

metajelo: 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 June 2018

Correspondence should be addressed to Lars Vilhuber. Email: lars.vilhuber@cornell.edu

The International Digital Curation Conference takes place on [TBC] in [TBC]. URL: <http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc>

Copyright rests with the authors. This work is released under a Creative Commons Attribution 4.0 International Licence. For details please see <http://creativecommons.org/licenses/by/4.0/>



Introduction

Reproducibility and replicability of scientific findings has been given great scrutiny in recent years (Camerer et al., 2016; Collaboration, 2015; Fanelli, 2018; Klein et al., 2014). Actual published individual reproductions or replications are traditionally not very common (in economics, see Bell & Miller, 2013; Duvendack, Palmer-Jones, & Reed, 2017). In part, this is because it often was difficult to find the materials required to conduct reproducibility or replication exercises (Dewald, Thursby, & Anderson, 1986; McCullough, McGeary, & Harrison, 2006; McCullough & Vinod, 2003).

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 (Stodden et al., 2016), and continue to explore novel and better ways of doing so. More and more journals are adopting “data and code availability” policies, though some doubt has been cast on their effectiveness (Höffler, 2017a; Stodden, Guo, & Ma, 2013; Stodden, Seiler, & Ma, 2018a). Some of the lack of replicability identified by recent studies (Camerer et al., 2016; Chang & Li, 2015, 2017; Höffler, 2017b; Stodden et al., 2018a) is despite the fact that journals have these policies. One issue is the lack of consistent, reliable metadata on the materials provided to journals, and in particular those provided through third-party locations.

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 when 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. 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. Both scenarios - early and third-party deposit of data and use of restricted-access data - make it difficult for journals and traditional archives to carry out their curation role. The resulting lack of transparency in data provenance is detrimental to the overall effort of increasing transparency in the sciences.

The approach outlined in this article proposes a metadata package, derived from existing metadata schemata where possible, that provides a lightweight approach to ameliorating this problem. In particular, the proposed metadata package, called *metajelo* (*metadata package for journals to support external linked objects*) documents

the key characteristics that journals care about in the case of supplementary materials that are held by third parties: existence, accessibility, and permanence. Our intent in defining the metadata package is two-fold. First, the package enables authors to provide the information as they submit articles to journals, allowing informed editorial decisions to be made. Second, at the time of publication, the information is made public, providing robust documentation on data provenance in an immutable package, in a compact fashion. The package allows for better documentation of any data, regardless of the difficulty of access. Thus the information provided for less accessible (non-public data) is improved by treating it symmetrically with open access data, therefore increasing the transparency of what up until now has been very opaque.

We start by providing some background. We describe the use case motivating our approach, with detailed use cases provided in the appendix. 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.

Background

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 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/Mendeley²), 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

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 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>, <https://www.nature.com/sdata/policies/repositories>

⁴ CoreTrustSeal, <https://www.coretrustseal.org/>

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 (e.g., the U.S. National Archives⁶ or the *Archives Nationales* in France⁷). In general, data stored in national archives is permanently archived; if it is not, this is clearly documented.⁸ 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 that returned by 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).

Use Case

We target a specific but very common use case. A researcher has written a paper with empirical content, and is required by the journal's data and code availability policy to prepare a "replication package." The journal's policy requires that the code and data be accessible to others, but does not require deposit of the materials as a "supplementary file," i.e., as a ZIP file on their website.⁹ However, in all cases, the journal wishes to ascertain three key attributes of the replication package or packages:

- the *existence* of the package
- the *access rules* to the package (license, terms of use)
- the *persistence* of the package

⁵ F1000Research (<https://f1000research.com/for-authors/data-guidelines>) 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>

⁷ <http://www.archives-nationales.culture.gouv.fr/>

⁸ 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).

⁹ In fact, some journals may not offer that option.

In an ideal scenario, the existence of the package can be easily ascertained in a reputable repository, it is made available under an well-specified (ideally open) license, and it is available “forever”. When the journal manages its own repository, these attributes are known. When the package is available elsewhere, these attributes need to be discovered. Furthermore, this needs to happen in a scalable, automated, and reusable fashion, as it should be feasible to do so for all articles, submitted to any journal.

Current Metadata Infrastructure and Use Cases

The current metadata infrastructure should be expected to work well for open-access data deposits. Deposits are encouraged in known repositories such as Inter-university Consortium for Political and Social Research (ICPSR), Zenodo, or the Open Science Framework¹⁰, which have been vetted according to certain criteria by the journals themselves.

But what if an author has deposited the information in a reputable but unlisted repository, for instance the Australian Data Archive¹¹? Emails are to be exchanged, and some case-by-case vetting of repositories, their reliability, and whether they assign Digital Object Identifier (DOI) is performed. FAIRsharing.org and re3data are invoked to ascertain their policies.

