A metadata package for journals to support external linked objects

Carl Lagoze
University of Michigan

Lars Vilhuber
Cornell University

Abstract

We propose a metadata package that is intended to provide academic journals with a lightweight means of registering, at the time of publication, the existence and disposition of supplementary materials. Information about the supplementary materials is, in most cases, critical for the reproducibility and replicability of scholarly results. In many instances, these materials are curated by a third party, which may or may not follow developing standards for the identification and description of those materials. As such, the vocabulary described here complements existing initiatives that specify vocabularies to describe the supplementary materials or the repositories and archives in which they have been deposited. Where possible, it reuses elements of relevant other vocabularies, facilitating coexistence with them. Furthermore, it provides an "at publication" record of reproducibility characteristics of a particular article that has been selected for publication. The proposed metadata package documents the key characteristics that journals care about in the case of supplementary materials that are held by third parties: existence, accessibility, and permanence. It does so in a robust, time-invariant fashion at the time of publication, when the editorial decisions are made. It also allows for better documentation of less accessible (non-public data), by treating it symmetrically from the point of view of the journal, therefore increasing the transparency of what up until now has been very opaque.

Submitted 18 June 2018

Correspondence should be addressed to Lars Vilhuber, 352 East Ives Hall, Ithaca, NY 14853, USA. Email: lars.vilhuber@cornell.edu

The 13th International Digital Curation Conference takes place on 4 - 7 February 2019 in Melbourne, Australia
URL: http://www.dcc.ac.uk/events/idcc19

Introduction

Reproducibility and replicability of scientific findings has been given greater scrutiny in recent years (REFS). Scientific journals, whether run by publishing companies (Springer, Elsevier, etc.) or learned societies (American Economic Association, Midwest Political Science Association, American Statistical Association, Royal Statistical Society, to name just a few in the social and statistical sciences), have been playing an important role in supporting these efforts for many years, and continue to explore novel and better ways of doing so. In particular, several journals have been hosting "supplementary materials" on their own journal websites or on affiliated repositories (e.g., Harvard Dataverse, Figshare) in support of reproducibility of the work described in published scientific articles. Data and code deposits are requested after authors' work has been (conditionally) accepted after peer review, or, less frequently, as part of the original manuscript submission process. In doing so, they assume for themselves (or delegate to a single trusted third party) the curation role for these materials, and can therefore know with certainty how long and how accessible these materials are to be preserved.

Authors are increasingly being encouraged and trained in reproducible methods from the outset of their research projects, rather than performing ex-post documentation. This includes carefully documenting provenance of third-party datasets being used, and properly curating generated datasets (surveys, collected data, etc.) in data archives as soon as possible. Furthermore, in at least some social sciences, the use of pre-existing but non-public data has increased substantially. Confidentiality and licensing constraints prevent authors from depositing such data in open archives. Both scenarios – early deposit and use of restricted-access data – make it difficult for journals and traditional archives to carry out their curation role. Journals must rely on an increasingly diverse cadre of data-holding institutions, not all of which are "archives" in the traditional sense, while satisfying increasing scrutiny of the provenance of the research results published by them.

The approach outlined in this article proposes a metadata package, derived from existing metadata where possible, that provides a lightweight approach to ameliorating this problem. In particular, the proposed metadata package documents the key characteristics that journals care about in the case of supplementary materials that are held by third parties: existence, accessibility, and permanence. The intention is that completion of the package occurs at the time of publication when the editorial decisions are made. It also allows for better documentation of less accessible (non-public data), by treating it symmetrically from the point of view of the journal, therefore increasing the transparency of what up until now has been very opaque.

We start by providing a detailed use case. We relate our approach to existing metadata, both in terms of structure as of content, and then describe the metadata package. We conclude by discussing some usability issues for three contributors or consumers of this information, and an outlook on a possible implementation.

Use Case

We target a specific but very common use case. In most applied sciences, it has become common publication practice to provide evidence of the statistical or laboratory data underlying the conclusions. This is done to support reproducibility and replicability of the scientific findings.[1] Journals with a data deposit policy have stored the supplementary

[1] There is considerable heterogeneity in the use of the terms "reproducibility" and "replicability". In this paper, we will adopt the following definitions: reproducibility is "the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the

materials on journal websites, often as simple web-based ZIP archives. While ensuring that the materials are preserved as long as the journal is active (permanence) and are accessible to any reader of the original article (subject to pay walls) (accessibility), certain shortcomings became apparent. Very large datasets and datasets with confidentiality concerns were nearly always out of scope.

