Components and Aspects of an Integrated Data Management Approach

Hela Mehrtens*, Pina Springer*, Dirk Fleischer*, Carsten Schirnick*, Kai Jannaschk**
*Leibniz Institute of Marine Sciences (IFM-GEOMAR), Kiel; **Department of Computer Science, Christian-Albrechts-Universität, Kiel

Kiel Data Management Infrastructure

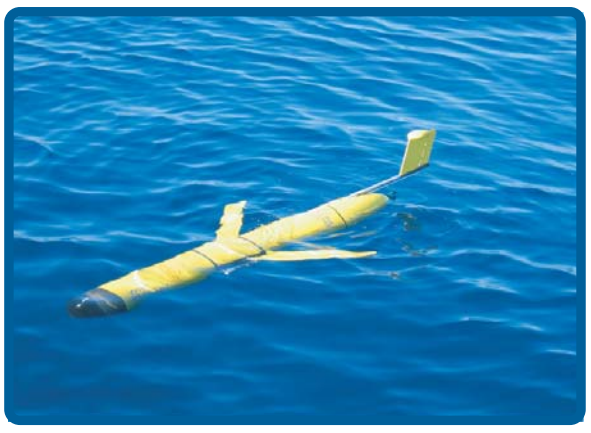

Fig. 1: Data aquisition: an autonomous instrument (glider) collecting marine chemical, physical and biological data over long time periods and distances (Source: Pierre Testor)

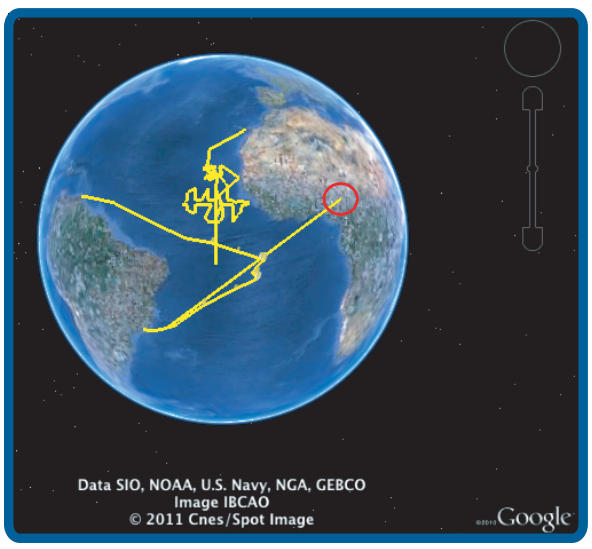

Fig. 2: An automatically generated KML file helps to identify errors within (meta)data, e.g. geographical position.

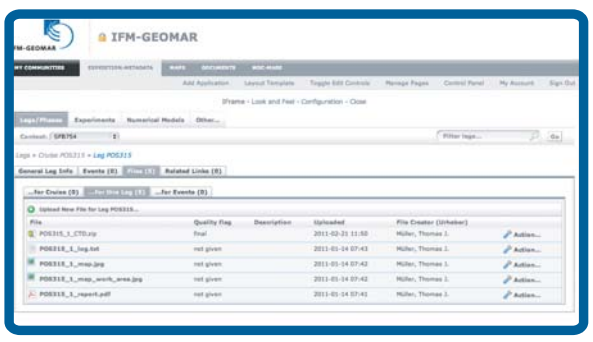

Fig. 3: List of files uploaded to a cruise with different access restrictions.

Today

Measurement

Scientific measurements in marine research involve instrumentations ranging from completely automated autonomous vehicles (Fig. 1) to manually controlled instruments on research vessels.
Assignment of standardized metadata to individual measurements requires common routines particularly when shipboard parties are used to assign their own identification labels to events, actions and samples aboard a research vessel. A major post cruise task for data managers is thus the correlation of official station books with the scientists' records which may even vary among different disciplines.

Quality control

Integrity of data and metadata is then maintained by individual scientists but not neccessarily homogenized between different groups thus requiring redundant work on the same issue. Poor book keeping and simplistic file sharing based data exchange during the cruises may even increase the amount of preliminary data and confusion about versioning. Visualization assists in error finding (Fig. 2).

