Geophysical Research Abstracts
Vol. 13, EGU2011-2385, 2011
EGU General Assembly 2011

Lessons learned in Chemistry: Building an Embedded Infrastructure for Primary Data from the Scratch – a Concept Study

Oliver Koepler, Janna Neumann, Irina Sens, and Jan Brase
German National Library of Science and Technology, Hannover, Germany

Over the last years publication of primary data has tremendously increased. The German National Library of Science and Technology (TIB) supports this new way of communicating scientific information. As a result the TIB became the first registration agency for primary data worldwide in 2005. Primary Data registered with a DOI becomes persistently citeable, accessible and searchable. In the beginning most projects of the TIB and their partners focused on domains with well established infrastructures for data publication and data archiving like Meteorology, Earth Observation, and Climate Simulation. We now report on a concept study in the domain of chemistry where no standards exists to persistently store chemical primary data in data centres and make it citable and linkable by the use of DOIs.
The concept study analyses the scientific workflow of chemical primary data in the scientific process. It identifies questions as well as requirements for an embedded infrastructure to publish and store primary data, and illustrates possible solutions for a prototypical implementation.
The scientific workflow and user requirements handling primary data were determined with a questionnaire about the current use of chemical data in daily work. While most researchers support the idea of long-term archiving their primary data and register datasets with DOIs, several challenges have been addressed. For once chemistry is divided into many sub disciplines with their own specific data formats; there is a lack of open, commonly used data formats to store data persistently. It is commonly agreed that archiving and publishing datasets should be easy to achieve and easy to implement in current workflows. An independent or separate publication of primary data is not a common practice in chemistry; data is almost always connected to an article publication. Therefore a vision of an extended publication process including primary data has been developed, including a proposed standard both for a common and a domain specific metadata scheme for the description of chemical primary data. Due to the diversity of the chemical domain, only a selected set of sub disciplines is considered in the first draft of a metadata scheme.
Technical aspects of storing chemical primary data are discussed in detail within the concept study, showing possible solutions regarding the challenge of format diversity, proprietary formats, the lack of open standards, and quality assurance of data integrity.