Meeting minutes

Meeting of the database managers of DFG research programs in biodiversity and ecological research

March 1st 2011, Rauischholzhausen/Marburg

Attendants	Project	eMail
Jie Zhang	FOR 1264 - Kilimanjaro	jie.zhang@uni-wuerzburg.de
Hela Mehrtens	SFB 574 & SFB 754; IFM-Geomar	hmehrtens@ifm-geomar.de
Pina Springer	SFB 574 & SFB 754; IFM-Geomar	pspringer@ifm-geomar.de
Dirk Fleischner	SFB 574 & SFB 754; IFM-Geomar	dfleischer@ifm-geomar.de
Thomas Lotz	FOR 816 - TMF Ecuador	thomas.lotz@uni-marburg.de
Maik Dobbermann	FOR 816 - TMF Ecuador	maik.dobbermann@uni-marburg.de
Karin Nadrowski	FOR 891 - BEF China	nadrowski@uni-leipzig.de
Constanze Curdt	SFB/TR32	c.curdt@uni-koeln.de
Dirk Hoffmeister	SFB/TR32	dirk.hoffmeister@uni-koeln.de
Sven Pompe	Jena Experiment	sven.pompe@uni-jena.de
Eleonora Petzold	Biodiv. Exploratories; BExIS	epetzold@bgc-jena.mpg.de
Dennis Heimann	Biodiv. Exploratories; BExIS	dheimann@bgc-jena.mpg.de
Birgitta König-Ries	Biodiv. Exploratories; BExIS	birgitta.koenig-ries@uni-jena.de
Andreas Ostrowski	Biodiv. Exploratories; BExIS	aostrow@bgc-jena.mpg.de
Christian Willmes	SFB 806 - Our way to Europe	cwillmes@uni-koeln.de
Thomas Nauß	FOR 816 - TMF Ecuador	thomas.nauss@staff.uni-marburg.de

Short introduction

Introduction of attendants, their individual projects and the data management systems used within.

Approval of agenda

With respect to the topics already surfaced during the introduction, the distributed agenda has been changed. The new agenda encompasses the following topics:

- Data licenses
- Data publication
- Long-term data storage
- Central infrastructure
- Regular data manager meetings

Data licenses

Most data managers are confronted with acceptance problems related to the individual and project based data repositories. It seems that quite a number of users still feel unpleasant when they have to supply their datasets to a centralized server. There are also concerns about the appropriate referencing of the own datasets if they are used/analyzed/published by the project partners.

The data managers suggest that for future projects a default data usage and publication policy agreement should be mandatory and the acceptance of the agreement should be the pre-requisite for the funding of the individual subprojects. To account for individual circumstances, the steering committees of the research units should have the opportunity to select one out of three default agreements. These agreements should also cover the license terms which are relevant for data utilization after the end of the research project (e. g. Creative Common License).

Data publication

In general, data publications should be equally regarded as scientific publications. The utilization of DOIs to reference certain datasets and versions is not used within many of the individual research projects.

The data managers suggest that DOI or alternative options for the publication of raw and derived datasets should be promoted within the database systems. For derived datasets, information on data provenance must be considered. This does not only include the implementation of workflows to transfer and publish the datasets within given archives (PANGAEA etc.) but also the possibility to publish raw datasets (the latter e. g. within central DFG data infrastructures, see below).

There should be a mandatory requirement for all projects to publish their datasets in the above ways not later than 8 to 12 month after the end of the project funding. Therefore, central repository infrastructures have to be installed at least for those datasets which cannot be published within existing third-party repositories.

Long term data storage

Long term data storage is a crucial task for the scientific community and at the same time almost impossible to guarantee by individual projects with a limited time of financial funding. Even if the datasets are still physically stored at one location, web-based query routines for the datasets etc. might be offline or no longer supported (server updates, migrations, etc.).

The data managers suggest the establishment of long-term metadata and data repositories. The metadata repositories can be centralized, distributed or mirrored; the data repositories can be centralized or distributed. The metadata repositories should provide interfaces and be connected to the internationally coordinated repository networks to allow the query of the metadata through these interfaces (e. g. dataONE, GenBank). The metadata should be linked to the central/distributed data repository to allow access to the individual datasets.

Central infrastructure

Most of the data management projects represented in the meeting have implemented their system from scratch or from a more general architectural framework (e. g. Apache Struts). So far only two projects are in close contact or have built upon database systems that have been developed within other research units. No common metadata standards etc. are used within the individual systems.

In this context, several ideas have been discussed. They are not related to the establishment of a central metadata and data repository for the long-term data storage which is regarded as necessary (see topic on long term data storage) but for the data management during the funding phase of the research units. The ideas are:

1. To allow a greater flexibility for the individual projects, local infrastructures should be used during the project phase and meta- as well as raw/derived datasets should be published/transferred to central repositories after the end of the project funding.
2. Established database systems developed within an individual research project should be transferred to some kind of template repository to ensure that future projects could still utilize these system architectures as their foundation.
3. Central data repository infrastructures should be installed which provide basic data hosting services including data ingestion, query, retrieval but at the same time allow for individual extensions or add-ons. A central help-desk which also acts as an adviser during the project setup should also be related to the central infrastructure.

There is no final suggestion on which way is preferred over another yet. In general, the utilization of centralized systems also during the actual project lifetime are not rejected but many issues regarding e. g. flexibility, options for add-ons, user acceptance etc. have to be discussed in more detail.

Further meetings

The data managers suggest further meetings on a regular, semi-annual basis (timeframe: 1 day). The meetings should be organized by the individual data management projects. The next meeting will be kindly organized by the BExIS team in Jena.

Minutes submitted by Thomas Nauß.