SciDataCon 2016

Conference paper

The relevance of documentation for the reuse of published datasets

Summary

Data publication is increasingly regarded as important scientific achievement and data publications are now fully citable in journal articles. This paper focuses on improving the reusability of data publications by providing comprehensive data descriptions complementary to standardized metadata. In this context, data reports proved to be a helpful tool to fill the gap between restricted 'README' information on one hand and preparing an extended peer-reviewed data article on the other hand.

Introduction

Data publication is increasingly regarded as an important scientific achievement and the number of data repositories is rising rapidly. Until May 2016, re3data.org, the registry of research data repositories, recorded 1587 data repositories across all scientific fields and 35 % of them are already assigning persistent identifiers (mainly DOI) to published datasets (Pampel et al., 2013).

The practice to cite datasets along with other sources is likely to achieve considerable momentum and is already becoming an important incentive for scientists to publish and share research data. A major step for the general acceptance of published datasets as important scientific outcome was the 'Statement of Commitment from Earth and Space Science Publishers and Data Facilities' (COPDESS, 2015, Hanson et al., 2015) that, in May 2016, has already been signed by 40 leading publishers and data facilities. All signatories have committed to regard datasets as legitimate products of research and allow the inclusion of dataset DOIs in reference lists of journal articles according to the "Joint Declaration of Data Citation Principles" (Data Citation Synthesis Group, 2014). COPDESS encourages scientists to cite dataset DOIs in their articles similar to the citation of research articles and other sources, i.e. to cite them in the text and include the full reference, including the DOI, in the reference list.

With the increasing number of data repositories and published datasets following the international requests by funding agencies and governments, it has become necessary to define quality standards for data repositories to make sure that the published datasets are accessible and intelligible for long-term reuse and validation of, e.g., results that have been derived from the data. The most popular among these certificates are the ICSU World Data System (WSD) and the Data Seal of Approval (DSA). Certified, or 'trusted', repositories guarantee professional and reliable long-term access to curated and quality-controlled research data and are undergoing a regular cycle of audit and certification.

Moreover, FORCE 11, a community of scholars, librarians, archivists, publishers, and research funders that originally arose to facilitate the change toward improved knowledge

DOI: http://doi.org/10.2312/GFZ.LIS.2016.002

creation and sharing, has recently developed the FAIR Principles, Guiding Principles for Findable, Accessible, Interoperable and Reusable Data Publishing (Wilkinson et al., 2016). These include FAIR data for machines (e.g. searchable, with persistent identifier, open protocols, standardized metadata, etc.), but also for people, which mainly involves the full documentation of the provenance and context of datasets (Hills et al., 2015), including a substantial quality control of the data and metadata by domain-experts (Lawrence et al., 2011).

Data reports – a "missing link" for data descriptions

In the following sections we focus on the 'R' of FAIR = reusable datasets. This has on the one hand a technical part (data format/standard/machine-readable) already addressed by Open Knowledge under the 'Frictionless Data' banner and the W3C and on the other hand a more narrative part for descriptive information that cannot be covered by standardized and machine-executable metadata. The latter is often the most time-consuming part in the data publication workflow for both the scientists documenting their data and the data curators involved in the data publications, especially whenever long-tail data is involved. It was already in 2012 when the Royal Society defined a set of requirements for 'intelligent openness' of research data, including that the datasets "must be accompanied by explanatory metadata for data discovery and reuse" (The Royal Society, 2012).

What is required to make data reusable? Where do we have to think beyond controlled vocabulary of standardized metadata? What are the best formats for data description? Data repositories always use controlled vocabulary to fulfil metadata standards for data discovery (e.g. Dublin Core, ISO19115, DataCite), which are machine readable, essential for data discovery, database interoperability, and metadata dissemination to portals, but often too strongly controlled to provide usable datasets if not complemented by additional documentation (Lawrence et al., 2011).

