Geophysical Research Abstracts
Vol. 13, EGU2011-3469, 2011
EGU General Assembly 2011

A generic data model - from files to database

Dirk Fleischer (1), Kai Jannaschk (2), Hela Mehrtens (1), Carsten Schirnick (1), and Pina Springer (1)
(1) Leibniz Institute of Marine Sciences (IFM-GEOMAR), Kiel, Germany (datamanagement@ifm-geomar.de), (2) Christian-Albrechts University, Department of Computer Sciences, Kiel, Germany

The Kiel Data Management Infrastructure (KDMI) started from a cooperation of three large-scale projects (SFB574, SFB754 and Cluster of Excellence The Future Ocean) and the Leibniz Institute of Marine Sciences (IFM-GEOMAR).
KDMI key features focus on the data provenance which we consider to comprise the entire workflow from field sampling or measurements through lab work to data calculation and evaluation. Managing the data of each individual project participant in this way yields the data management for the entire project and warrants the reusability of (meta)data. Accordingly scientists provide a workflow definition of their data creation procedures resulting in their target variables.
The central idea in the development of the KDMI presented here is inspired by the object oriented programming concept which allows to have one object definition (workflow) and infinite numbers of object instances (data). Each definition is created by a graphical user interface and produces XML output stored in a database using a generic data model. On creation of a data instance the KDMI translates the definition into web forms for the scientist, the generic data model then accepts all information input following the given data provenance definition. An important aspect of the implementation phase is the possibility of a successive transition from daily measurement routines resulting in single spreadsheet files with well known points of failure and limited reusability to a central infrastructure as a single point of truth.
An interim system allows users to upload and share data files from cruises and expeditions. It relates files to metadata such as where, when, what, who etc. As a proof of concept we use a 'truncated workflow' to migrate a selection of marine chemical data files and their structured metadata into the generic data model. A web application will allow data extraction for selectable parameters, time and geocoordinates. The availability of these widely used data is expected to motivate more scientists to design their own workflows for their upcoming work and their resulting data.
This data provenance approach in terms of human workflows has several positive side effects: (1) the scientist designs the extend and timing of data and metadata prompts by workflow definitions oneself while (2) consistency and completeness (mandatory information) of metadata in the resulting XML document can be checked by XML validation. (3) Storage of the entire data creation process (including raw data and processing steps) provides a multidimensional quality history accessible by all researchers in addition to the commonly applied one dimensional quality flag system and thus (4) improves the reuseability of the data. (5) The KDMI concept focuses on bringing data management infrastructure into the daily measurement routines instead of the final data management hassle at the end of each project. (6) The KDMI can be extended to other scientific disciplines or new scientific procedures by simply adding new workflow definitions. The data input can start from this point while domain specific outputs with the newly added data instances will be created by the KDMI-Team.
The KDMI is inspired by social network systems but instead of sharing privacy or making friends the KDMI communities (projects, working groups, etc.) are all about sharing daily scientific work and data on a collaborative working platform for project partners.