

Some explorations

Lars Vilhuber

October 22, 2018

Unreliability of metadata

We picked a few DOIs where we know that reasonable data exists.

Archive 1

The first one is an openICPSR deposit:

Dataset:

McKinney, K. L., Green, A. S., Vilhuber, L., Abowd, J. M., & Abowd, J. M. (2017). Replication data: Total Error and Variability Measures for QWI and LODES [Data set]. ICPSR - Interuniversity Consortium for Political and Social Research. <https://doi.org/10.3886/e100590v1>

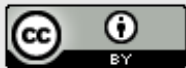
We note that the archive is listed as

Attribute	Value
container-title	ICPSR - Interuniversity Consortium for Political and Social Research
data-center-id	gesis.icpsr

neither of which completely describe the particular sub-repository within the ICPSR universe. More worryingly for our purposes, the license field (optional on DataCite) is empty:

Attribute	Value
license	NULL

even though the website lists a license:



This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

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.

Figure 1: openICPSR license display

Archive 2

The second deposit was picked because it is the most downloaded dataset as of 2018-10-22:

Dataset:

Harris, K. M., & Udry, J. R. (2008). National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use] [Data set]. Inter-University Consortium for Political and Social Research. <https://doi.org/10.3886/icpsr21600.v21>

The archive's name is again listed as

Attribute	Value
container-title	Inter-University Consortium for Political and Social Research
data-center-id	gesis.icpsr

The license is again listed as

Attribute	Value
license	NULL

Clearly, ICPSR uses a common registration mechanism both for uncurated open deposits (openICPSR) and formal curated archives (ICPSR). So let's explore a different archive.

Archive 3

The third deposit was picked because it is in a different archive (and we know the creators):

Dataset:

Vilhuber, L. (2018). Bibliography Of The Nsf-Census Research Network. <https://doi.org/10.5281/zenodo.1306968>

The archive's name is listed as

Attribute	Value
container-title	Zenodo
data-center-id	cern.zenodo

The license is listed as

Attribute	Value
license	https://creativecommons.org/licenses/by-nc-nd/4.0/

which corresponds to the license listed on the website (there is no default license on Zenodo).

Archive 4

The fourth deposit was picked because it has a restrictive license (and we know the creators):

Dataset:

Sexton, W., Abowd, J. M., Schmutte, I. M., & Vilhuber, L. (2017). Synthetic Population Housing And Person Records For The United States [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.556121>

The archive's name is listed as

Attribute	Value
container-title	Zenodo
data-center-id	cern.zenodo

The license is listed as

Attribute	Value
license	NULL

When a Zenodo deposit is restricted, it has no license attribute (and none is shown on the website).