The 'classical format' for this additional data description is to directly attach short technical remarks as README files to the datasets, ideally as data package. The information provided this way is quite limited and cannot contain images. In addition, the packages have to be fully downloaded to access the documentation file.

Another format for data documentation is the publication of data-description articles in one of the newly developed Data Journals (like, e, g., Earth System Science Data ESSD, Scientific Data, CODATA Data Science Journal). Data journals publish peer-reviewed articles describing original research datasets, collections or databases with a strong focus on the methodology and data provenance and without an interpretation of the data itself. An increased popularity of data publications makes data journals an attractive publishing platform for scientists. The first data journals are now being indexed, for example, in tools like Thomson-Reuters' Web of Science and have already achieved impressive impact factors (e.g., Earth System Science Data, http://www.earth-system-science-data.net/about/news_and_press/2016-06-17_first-impact-factor-for-essd.html).

Seeing both documentation formats described above, it becomes clear that there must be something in between. README files are often not suitable for long data descriptions or explanatory figures and not every scientist wants to make the effort of writing a data article including the peer-review process, which is as much work as any other scientific article. It is

DOI: http://doi.org/10.2312/GFZ.LIS.2016.002

also in question if detailed technical descriptions of datasets, e.g. product guides of higher remotely sensed data products (Bartsch et al., 2011) or pages mainly filled with the definition of variables given in the data files (Lorenz et al., 2015) will be suitable for an article in a data journal. To fill in this gap, data reports may act as 'missing link', especially for comprehensive technical descriptions of datasets. Reports have a long tradition, but often with a bad reputation as 'grey' literature (i.e., they were not peer reviewed and only printed in small editions, hence difficult to access). However: they were and are an important additional source of information. Today, they are published online with an assigned persistent identifier, are world-wide searchable through their additional metadata and they are often internally reviewed. With this, they have lost much of their 'greyness'.

The GFZ German Research Centre for Geosciences is the national laboratory for solid-earth geosciences in Germany and part of the Helmholtz Association, Germany's largest scientific organization. The datasets, archived in and published by the GFZ Data Repository cover all geoscientific disciplines and range from large dynamic datasets deriving from data intensive global monitoring networks with real-time data acquisition (seismic, geodetic, geomagnetic, etc.), to automatically generated data publications of climate station data and the full suite of long-tail data in form of, e.g., remotely sensed satellite products, complex model results, geochemical analyses from various labs, individual field observations, and many more. The data repository of GFZ Data Services (Ulbricht et al., 2016) has a large focus on the DOI-referenced publication of these small and highly variable datasets with the aim to reach a high grade of reusability through a comprehensive data description.

For the reasons given above, GFZ has opened his traditional report series 'Scientific Technical Reports – STR' for the technical description of datasets published through the GFZ Data Repository. These 'Scientific Technical Report STR – Data'. (STR – Data) have individual DOIs and are linked to the datasets via their metadata (see below). STR – Data are internally reviewed by discipline experts and offer a full and consistent overview and description to all relevant parameters of a linked published dataset (Foerster et al., 2015a, Lorenz et al., 2015). To facilitate reuse, GFZ Data Services is developing standardized templates for each discipline (for contents and layout as appropriate, e.g. EnMAP Technical Reports, http://search.datacite.org/data-centers/tib.gfzbib?query=enmap). Data reports have been proven to be a helpful tool to fill the gap between basic metadata and restricted README information on one hand and preparing extended peer-reviewed data articles on the other hand.

