IDCC19 | *Research Paper*

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 December 2018

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

The 14th International Digital Curation Conference takes place on 4–7 February 2019 in Melbourne. URL: http://www.dcc.ac.uk/events/idcc19/

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).[1] Actual published individual reproductions or replications are traditionally not very common (in economics, see Bell & Miller, 2013; Duvendack, Palmer-Jones & Reed, 2017). One possible cause has been attributed to the difficulty of obtaining the materials required to attempt verification of the reproducibility or replicability of empirical articles (Dewald, Thursby & Anderson, 1986; McCullough, McGeary & Harrison, 2006; McCullough & Vinod, 2003). In response, journals and learned societies started adopting what are called "data availability policies," though they also generally define rules regarding the availability of code and software. We refer to them throughout as data and code availability policies (DCAPs). In the social sciences, the major economics and political science journals published DCAPs in the mid-2000s (American Economic Association, 2008; nicholaseubank, 2014). However, doubts have been cast on the effectiveness of such policies (Chang & Li, 2017a; Höffler, 2017a; Stodden, Guo & Ma, 2013; Stodden, Seiler & Ma, 2018), leading to renewed calls for better reproducibility (Stodden et al., 2016), broad efforts to better define DCAPs (Center for Open Science, 2016; Hrynaszkiewicz et al., 2017), and increased enforcement of DCAPs (Duflo & Hoynes, 2018; Jacoby, Lafferty-Hess & Christian, 2017; Lars Vilhuber, 2019).

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.

Some of the lack of replicability identified by recent studies (Camerer et al., 2016; Chang & Li, 2015, 2017b; Höffler, 2017b; Stodden et al., 2018) occurs despite the fact that journals have policies that encourage the provision of replication packages. Evaluating compliance with policies as well as quality and utility of replication packages is arduous, if not impossible, due to a lack of consistent, reliable metadata on the materials provided to journals. In many cases, while a replication package is provided to the journal, the underlying data are not available within the replication package, due to a mix of non-compliance, legal, and ethical constraints on redistribution of the data.

Authors are increasingly being encouraged and trained in reproducible methods from the outset of their research projects (Christensen, Freese & Miguel, 2019; Wilson et al., 2016), rather than describing their data and code much later, *i.e.* after submission to journals. This includes carefully documenting provenance of third-party datasets being

[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). See also National Academies of Sciences, Engineering, and Medicine, 2019 for a similar definition.

used, and properly curating generated datasets (surveys, collected data, etc.) in data archives as soon as possible. Such early deposit allows more time for curation, potentially improving the quality of deposits. However, it conflicts with some (but not all) journal workflows, which integrate data deposit into the article submission process. Prior deposits may not be captured by the same metadata as in-workflow deposits.

Furthermore, in at least some social sciences, the use of pre-existing but non-public data has increased substantially (Chetty, 2012) and remains high: Kingi, Stanchi, Vilhuber and Herbert, 2018 show about 40% of economics articles using restricted-access data. Confidentiality and licensing constraints prevent authors from depositing such data in open archives. Data citation (Data Citation Synthesis Group & Martone, 2014) of such data is often challenging. Journals must rely on an increasingly diverse cadre of data-holding institutions, not all of which are or perceive themselves as 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 authors and journals to document the full provenance of the data underlying the scientific results in published articles. The resulting lack of transparency in data provenance is detrimental to the overall effort of increasing transparency in the sciences, in particular FAIRness of data access (Hagstrom, 2014).[2]

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 some of the key characteristics that journals care about in the case of supplementary materials that are held by third parties, within the context of FAIR: existence, accessibility, and permanence. Our intent in defining the metadata package is three-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. Finally, by providing the information in a machine-readable format, the evaluation of compliance with DCAPs can be more easily assessed systematically. Overall, metajelo aims to increase 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 and 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.

[2] We note that restricted access to data is not inherently incompatible with FAIR, as long as there exists metadata that is FAIR.

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. 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 (*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[3]), or have allowed for data and code to be stored more generally on any of a curated list of trusted[4] or approved whitelist of third-party repositories.[5] Each of these alternatives rely on a journal or publisher vetting the repositories and ascertaining that it meets some set of criteria, or relying on third-party vetting of repositories exists, such as CoreTrustSeal.[6] For instance, Nature Scientific Data, 2019 assesses relevance to the community, cost to researchers, data access conditions, repository longevity, data persistence and versioning. CoreTrustSeal, 2017 assesses similar criteria, as well as policies surrounding a set of requirements. The presence on a list of recommended data repositories, or a successful CoreTrustSeal certification, are strong indicators of robust and persistent archives.

