SPATIAL DATA INFRASTRUCTURE FOR SOIL-VEGETATION-ATMOSPHERE MODELLING: SET-UP OF A SPATIAL DATABASE FOR A RESEARCH PROJECT (SFB/TR32)

C. Curdt [a,] *, D. Hoffmeister [a], G. Waldhoff [a], G. Bareth [a]

[a]Dept. of Geography (GIS & RS), University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany – (c.curdt, dirk.hoffmeister, guido.waldhoff, g.bareth)@uni-koeln.de

Working Group IV/1

KEY WORDS: Spatial Data Infrastructure, Data Management, Computer Vision, Interoperability, Environmental Monitoring, Spatial Database, Internet GIS, Metadata

ABSTRACT:

Data storage and data exchange is a key issue of interdisciplinary research projects which focus on environmental field studies and regional modelling. The overall success of such projects depends on the well organized data management and data exchange between all involved sub-projects. This includes the organization of data, the implementation of a database for and the maintenance of such a system for intensive data exchange between the project sections. The project database ensures the sustainable use of collected measurement data and research results of long-term research projects. Especially, projects which focus on spatial modelling of soil, vegetation and atmosphere interactions rely on data exchange of geo and attribute data. In this contribution, the design and set up of a spatial data infrastructure for a research project (TR32) that focuses on soil-vegetation-atmosphere modelling is presented. The introduced data management design enables web-based (i) up- and (ii) download of data, implementation of different (iii) user views, interactive input of (iv) metadata and the integration of a (v) WebGIS.

1. INTRODUCTION

Spatial data management including data storage and data exchange (between several project sections) is particular important for interdisciplinary research projects which focus on environmental field studies and regional modelling (Mückschel and Nieschulze 2004). Especially Transregional Collaborative Research Centers (TR) which focus on spatial data modelling need a well organized data management. They are characterized as research projects that are based at separate locations, operate for up to 12 years and combine cross-disciplinary research interests and material resources. Therefore, it is essential to store and backup the multiplicity of different interdisciplinary project data and the huge amount of data gathered during the project phases in a well organized structure (Mückschel et al. 2007). These research projects are funded by the German Research Foundation (DFG) and have the requirement to contain a project section (SP) that is responsible for data management. The TR has the duties and responsibilities in terms of 'Good Scientific Practice' to store, manage, maintain and backup the whole research data in a permanent, sustainable and stable system in cooperation with the local computing center. Project data has to be stored during the project activities and up to 10 years after the project is finished (DFG 1998).

In this context, we introduce the data management approach of the inter- and multidisciplinary research project "Transregional Collaborative Research Centre 32: Pattern in Soil-Vegetation-Atmosphere Systems: Monitoring, Modelling, and Data Assimilation" (TR32) funded by the DFG (http://www.dfg.de/en/). The TR32 is a joint project between the Universities of Aachen, Bonn, Cologne, and the Research Centre Jülich. Now the TR32 is situated in the second year of the first of the three phases, each running for four years. The research area of the TR32 is the watershed of the river Rur which is situated in Western Germany and partly in Belgium and the Netherlands. In the first phase, the field research is focused on three sub watersheds that represent three typical land use forms (forest, arable, and grass land).

The TR32 works on exchange processes between the soil, vegetation, and the adjacent atmospheric boundary layer (SVA). The overall project is subdivided into four project areas (clusters). The clusters (A, B, C, and D) differ by the subsystem (SVA) on which they concentrate and also by the spatial scale range, they deal with (laboratory - region). The clusters are split up into project sections. Within the whole TR32, 13 SPs work on research. Furthermore, cluster overlapping cross-cutting-groups were set up to arrange the exchange of information between the clusters.

The overall research goal is to yield improved numerical SVA-models for the prediction of water-, CO_2- and energy-transfer by accounting for the patterns occurring at various scales. The hypothesis of the TR32 covers the explicit consideration of patterns and structures which lead to a common methodological framework. This will increase our understanding and our capability of describing and predicting the SVA system in a comprehensive manner. The research partners of the various participating affiliations are from the fields of soil and plant science, remote sensing, hydrology, meteorology, and mathematics. They will approach the SVA continuum under this new paradigm (TR32-Wiki 2007).

* Corresponding author.