In Appendix , we demonstrate for three cases that this infrastructure - DataCite, re3data, and FAIRsharing - will fail on even simple scenarios. In all cases, we attempt to ascertain *existence*, *access rules* (terms of use and licenses), and *persistence* (preservation policies) via machine-readable metadata. We fail to collect complete information in all cases. Furthermore, as of the writing of this article, and presumably for some time yet, this infrastructure simply cannot support scenarios that use broadly available restricted-access data. By “broadly available restricted-access”, we mean that a non-trivial fraction of a research community can be granted access to these data, which are restricted-access only for reasons of confidentiality. This scenario is quite common - it applies to clinical data in psychology as much as demographic data collected by national statistical agencies in every country in the world.

The three cases are as follows. First, we show that a user-initiated data deposit (the object) at ICPSR, more specifically: openICPSR¹², properly recorded in DataCite (*existence*), cannot reveal the remaining metadata through queries to the infrastructure. A customized parser can ascertain the *license* by querying the landing page of the object. Queries to DataCite fail to elicit the license because it is optional. Queries to re3data fail because a record cannot be found using information available through the DOI, in particular, the name of the repository. Cheating somewhat, when we force a query to re3data’s entry for ICPSR (Re3data.Org, 2013), it fails to yield correct information, presumably because the record is not maintained by ICPSR staff, and does not hold information on openICPSR policies. We fail to ascertain the preservation policy through queries to all sources, and only subject-matter expertise can find the information on ICPSR’s website.

The second query is for the Panel Study of Income Dynamics (PSID) Geospatial Data (PSID, n.d.-b). The PSID is a longitudinal household survey conducted by the University

¹⁰ <https://osf.io>

¹¹ <https://ada.edu.au/>

¹² <https://www.openicpsr.org>

of Michigan, which began in 1968. More than 4000 peer-reviewed publications have used the data (PSID, n.d.-a). The data are available without cost to researchers - but they do require that terms of use be agreed to before downloading, through registration. This is accurately reflected in the r3data entry for the PSID (Re3data.Org, 2017). However, we are considering the Geospatial Data, which is restricted data. re3data fails to record any information for this access mechanism. Furthermore, although PSID has acted as a data curator for its own data for 50 years, it does not assign persistent identifier (PID) to the data. DataCite has no information on any PSID data holdings, which are only available through the PSID website. Until recently, both non-restricted and restricted data could not be deposited at journal websites or other repositories, as per the terms of use.¹³ Finally, although the PSID has, of course, a 50-year track record, no statement can be found on the website attesting for preservation plans, or for versioning of data (preservation of prior versions).¹⁴

The third example is a confidential dataset made available by a National Statistical Organization (NSO), in this case the U.S. Census Bureau, although it is typical of microdata holdings by NSO around the world. The Longitudinal Business Database (LBD) (Miranda & Jarmin, 2002; U.S. Census Bureau, 2017) is one of the most widely used microdata files in the Federal Statistical Research Data Center (FSRDC) system. The FSRDC system is used by nearly 700 researchers at 29 locations around the United States (U.S. Census Bureau, 2018). As with the PSID, entries for the U.S. Census Bureau exist on re3data (Re3data.Org, 2018), but have no information on the FSRDC. No PID have yet been assigned to any datasets. Furthermore, no data can be removed from the FSRDC. Researchers must thus rely on the U.S. Census Bureau for preservation. In addition to the LBD itself, which is presumably covered by a record schedule, detailing its preservation period, researchers also need to consider the preservation of any derivative files they wish to make available as part of their research. If these are aggregated results (model coefficients, etc.), they are released by the U.S. Census Bureau to the researcher. Microdata cannot be released. Most of this information is provided to researchers when they obtain access, but cannot easily be communicated to journal editors or readers of articles. Nevertheless, as we have argued (Lagoze & Vilhuber, 2017) and experienced in our own research (Abowd, McKinney, & Vilhuber, 2009; Abowd & Vilhuber, 2005; McKinney, Green, Abowd, & Vilhuber, 2017), it is definitely feasible to do reproducible research in this environment. The difficulty consists in communicating that information, in a reliable fashion, to editors, referees, and readers.

Common Denominator

We have chosen three types of datasets – public-use, restricted-access with light restrictions, restricted-access with strong restrictions –, curated by three different institutions – an open repository, a panel survey provided by a recognized leader in the field, and confidential business microdata provided by one of the largest and oldest NSO in the world – all with impeccable data curation reputations. The choice is idiosyncratic, but it presumably is symptomatic of the still young state of the metadata infrastructure.

¹³ This has changed recently with the introduction of an openICPSR-hosted PSID repository, but see the issues above.

¹⁴ Personal communication in November 2018 with David S. Johnson, at the time Director of the PSID, indicates that all versions of non-restricted and restricted data are preserved in a dark archive.