Data exchange within a project

At present we offer a web based solution which allows scientists to upload bulk data files in the context of cruises, expeditions or models with allowance of any file format and structure of its content. Access control for a file is primarily based on the community context it was uploaded within but may be restricted by the file's owner (Fig. 3). There is currently no way to merge data and search for individual parameters or regions but at least the metadata of what, when, where and who are well documented for a file's content (i.e. data) and aid homogenization of the metadata.

Publication

In the process of publication data usually undergo a new quality control. New subsets are created and additional data are taken for comparison. The review process takes its time and shortly before final submission the idea of a supplementary data publication arises. At this stage it is difficult for a data curator to start collecting the necessary metadata such as stations, gears and parameter names which are mandatory for publication e.g. at WDC-MARE. That's why we go for an integrated data management approach!

Tomorrow

Workflow Editor

In advance of data collection the measurement procedure with all parameters has to be defined by the scientist (Fig. 4). This ensures the collection of all the necessary information during the data creation process.
Data input can be a file import or provided by hand into a web formular (Fig. 5) (replacing excel sheets).
The underlying datamodel is capable to store data from very different disciplines. New procedures can be included just by defining a corresponding workflow.

XML Validation

Given an XML workflow definintion the "filled" workflow can be validated via an XSD schema. This includes consistency and completeness of metadata, data types (geographical position, datetime, number and text) and data values with ranges. The database can serve as "single point of truth".

Data retrieval

Looking for chlorophyll data at the Capverde station? Use a map and find data from that region, certainly there are oxygen concentrations, as well. And what about the fish abundance, maybe compare those data? The database provides personalized or project based data retrieval by region or parameter in a common search interface. You consider a value anomalous high? Look at its provenance, check calibration and method for data quality.
All information at your fingertips (or mouse click) because someone defined the entire workflow and included all relevant worksteps.

Repositories

A repository for fulltext print publications has been set up at IFM-GEOMAR. The repository is connected to the datamanagement, authors and their data can thus be linked. The connection of data and print publication will be in both systems. You may find a publication in the repository and get a datalink, or you find a datacollection in the database, giving references to publications (Fig. 6). Especially projects are used to have a publication list as their final outcome, it is possible to have a linked data list, as well.

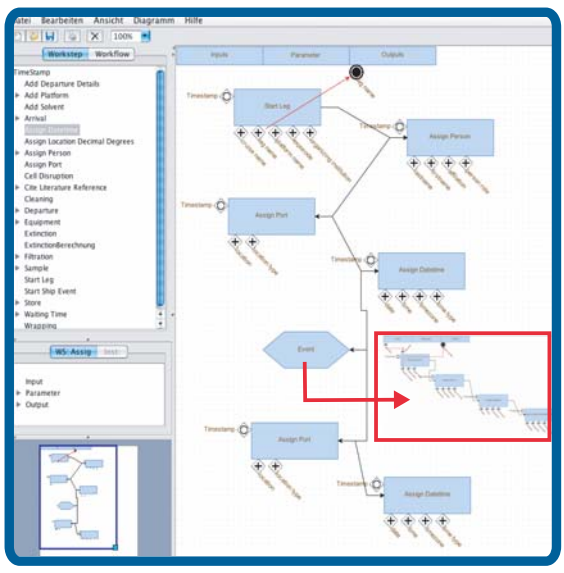

Fig. 4: Graphical work- and dataflow editor. This is an example of a ship cruise with repeatable sampling events.

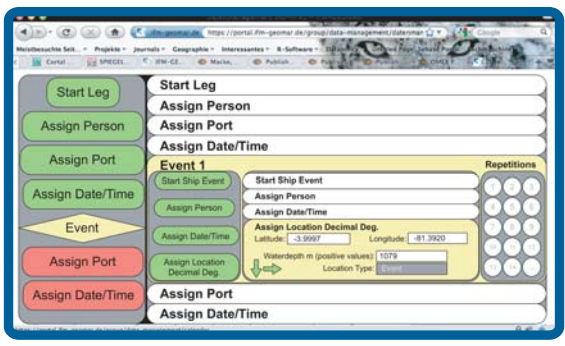

Fig. 5: A dynamically generated web formular based on the workflow definition allows to enter data into the generic database model.

Fig. 6: General metadata of georeferenced sampling locations with information on data files, literature and related links can be presented up to date in GoogleEarth via network links.

Publish Data in Data Centers

Contact: Data Management Team - Leibniz Institute of Marine Sciences (IFM-GEOMAR) - Wischhofstr. 1-3, 24148 Kiel