

A path to filled archives

Dirk Fleischer and Kai Jannaschk

Reluctance to deposit data is rife among researchers, despite broad agreement on the principle of data sharing. More and better information will reach hitherto empty archives, if professional support is given during data creation, not in a project's final phase.

Professionally managed, permanent data archives are essential to ensure the preservation and reusability of data. The idea of data deposition is supported by publishers and funding agencies around the world. Scientists, too, are generally in favour: they appreciate acknowledgement of (and future reference to) their hard-won data. Yet many repositories are almost empty¹. Clearly, there are discrepancies between scientists' attitudes to the principle of data sharing and their actions when it is time to deposit their data.

There are reasons for this gap¹. When the time of deposition comes — usually either at the end of a project or on publication of the results — the data are often scattered between various storage media, not uniformly formatted, and insufficiently tagged with metadata. As a result, deposition requires a substantial amount of effort, at a time when scientists really want to think about their next research question. Ongoing activities on data discovery and access that aim to innovate data reusability in the geosciences, such as the Data Observation Network for Earth project², do not appropriately address the issue of capturing data.

We argue that most scientists view data deposition in remote archives as a burden³, because it is too far removed from their daily routine. Scientists need and want professional and locally supported systems to store their data in a structured and reusable form. Support for scientists in this way, right at the beginning of the data life cycle, can avoid the discrepancy between the principles and actions of data sharing. If raw data and their derivatives are recorded in a professional manner during collection and analysis, the task of data deposition can be automated. It will then need only a mouse click by the scientist to initiate formal deposition, and not the laborious work of days. In such an environment, local data



managers become data navigators, rather than curators.

Scaling up

From the point of view of funding agencies and publishers — the main parties interested in data reusability and accessibility — data deposition at the end of the project or at the time of publication is sufficient. But data sharing is likely to evolve into a mandatory part of the research and publication process in the near future^{4,5}.

If so, data pathways must be organized in a way that can be scaled up without a vast drain on resources. In the present system, projects and their data managers are focused on one dedicated data repository. As a result, data managers provide individual support to scientists who wish to deposit their data, for example, by converting scientific data files into the format required by the chosen repository. This kind of curation is

inflexible and very time consuming⁶, and it requires personal communication between the scientist, data manager and repository staff for quality assurance of the metadata.

The human interaction in the data pathway creates unacceptable bottlenecks: only an automated process can turn around the full quantity of data that are generated and published. The curation system simply will be overwhelmed if all data are to be submitted.

The nagging problem

An analysis of data management requirements within the Cluster of Excellence: Future Ocean in Kiel⁷ revealed that researchers strongly desire reliable personal communication with local data curators³. They do not favour support by remote data managers: scientists like to be in charge of their data. (Think about it: would you give your children to someone you barely know?) Our survey, confirmed by an independent study³, included personal

interviews and revealed that the intensifying efforts on data management over the past ten years are viewed as a “decade of nagging and annoyance”, but have not provided a clear advance in data availability.

In our survey, scientists from different geoscience disciplines consistently agreed that there is a temporal gap between data creation and preservation efforts. Apart from making data deposition cumbersome, this time lag results in a loss of the accuracy of metadata⁶. Capturing data at the time of creation and the point of origin is therefore the most effective approach to data preservation^{6,8}.

The early part of the data life cycle is not currently managed by remote data archives and probably never can be. Instead, we propose that each scientific institution should support its scientists in the form of local data navigators, in combination with structured data storage. If the history and provenance of data are stored together in an institutional repository the data pathways can be automated. Long-term preservation or publication of the most valuable data in data archives can follow at a later stage, by a simple mouse click.


A new role for data managers

We argue that the role of data managers needs to change. Instead of gathering domain-specific knowledge from the scientists that is necessary for file conversion⁹, data managers should capture information in a structured database. Specifically, we propose that the task of capturing data should be focused around the human activities that transform raw measurements into a processed data set,

and it should record the algorithms of any transformations and error sources as well as data quality along the way. Each step in the data creation process by humans must be defined in close cooperation between the data navigator and the scientist before the measurement campaign or experiment starts, and stored in a local workflow repository. Here, some duplication of effort can be avoided by recording data not by project, but by scientist: the specific expertise of scientists and their methods often do not change much between projects, whereas the context (such as the geographical region) can be very different.

Once the workflow is defined, a software system along those lines generates the necessary input forms that capture the data for this process. As a result, scientists can measure and store their data much more easily. And a systematic collection of data according to pre-agreed workflows from each scientist in a collaborative project provides a functional working platform for a community of researchers.

If these basic principles are adhered to in the early stages of the data life cycle, data-sharing requirements of publishers and funding organizations can be met very easily: the publication of an article or the end of the funding period could be used as an automated trigger to submit the relevant portion of the data into public long-term data archives. Because the scientists helped organize the available information at the stage of workflow definition — before measurements even started — such an automated data submission should not require much input from the scientist.

As an added benefit, data submission to a long-term archive through an automated submission process can link the publically available data sets with the locally stored, working data sets to keep track of ongoing data cleaning or quality improvements. In Kiel, we are working on a data management infrastructure along these recommendations that will include an automated data publication system. The system is not yet fully functional, but it will eventually provide a test case for the system for archiving data provenance as described here. 

*Dirk Fleischer is a member of the Kiel data management infrastructure at the Leibniz Institute of Marine Sciences IFM-GEOMAR, Wischhofstrasse 1-3, 24148 Kiel, Germany; Kai Jannaschk is at Information Systems Engineering, Christian-Albrechts-University at Kiel, Christian-Albrechts-Platz 4, 24118 Kiel, Germany.
e-mail: dfleischer@ifm-geomar.de;
jannaschk@is.informatik.uni-kiel.de*

References

1. Nelson, B. *Nature* **461**, 160–163 (2009).
2. Michener, W., Vision, T. & Cruse, P. *D-Lib Mag.* <http://dx.doi.org/10.1045/january2011-michener> (January/February 2011).
3. Tenopir, C. *et al.* *PLoS ONE* **6**, e21101 (2011).
4. Hey, T., Tansley, S. & Tolle, K. *The Fourth Paradigm: Data intensive Scientific Discovery* (Microsoft Research, 2009).
5. Nielsen, M. *Phys. World* **22**, 30–35 (May 2009).
6. Treloar, A. & Wilkinson, R. in *Proc. 4th IEEE Int. Conf. eScience* <http://dx.doi.org/10.1109/escience.2008.41> (2008).
7. <http://www.ozean-der-zukunft.de/english/>
8. Beagrie, N., Chruszcz, J. & Lavoie, B. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities* (HEFCE, 2008). Available via <http://go.nature.com/4pERCH>
9. Treloar, A., Groenewegen, D. & Harboe-Ree, C. *D-Lib Mag.* <http://dx.doi.org/10.1045/september2007-treloar> (September/October 2007).

Published online: 21 August 2011