We don't believe these examples are exceptions - similar institutions exist all over the world, and we could as easily have done such examples with data from Australia (Department of Social Services, 2018), Germany (Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)). Presumably, counterexamples can be given. But journal editors and authors need such mechanisms to be broadly feasible if they are to use them. At present, that is not the case.

We set out to accomplish this by designing a metadata package, drawing on existing schema used within the infrastructure, but populating it in a decentralized fashion, at the point of first use: the journal submission system, or if the researcher uses a reproducible workflow, at data acquisition by the researcher. An associated application can leverage the metadata infrastructure where it does provide information, and pre-fill any fields. However, when ambiguous responses are obtained, or no information is available, the researcher can provide guided or verbatim answers. At both points in time, the researcher has the best incentives to provide the information accurately - the acceptance of the submission may depend on the accuracy of the information - and the most timely recollection of where to obtain the information.

Related Metadata and Efforts

A number of initiatives address the issue of reusability of research objects and replicability of science, some of them through proposed metadata standards. None of these efforts can completely provide the information and benefits our proposed metajelo package (described in more detail in Section) provides. Nevertheless, we have endeavoured to leverage these efforts when possible (i.e., when semantics of tags overlap with our goals and when their XML schema can be cloned for interoperability).¹⁵ 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.

DataCite

The most related metadata vocabulary comes from DataCite¹⁶, 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

¹⁵ We originally attempted to reuse other schema by reference, import, and use of name spaces. However, we encountered multiple problems. Name spaces were not handled consistently across parsers. The schemas we intended to re-use were not designed for that purpose. We thus reverted to "re-use by cloning", for lack of robust alternatives.

¹⁶ <https://www.datacite.org/>

to register the placement of those products in a repository and ancillary information about that placement. While the DataCite schema has a `license` field, it is optional, and often empty. There is no information on more complex access policies, and no information on preservation.

Re3data

The goal of describing repositories and archives for data curation is directly addressed by the Re3data initiative (Re3data.Org, 2015; Rücknagel et al., 2015). 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 Re3data schema where appropriate and possible. We will describe the details of this reuse later in this paper.

CrossRef

CrossRef¹⁷ 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) among the named objects (CrossRef, n.d.). 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 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.

Scholix

The Scholix effort (Burton, Fenner, Haak, & Manghi, 2017) is also closely related to our proposed package. However, while it may lay the groundwork for the information here, it fundamentally does not have rich enough information about the linked objects to fulfill our core purpose.

¹⁷ <https://www.crossref.org/>

CoreTrustSeal

Two additional related initiatives are worthy of mention. The Core Trustworthy Data Repository Requirements (CoreTrustSeal, n.d.) 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 is 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. However, as the CoreTrustSeal does not provide an Application Programming Interface (API), the information embedded within the certification cannot be re-used. Furthermore, as noted for re3data, an institution may have multiple policies, and it may not always be easy to attribute a particular policy to a particular object.

JATS

The JATS (Journal Article Tag Suite)¹⁸, led by the NCBI (National Center for Biotechnology Information) aims to develop specifications for standardized (XML) markup for scholarly articles. The effort grows out of work done on so-called “NLM DTDS”, which modelled tag sets for scholarly document structuring. JATS4R¹⁹ (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). 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.

Data Accessibility Statements

The Belmont Forum has recently started a project²⁰ to standardize a Data Accessibility Statement (DAS). Its goals seem to be quite similar to our project, and while independently developed, we look forward to seeing their suggestions, and will collaborate in moving that forward.

Metadata Package

The high-level structure of our proposed metadata package is illustrated in the Figure 1 (produced by OxygenXML). 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

¹⁸ <https://jats.nlm.nih.gov/>

¹⁹ <https://jats4r.org/>

²⁰ <http://www.bfe-inf.org/resource/belmont-forum-data-publishing-policy-workshop-report-draft>

