications are needed, and build a solution that relies on automated data exchange from the outset. While creating standard terminologies for data repositories, journal publishers, and other stakeholders to employ when establishing data-article links, we also look to set expectations that platform providers of all sizes can aspire to meet.

This project does not look to identify positive or negative links/associations between datasets and articles, such as whether a citing article agrees or disagrees with a particular set of conclusions. Context information such as license and how to use data is expected to be addressed by metadata standards. We do not look to recreate existing work and solutions, such as Scholix and various types of Persistent Identifiers (PIDs).

Deliverable:

A NISO Working Group, consisting of volunteers from relevant stakeholder communities will create a Recommended Practice describing a bidirectional environment and a proof of concept, where information

is exchanged about data objects between repository and publishing community (different skills to develop schema and UI):

- Metadata focused, not object focused (minimal metadata to describe associations)
- At earliest possible opportunity in the publication workflow
- Build terminology of events - what could change over time, versioning, citation patterns, how the objects interrelate, how events may happen to an object and how to communicate the change of status to others in the community
- Define vocabulary to describe the events. Metadata to communicate the events, with possible schemas to incorporate the interchange
- Data origin, data validity status, version history, use (status)
- Build terminology for citation/link types for data/article relationships, to resolve reporting issues with links to data availability statements vs. reference sections and others
- What information needs to be exchanged and in what direction?
- Set expectations that platform providers of all sizes can aspire to meet.
- Scope potential next steps for implementation in workflows

To be successful, the solution needs to be easy, simple, scalable, and automated, to enable widespread adoption by systems providers such as repositories to build into their platforms. We anticipate that ROI for particular parties will include:

Beneficiary	**Benefit(s)**
Researchers	Increased ease of data sharing, credit for their work, ease of access, discoverability, impact, reusability, reduction of duplication, easier compliance with funders' policies, comprehensive coverage
Research Institutions	Aggregated credit for researchers' work, discoverability, impact, monitoring compliance with institutional and funders' policies, comprehensive coverage
Publishers	Help build trust, easier for authors, consistent metadata, automated data availability statements
Reviewers	Data needed for thorough scientific assessment is available
Software providers	Clear requirements, greater interoperability, consistency
Indexers	Standards, completeness, more information and links, discoverability, rich outputs, improve internal workflows
Data repositories	Consistent metadata, Improved gauges of repository impact, ease of deposit, more reliable workflows, building knowledge of network events
Aggregators	Rich outputs, faster accessibility, automated collection/distribution, disambiguation

Funders	Demonstrated impact of investment, supports tracking, indication of success of policy

Engagement Plan:

To be successful, both ends of the bidirectional links connecting data and other research objects with literature must be in place, with participation on both sides. Publisher and repository workflows that already exist will need to be adapted; however, the project aims to avoid overcomplicating the publications process so that small publishers and repositories may continue to provide competitive services.

Publisher participation, funder mandates, policy interventions, and support from scientific societies/unions would all encourage adoption. Reporting from data producers to show the impact of downstream use of data would help to find reviewers familiar with specific datasets. Researchers will benefit from the recognition of their work, encouraging a bottom-up approach to adoption, and will be incentivized to deposit datasets in repositories where they will be sustained and linked. Universities and institutions will benefit as they more easily identify the contributions of their researchers, better enabling such initiatives as the Research Excellence Framework (REF, https://www.ref.ac.uk/) and the Declaration on Research Assessment (DORA, https://sfdora.org/). Agencies that make data available will benefit from increased impact and measurement of impact.

An aim of the project will be to establish calls for action around the terminologies that will inspire adoption by stakeholder groups and promotion by partner organisations. Working prototypes, if possible, may help to encourage adoption.

Partners and Participation:

This project should seek to build upon what already exists and collaborate to establish consensus. The following organizations and working groups have done work that is related and should be encouraged to participate in this project:

- DataCite
- Crossref
- Curation Credit Taxonomy
- EuropePMC
- GOFAIR initiative/community work
- NISO JAV (Journal Article Versions) RP
- NISO CRediT taxonomy
- Force11 (https://www.force11.org/datacitationprinciples)
- FORCE11-RDA FAIRSHARING WORKING GROUP
- Preprint supporters (*Building trust in preprints: recommendations for servers and other stakeholders*, 2020)
- STM Association
- Research Data Alliance and its subgroups on:
    - RDA Data Policy Framework

    - RDA/WDS Publishing Data Interest Group (and various of its Working Groups - Workflows)
    - RDA/WDS Scholarly Link Exchange Working Group (Scholix)
- Persistent identifier providers, including DOI / handle, archival resource key (ARK), and others

In general, the group should include representatives from publishers, data repositories, research institutions and funders.

Timeline:

- Month 1: Appointment of working group
- Month 2: Approval and publication of charge and initial work plan (including final determination of scope/requirements)
- Months 3-9 (Phase 1): Completion of information gathering, including (potentially) examination of related work, interviews/surveys with stakeholders, analysis of necessary workflows, strategizing for future adoption
- Months 10-13 (Phase 2): Completion of initial draft recommended practices document; framing calls to action; charting of prototypes
- Months 14-16: Public comment period
- Month 18: Responses to comments and publication of final NISO Recommended Practice (target March 2023)

Funding:
TBD. For now we are planning that this NISO project will be accomplished with industry volunteers meeting via conference call and asynchronous work.