Supplementary Materials

Lars Vilhuber and Carl Lagoze

December 2, 2018

1 Scenario 1: Public-use information at openICPSR

In the first case, the researcher has used public-use data, and identifies a Digital Object Identifier (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 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 (Figure 1). 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 [1], 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 [3]. We note, however, that the rules and policies for openICPSR may differ from ICPSR[1]. The re3data record lists three types of data access. Furthermore, three data licenses are listed: two `other` and one `copyright`.

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

```
<?xml version="1.0" encoding="UTF-8"?>

  <doc>

    <str name="datacentre">GESIS.ICPSR - ICPSR</str>

    <str name="doi">10.3886/E100590V1</str>

    </arr>

    <str name="publisher">ICPSR - Interuniversity Consortium for
         Political and Social Research</str>
```

Figure 1: Select lines from DataCite query for DOI 10.3886/E100590V1
The full query response can be found in the appendix.

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 on through CoreTrustSeal, for three reasons. First, searching again, as we

```
<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/licenses/b
        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 2: Use Case 1, Encoding of license in HTML of landing page

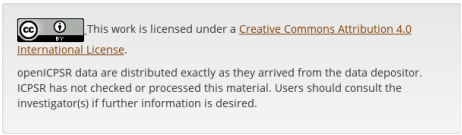

Figure 3: Use Case 1, license as displayed on website on 8 October 2018

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.[2] 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 2) with an associated displayed license badge (Figure 3). The second instance is encoded in the JSON-LD payload,

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

Both provide the same information about the license.

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

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" [2], 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.

2 Scenario 2: restricted-access PSID

The Panel Study of Income Dynamics (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`, 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.

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

The Longitudinal Business Database[3] (LBD) data at the U.S. Census Bureau is one of the most requested datasets in the Federal Statistical Research Data Center (FSRDC) network. Access procedures are described at various locations, including here[4] and here[5]. 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
- While there is a re3data page at https://www.re3data.org/repository/r3d100010200, 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.

References Cited

[1] DataCite Metadata Working Group. "DataCite Metadata Schema for the Publication and Citation of Research Data v4.1". In: (2017). Ed. by Jan Ashton et al. DOI: 10.5438/0015. URL: https://schema.datacite.org/meta/kernel-4.0/index.html (visited on 06/28/2018).

[2] Inter-university Consortium for Political and Social Research. *Digital Preservation Policies and Planning at ICPSR*. URL: https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/index.html.

[3] Re3data.Org. *Inter-university Consortium for Political and Social Research*. 2013. DOI: 10.17616/r3bc8q.

[3] https://www.census.gov/ces/dataproducts/datasets/lbd.html
[4] https://www.census.gov/ces/rdcresearch/index.html
[5] https://www.census.gov/ces/rdcresearch/howtoapply.html