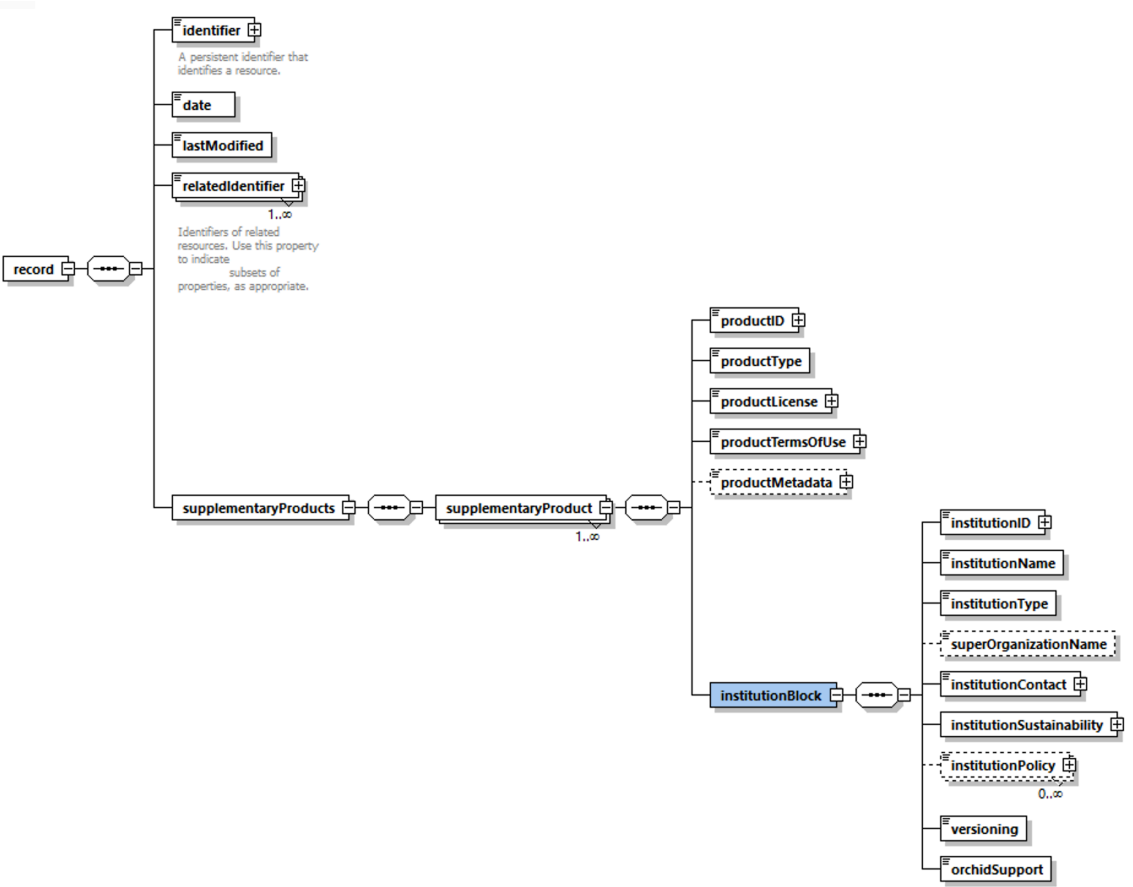


Figure 1. High-level structure of proposed package

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. Each supplementaryProduct has an associated location block, which contains information about the institutional archive at which the respective supplementaryProduct is located. Finally, for each institution, the set of possible policies are listed, with a boolean designation of the applicability of a policy to the respective supplementary object. The full annotated schema is available for examination online at github.com/labordynamicsinstitute/metajelo.

Table 1: metajelo Description

ID	Field	Definition	Notes	
1	record	Wrapper element representing the linkage between a publication and supplementary products.		
1.1	identifier	Persistent ID assigned to this metajelo record (ID types attribute: DataCite)		
1.2	date	Date of original creation of this record		
1.3	lastModified	Date of most recent modification of this record		
1.4	relatedIdentifier	Identifiers (links) to the object (article) with which the supplementary products are associated (ID types attribute: DataCite) This could be a link to the object itself or to metadata. Note presence of attribute indicating relationship type, controlled vocabular of datacite:relationTypes .	repeatable	DataCite
1.5.1	supplementaryProduct	Wrapper element around the metadata for an individual supplementary product.	repeatable	
1.5.1.1	resourceID	Persistent ID of supplementary product (ID types attribute: DataCite)	Optional (some products will not have PID)	
needs own number	basicMetadata	Bibliographic metadata (Creator, Title, PublicationYear [DataCite]). Note that Publisher is drawn from institutionName.	Required	DataCite
1.5.1.2	resourceType	Description of supplementary product		DataCite
1.5.1.2.1	resourceTypeGeneral	General type of the supplementary product.	Controlled vocabulary	DataCite
1.5.1.3	Format	Technical format of the product (e.g., stata, csv, zip, gzip). Use file extension or MIME type if possible.	optional, repeatable	DataCite

(cont)

Table 1 : metajelo Description

ID	Field	Definition	Notes	
1.5.1.4	resourceMetadataSource	Handle referencing metadata describing this supplementary product (e.g., in DataCite) (ID types attribute: DataCite)	optional	
1.5.1.5	location	Wrapper element for metadata describing location of supplementary product		
1.5.1.5.1	institutionID	Persistent ID assigned to the institution holding this object (ID types attribute: DataCite). This should be the same as in re3data, CoreTrustSeal, etc.	optional, repeatable	re3data
1.5.1.5.2	institutionName	Free text name of location. Maps to datacite:Publisher and re3data:repositoryName or re3data:institutionName. Used for bibliographic citation.		re3data, DataCite
1.5.1.5.3	institutionType	Organization type of the location.	Controlled vocabulary based on re3data, possibly expanded	re3data
1.5.1.5.4	superOrganizationName	Broader organization context of the location	Optional	
1.5.1.5.5	institutionContact	Email address or online contact form of contact party	Suggested	re3data
1.5.1.5.6	institutionSustainability	Wrapper element for assertions concerning institution sustainability		
1.5.1.5.6.1	missionStatementURL	Link to institution mission statement	Optional	re3data
1.5.1.5.6.2	fundingStatementURL	Link to documentation of institution funding	Optional	
1.5.1.5.7	institutionPolicies	Wrapper element for set of policies at institution regarding access, archiving, etc)		
1.5.1.5.7.1	institutionPolicy	A policy of the institution.	Repeatable.	re3data
		Attribute re3data:policyType specifies the nature of the policy (access, collection, etc.). We also include License here.	At least one of Preservation policy and (Terms of Use or License) are required.	
		Attribute appliesToProduct is true for policies that apply to this product.		
1.5.1.5.7.1.1	institutionPolicyFreeText	A free text explanation of the policy. This may be scraped from the URI if not a PID.	Choice	