2. METHODS

2.1 Data storage (at the ZAIK/RRZK)

A persistent, sustainable, secure storage of data is very important for all kinds of projects that handle data. This has to be organized in a clearly arranged and ordered structure. Moreover only authorized users should be permitted to access the data storage. Furthermore, it is essential to backup or archive data to ensure the recovery of destroyed data. Hence e.g. the German Federal Office for Information Security (BSI) follows the just given principles for the storage and management of data (BSI 2008). Considering the mentioned facts and guidelines of the DFG, the cooperation with a computing center is indispensable for an interdisciplinary project.

The storage of data at the Regional Computing Center (RRZK)/the Center for Applied Computer Sciences (ZAIK) of the University of Cologne is organized in different stages. These are the Storage-Area-Network (SAN) based Disk Storage, the Andrew File System (AFS) and the High Performance Computing (HPC) - File System (see Figure 1).

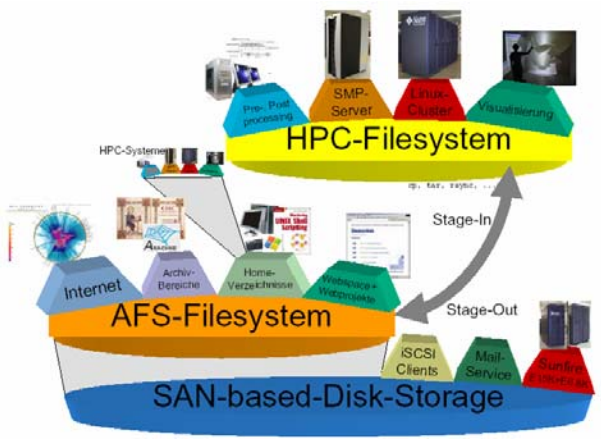

Figure 1. Data storage at the ZAIK/RRZK (Kalle 2005)

In this contribution, the focus will be on AFS, a distributed networked file management system (file database). Besides sharing and synchronizing federated file storage, main advantages of the AFS are aspects like security, scalability, and user administration by an Access Control List (ACL). Furthermore, AFS supports location inference, cross platform access (e.g. UNIX, Linux and Microsoft Windows) and simple archive and data backup (AFS 2008).

2.2 Databases

The persistent, consistent and efficient management and storage of a huge amount of data sets is the central duty of a Database Systems (DBS). A DBS is a combination of a database (DB) and a Database Management System (DBMS) (Elmasri and Navathe 2005). The real storage of data operates at the DB. The DBMS is the software that enables the management of data in a specified data model. The characteristic properties or rather requirements of a DBS are data independence, central data management, multiuser access, data security, data integrity (redundance-free data storage), access control, processing of requests, and operating system independence (Brinkhoff 2008).

Data models are described in several structures. In this context relational and object-relational models and databases are important. Relational database models are identified with connected database tables that store the data like IBM DB2, Microsoft Access or MySQL. Object-relational database models manage objects that are stored in linked relational data tables like Oracle Spatial or PostgreSQL (Türker and Saake 2005). They are qualified to be geodatabase systems on condition that they have a corresponding extension (Brinkhoff 2008).

2.3 Metadata

The term metadata describes data that contain information about data e.g. quality, author, location, and year of publication for a book. The common use of metadata is to describe data and enable a better search. For more explanations on metadata and the importance of see Yeung (2007).

The metadata structure should follow recent standards and principles. These are basically spatial data standards from the Open Geospatial Consortium (OGC), the International Organization for Standardization (ISO), the Dublin Core Metadata Initiative (DCMI), and the World Wide Web Consortium (W3C) (Bartelme 2005 and Nogueras-Iso et al. 2005). For example the geo-metadata follows the guidelines of ISO Norm 19115 (ISO19115 2003).

2.4 Web-Interface and WebGIS

The main duty of a web-interface is the representation of information at the World Wide Web. The implementation of a web-interface is predominantly carried out with the platform independent markup language HTML/XHTML ((Extensible) Hyper Text Markup Language). In addition the client-side scripting language PHP (Hypertext Preprocessor) is used for dynamic web pages as well as CSS (Cascading Style Sheets) to describe the style of the web-interface. A combination of PHP and SQL (Structured Query Language) is used to customize the interaction between the database and the web-interface. The encrypted SSL (Secure Sockets Layer) technology including a digital certificate ensures a secure communication via internet e.g. for web browsing or data exchange (Open SSL 2008).