More recently, journals have leveraged either dedicated, journal-branded views onto larger archives (e.g, Dataverse, Figshare), built their own data archive infrastructure (Elsevier/Mendeley[2]), or have allowed for data and code to be stored more generally on any of a curated list of "trusted" or "approved" whitelist of third-party repositories[3]. Each of these alternatives rely on a journal or publisher "vetting" the repositories and ascertaining that it meets some set of criteria. While some third-party vetting of repositories exists[4], it is far from being universally accepted at this time. In all cases known to us, the support for restricted-access repositories is quite limited. Thus, most of the known support for third-party repositories does not provide much information about accessibility (the presumption is that access is open), nor about the permanence of the repositories - this is presumably one of the evaluation criteria that journals and publishers use, but is not clearly defined as such. In fact, at least one of the consulted publishers explicitly allows for quite transitory repositories for code, without clearly distinguishing that from archives that are more permanent.[5]

Nevertheless, much of the information about persistence of archives and materials stored within those archives is available, albeit in idiosyncratic and non-machine readable form. Consider only the case of national archives[6]. In general, data stored in national archives is permanently archived; if it is not, this is clearly documented[7]. Furthermore, access is generally not restricted – if it is, this is clearly documented. However, materials in national archives do have certain restrictions – they may require sending in a written request, or a physical visit to a location with copies of the data. Thus, while the information may satisfy the publication requirements of even the most open journal, there is no robust and standardized way of documenting the additional restrictions on access that persist.

In proposing the metadata package outlined in this article, we attempt to improve on this situation. By providing a sparse but sufficient encapsulation of the information collected from authors, archives, and other third-parties, we create greater transparency about the data supporting the research. By relying on existing metadata schemas and metadata content, we minimize the effort by all parties involved, increasing the likelihood of adoption. And by intrinsically addressing the possibility that the information obtained at the time of publication may differ from later requests for the same information, we provide the tools to journals, publishers, and their editors to document that the decision to publish was based on adequate information at the time of the publication (or acceptance decision).

---

original investigator," (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015) whereas replicability differs in that "new data are collected." (ibidem).

[2] https://www.elsevier.com/authors/author-services/research-data

[3] https://f1000research.com/for-authors/data-guidelines#hosting, https://www.nature.com/sdata/policies/repositories

[4] https://www.coretrustseal.org/

[5] https://f1000research.com/for-authors/data-guidelines#hosting allows for code deposits through github.com, which has no mandate to preserve, and allows code owners to delete materials at any time without restrictions.

[6] https://www.archives.gov/dc/researcher-info, http://www.archives-nationales.culture.gouv.fr/

[7] For instance, the program code for the Business Register is destroyed when a new system is put in place – they are never kept (U.S. Census Bureau, 2009). Unedited master files for the American Community Survey are destroyed 6 years after the Edited master files are verified, unless still needed "for Census operations" (U.S. Census Bureau, 1999).

Related Metadata

A number of other initiatives address the issue of reusability and replicability, some of them through proposed metadata standards. We have endeavoured to leverage these efforts when possible (i.e., when semantics of tags overlap with our goals and when their XML schema are designed for reuse). Our hope is that this makes both interoperability with those efforts as easy and possible, and that the use of already established and perhaps familiar tags and attributes decreases the learning curve for use of our proposed schema. In the remainder of this section we describe related initiatives and the influence they have on our metadata design.

The most related metadata vocabulary comes from DataCite[8], which provides infrastructure to locate, identify, and cite research data. Identification is done via the DOI infrastructure for persistent identification, which has emerged as the standard for naming scholarly objects. The DataCite metadata schema (DataCite Metadata Working Group, 2017a, 2017b) specifies elements and attributes to describe data resources for the purpose of citation, location and retrieval. Because of the notable overlap in the purpose of DataCite and our proposal, we make use of multiple parts of this schema. Note, however, that DataCite is targeted as describing the data products themselves, where our concern is to register the placement of those products in a repository and ancillary information about that placement.

The goal of describing repositories and archives for data curation is directly addressed by the Re3data (Re3data.Org, 2015; Rücknagel et al., 2015) initiative. The goal of Re3Data is to support an online registry of research data repositories. The mechanics underlying this is to establish a common metadata standard for describing such repositories, This metadata is then used to power a search interface. The registry and search interface are targeted at researchers searching for the appropriate repository in which to store their data.

A primary technical output of the work of re3data is a "Metadata Schema for Description of Research Data Repositories" now in its 3rd version and expressed as an XML schema. The schema addresses repository characteristics such as identification, language, administrative contacts, subject focus, funding basis and the like. Our work addresses repository characteristics and reuses semantics from the Re3 schema where appropriate and possible. We will describe the details of this reuse later in this paper.