(cont)

Table 1 : metajelo Description

ID	Field	Definition	Notes
1.5.1.5.7.1.2	institutionPolicyURL	URI of policy. Attribute re3data:policyURL points to an online copy of the policy. (ID types attribute: DataCite)	Suggested re3data
1.5.1.5.7.2	versioning	Does the location support versioning of content	Allowed values: yes, re3data no, unknown

Usability Notes

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. Systems that do know how to ingest the information should do so, but are able to collect the information more efficiently. 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 Human-Computer Interaction (HCI) testing would need to be done to determine the optimal structure. However, the starting point is the DOI of the object being described, if available, or a bibliographic record, otherwise. 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 Content Management System (CMS).

Decentralizing the linkage architecture

A final point is worth highlighting. When journals iadopt the metajelo package, then much information will be made available at a key point in the scientific cycle, when incentives are aligned: at the point of publication. By having authors themselves, possibly with help from the editors, create the linkage information (linking data and code archives to articles), having them describe what they know of access and retention policies at archives, creates information on thousands of articles every year, across hundreds of journals. This information can be harvested. Clearly, not all of the information will be accurate or consistent - but neither is the information currently being curated in centralized repositories of such information. Disambiguation algorithms will need to be deployed, and aggregation needs to allow for multiple (non-authoritative) answers. Facilities like Re3data will become aggregators instead of creators of such metadata. Our hypothesis is that the error rate in metadata will decline, but not disappear.

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 either the American Economic Association or the National Science Foundation.

References

- Abowd, J. M., McKinney, K. L., & Vilhuber, L. (2009). The link between human capital, mass layoffs, and firm deaths. In T. Dunne, J. B. Jensen, & M. J. Roberts (Eds.), *Producer dynamics: new evidence from micro data*. University of Chicago Press. Retrieved from <http://www.nber.org/chapters/c0497/>
- Abowd, J. M. & Vilhuber, L. [Lars]. (2005). The sensitivity of economic statistics to coding errors in personal identifiers. 23(2), 133–152. Retrieved from <http://www.jstor.org/stable/27638803>
- Bell, M. & Miller, N. (2013). How to persuade journals to accept your replication paper. Retrieved October 8, 2014, from <https://politicalsciencereplication.wordpress.com/2013/09/11/guest-blog-how-to-persuade-journals-to-accept-your-replication-paper/>

- 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*. National Science Foundation. Retrieved May 20, 2018, from https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf
- Burton, A., Fenner, M., Haak, W., & Manghi, P. (2017). Scholix Metadata Schema For Exchange Of Scholarly Communication Links. doi:10.5281/zenodo.1120265
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, aaf0918. doi:10.1126/science.aaf0918. pmid: 26940865
- Chang, A. C. & Li, P. (2015). *Is economics research replicable? sixty published papers from thirteen journals say "usually not"* (Finance and Economics Discussion Series No. 2015-83). Board of Governors of the Federal Reserve System (U.S.) Retrieved from <https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf>
- Chang, A. C. & Li, P. (2017). A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review*, 107(5), 60–64. doi:10.1257/aer.p20171034
- Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716–aac4716. doi:10.1126/science.aac4716
- CoreTrustSeal. (n.d.). Data Repositories Requirements. Retrieved June 14, 2018, from <https://www.coretrustseal.org/why-certification/requirements/>
- CrossRef. (n.d.). Relationships between DOIs and other objects. Retrieved October 22, 2018, from <http://support.crossref.org/hc/en-us/articles/214357426-Relationships-between-DOIs-and-other-objects>
- DataCite Metadata Working Group. (2017a). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.1. doi:10.5438/0014
- DataCite Metadata Working Group. (2017b). DataCite Metadata Schema for the Publication and Citation of Research Data v4.1. doi:10.5438/0015
- Department of Social Services. (2018). The household, income and labour dynamics in australia (hilda) survey, general release 17 (waves 1-17). Melbourne Institute of Applied Economic and Social Research. doi:10.26193/PTKLYP
- Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *American Economic Review*, 76(4), 587–603.
- Duvendack, M., Palmer-Jones, R., & Reed, W. R. (2017). What is meant by replication and why does it encounter resistance in economics? *American Economic Review*, 107(5), 46–51. doi:10.1257/aer.p20171031

- Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11), 2628–2631. doi:10.1073/pnas.1708272114. pmid: 29531051
- Höffler, J. H. (2017a). Replication and economics journal policies. *American Economic Review*, 107(5), 52–55. doi:10.1257/aer.p20171032
- Höffler, J. H. (2017b). ReplicationWiki: improving transparency in social sciences research. *D-Lib Magazine*, 23(3/4). doi:10.1045/march2017-hoeffler
- Inter-university Consortium for Political and Social Research. (n.d.). Digital preservation policies and planning at icpsr. Retrieved from <https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/index.html>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahnk,., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, 45(3), 142–152. doi:10.1027/1864-9335/a000178
- Lagoze, C. & Vilhuber, L. [L]. (2017). Making confidential data part of reproducible research. *Chance*. Retrieved from <http://chance.amstat.org/2017/09/reproducible-research/>
- McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Lessons from the JMCB Archive. *Journal of Money, Credit and Banking*, 38(4), 1093–1107. Retrieved from <http://ideas.repec.org/a/mcb/jmoncb/v38y2006i4p1093-1107.html>
- McCullough, B. D. & Vinod, H. D. (2003). Econometrics and software: comments. *Journal of Economic Perspectives*, 17(1), 223–224. Retrieved from <http://EconPapers.repec.org/RePEc:aea:jecper:v:17:y:2003:i:1:p:223-224>
- McKinney, K. L., Green, A. S., Abowd, J. M., & Vilhuber, L. (2017). Total error and variability measures with integrated disclosure limitation for Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in OnTheMap. *submitted*.
- Miranda, J. & Jarmin, R. (2002). *The Longitudinal Business Database* (Discussion Paper No. CES-WP-02-17). U.S. Census Bureau, Center for Economic Studies. Retrieved from <http://ideas.repec.org/p/cen/wpaper/02-17.html>
- PSID. (n.d.-a). Home page: panel study of income dynamics (PSID). Retrieved December 6, 2018, from <https://psidonline.isr.umich.edu/default.aspx>
- PSID. (n.d.-b). Panel study of income dynamics (PSID) geospatial data. Retrieved December 6, 2018, from <https://simba.isr.umich.edu/restricted/Geospatial.aspx>
- Re3data.Org. (2013). Inter-university consortium for political and social research. re3data.org - Registry of Research Data Repositories. doi:10.17616/r3bc8q
- Re3data.Org. (2015). Re3data.org Metadata Schema 3.0 XML Schema. re3data.org. doi:10.2312/re3.009

- Re3data.Org. (2017). Panel study on income dynamics. re3data.org - Registry of Research Data Repositories. [doi:10.17616/R3503G](https://doi.org/10.17616/R3503G)
- Re3data.Org. (2018). United states census bureau. re3data.org - Registry of Research Data Repositories. [doi:10.17616/R3SP4B](https://doi.org/10.17616/R3SP4B)
- 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 German Research Center for Geosciences. [doi:10.2312/re3.008](https://doi.org/10.2312/re3.008)
- Stodden, V., Guo, P., & Ma, Z. (2013). Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLoS ONE*, 8(6), e67111. [doi:10.1371/journal.pone.0067111](https://doi.org/10.1371/journal.pone.0067111)
- Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., ... Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, 354(6317), 1240–1241. [doi:10.1126/science.aah6168](https://doi.org/10.1126/science.aah6168)
- Stodden, V., Seiler, J., & Ma, Z. (2018a). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 201708290. [doi:10.1073/pnas.1708290115](https://doi.org/10.1073/pnas.1708290115). pmid: 29531050
- Stodden, V., Seiler, J., & Ma, Z. (2018b). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11), 2584–2589. [doi:10.1073/pnas.1708290115](https://doi.org/10.1073/pnas.1708290115)
- U.S. Census Bureau. (1999). Records Control Schedule: American Community Survey Records. National Archives and Records Administration. Retrieved from https://www.archives.gov/files/records-mgmt/rcs/schedules/departments/departments-of-commerce/rg-0029/n1-029-98-001_sf115.pdf
- U.S. Census Bureau. (2009). Records Control Schedule: Business Register. National Archives and Records Administration. Retrieved from https://www.archives.gov/records-mgmt/rcs/schedules/departments/departments-of-commerce/rg-0029/n1-029-10-002_sf115.pdf
- U.S. Census Bureau. (2017). *Longitudinal business database (LBD)*. U.S. Census Bureau [distributor]. Washington,DC. Retrieved from <https://www.census.gov/ces/dataproducts/datasets/lbd.html>
- U.S. Census Bureau. (2018). *Center for Economic Studies and Research Data Centers Research Report: 2017*. Research and Methodology Directorate. Retrieved December 7, 2018, from https://www.census.gov/ces/pdf/2017_CES_Research_Report.pdf

```
1 <?xml version="1.0" encoding="UTF-8"?>
9 <doc>
10 <str name="datacentre">GESIS.ICPSR – ICPSR</str>
11 <str name="doi">10.3886/E100590V1</str>
22 </arr>
23 <str name="publisher">ICPSR – Interuniversity Consortium for
24 Political and Social Research</str>
```

Figure 2. Select lines from DataCite query for DOI 10.3886/E100590V1
The full query response can be found in the appendix.