In recent years the presentation of GIS and cartographic functionalities (web mapping) in the internet became more and more important (Asche and Herrmann 2003). Hence a WebGIS is absolutely essential to visualize, manage, and analyze spatial geodata within a web-interface. ESRI's ArcGIS Server connected with ArcSDE (Spatial Database Engine) provides an opportunity to realize a WebGIS (ESRI 2008).

3. TR32 DATABASE DESIGN

The essential duty of the TR32 data management SP is to enable data storage and exchange for different SPs. In addition backup functionalities have to be implemented as well as corresponding metadata for the project data. With regard to the project duration, it is important to design a stable and sustainable system, called the 'TR32 Database' (TR32DB).

As a result of the interdisciplinary background of the TR32, the TR32DB stores a multiplicity of different data: measured and purchased project (geo-) data. The measured project data comprise data that are collected of the various TR32-SPs. The

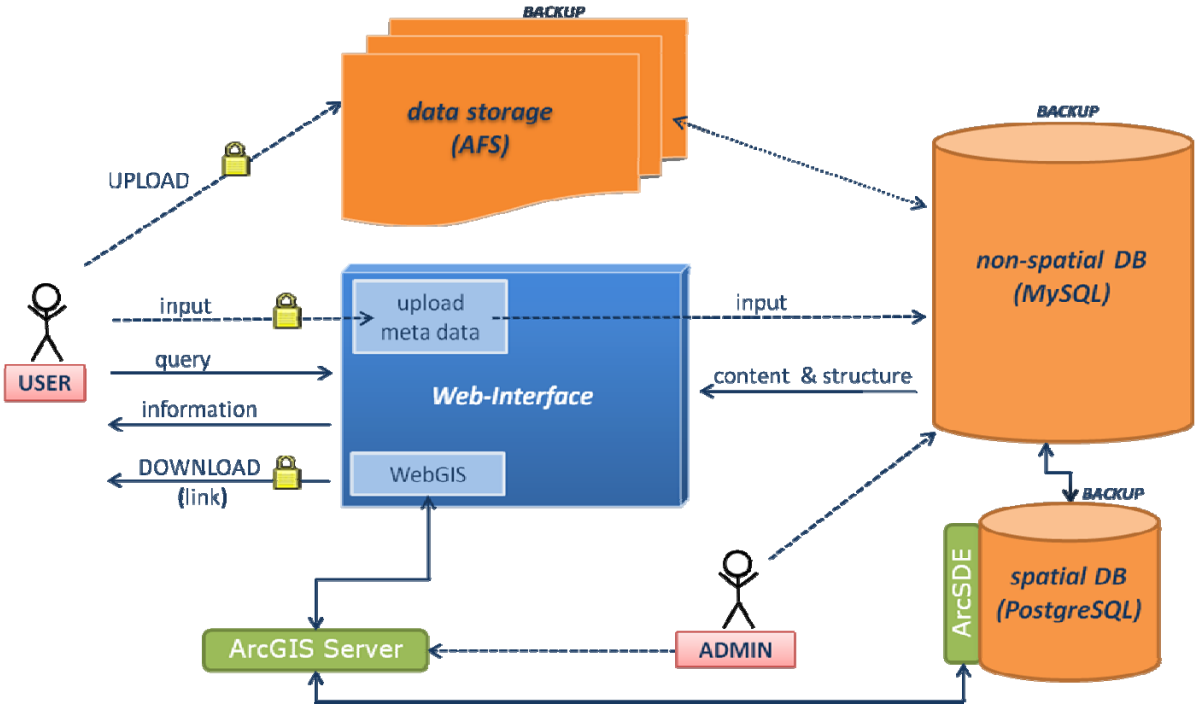

Figure 2. TR32DB Design

purchased data include geodata like topographic, elevation, remote sensing, soil, land use or weather data in different scales. Besides collected measured data and purchased data, further data e.g. publications, reports, videos, and pictures of the TR32 participants need to be managed with their corresponding metadata.

The TR32DB-Structure (see Figure 2) is basically subdivided in three components: (i) data storage, (ii) databases, and (iii) web-interface. The data storage is the system component where the project data is physically stored (upload of data). The database stores associated metadata and administrative data. Finally the web-interface provides uploaded project data of (download of data) and information about the TR32. A backup of the whole system is continuously done. In the following paragraphs, each core component of the TR32DB system is described in detail.