However, in our experience (Kingi et al., 2018; Lars Vilhuber, forthcoming), only a few of the holders of restricted-access data appear on such lists. Large survey institutions, many national statistical offices, and nearly all private-sector holders of restricted-access data provide some information about accessibility, but nearly no (publicly accessible) information about data persistence, versioning, or citability of their data assets. While publishers and (some) funders expect that repositories support researchers in making data FAIR, many data providers have yet to respond. In some cases, data access by researchers is incidental, and data providers are not responsive to FAIR considerations, in particular for private sector and sub-national providers of administrative data. Even at the national level, regulations in various countries that aim to improve access to and preservation of data assets for research are very recent (Digital Economy Act of 2017 in the UK, Loi pour une République Numérique 2017 in France, and the Foundations for Evidence-Based Policymaking Act of 2018 in the United States, to mention only a few examples), and have yet to make a measurable impact.

Even when data are public-use, or even when the repository is indexed in re3data,[7] information about accessibility and permanence are incomplete or wrong. Institutions are also able to list multiple access and preservation policies, leaving it open which policy applies to a particular data object. See the Appendix for additional details.

[3] https://www.elsevier.com/authors/author-services/research-data
[4] CoreTrustSeal, https://www.coretrustseal.org/
[5] https://f1000research.com/for-authors/data-guidelines, https://www.nature.com/sdata/policies/repositories
[6] https://www.coretrustseal.org/
[7] https://www.re3data.org

To a large extent, the onus on reporting on these facets of data archives falls onto the researcher who uses these data, and will continue to do so for considerable time. 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 Archives[8] or the *Archives Nationales* in France[9]). In general, data stored in national archives is permanently archived; if it is not, this is clearly documented.[10] 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.[11] 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

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

[8] https://www.archives.gov/dc/researcher-info

[9] http://www.archives-nationales.culture.gouv.fr/

[10] 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.CensusBureauRecordsControlSchedule2009**). 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.CensusBureauRecordsControlSchedule1999**).

[11] In fact, some journals may not offer that option.

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 Framework[12], 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 Archive[13]? 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 the 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 of a digital object at openICPSR[14], properly recorded in DataCite, can at best reveal *existence*, but cannot reveal the remaining attributes (*access rules* and *persistence*) 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, 2018b). The PSID is a longitudinal household survey conducted by the University of Michigan, which began in 1968. More than 4000 peer-reviewed publications have used the data (PSID, 2018a). 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

[12] https://osf.io
[13] https://ada.edu.au/
[14] https://www.openicpsr.org

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 a 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.[15] 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).[16]

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) (Jarmin & Miranda, 2002; U.S. Census Bureau, 2017) is one of the most widely used microdata files in the Federal Statistical Research Data Centers (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. 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

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

[16] 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.

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 that our proposed metajelo package (described in more detail below) 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).[17] 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, attributes, and controlled vocabularies 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 DataCite[18], 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. 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.

[17] We originally attempted to resuse 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.

[18] https://www.datacite.org/

Re3data

The Re3data initiative (Re3data.Org, 2015; Rücknagel et al., 2015) addresses the goal of describing repositories via an online registry of research data repositories based on a common metadata standard 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

CrossRef[19] 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.

CoreTrustSeal

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

[19] https://www.crossref.org/

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. 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)[20], 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. JATS4R[21] (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 project[22] 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 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

[20] https://jats.nlm.nih.gov/
[21] https://jats4r.org/
[22] http://www.bfe-inf.org/resource/belmont-forum-data-publishing-policy-workshop-report-draft