Appendix: Detailed Use Cases

Scenario 1: Public-use information at openICPSR

In the first case, the researcher has used public-use data, and identifies a DOI to the journal (<http://doi.org/10.3886/E100590V1>). From this DOI, the journal will attempt to identify the three attributes outlined above, using automated mechanisms. We thus start with the DOI, which resolves to the following citation:

McKinney, Kevin L., Green, Andrew S., Vilhuber, Lars, and Abowd, John M. Replication data: Total Error and Variability Measures for QWI and LODS. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2017-12-15. <https://doi.org/10.3886/E100590V1>

DataCite

We first query the DataCite API (Figure 2). The query reveals the identity of the datacentre and the publisher. However, there is no information on the license under which the object is made available, no copyright, license, or terms of use information, nor any information on persistence of the data. The `license` attribute is optional as per DataCite Schema (DataCite Metadata Working Group, 2017b), and is empty here.

re3data

We turn to re3data for further information, and find two possible problems. A lookup for the contents of the datacentre field yields 0 results. A search for the contents of the publisher field yields a wrong result (`<odesi>`). We applied human judgment to find a re3data record for ICPSR: <https://www.re3data.org/repository/r3d100010255> (Re3data.Org, 2013). We note, however, that the rules and policies for openICPSR may differ from ICPSR²¹. The re3data record lists three types of data access. Furthermore,

²¹ <https://www.openicpsr.org/openicpsr/faqs>

three data licenses are listed: two other and one copyright.

Data access (3)

Type of access to data	closed
------------------------	--------

Type of access to data	open
------------------------	------

Type of access to data	restricted
------------------------	------------

Data access restriction type(s)	other registration
---------------------------------	-----------------------

Data licenses (3)

DataLicense	other
-------------	-------

URL	http://www.icpsr.umich.edu/icpsrweb/membership/support/faqs/2009/01/what-are-icpsrs-terms-of-use
-----	---

DataLicense	other
-------------	-------

URL	http://www.icpsr.umich.edu/files/ICPSR/access/restricted/all.pdf
-----	---

DataLicense	Copyrights
-------------	------------

URL	http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/details.html
-----	---

Thus, while re3data does contain entries of *possible* licenses, we have no information on which one applies to the replication package above. Furthermore (not displayed here), there is no machine-readable information on persistence. While knowledgeable data archivists and librarians, as well as many social scientists, “know” that ICPSR is a reputable archive with a long history and presumably a long future, this is not encoded anywhere where non-domain experts could ascertain it.

CoreTrustSeal

We do not investigate whether this information is available on through CoreTrustSeal, for three reasons. First, searching again, as we did, through the website, neither of the search terms that the DataCite record provides yield findable results. Second, when we manually identify ICPSR on the website’s map of institutions, we observe that ICPSR had a “Data Seal of Approval” (the predecessor to CoreTrustSeal), but that it expired in 2017, which may explain the lack of search results. Finally, the CoreTrustSeal certification is encapsulated in PDFs, and does not provide an API to search for attributes of a certified repository. While it may be feasible for a human to track down the relevant information, it is not scalable.

Data publisher website

Finally, we attempt to obtain metadata directly from the landing page indicated by the DOI.²² The page offers five types of metadata: the in-page metadata in XML format, in-page metadata encoded as JSON-LD, a link to a OAI-PMH record, a link to a DDI 2.5 record, and a link to a DDI 3.1 record. The webpage provides two instances of license information. The first instance is within the `re1` identifier within the `a` link field (Figure 3)

²² The query was run on 8 October 2018.

```

1 <div class="well">
2   <p>
3     <a rel="license" href="http://creativecommons.org/licenses/by/4.0/"
4       target="_blank">
5       
7     </a>
8     This work is licensed under a <a rel="license" href="http://creativecommons.org/licenses/by/4.0/"
9       target="_blank">
10      Creative Commons Attribution 4.0 International License</a>.
11   </p>
12   <p>openICPSR data are distributed exactly as they arrived from
13     the data depositor. ICPSR has not checked or processed this
14     material. Users should consult the investigator(s) if further
15     information is desired.</p>
16 </div>

```

Figure 3. Use Case 1, Encoding of license in HTML of landing page

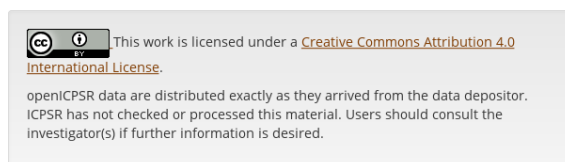


Figure 4. Use Case 1, license as displayed on website on 8 October 2018

with an associated displayed license badge (Figure 4). The second instance is encoded in the JSON-LD payload,

```

1 "license": "https://creativecommons.org/licenses/by/4.0/deed.en_US"

```

Both provide the same information about the license.