3.1 TR32DB Data Storage (upload of data)

The essential data storage of the TR32DB (upload of data) is carried out in the already established AFS. The latter was chosen in cooperation with the ZAIK/RRZK due to the mentioned reasons under 2.1 and additionally due to the support opportunities provided by the ZAIK/RRZK. Furthermore, the cooperation with a local computing center is requested by the DFG.

In detail, the TR32DB file database is organized in a folder system. Due to the fact that the TR32DB contains a multiplicity of several project data, the data is arranged in different folders. This is designed in a specific hierarchy according to the mentioned structure of the TR32. The folder structure is therefore ordered by: (i) cluster, (ii) project sections and (iii) data type. The data type is divided into: data, publications, pictures, presentations and reports. Only users who are participants of the TR32 are authorized to store their data in the TR32DB.

The upload of data within the AFS operates as follows. Users who are authorized to the AFS system own a specific view in their home directory. They have the opportunity to store their data directly via copy and paste in their particular project section folder sorted by the five data types. The actual project section folder is joined by a symbolic data linkage with a directory in the TR32DB. Consequently, uploaded user data is immediately located in the TR32DB system. The data is moved into the final folder directory by a script that is developed in cooperation with the ZAIK/RRZK. The developed solution has several advantages. These are basically the immutability and documentation of the moved data files. A database entry is automatically generated by the script for each moved file that contains e.g. the file name, the file extension, and the final file storage directory. Only the administrators are permitted to edit or delete documented information in the database as well as the data file directory in the AFS.

The announced system has some restrictions. Some are set by the actual AFS version like a single file size limit of 2GB (which will be increased in near future) and others are established by the administrators for safety aspects. These are limitations like access restriction for other project section folders due to the user AFS-account. This has the advantage that the administrators are able to define the exact folders where users are permitted to store their data. Users are required to label their data files in a well-defined manner and compress them if possible. Furthermore, the current folder size of the data types is limited to 8GB (which will be extended in near future, too). Therefore, each SP is only able to upload 8GB per data type in their specific folders per day since the script operates over night. In exceptional cases a movement of data managed by the administrators is possible.

3.2 TR32DB Database

The current TR32DB database is subdivided in two parts (see Figure 2). The first component manages non-spatial data of the TR32 including project data, metadata, and administration data. The other component manages the spatial related data of the TR32.

Purchased geodata and future spatial-related research results of the TR32 will be stored in the open-source object-relational database system PostgreSQL which is extended by PostGIS (see Figure 2). Main reasons for this database system are the possibility to manage geodata and the support of ESRI's ArcGIS Server with the ArcSDE technology to present those data in a WebGIS.

The non-spatial TR32 data is managed by a multithreaded MySQL database that is implemented in cooperation with the ZAIK/RRZK. The multiuser open-source-system is popular for web applications in combination with PHP (MySQL 2008). In fact the TR32DB database structure (simplified in Figure 3) only contains references to the uploaded project data (by the file directory) that is stored in the AFS. In addition, the database includes the corresponding metadata of these project data. The quantity of metadata attributes depends on the type of data.

Different data types require attributes that are not important for other data types e.g. details about a measuring instrument which is important for measured data, but not for publications. Final metadata attributes for the different data types still have to be discussed with the individual SP staff in order to customize the needs and requirements of the project participants. The input of metadata into the database operates by the TR32DB-web-interface (see Figure 4).

Furthermore, administration data is saved in the TR32DB structure that is implemented by the MySQL database, e.g. the authorization of TR32DB-users and related user views. Finally, the structure and content of the TR32-web-interface is arranged by the database. This contains information about e.g. users, project sections or cross cutting groups.

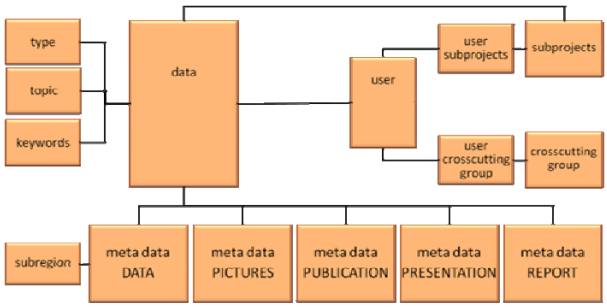

Figure 3. TR32DB Database structure

3.3 TR32DB Web-Interface with integrated Web-GIS (download of data)

The current TR32DB web-interface (see Figure 4) is located at http://www.tr32db.uni-koeln.de. Basic functions of the web-interface are representation, search and download (only TR32 participants) of non-spatial project data and temporary download of purchased geodata. The input of corresponding metadata is available, and spatial data is accessible via the integrated Web-GIS. Non-spatial project data of the TR32 as well as content and structure of the TR32DB web-interface are in most instances managed and filled due to access to data from the TR32DB-file-system or the MySQL-database. Furthermore, the web-interface offers information about the TR32.

