Metadata for research data discovery and management

Lizz Jennings, Research Data Librarian (Systems), University of Bath

Background

Requirements for sharing research data have increased in recent years, partly in response to the open science agenda and partly as a means to make better use of data generated using public money. Research funders increased their expectations of researchers in relation to research data sharing, and the Engineering and Physical Sciences Research Council (EPSRC) was the first Research Council to put the onus on institutions to provide support for this.

In response to the EPSRC expectations[1], the University of Bath developed a Research Data Archive[2] for researchers to archive and share their data. This was developed using the institutional repository software, EPrints.[3] EPrints has a flexible approach to metadata, so developing a metadata schema for use in the Archive was a significant task.

Priorities

The University's Roadmap for EPSRC[4] set out several requirements for metadata, such as:

- making data discoverable within the University and through external data discovery services;
- relating data to publications and other information stored in the research information system, Pure;
- using external metadata standards such as DataCite[5] to represent the data and provide unique identifiers;
- providing scope to include documentation with the data; and
- automating data and metadata capture where possible.

In addition to the discovery requirements, we had to ensure that research data uploaded to the system had sufficient management metadata to provide for its future care.

Developing a metadata schema

We decided to align with the DataCite metadata schema as we intended to use their service to create Digital Object Identifiers (DOIs) for datasets, and any metadata would be included in their global search facility. We also worked with the Pure datasets working group to provide feedback on their approach for dataset metadata. This was also based on the DataCite schema, and it was important to align with Pure to provide connections between projects, publications and equipment related to datasets.

The initial phase of work was overseen by a working group, and sign off for required metadata was essential. Some of the required metadata was specified by DataCite:

- Identifier (the DOI itself)
- Title
- Creator(s)
- Publisher (University of Bath)
- Publication Year

Although some institutions felt this was sufficient, we had concerns that making data available with this limited set of descriptors would not further the aims of making the data discoverable or understandable. We therefore added the following:

- Abstract (a description of the dataset to help users decide whether to access the data)
- Department
- Research Funder
- Methodology (a detailed explanation of what the data are and how they were created)
- Rights Holder(s)
- Contact Person

In practice all datasets have the following fields set automatically, although they can be overridden:

- Version
- Language
- Resource type (introduced as a required field in version 4 of the DataCite metadata schema)

Implementing the metadata schema

For each of the metadata fields we used, we had to decide on whether we wanted to follow the DataCite schema, the Pure schema, or our own implementation. Generally, we preferred the DataCite schema where possible, as this was most compatible with other databases.

Where a format was not specified by DataCite, we considered the Pure implementation. In some cases, we felt the Pure implementation was not effective, such as with contact information. There was a conflict of use cases for this field. For researchers, this was effectively a chance to specify a corresponding author. For us, this needed to be an institutional rather than a personal contact. We were able to implement both use cases in the Archive by enabling each Creator to have an additional “contact person” field. We asked for them to nominate at least one person, and made the first Creator listed a default contact if they did not specify. This also helped us to record the primary contact for future reference, should there be a query with the dataset. For external users, there is a request access button on the record which appears if the data are not publicly downloadable.

However, Pure introduced useful metadata fields for capturing legal or ethical restrictions for internal use. This meant we were able to design EPrints to mirror the metadata included in Pure. The tick box fields mapped exactly, but we decided to have a single text box for the text information about ethics. This gave people the freedom to include further information beyond the main categories. For this field, we concatenated the text from the separate text boxes in Pure.

We also implemented fields which were unique to our repository. In these cases, we considered the information we were trying to capture and how structured that needed to be. In some cases, we decided to capture the metadata in a range of formats. For example, we expected researchers to include documentation. We provided metadata fields to support describing what the data were and how they were created. We also allowed researchers to link to external documentation or upload readme files. As a result, our guidance for using the fields has many caveats and options, which actually makes it harder for researchers to use the system. If we were to review this section, we would favour providing template readme documents and limiting the format to an uploaded file. This would make guidance clearer and enable some automatic validation.

Using the metadata

Import

Researchers start the process by registering their datasets in Pure. The Archive has a custom script that transforms Pure metadata into EPrints metadata. At first, this process was manual. We downloaded metadata from Pure as XML, then uploaded it into EPrints. However, I improved the import to enable direct fetching of XML from the Pure web service. This not only made the process faster, but also enabled us to capture metadata from related sections such as people and projects. Although we copied these metadata manually before this script was updated, the change has affected our researcher engagement. We now encourage researchers to link to projects when adding metadata to Pure as it results in more consistent metadata in the Archive.

I also developed an import script to capture metadata from the UK Data Service (UKDS).[6] This is the standard location for our ESRC-funded researchers to deposit their data. Initially, we intended this to form the basis of an import from DataCite metadata. However, it soon became clear that the UKDS DataCite metadata were minimal, but richer metadata were kept in Data Documentation Initiative (DDI)[7] format. I decided to write a specific DDI import script. This was more challenging than the Pure import, as we had not mapped our metadata to DDI fields, and had to make decisions about which fields most closely matched our schema. We were able to use the same technical process as for Pure to provide an import from the accession number, rather than needing to save the metadata and upload it.

