IJDC | *Conference Pre-print*

Access Some Areas: Reforming Access Categories for Data in a Social Science Data Archive

Laurence Horton
Faculty of Information,
University of Toronto

Anja Perry
GESIS,
Leibniz Institute for the Social Sciences

Abstract

In this paper we outline the process of revising data access categories for research data sets in GESIS – a large European social science data archive based in Germany. The challenge is to create a minimal set of workable access conditions that cope with a) facilitating as "open as possible, closed as necessary" expectations for data reuse; b) map on to existing legacy access categories and conditions in a data archive.

The paper covers the work done in gathering data on data access categories used by data archives in their existing data catalogues, the choices offered to depositors of data in their user agreements, and work done by other data reuse platforms in categorising access to their data. Finally, we talk through the process of refining a minimal set of data access conditions for the GESIS data archive.

Submitted 16 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to Anja Perry, Unter Sachsenhausen 6-8, 50667, Koeln, Germany. Email: anja.perry@gesis.org

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

One of the hands that "make data work" are data archives. Data archives have a long history (over fifty years in some cases) of curating research data for preservation and reuse, often enhancing data by collecting documentation and metadata that can be used to create catalogue records for data discovery. A critical element of reuse is information on data accessibility: who can use the data and what, if any, restrictions are placed on use?

The expectation of the open science movement is for transparency throughout the research process, which, as one of its principles, includes an expectation for research data to be as open (Nielsen, 2011). While the open data movement expects minimal (attribution) to no restrictions on either audience or purpose of reuse, social science data archives operate in an environment where most data in their collections cannot be made available to everyone for any purpose. For example, our research finds that 98 percent of UK Data Service's datasets have some level of access or reuse restriction, while 85 percent of GESIS – Leibniz Institute for the Social Sciences is also restricted in some form. The reasons for closed as necessary are often built around data protection legislation, research ethics norms of care for research participants, or agreed expectation of confidentiality (Corti et al., 2014). They operate in an attitude of "as open as possible, as closed as necessary" (European Commission, 2016) with restrictions on who can access research data, how they access data, and the scope for reuse. Clear licencing is a requirement of findable, accessible, interoperable, and reusable (FAIR) metadata (Wilkinson et al., 2016). But while social science data archives have licences attached to data sets in their collection, there remains variation on the types and content of those licences – especially around the question of "commercial" use of data and scope of potential reusers.

In this paper we outline the process of revising data access categories for research data sets in GESIS – a large European social science data archive based in Germany. The challenge is to create a minimal set of workable access conditions that cope with a) facilitating as "open as possible, closed as necessary" expectations for data reuse; b) map on to existing legacy access categories and conditions in a data archive.

The paper covers the work done in gathering data on data access categories used by data archives in their existing data catalogues, the choices offered to depositors of data in their user agreements, and work done by other data reuse platforms in categorising access to their data. Finally, we talk through the process of refining a minimal set of data access conditions for the GESIS data archive.

We believe the results of this work will be of interest to the digital curation community. This is work that can contribute to those with an interest in standardisation of access and reuse categories for social science data archives. Specifically, those working in setting up data archives and repositories (especially if they contain data that has been pseudonymised from identifiable personal data), but also those working in established data archives that are looking to refresh their data access categories for an age of digital data on demand. The paper might also be of interest to copyright advisers on how to capture the challenges of protecting intellectual property and appropriate reuse for research data outside of Creative Commons templates (which are not specifically designed for data sets and not sensitive research data).

Methods

To do this work we decided to look at three areas of data. First, what currently exists in data archive collections? Second, what do data archives offer potential depositors in terms of licence agreements for data reuse? And finally, what work is being done on access classifications for sensitive research data, and on other potential restrictions on reuse like limiting usage to academic research projects only or excluding "commercial" usage?

