Citation for published version:
Ball, A 2020, 'Metadata for Better Data', *Catalogue and Index*, vol. 198, pp. 17-20.

Publication date:
2020

Link to publication

Metadata for better data

Alex Ball, Research Data Librarian (Systems), University of Bath

For much of the history of scholarly communication, academic research was operating in what we might call an era of small data. Papers were published reporting on just a handful of hard-won data points. At that scale, it is a realistic prospect to print the data within the paper, fully contextualised and described, and a tractable task for someone to copy them out by hand to use in calculations (Boulton, 2012).

At some point, all domains hit a point where the questions became too big to answer with such small datasets, and advances in technology allowed vastly more data to be generated, processed and analysed in a reasonable amount of time. To save space, full datasets in papers were replaced with summary tables, charts and graphs. In a handful of cases, domain repositories were established to curate the underlying data but in most cases the data points were lost to the literature. This lead to absurd situations such as scientists using rulers and magnifying glasses to extract data from graphs, but more seriously it resulted in a loss of transparency, reproducibility, and reusability. The potential for fraud, undetected error and wasted effort was enormous.

The last twenty years have seen a mounting backlash against this phenomenon. The idea that data is a lasting resource or asset instead of a throwaway by-product has spread from a handful of disciplines to governments, funding bodies and mainstream academia. In 2016, community discussions on the topic were crystallised into the FAIR Data Principles (Wilkinson et al., 2016), which unpack (to a degree) what it means for data to be Findable, Accessible, Interoperable and Reusable. Since then the FAIR Data Principles have been widely adopted, most notably within funder data policies such as the open data stipulations of the European Commission’s Horizon 2020 and Horizon Europe programmes (SPARC Europe, 2019).

Research Data Management is now a major concern in the research community, and supplementing the established domain repositories are a new breed of institutional repositories dedicated to research data. As a research data librarian in charge of one of these, one of my roles is to help researchers archive their data in a way that is FAIR, partly as a matter of compliance but first and foremost for the benefit of the scientific enterprise as a whole. Looking at the FAIR Data Principles in detail, it is striking that thirteen out of fifteen explicitly refer to metadata; the other two relate to the protocol for retrieving metadata. It is therefore vital that we collect the right metadata when ingesting datasets into our repository.

When setting up the University of Bath Research Data Archive in 2015, my predecessor Lizz Jennings had a degree of latitude in deciding what metadata we should collect. Her starting point was the scheme that came with the ReCollect plugin for EPrints, released by the Research Data @ Essex project in 2013 (Ensom, 2014), but she adapted it for our local needs and I have since added to it as our needs have evolved.

According to the FAIR Data Principles, for data to be Findable, it must have rich metadata, including a globally unique and ‘eternally persistent’ identifier, that is registered or indexed in a searchable resource. Of course, taken literally eternal persistence is impossible to judge, but on a practical level it means an organisational commitment that the identifier will never be re-used for a different concept or resource and that it will always be possible to look up the metadata using that identifier. In fulfilment of those criteria, we assign DOIs to the datasets we publish through the Archive. A condition of that process is that we provide a metadata record compliant with the DataCite Metadata Schema (DataCite Metadata Working Group, 2019). pertinent information as possible.

The minimum requirements of that schema are rather basic – just the typical elements of a reference list entry, such as creator and title – but in order to enrich our metadata as much as possible we collect and syndicate the recommended and optional elements of the schema as well. This is not to say all our records are complete according to the schema – some elements may not apply, or the depositor may be unable to provide the information – but as part of our pre-publication review process we work with depositors to collect as much pertinent information as possible.

The Archive itself is a searchable resource but in order for the data to be truly discoverable the metadata has to be available in systems where potential users would search habitually. Another advantage of registering DOIs is that the metadata we upload can then be queried in DataCite's own metadata search service, and harvested from there by other data discovery services. Indeed, ResearchFish,[1] a service that registers outputs of funded research projects, is collecting information about datasets from DataCite on a trial basis. But for greatest visibility, our datasets need to be findable using generic search engines, and to that end we translate our metadata records into Schema.org metadata for datasets in JSON-LD format and embed the result into the our dataset landing pages.[2] In combination with a machine-readable sitemap, this means our datasets show up in Google Dataset Search and Google Scholar among other places.

The FAIR concept of Accessible data is really an extension of that idea of eternal persistence discussed above: data and metadata must be retrievable by their identifier using an open standard protocol, and metadata must be accessible even after the associated data has been withdrawn. The first criterion is trivial to satisfy for metadata by making records available over HTTP(S) using URLs derived from the identifier, as is the case with DOIs, but sometimes it is good to take a broader view. For example, as well as direct access to landing pages we also make our records available via OAI-PMH (Open Archives Initiative, 2015), and I hope one day we may also support the newer ResourceSync protocol (Open Archives Initiative, 2017). Making data files themselves retrievable by identifier is a more debatable point; a dataset is often recorded as a set of files, and may not be suitable for network transmission if it is too large or in a non-digital medium, so some intermediate stops are to be expected in many cases.

Regarding the second criterion, we have an organisational commitment to provide a persistent landing page for each DOI we assign, and in the summer of 2019 we adopted a Tombstone Protocol which covers, among other things, procedures for withdrawing datasets which have exceeded their retention periods (Ball, 2020). In such cases, the data files are removed from the record and instead a non-availability notice ('epitaph') is added to the record to explain what has happened. The remaining metadata and documentation files are retained for inspection.