Technical requirements

In addition to the data description by metadata and textual documentation, data reuse is also strongly dependent on the data format and the cross-reference between the datasets and the data description. The formats of the datasets published by the GFZ Data Repository are as variable as the geoscientific disciplines and follow community-specific standards. We hereby avoid proprietary formats (e.g. Matlab, Excel) wherever possible or provide them in addition to non-proprietary formats (e.g. ASCII, csv, NetCDF). Datasets and README files are combined in one 'data package' (e.g. zip folder) and DataCite metadata (DataCite, 2014) is used to cross-reference through 'related identifiers' this package, a data report, a data article, or a scientific article describing or using the dataset. Furthermore, we make use of the 'relation types' of the DataCite metadata schema to classify the related material in dataset

DOI: http://doi.org/10.2312/GFZ.LIS.2016.002

documentation, supplement to a journal article, and material for further reading. On the DOI Landing Pages of the datasets, the key references for the data description (i.e. the data report, the data article and/ or the scientific article) are prominently highlighted as 'Data Description' (Foerster et al., 2015b).

Competing Interests

The authors declare that they have no competing interests.

References

Bartsch, A; Naeimi, V; and Melzer, T: 2011 *ESA DUE Permafrost - ASCAT Surface Soil Moisture/Freeze-Thaw V1 product guide*, Vienna, http://hdl.handle.net/10013/epic.38769

COPDESS 2015 Coalition on Publishing Data in the Earth and Space Sciences: Statement of Commitment from Earth and Space Science Publishers and Data Facilities Available at http://www.copdess.org/statement-of-commitment/ [Last accessed: 25 June 2016].

Data Citation Synthesis Group 2014 Joint Declaration of Data Citation Principles Available at https://www.force11.org/datacitation [Last accessed: 25 June 2016].

DataCite: 2014 *DataCite Metadata Schema 3.1 XML Schema*, http://doi.org/10.5438/0011

Foerster, S; Brosinsky, A; Wilczok, C; and Bauer, M: 2015a *Isábena 2011 - An EnMAP Preparatory Flight Campaign*, Potsdam, Germany: GFZ Data Services, http://doi.org/10.2312/enmap.2015.007

Foerster, S; Brosinsky, A; Wilczok, C; and Bauer, M 2015b Isábena 2011 - An EnMAP Preparatory Flight Campaign (Datasets), GFZ Data Services, http://doi.org/10.5880/enmap.2015.007

Hanson, B; Lehnert, K; and Cutcher-Gershenfeld, J 2015 Committing to Publishing Data in the Earth and Space Sciences, *Eos*, 96, http://doi.org/10.1029/2015eo022207

Hills, D; Downs, R R; Duerr, R; Goldstein, J; Parsons, M; and Ramapriyan, H 2015 The Importance of Data Set Provenance for Science, *Eos*, 96, http://doi.org/10.1029/2015eo040557

Lawrence, B; Jones, C; Matthews, B; Pepler, S; and Callaghan, S 2011 Citation and Peer Review of Data: Moving Towards Formal Data Publication, *International Journal of Digital Curation*, 6, 4-37, http://doi.org/10.2218/ijdc.v6i2.205

Lorenz, H, et al.: 2015 *COSC-1 operational report: explanatory remarks on the operational data sets*, Potsdam, Germany: GFZ German Research Centre for Geosciences, 26 pp., http://doi.org/10.2312/ICDP.2015.001

Pampel, H, et al. 2013 Making research data repositories visible: the re3data.org Registry, *PLoS One*, 8, e78080, http://doi.org/10.1371/journal.pone.0078080

The Royal Society 2012 *Science as an open enterprise*, London, UK, The Royal Society,

Ulbricht, D; Elger, K; Bertelmann, R; and Klump, J 2016 panMetaDocs, eSciDoc, and DOIDB—An Infrastructure for the Curation and Publication of File-Based Datasets for GFZ Data Services, *ISPRS International Journal of Geo-Information*, 5, 25, http://doi.org/10.3390/ijgi5030025

Wilkinson, M D, et al. 2016 The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, 3, 160018, http://doi.org/10.1038/sdata.2016.18

Authors:

Kirsten Elger, Damian Ulbricht
GFZ German Research Centre for Geosciences, Potsdam, Germany
kelger@gfz-potsdam.de

DOI: http://doi.org/10.2312/GFZ.LIS.2016.002