CrossRef[9] sits functionally between our work and the two initiatives described above. It was conceived by publishers as a DOI registry that, in addition to providing the resolution of those DOIs, stores metadata for the corresponding scholarly object. An important aspect of this metadata are cross-references (citations)[10] among the named objects. In that sense, CrossRef acts as a "switchboard", documenting linkages between scholarly objects. Originally, the linkages were citations between journals, but with increasing interest in data these linkages have been expanded to include these supplementary materials. In this context, CrossRef collaborates and interoperates with DataCite, with the former focusing on registration and description of journal articles and conference papers, and the latter on data and other supplementary artifacts . The CrossRef schema is a relatively complex tag set for describing articles. As our intention is

[8] https://www.datacite.org/
[9] https://www.crossref.org/
[10] https://support.crossref.org/hc/en-us/articles/214357426-Relationships-between-DOIs-and-other-objects

to promote a lightweight approach (not necessarily exclusive but perhaps in tandem with CrossRef), we have not directly borrowed from their schema. Also, our focus is linking to repositories or archives that contain supplementary material, as opposed to the object itself.

Two additional related initiatives are worthy of mention. The Core Trustworthy Data Repository Requirements (CoreTrustSeal, 2017) are the result of work within the Research Data Alliance to establish standards for so-called "trustworthy" repositories. These are repositories that meet a set of criteria that deem them dependable for the long-term curation of data. The criteria are a mixture of technical, administrative, financial, and personnel characteristics. The criteria are not as of yet, or planned to be, encoded in a machine-readable schema. Instead, repositories apply for trusted status through a form that his reviewed by a human board of review. Our proposed metadata format allows for the attribution of a repository as "trusted" and thus integrates minimally with the CoreTrustSeal effort.

The JATS (Journal Article Tag Suite), led by the NCBI (National Center for Biotechnology Information) aims to develop specifications for standardized (XML) markup for scholarly articles[11]. The effort grows out of work done on so-called "NLM DTDS", which modelled tag sets for scholarly document structuring. JATS4R (JATS for reuse) is a follow-on effort, designed to reuse and extend XML models defined by JATS, with the primary goal of facilitating reuse of existing scholarly material (publications and supplementary data)[12]. The result is a set of models specifying document structure, rather than simply metadata. The structural elements address issues such as how to mark-up authors and affiliations, citations, data citations and the like.

