Geophysical Research Abstracts
Vol. 13, EGU2011-6164, 2011
EGU General Assembly 2011
© Author(s) 2011

Acquiring High Quality Research Data for Publication

Andreas Hense and Florian Quadt
Bonn-Rhine-Sieg University oAS, 53757 Sankt Augustin, Germany

Acquiring High Quality Research Data for Publication

In all research endeavours where data play a central role the expectations of verifiability of experiments have grown and the need for reusing and recombining existing data sets for further investigations is growing, too. The publication of data is still under development, and the state of the art differs a lot in the research communities and disciplines.

In comparison to traditional text publications, data publications present some new challenges such as
. a variety of file formats for data, which are optimised for the needs of a certain domain,
. different characteristics of data files and research texts,
. varying procedures for scientific and formal quality assurance,
. different approaches for indexing data files,
. high requirements concerning storage sites and repositories.

How can a data repository acquire high quality research data despite the challenges mentioned above? The answer is simple: the repository has to be attractive for researchers, both in the externalisation (publishing) and the internalisation (learning) phase of their individual knowledge-spiral. Only if a repository attracts enough high quality submissions, a thorough selection is possible, and many researchers will use the data sets in a repository if quality and quantity are right.
The attractiveness of a repository can be seen as a first high-level key factor. Attractiveness itself depends on the following three key factors:
. Good reputation of the repository in the research community,
. high reliability of the technical infrastructure,
. user-friendly submission and quality assurance processes.

Based on the knowledge-spiral, we have developed a data publication cycle. According to this data publication cycle, the acquisition of high quality data mainly happens in an externalisation stage (publishing phase). The most important activities in this context are interpreting and annotating data, performing quality assurance, and publishing and exposing of the data.

Within these activities, we present three examples on measures how to improve user-friendliness and visibility of repositories. Two examples are from Australia, one is from our own project.

From conversations with other researchers we have learned that they often do not know where to store their data permanently and that data publication appears difficult and costly to them. We found a very promising approach in the e-Research community of Australia called Data Fabric. This service provides researchers with free data storage, which can also be shared and offers support for publishing data. For solar, geophysical and related environmental data there exists the World Data Center System.

Scientific quality assurance on data needs some kind of computer support in order to deal with huge amounts of data. From our experience, quality assurance of the primary data is mostly done by the author himself. There are efforts to support the author in this process by providing software tools which inspect the data, visualise them and hint at abnormalities. We are currently developing a standalone software package for reviewing meteorological data. The reports generated by this software together with the data can be submitted to the long-term archive to document the quality measures. Our software also supports the process of metadata review. A web-based software reads existing metadata from the long-term archive and presents the metadata as a series of forms that the user can traverse, similarly to a software installation wizard.

Regarding publishing data and exposing them to catalogues and search engines we demonstrate services of another Australian national initiative called Australian National Data Service (ANDS) and present the international initiative DataCite.