Geophysical Research Abstracts
Vol. 19, EGU2017-8594, 2017
EGU General Assembly 2017

Challenges and Best Practices for the Curation and Publication of Long-Tail Data with GFZ Data Services

Kirsten Elger, Damian Ulbricht, and Roland Bertelmann

GFZ German Research Centre for Geosciences, Potsdam, Germany (kirsten.elger@gfz-potsdam.de)

Open access to research data is an increasing international request and includes not only data underlying scholarly publication, but also raw and curated data. Especially in the framework of the observed shift in many scientific fields towards data science and data mining, data repositories are becoming important player as data archives and access point to curated research data. While general and institutional data repositories are available across all scientific disciplines, domain-specific data repositories are specialised for scientific disciplines, like, e.g., bio- or geosciences, with the possibility to use more discipline-specific and richer metadata models than general repositories.

Data publication is increasingly regarded as important scientific achievement, and datasets with digital object identifier (DOI) are now fully citable in journal articles. Moreover, following in their signature of the "Statement of Commitment of the Coalition on Publishing Data in the Earth and Space Sciences" (COPDESS), many publishers have adopted their data policies and recommend and even request to store and publish data underlying scholarly publications in (domain-specific) data repositories and not as classical supplementary material directly attached to the respective article.

The curation of large dynamic data from global networks in, e.g., seismology, magnetics or geodesy, always required a high grade of professional, IT-supported data management, simply to be able to store and access the huge number of files and manage dynamic datasets. In contrast to these, the vast amount of research data acquired by individual investigators or small teams known as 'long-tail data' was often not the focus for the development of data curation infrastructures. Nevertheless, even though they are small in size and highly variable, in total they represent a significant portion of the total scientific outcome. The curation of long-tail data requires more individual approaches and personal involvement of the data curator, especially regarding the data description. Here we will introduce best practices for the publication of long-tail data that are helping to reduce the individual effort, improve the quality of the data description.

The data repository of GFZ Data Services, which is hosted at GFZ German Research Centre for Geosciences in Potsdam, is a domain-specific data repository for geosciences. In addition to large dynamic datasets from different disciplines, it has a large focus on the DOI-referenced publication of long-tail data with the aim to reach a high grade of reusability through a comprehensive data description and in the same time provide and distribute standardised, machine actionable metadata for data discovery (FAIR data). The development of templates for data reports, metadata provision by scientists via an XML Metadata Editor and discipline-specific DOI landing pages are helping both, the data curators to handle all kinds of datasets and enabling the scientists, i.e. user, to quickly decide whether a published dataset is fulfilling their needs. In addition, GFZ Data Services have developed DOI-registration services for several international networks (e.g. ICGEM, World Stress Map, IGETS, etc.). In addition, we have developed project-or network-specific designs of the DOI landing pages with the logo or design of the networks or project