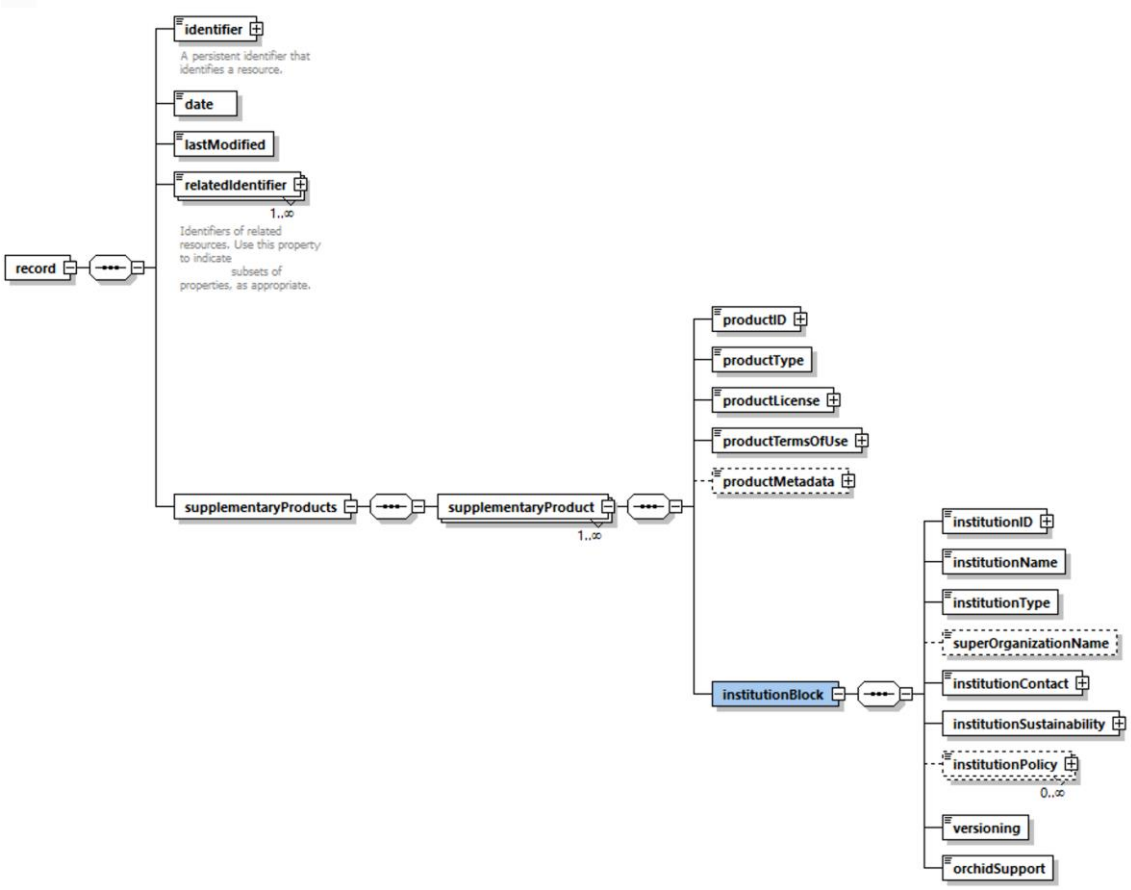
Metadata package

The high-level structure of our proposed metadata package is illustrated in the figure (produced by XMLSpy) below. As shown, each package is structured as a `record`, which conceptually models a linkage between a publication and its supplementary materials. As shown, a record has an identity (DOI), a date created, a last modified date, and the identity (DOI) of the research objects (papers) that are associated with the supplementary products.

Each `record` then can describe an unlimited number of `supplementaryProducts`. Each product has an identifier, a description of its type, licensing information, and linkages to full metadata available elsewhere that fully describes the product. Finally, each `supplementaryProduct` has an associated `institutionBlock`, which contains information about the institutional archive at which the respective `supplementaryProduct` is located.

The full annotated schema is available for examination online at (URL TBD).

[11] https://jats.nlm.nih.gov/
[12] https://jats4r.org/



Usability Notes and Outlook

Academic publishing outsources much of the content-related work to authors and subject matter editors. In order to be useful, the proposed package needs tools around it. We sketch out two such tools, and also address the role archives and repositories themselves play.

Metadata ingest

We envision that the package be provided as a single file during the manuscript submission process by the author. This ensures that existing editorial workflow packages can seamlessly track the package, without needing upgrades to understand the content. The package can be inspected by curation specialists and data editors and made available to reviewers as needed, and will follow the main document throughout the review process.

Creation by authors

In order to create the package, we envision a simple website, which helps authors fill in the required information. Appropriate HID testing would need to be done to determine the optimal structure. However, the starting point is the DOI of the object being described. From the DOI, a backend query to DataCite or CrossRef can reveal the hosting institution's institutionID. In turn, lookup in re3data or fairsharing.org will reveal elements of the institutional policies with regards to general access or preservation. Institutions often have multiple access policies and licenses, and which one applies to the object identified

by the DOI may be hard to determine automatically. The author will be able to choose the appropriate license she consented to from a set of choices appropriate for the object and its hosting institution. In theory, all such information is provided through re3data, but failing to look up complete or accurate information, the author can also fill in the information manually.

Hosting by journals

Journals are expected to post the package on their website, on the same landing page as the article itself. By doing so, the package itself can be parsed by appropriate in-page Javascript (provided through a open source library), and displayed with appropriate CSS (also provided through an open source library). Naturally, more complex journal websites can include the contents in the page source code or in their CMS.

Acknowledgements

Lagoze and Vilhuber acknowledge funding received from NSF Grant #1131848 (NCRN) and the American Economic Association. The views and recommendations expressed herein are those of the authors, and not those of the American Economic Association or the National Science Foundation.

References

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science* (Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences). National Science Foundation. Retrieved from https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

CoreTrustSeal. (2017, January 2). Data Repositories Requirements. Retrieved June 14, 2018, from https://www.coretrustseal.org/why-certification/requirements/

DataCite Metadata Working Group. (2017a). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.1. https://doi.org/10.5438/0014

DataCite Metadata Working Group. (2017b). DataCite Metadata Schema for the Publication and Citation of Research Data v4.1. https://doi.org/10.5438/0015

Re3data.Org. (2015). re3data.org Metadata Schema 3.0 XML Schema. re3data.org. https://doi.org/10.2312/re3.009

Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D., … Kirchhoff, A. (2015). Metadata Schema for the Description of Research Data Repositories. GFZ Germans Research Center for Geosciences. https://doi.org/10.2312/re3.008

U.S. Census Bureau. (1999, June 4). Records Control Schedule: American Community Survey Records. National Archives and Records Administration. Retrieved from https://www.archives.gov/files/records-mgmt/rcs/schedules/departments/department-of-commerce/rg-0029/n1-029-98-001_sf115.pdf

U.S. Census Bureau. (2009, October 30). Records Control Schedule: Business Register. National Archives and Records Administration. Retrieved from https://www.archives.gov/records-mgmt/rcs/schedules/departments/department-of-commerce/rg-0029/n1-029-10-002_sf115.pdf