IASSIST Session 5P Summary: Big Picture Metadata, June 5, Toronto, CA

by San Cannon [1]

Abstract

This paper is a summary of three presentations given on metadata and related topics at IASSIST 40 in Toronto. Some of the text is adapted from the presenter's published abstracts; other parts were contributed by the session chair.

Keywords: Metadata, identifiers, ORCiD, DOI, DataCite, Blaise, MQDS, XML, DDI, best practices, open journal..

Using identifiers to connect researchers, authors and contributors with their research data

IThe first speaker in the session was Elizabeth Newbold from the British Library who was filling in for a colleague who had proposed the paper but then took another position elsewhere. She gave an update on the ORCiD and DataCite Interoperability Network (ODIN) project which was the subject of a session at IASSIST 2013. The project is collaboration between The British Library, CERN, ORCiD, DataCite, Dryad, arXiv and the Australian National Data Service with the aim of using persistent, open and interoperable identifiers for people and for datasets to connect researchers, authors and contributors with their research data.

These best practice descriptions will be modular with a homogeneous format, allowing reorganization in multiple ways

The presentation outlined some key results from the first year of the project including the proofs of concept (POCs) in the Humanities and Social Sciences (HSS) and High Energy Physics (HEP). They faced challenges in a few areas: access, discoverability, interoperability and sustainability. The project allows for the identification of contributors as well as authors. For the POCs, there was an extreme dichotomy between the HSS and HEP communities: there can be as many as 100 names associated with a paper in HEP and in some cases an entity and not an individual is associated with a paper. These practices are quite foreign to the HSS world. The project is, however, looking to outline commonalities between the extremely different disciplines..

The first year of the project centered on building the conceptual model for connection creators, curators, contributors, and data sets. This approach was more straightforward for new data being cataloged but much harder for data already being held. People are even harder to retrofit: assigning identifiers to inactive researchers may be problematic, especially if they are dead. And when researchers change institutions, assigning a new ID to a researcher that affiliates them with their current institutions causes problems if the institution they were at when the research was done still wants credit.

The second year focus was on identifying generic workflow and how to use it as a framework for implementing workflows for assigning DOIs and ORCiDs. When assigning metadata, there is no equivalent to the Data Documentation Initiative (DDI) in HEP. In order for the assignment of ORCiDs and DataCite to work well, there needs to be interoperability. The project includes a tool for claiming datasets within your ORCiD file. Work is also being done to link International Standard Name Identifiers (ISNI) to ORCiDs. Other work includes interacting with many stakeholders, including funders and policymakers. Another update is planned for after the fourth plenary of the Research Data Alliance (RDA) in September 2014. The Humanities and Social Science Proof of Concept report (*http://dx.doi.org/10.6084/m9.figshare.824317*) by John Kaye (BL), Tom Demeranville (BL), Steven McEachern (ADA) was published in July 2013.

Rich Metadata from Blaise

The second presentation was by Beth-Ellen Pennell of the Institute for Social Research and University of Michigan, to which Gina Cheung also contributed. This presentation focused on the process and challenges faced during the harmonization and preparation of the metadata and data files of the Collaborative Psychiatric Epidemiology Surveys (CPES) *http://www.icpsr.umich.edu/CPES/index.html*.

The CPES joins together three nationally representative surveys of adults living in the United States: the National Comorbidity Survey Replication, the National Survey of American Life, and the National Latino and Asian American Study. These data were collected face-to-face using the Blaise software, a product from Statistics Netherlands often used by statistical agencies and others.

The Michigan Questionnaire Documentation System (MQDS) was used to extract metadata from Blaise using DDI standards. This system is free to Blaise users and allows for data transformation from the Blaise format to other such as SAS, SPSS, and SQL. In this case, the Blaise data were transformed into XML capturing the rich metadata available in Blaise data models. The combined CPES dataset contains approximately 20,000 interviews. The initial combined dataset had 9.400 raw variables distributed over 92 sections of the three surveys which needed to be harmonized across the datasets. A survey instrument crosswalk was needed because the same instrument wasn't used in each survey. The sections appeared in different orders and even within a section, question order may vary. Initial cleanup of the data was required before documentation could be created.

The final dataset contains approximately 5,600 harmonized variables, 400 constructed variables and 14 separate weights. The website contains rich metadata including an interactive cross-walk of all harmonized variables with question text in 5 languages, response options, missing data codes, descriptive statistics (frequencies, etc.), universes, detailed documentation of all constructed variables, and descriptive statistics of all variables, among a wide variety of other products. Future work includes a move to DDI3.2 to cover the full survey lifecycle and dealing with the release of Blaise 5.

DDI Handbook – Overview and Examples of Recommended Best Practices

The final speaker in the session was Joachim Wackerow from GESIS – Leibniz Institute for the Social Sciences. This discussion introduced the DDI Handbook Project. The use of DDI is increasing both the number of users and producers of DDI materials as well as the number of new projects at GESIS that use DDI. The use of DDI is heterogeneous and many users find DDI to be complex because the subjects are complex. There is some documentation already available but coverage is incomplete, some documents are outdated, and for others the understanding has changed.

Building upon previous efforts at various DDI workshops, the project plans to produce a collection of best practices on using DDI using a community approach. The goal is compile a set of best practices aimed at a broad audience so that the output is useful and accessible information providing a balance between a book and a list of frequently asked questions. These best practice descriptions will be modular with a homogeneous format, allowing reorganization in multiple ways. The primary structure for the collection will be organized in alignment with the DDI Lifecycle. A goal will be to involve the DDI community in producing a shared body of resources for all organizations and individuals using the DDI specification.

The format is to create an independent open access journal that is published twice a year in coordination with, but not published by, the DDI Alliance. The platform will use the Open Journal System which provides online presentation and editorial management workflow. The system will provide a structured paper template and publish under the Creative Commons ShareAlike license. Multiple formats will be available for reuse (DocBook or DITA) and additional materials can be provided as XML files.

The submitted best practice documents will be reviewed by a team of editors and reviewers and published on a dedicated website. The editorial board is now being formed with 4-5 members who will set policies and make final decisions on papers. There will be an open peer review process, partly because the DDI community is small and partly because open peer commentary could result in a new article or new version of the original article. In addition, the content could be used as the basis for tutorials or other teaching materials.

Some initial topics may include guidelines for archives introducing DDI into their workflow and other institutions already using DDI Codebook and shifting some of their workflow to DDI Lifecycle. Another area of interest will be utilizing DDI for data discovery. The project is looking for outside involvement from the community in the form of comments, papers, and reviewers.

One audience member asked why the user community had to write their own documentation instead of having experts write it for them. The discussion centered on how the broader community can provide a larger selection of use cases from which others can learn. A suggestion to include "what not to do" or "lessons learned" articles as well as best practices was well received.

NOTES

1. .San Cannon, Deputy Chief Data Officer, Federal Reserve Board, Washington, D.C. 20551 *sandra.a.cannon@frb.gov*