Conclusion on Use Case 1

We note that re3data did not provide additional information about accessibility, even though ICPSR does provide data with more restrictive access rules, for instance, through secure cloud instances. Furthermore, no information is provided about persistence. The openICPSR FAQ contain such information, but do so somewhat obliquely, and do not point to a policy. Browsing the website, one might encounter the “[Digital Preservation Policies and Planning at ICPSR](#)” (Inter-university Consortium for Political and Social Research, n.d.), which lays out the policies.

We note that DataCite, while providing a means to communicate the license, did not do so at this time. DataCite does not provide a means to convey access rules or persistence, nor does it provide a means to point to specific policies on re3data. Re3data, in turn, lists three possible licenses, none of which apply in the present case, possibly because it lists information on the main ICPSR repository, and not on the associated but distinct openICPSR instance.

In this relatively straightforward case, we would need to query the user about which access policy applies to the particular dataset at hand.

Scenario 2: restricted-access PSID

The PSID has published data for several decades, and is widely used (several thousand articles). Currently, researchers access the data by downloading them from the PSID website, if the data is public-use. PSID also provides some restricted access files, for instance with more detailed geocodes. Access procedures are described at <https://simba.isr.umich.edu/restricted/ProcessReq.aspx>. The PSID has not assigned DOI to any of its data products. Personal communication reveals that both public-use and restricted-access data are versioned internally, and that the data themselves contain a variable with the versioning information; there is, however, no metadata on the website listing the available past datasets, only the most current one. There is no explicit retention information on the website.

In this scenario,

- CrossRef or DataCite offer no information on the data
- While there is a re3data page at <https://www.re3data.org/repository/r3d100011131> (Re3data.Org, 2017), it does not provide information on the restricted access conditions
- the product page offers some unstructured information

We also note that even if re3data had the correct access policy for 2018, it is difficult to obtain information on past access policies. The PSID used to provide restricted-access data via shipment of CDs to researchers, who would put the data on computers that were not connected to networks, secured in a locked room. Authors are still publishing articles today that rely on data obtained through the outdated access mode.

Scenario 3: Restricted access at the U.S. Census Bureau

The Longitudinal Business Database²³ (LBD) data at the U.S. Census Bureau is one of the most requested datasets in the FSRDC network. Access procedures are described at various locations, including here²⁴ and here²⁵. The LBD data, as most business data at the U.S. Census Bureau, contain Federal Tax Information (FTI); however, this is not noted on the product description page. In contrast to many person or household data, which are archived at the National Archives as per a published Records Schedule, the business data are not sent to the National Archives, due to the presence of said FTI. It is quite difficult to find information on this. In fact, the Center for Economic Studies is the official archiver, and maintains these files in perpetuity. The Census Bureau has not assigned DOI to any of its data assets as of 2018.

In this scenario,

- CrossRef or DataCite offer no information on the data

²³ <https://www.census.gov/ces/dataproducts/datasets/lbd.html>

²⁴ <https://www.census.gov/ces/rdcresearch/index.html>

²⁵ <https://www.census.gov/ces/rdcresearch/howtoapply.html>

- While there is a re3data page at <https://www.re3data.org/repository/r3d100010200> (Re3data.Org, 2018), it does not provide any information on the FSRDC (the entry has several other issues as well, regarding license information, but those are not relevant here)
- the product page offers no structured information, and policy information is scattered throughout the website.

Appendix: Full query response for Use Case 1

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <response>
3
4 <lst name="responseHeader">
5   <int name="status">0</int>
6   <int name="QTime">10</int>
7 </lst>
8 <result name="response" numFound="1" start="0">
9   <doc>
10     <str name="datacentre">GESIS.ICPSR – ICPSR</str>
11     <str name="doi">10.3886/E100590V1</str>
12     <arr name="creator">
13       <str>McKinney, Kevin L.</str>
14       <str>Green, Andrew S.</str>
15       <str>Vilhuber, Lars</str>
16       <str>Abowd, John M.</str>
17       <str>Abowd, John M.</str>
18     </arr>
19     <arr name="title">
20       <str>Replication data: Total Error and Variability Measures
21         for QWI and LODES</str>
22     </arr>
23     <str name="publisher">ICPSR – Interuniversity Consortium for
24       Political and Social Research</str>
25     <str name="publicationYear">2017</str></doc>
26 </result>
27 </head></response>
```