Current archive Collections

Using the API OAI-PMH feeds provided by a range of European social science data archives, we gathered data on the contents of their collection through their catalogue records using either DDI (Data Documentation Initiative) or Dublin Core metadata schemas. GESIS, UK Data Service, and a range of other CESSDA data archives were chosen as they cover a range of European Social Science data archives who make their catalogue available through an API – either directly or through a DOI registration agency. R was used to import data catalogue information, clean (to control for missing information), and extract access condition information, including the prevalence of Creative Commons or template licence equivalents (Open Data Commons, Open Government). Data gathered through the API and replication code is being made available in Zenodo under a CC0 licence (Horton, 2019)

Current Data Depositor Licence Options

The next stage was to review licence options offered by fourteen European social science data archives (plus two self-deposit platforms offered by those archives) and 29 German academic research centres to potential depositors. In line with CoreTrustSeal requirements (CoreTrustSeal, 2016), data archives often provide information on the access categories they offer people who wish to deposit research data. Our aim here was to capture that information to see what options are presented to depositors and then compare across archives for commonalities and differences, but also compare, where possible, to the actual content of their data catalogues.

DataTags and Creative Commons Templates

The final stage was to look at work done in other data services on classifying data according to its sensitivity and at other data licencing initiatives on template approaches to data licencing. One significant source was the DataTags work produced by Dataverse (Sweeny et al., 2015) to provide criteria and categories for classifying research data according to the sensitivity of the data and how that can be applied to a data collection of around 1800 data sets, likewise guidance on licencing research data (Ball, 2014) contributes to our project. Work done (Creative Commons, 2009; Klimpel, 2012) on defining and clarifying the concept of "commercial use" is also used to address the common, but not particularly clear, restriction on "commercial use". We use this to clarify the difference between user and usage, to permit usage by non-academic research users if usage is not intended for monetary gain or commercial advantage. Finally, we looked at data orientated template licences created by Open Data Commons (2019).

Result

The result of our work was a set of three access categories for data sets: "Open" which uses CC BY as an open licence that meets the criteria for data available for any person for any use and can be applied to data where identifying information about individuals has either been removed or was never collected. "Accountable" covers the bulk of data sets in the archive. Here data has been pseudonymised to remove significant risk of re-identification, but does not remove all risk so mediation between the potential user and access to the data is required. This category retains a requirement to register and sign a user contract with the data service before being permitted to access the data. The default option in this category has also been changed to permit commercial usage using clearer language: "for use that includes commercial advantage or monetary compensation." And finally, "restricted" covers data where there is a significant risk of breaching data protection unless identity of users is verified, approved, and agreements on handling and use of the data are in place.

To address legacy issues and in the case of specific requests from depositors the archive feels should be met, we propose a set of atypical alternate categories that can slot in additional requirements for the controlled and restricted levels of data - like preventing commercial research organisations or upholding depositor approval for the use of research data.

Acknowledgements

The authors would like to thank Libby Bishop for guidance in this project.

References

Ball, A. (2014). How to license research data. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

Corti, L., Van den Eynden, V., Bishop, E., Woollard, M. (2014). Managing and sharing research data: A guide to good practice. London: Sage

Creative Commons. (2009). Defining "noncommercial": A study of how the online population understands "noncommercial use". San Francisco: Creative Commons Corporation. Available online: https://mirrors.creativecommons.org/defining-noncommercial/Defining_Noncommercial_fullreport.pdf

Data Seal of Approval/ICSU World Data System. (2016). Core Trustworthy Data Repositories Requirements v.01.00. Available online: https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

European Commission Directorate-General for Research & Innovation. (2016). Guidelines on FAIR Data Management in Horizon 2020 v.3.0 Brussels: European Commission. Available online: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Horton, L. (2019). GESIS – Leibniz Institute for the Social Sciences data access categories v.1.0) [data set]. Zenodo. doi:10.5281/zenodo.2566353

Klimpel, P. (2013). Free knowledge based on Creative Commons licenses: Consequences, risks and side-effects of the license module "non-commercial use only – NC". Berlin: Wikimedia Germany. Available online: https://openglam.org/files/2013/01/iRights_CC-NC_Guide_English.pdf

Nielsen, M. A. (2011). Reinventing discovery: The new era of networked science. Princeton. N.J: Princeton University Press.

Open Data Commons. (2019). Licences. Available online: http://opendatacommons.org/licences/index/html

Sweeney L., Crosas M., Bar-Sinai M. (2015). Sharing sensitive data with confidence: The Datatags system. Technology Science, 2015101601. Retrieved from https://techscience.org/a/2015101601

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3. doi:10.1038/sdata.2016.18