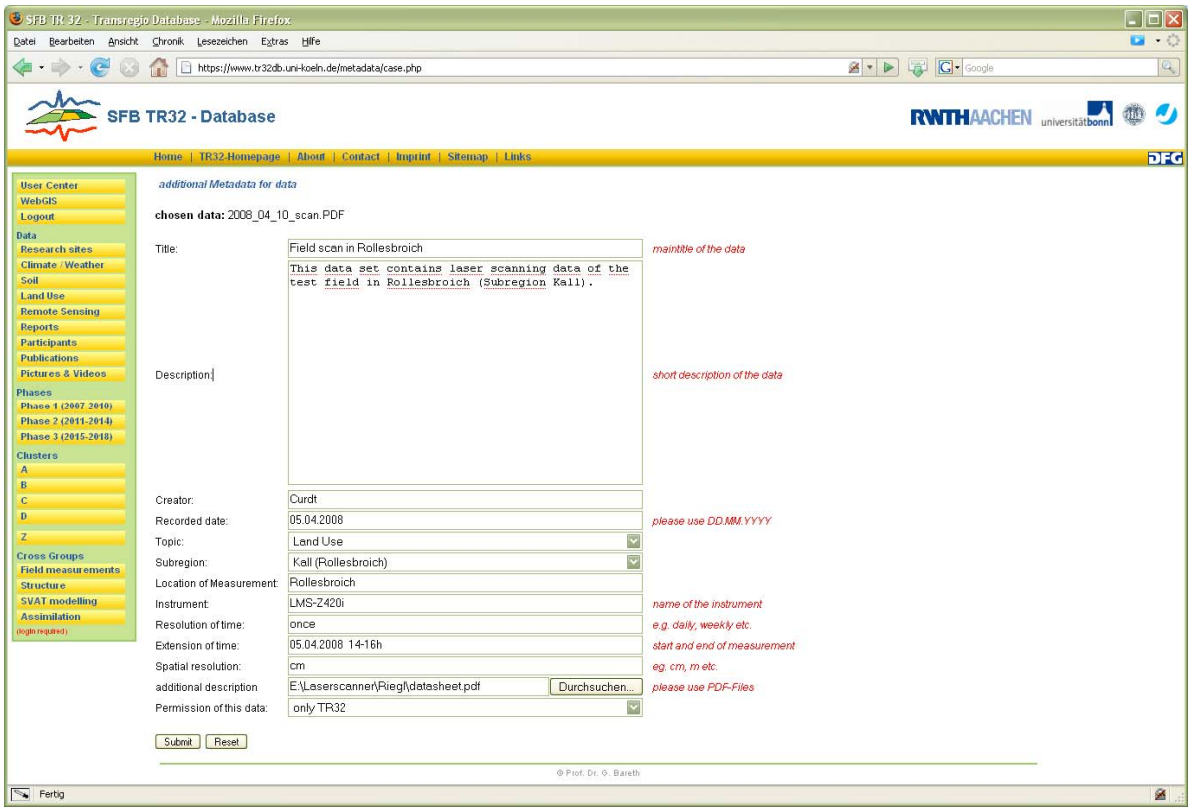

Figure 4. TR32DB Web-Interface (for metadata input)

With regard to licensed data and the possibility to download large data sets, it is essential to develop a safe and stable web-interface (see 2.4). Therefore the web-interface provides several features depending on the authorization of the TR32DB visitors. Guests (not authorized TR32DB users) of the TR32DB web-interface are only able to view general project information that contain the cross cutting groups and clusters including project sections, their duties and participants. They are able to view and search some project data that is disposable for every web-interface visitor.