The criteria for Interoperable data are aspirational to some extent. The vision is that all data and metadata should be machine-actionable, which means at a minimum that the information must be structured in a format that a machine can easily read (e.g. XML, JSON); the data elements must be drawn from a known public scheme; and data values should where possible be given in a standard format (e.g. ISO 8601 for dates) or using a Semantic Web-compatible controlled vocabulary. As an institutional data repository, there is not much we can do to enforce this vision for data files, though of course we do run training for researchers on how to make their data amenable to automated processing. But we can and do ensure our metadata records are available in a range of standard formats. As well as the mappings already discussed to DataCite and Schema.org, the underlying ReCollect scheme was designed to be compatible with social science standard DDI (DDI Alliance, 2012) and the geospatial standard UK GEMINI.[3] Our subject headings are drawn from the RCUK Subject Classification scheme, and we record identifiers where available for creators and contributors (ORCID), funders (FundRef DOIs) and projects.

[1] https://researchfish.com/researchfish/

[2] https://schema.org/

[3] https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini

Going beyond FAIR, we have some immediate and practical needs for interoperability ourselves. Information on datasets is also collected by our Current Research Information System (CRIS), and so to avoid re-keying information, we have established two-way communication between the EPrints-based Archive and the Pure CRIS. We maintain an import plugin for EPrints that allows us to look up dataset records in Pure using its REST API and copy the information into a new EPrints record.

For the reverse operation, the Archive generates a list of the 20 most recently published dataset records translated into Pure's bulk import XML format. Pure monitors this page and imports any new records it finds. For this to work, we have had to align certain fields in EPrints to the conventions used by Pure, for example how issues of sensitivity are recorded. Happily, Pure's metadata is also aligned with the DataCite Metadata Schema, so we did not need to translate the controlled vocabulary used for contributor roles, for example. But in many other cases we have to maintain mappings between entities and values in Pure and those in EPrints, most especially people, projects, organisational divisions and scientific instruments. The situation is complicated further by the fact that the format used to present information through the Pure API is quite different to Pure's bulk import format, so the two transformations have to be maintained independently: they are not the inverse of each other.

Finally, for data to considered Reusable, it must have a clear licence, its provenance must be known, and it must meet relevant community standards. To assist with this, we explicitly release all our metadata under a Creative Commons Zero Universal Public Domain Dedication. At the point of reviewing deposited datasets, we strongly encourage depositors to apply licences to the data files they upload; we recommend using a Creative Commons Attribution v4 licence for open access data, since this permits the greatest flexibility of use while protecting the right of those involved in its creation to receive due credit for their work.

Regarding provenance, we rely on the facilities built into our EPrints repository software for tracking the version history of the metadata record, and once published the data files cannot be altered. Regarding community standards, we recognise that while the DataCite Metadata Schema meets generic expectations for discovery metadata, it is not sufficient for domain-specific needs. As a generalist repository, it would not be practical for us to design a native scheme that covered all domain needs, so instead we provide a facility to attach metadata files to a record, so that alternative records conforming to specific domain standards can be included alongside the data and main record.

While we do strive for FAIRness with our data, we have to be realistic and acknowledge that as an institution we need to learn to walk before we can run. Fully machine-actionable data and metadata is an ultimate goal, but the fact is that existing data structures are not so granular or widespread that they can convey all that is needed to communicate the meaning of the data in a fully machine-actionable way, and for that matter the software that would be needed to read it is not universally available either. Our more immediate task is to try to ensure that data is accompanied by enough documentation that a peer researcher could understand it, much as it used to be in earlier scholarly communications. We encourage structure as much as we can – we divide our documentation elements into collection methodology, data preparation (e.g. cleaning or anonymisation processes), technical requirements and notes on encoding, and we provide structured README templates for more complex cases – but I foresee it will be a very long time before we can dispense entirely with natural language, human-to-human (or human-to-AI?) text for communicating the nuances of meaning in the data.

References

Ball, A. (2020, February 18). *The Tombstone Protocol: an undertaking for unfortunate events*. 15th International Digital Curation Conference, Dublin, Ireland. https://doi.org/10.5281/zenodo.3674987

Boulton, G. (2012, October 30). *Why does open data matter and how can we make it a reality?* 23rd International CODATA Conference, Taipei, Taiwan.

DataCite Metadata Working Group. (2019). *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data* (Version 4.3). DataCite e.V. https://doi.org/10.14454/7xq3-zf69

DDI Alliance. (2012). *DDI-Codebook 2.5*. https://ddialliance.org/Specification/DDI-Codebook/2.5/

Ensom, T. (2014, September 12). *ReCollect: a research data plugin for EPrints*. https://researchdataessex.wordpress.com/2014/09/12/recollect-research-data-plugin-for-eprints/

Open Archives Initiative. (2015). *The Open Archives Initiative Protocol for Metadata Harvesting, Version 2*. http://www.openarchives.org/OAI/openarchivesprotocol.html

Open Archives Initiative. (2017). *ResourceSync Framework Specification*. http://www.openarchives.org/rs/1.1/resourcesync

SPARC Europe. (2019, May 7). *Open Science essential for new Horizon Europe funding programme*. https://sparceurope.org/open-science-essential-for-new-horizon-europe-funding-programme/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L. B. da S., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, Article 160018. https://doi.org/10.1038/sdata.2016.18