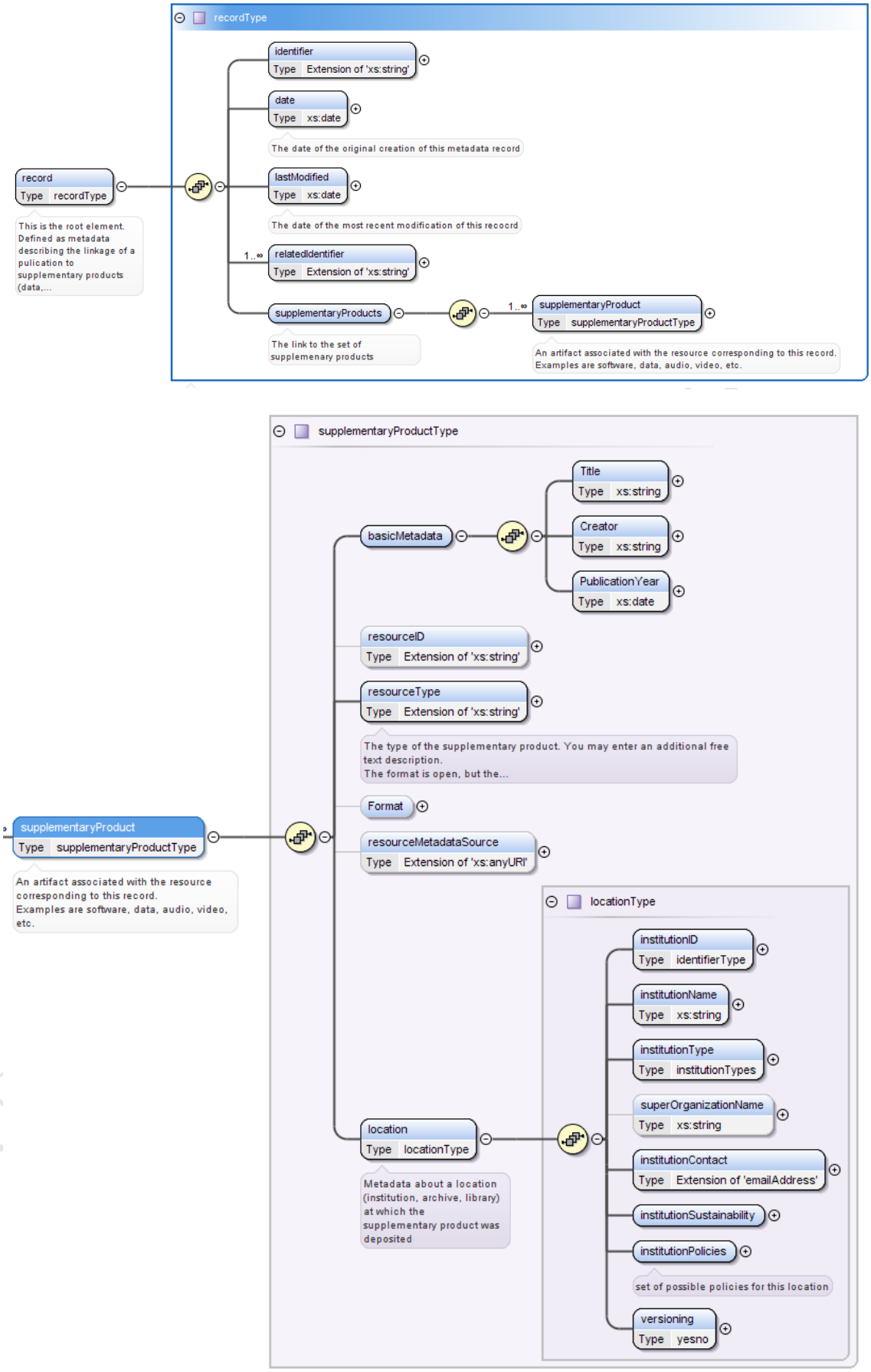

Figure 1. High-level structure of proposed package

Table 1. metajelo Description

[about here]

applicability of a policy to the respective supplementary object. The full annotated schema is available for examination online at github.com/labordynamicsinstitute/metajelo.

We highlight a few key elements. First, much of the information about the object itself mirrors the DataCite schema (DataCite Metadata Working Group, 2017a, 2017b), even if no DOI exists. The bibliographic metadata schema is based on DataCite for simplicity, and is always required. Much of the information on the institution, including its policies, mirrors the re3data schema (Re3data.Org, 2015; Rücknagel et al., 2015), with much simplification. In particular, we are interested primarily in policyType="Preservation Policy" and policyType="Terms of Use". In contrast with the re3data schema, we have merged licenses into the same repeatable element, so that policyType="License" is a valid option. In all cases, we also allow for verbatim capture of the text of the policy, since policies posted on websites, and not versioned, may change over time. We envision either manual entry by the researcher, or webscraping of the provided policy URL to populate this field.

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

American Economic Association. (2008). Data availability policy. Retrieved September 21, 2019, from https://web.archive.org/web/20180927113622/https://www.aeaweb.org/journals/policies/data-availability-policy

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

Center for Open Science. (2016). *TOP Guidelines summary table*. Center for Open Science. Retrieved November 19, 2019, from https://osf.io/kgnva/

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. (2017a). 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

Chang, A. C. & Li, P. (2017b). 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

Chetty, R. (2012). Time Trends in the Use of Administrative Data for Empirical Research. Retrieved July 19, 2018, from http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf

Christensen, G. S., Freese, J. & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. Oakland, California: University of California Press.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. doi:10.1126/science.aac4716

CoreTrustSeal. (2017). 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

Data Citation Synthesis Group & Martone, M. (2014). *Joint Declaration of Data Citation Principles*. Force11. doi:10.25490/a97f-egyk

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). 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.

