IN53B
1632

Where are the gaps in the data lifecycle? Developing policies and workflow tools for digital preservation of research data in the Geosciences

Wolfgang Peters-Kottig[1], Jens Klump[2], Ingo Kirchner[3], Roland Bertelmann[2], Beate Rusch[1], Martin Wattenbach[2], Damian Ulbricht[2], Petra Gebauer[3]

[1] Zuse-Institut Berlin / KOBV, Germany [2]Deutsches GeoForschungsZentrum Potsdam, Germany
[3]Institut für Meteorologie, Freie Universität Berlin, Germany

BACKGROUND

Although the geosciences compete in the upper ranks of scientific disciplines when it comes to curation and long-term digital preservation of research data, there are gaps in the workflows throughout the data lifecycle, particulary at the curation boundaries (see Fig. 1).

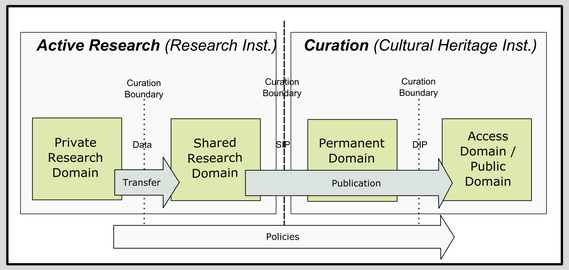

Figure 1. *Data Flow between Domains in the Curation Continuum (modified after Treloar & Harboe-Ree 2008, Klump 2011)*

At present, it is still too complicated to upload data into a repository. The workflow tools that allow for identification and validation of file format information and extraction of metadata are incomplete and should be easier to use. Tools should make use of contextual information.
Another issue that needs to be addressed: Training in research data management skills as (future) key competence for students in a higher education setting.
In order to address these problems, project EWIG† was initiated as a cooperation between an infrastructure facility (Zuse Institute) and two different data producers from the field of geosciences (GFZ and FU Berlin Institute for Meteorology).

†*EWIG is an acronym (in German) for »development of workflow components for long-term preservation of research data in the geosciences«. The project is funded by the German Science Foundation DFG and involves a team of currently eight researchers from the geosciences and library and information sciences.*

AIMS & METHODS

- **Identify remaining gaps and develop currently missing workflow components in the research data lifecycle** (e.g., mapping JHOVE2 onto FITS output; evaluating certification protocols of »certified repositories«; …and there are as yet no validation tools for movie files)
- **Develop digital preservation policies and a blueprint for dLTP policies in the geosciences**
- **Perform tests on the re-useability of produced research data packages (DIP, SIP)**. Scientists, graduates and students will check the semantic context of information packages. Experiences from the test procedures will be used to enhance the quality of metadata creation procedures in an iterative process
- **Design a university lecture/seminar series** based on the test procedures, which will contribute to raise awareness for data curation issues among students and graduates in the Geosciences

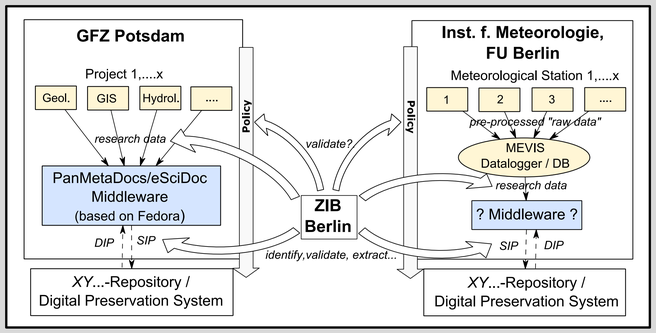

Figure 3. *Data infrastructure architecture of GFZ and Inst. f. Meteorologie (for information on PanMetaDocs see abstract No. IN23C-1461)*

GFZ Potsdam and FU Berlin Institute for Meteorology produce and deliver test data from finished an ongoing projects that will be packed into »repository ready« information packages as source material for validation and re-usability testing (Test data: weather data from meteorological stations in Berlin [Figure 2] and various research data from a lake drilling project in Siberia).
Testing procedures depend on the data infrastructure architectures that differ significantly between GFZ and FU Berlin Institute for Meteorology (Figure 3).

Figure 2. *Meteorological Station at Botanical Garden Berlin.*

The meteorological network in the city of Berlin includes several stations, measuring air temperature, humidity, precipitation, soil surface and ground temperature. At some stations additional wind direction and wind speed, sunshine duration, precipitation, barometric pressure and radiation quantities are recorded. The measurement interval is one minute, resulting in more than 1,000,000 values reported per day.

TAKE HOME MESSAGE

- Do you know of any specific gaps in the geoscience data life cycle that are worth to be filled? Tell us – we will try and do our best to fill them!
- Do you have an institutional policy for data management that also includes long-term preservation? Maybe we could learn from you!
- Do you teach a university lecture on long-term preservation and/or research data management in your earth science department? Lets exchange knowledge and experience!

REFERENCES

Treloar, A., & Harboe-Ree, C. (2008): Data management and the curation continuum: how the Monash experience is informing repository relationships. VALA2008, Melbourne, Australia. *http://www.laconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf*

Klump, J. (2011): Langzeiterhaltung digitaler Forschungsdaten. In: Büttner, S., Hobohm, H.-C., & Müller, L. (ed.). Handbuch Forschungsdatenmanagement, pp. 115-122. Bock + Herchen, Bad Honnef. *urn:nbn:de:kobv:525-opus-2412*

Presented at the AGU Fall Meeting, San Francisco, December 2011

Funded by DFG