Export

EPrints has a wide range of export formats as standard. At first the most important format for us was the DataCite schema. The DataCiteDoi plugin[8] we use to create DOIs, included version 2.2 of the schema, but only transformed a minimal set of fields. We updated and extended the default schema to version 3. This enabled us to make use of new relations, such as linking to additional metadata, and to list rights in detail, rather than having to summarise for the whole dataset.

We decided to include as much metadata as possible in our DataCite export. The aim was to make our datasets findable in their searchable metadata store. We realised that not all researchers would want full details being shared in this way, so developed the export to be able to exclude groups of metadata from the export file. All datasets still had to export the DataCite required metadata, plus some additional fields which were not considered open to unwanted use, such as version and language. Adding this flexibility had unintended benefits in practice, as we can exclude fields which meet our Archive metadata quality requirements, but do not work with the DataCite system. For example, if a researcher has included an ampersand in a file name, this caused an error when creating the DOI, but to escape the character to make it compatible might have negative consequences for future linking.

More recently, I introduced an export to JSON Linked Data. This embeds structured JSON metadata that Google understands in abstract page headers. Google have produced a draft content type for datasets.[9] We were interested in whether this worked for our diverse data types. The structured data testing tool[10] was important for developing this export format, as it provided a level of validation, and made the metadata human -readable. Of the required fields, only ‘variableMeasured’ was problematic. Although the description claimed the content type applied to any kinds of data, this field is only applicable to datasets providing quantitative data. We hope that this will make our research datasets discoverable outside of the academic bubble.

Reviewing and updating the schema

There have been two main drivers for updates we have made to the metadata in our Archive: external schema changes, and experience of how the system is used in practice.

The update of the DataCite schema to version 4 prompted a metadata review. Some of the new fields mapped closely to our existing metadata, such as the fields to capture detailed funding information. However, the new schema introduced a more detailed geographical polygon field. Previously we had intended to include values for all relevant fields in the Archive. However, we had not seen extensive usage of any geographical coordinate fields in any of our datasets. We decided not to implement this field in the Archive as it would have been a significant piece of development, and we did not feel there was demand to represent geographical information to this level of detail.

We have added new administrative fields in response to our improved understanding of how the system is used. For example, we needed to know different information about metadata-only records. We added metadata to capture information about where the data were held. We can now run reports to identify data which needs to be 'rescued' should another repository go out of business. We have also extended metadata in response to requests for data. We expect researchers to include a data access statement in their papers to explain how to access the underlying data. We added a new field to capture these statements, as this is already checked as part of our review process. This provides quantitative data on compliance, which is challenging to monitor in isolation. It has also given us a body of examples which we can use to inform guidance.

Conclusion

Metadata has been at the heart of developments of our research data archiving solution. The need to interoperate with internal and external systems has driven the direction of our developments. However, this has also been balanced by an increased understanding of the management and operational metadata that enable us to curate and preserve the data we hold.

Supporting Information

The metadata schema referred to in this article is available from https://doi.org/10.15125/BATH-00374.

References

1. EPSRC, 2011. Expectations. Available from: https://www.epsrc.ac.uk/about/standards/researchdata/expectations/ [Accessed 16 May 2017].
2. University of Bath, 2017. Research Data Archive. Available from: http://researchdata.bath.ac.uk/ [Accessed 16 May 2017].
3. Electronics & Computer Science, University of Southampton, 2017. EPrints. Available from: http://www.eprints.org/ [Accessed 16 May 2017].
4. Lyon, L. and Pink, C., 2012. University of Bath Roadmap for EPSRC: Compliance with Research Data Management Expectations. University of Bath. Available from: http://opus.bath.ac.uk/31279/ [Accessed 16 May 2017].
5. DataCite, 2016. DataCite Metadata Schema. Available from: http://schema.datacite.org/ [Accessed 16 May 2017].
6. ESRC, University of Essex, University of Manchester and Jisc, 2017. UK Data Service. Available from: https://www.ukdataservice.ac.uk/ [Accessed 16 May 2017].
7. DDI Alliance, 2014. DDI-Codebook 2.5. Available from: http://www.ddialliance.org/Specification/DDI-Codebook/2.5/ [Accessed 16 May 2017].
8. DataCiteDOI plugin. Available from: https://github.com/eprintsug/DataCiteDoi [Accessed 16 May 2017].
9. Google, 2017. Datasets. Available from: https://developers.google.com/search/docs/data-types/datasets [Accessed 16 May 2017].
10. Google, 2017. Structured Data Testing Tool. Available from: https://search.google.com/structured-data/testing-tool [Accessed 16 May 2017].