Duflo, E. & Hoynes, H. (2018). Report of the search committee to appoint a data editor for the AEA. *AEA Papers and Proceedings*, *108*, 745. doi:10.1257/pandp.108.745

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

Hagstrom, S. (2014). The FAIR Data Principles. Retrieved May 20, 2018, from https://www.force11.org/group/fairgroup/fairprinciples

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

Hrynaszkiewicz, I., Birukou, A., Astell, M., Swaminathan, S., Kenall, A. & Khodiyar, V. (2017). Standardising and Harmonising Research Data Policy in Scholary Publishing. *International Journal of Digital Curation*, *12*(1), 65–71. doi:10.2218/ijdc.v12i1.531

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

Jacoby, W. G., Lafferty-Hess, S. & Christian, T.-M. (2017). Should Journals Be Responsible for Reproducibility? Retrieved July 22, 2018, from https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility

Jarmin, R. & Miranda, J. (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

Kingi, H., Stanchi, F., Vilhuber, L. & Herbert, S. (2018). *The Reproducibility of Economics Research: A Case Study*. Berkeley, CA. Retrieved from https://osf.io/srg57/

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., 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. (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*.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. doi:10.17226/25303

Nature Scientific Data. (2019). Scientific Data Repository Questionnaire. Retrieved December 22, 2019, from https://www.nature.com/documents/scidata-repository-questionnaire.docx

nicholaseubank. (2014). A Decade of Replications: Lessons from the Quarterly Journal of Political Science. Retrieved December 20, 2019, from https://thepoliticalmethodologist.com/2014/12/09/a-decade-of-replications-lessons-from-the-quarterly-journal-of-political-science/

PSID. (2018a). Home page: Panel Study of Income Dynamics (PSID). Retrieved December 6, 2018, from https://psidonline.isr.umich.edu/default.aspx

PSID. (2018b). 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. doi:10.17616/r3bc8q

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

Re3data.Org. (2017). Panel Study of Income Dynamics. doi:10.17616/R3503G

Re3data.Org. (2018). United States Census Bureau. doi: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. doi: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

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

Stodden, V., Seiler, J. & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 201708290. doi:10.1073/pnas.1708290115

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

Vilhuber, L. [Lars]. (forthcoming). Report by the AEA Data Editor (2020). *AEA Papers and Proceedings*, *110*.

Vilhuber, L. [Lars]. (2019). Report by the AEA Data Editor. *AEA Papers and Proceedings*, *109*, 718–729. doi:10.1257/pandp.109.718

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. & Teal, T. K. (2016). Good Enough Practices in Scientific Computing. *arXiv:1609.00037 [cs]*. arXiv: 1609.00037 [cs]. Retrieved December 8, 2018, from http://arxiv.org/abs/1609.00037

Appendix: Detailed Use Cases

In all use cases, we attempt to identify the three attributes outlined in the main text, using automated mechanisms.

Use Case 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). 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 LODES. 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. A subset of the response is depicted in Figure 2 (emphasis added). 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 (**DataCiteMetadataWorkingGroupDataCiteMetadataSchema2017**), 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 ICPSR[23]. The re3data record lists three types of data access. Furthermore, three data licenses are listed: two other and one copyright.

```
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

[23] https://www.openicpsr.org/openicpsr/faqs

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 knowledgable 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 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.[24] 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 rel identifier within the a link field (Figure 3) with an associated displayed license badge (Figure 4). The second instance is encoded in the JSON-LD payload,

[24] The query was run on 8 October 2018.

```
<div class="well">
  <p>
    <a rel="license" href="http://creativecommons.org/licenses/by/4.0/"
       target="_blank">
      <img src="/openicpsr/resources/images/cc.png"
       alt="Creative Commons License"/>
    </a>
    This work is licensed under a <a rel="license" href="http://creativecommons.org/li
       target="_blank">
       Creative Commons Attribution 4.0 International License</a>.
  </p>
  <p>openICPSR data are distributed exactly as they arrived from
       the data depositor. ICPSR has not checked or processed this
       material. Users should consult the investigator(s) if further
       information is desired.</p>
</div>
```

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

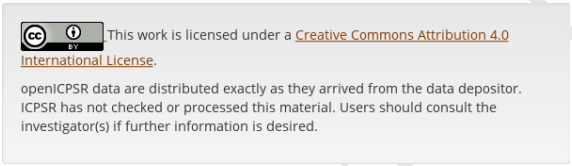

Figure 4. Use Case 1, license as displayed on website on 8 October 2018

```
"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.

Use Case 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.

Use Case 3: Restricted access at the U.S. Census Bureau

The LBD data (Jarmin & Miranda, 2002; U.S. Census Bureau, 2017) 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 here[25] and here[26]. 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

[25] https://www.census.gov/ces/rdcresearch/index.html
[26] 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.