Authorized users (only project participants) of the TR32DB web-interface have further opportunities than guests have. They are permitted to search for explicit project data and their metadata, the project phases, or topics (climate, soil, land use or publications). They are allowed to obtain more detailed, extended, internal project data and research results. Furthermore, they are able to download these project data. The web-interface also provides temporary download of purchased geodata (limited by license agreements) with password protection. A principal duty of the web-interface is to enable the input of corresponding metadata to project data for authorized TR32DB users. Each authorized user is able to view all data of his project section that is stored in the AFS and has no metadata yet. By selection of a data set the user is able to feed metadata into the database. Depending on the type of data there will be a particular metadata input wizard. Users are able to enter metadata like title, description, instrument or creator as well as the access restriction of the data set. Finally, there is an opportunity to add an explanation document.

The integrated WebGIS of the web-interface is developed to access the spatial geodatabase with the purpose to illustrate the purchased geodata and the prospective spatial-related research results. Another functionality of the WebGIS is to generate uniform map layouts for all TR32 participants that may be used for publications. Further GIS-analysis-functions e.g. overlay, statistics will be included in the WebGIS.

4. DISCUSSION AND CONCLUSION

Spatial data infrastructure or rather data management in huge interdisciplinary research projects are becoming more and more important in context of the current discussion about the Infrastructure for Spatial Information in Europe (INSPIRE) and consequential for the national geodata infrastructure (GDI) of the European Union member states including metadata management (Bernard et al. 2005 and Yeung et al. 2007).
The central data infrastructure and management within the scope of a huge interdisciplinary research project involves different problems that contain or rather have the requirement to store heterogeneous data sets in a explicit and ordered structure including their metadata (Mückschel et al. 2007). In the context of the computer-based developing and enhancements of e.g. data standards there are different process approaches. Data management systems are also advanced due to the exchange of administration experiences.

Examples of research projects with a central project data management in Germany are: the Collaborative Research Centre (SFB) 552 (http://www.storma.de) and SFB 299 (http://www.sfb299.de/) as well as the Research Union (RU) 402 (http://www.bergregenwald.de/) and RU 816 (http://www.tropicalmountainforest.org/).

The project data management structure of the SFB 552 and the SFB 299 offer a similar structure. The SFB 552 benefits from the gathered experiences of the SFB 299. Both use open-source-software that includes the LAMP components (Linux, Apache, MySQL and PHP/Perl) and the open-source Content-Management-System TYPO3 as well as the UMN Mapserver and PostgreSQL with PostGIS to visualize and store geodata (Mückschel et al. 2007).

By contrast the developments of the RU 816 just benefit in some basic parts from the RU 402; the RU 816 is rather a further technical development of the RU 402. The RU 402 is a combination of file system with a corresponding relational metadatabase. The RU 816 is more a complex ‘data warehouse system’ approach where all project data including actual project data with corresponding metadata is stored directly in a database. The RU 816 is implemented by a MySQL database and a realization with java server pages. The difference between theses systems is the opportunity to store, search and download single data files. The RU 402 approach is more appropriate to interdisciplinary projects in contrast to the RU 816 approach that is more qualified for interdisciplinary to integrative orientated research projects (Nauss et al. 2007).

There are different projects overseas which realize similar data management approaches. One example is the Canadian Carbon Program (CCP) which is represented by the internet-based Data Information System (DIS) (http://fluxnet.ccrp.ec.gc.ca/). The project data storage is realized in a clearly structured folder system. It is just permitted to save project data files in ASCII format. The upload and submission of project data and corresponding metadata is implemented with FTP. The data access operates with FTP or seamless links from the DIS (Fluxnet Canada 2004). Some developments of the CCP are not able to enhance within the TR32. The application of FTP is for example not permitted within the TR32 due to security reasons at the ZAIK/RRZK. Furthermore, the use of ASCII is not realizable. Saving all data in this common file format has the advantage of being widely accessible and useable. The disadvantage is that not all data can be formatted in such a file format and that those data files tend to be very large. Without the usage of a database it is more difficult to administrate and enquired the huge amount of project data.

In contrast to the mentioned examples above, the TR32 follows other interdisciplinary research goals and tries to combine the advantages of the described systems. With regard to different scales of data (SVA) and a prognosticated huge amount of big data files the TR32DB is developed in a secure, stable, sorted, and well organized structure within the environment of an institution (ZAIK/RRZK). This will ensure the availability of the database after the end of the project. The structure enables upload, download, and high sophisticated backup of project data as well as metadata. The whole system is developed in cooperation and with the agreement of the local computing centre.

REFERENCES

AFS, 2008, http://www.openafs.org (accessed 14. April 2008).

Asche, H. and Herrmann, C., 2003, *Web.Mapping 2, Telekartographie, Geovisualisierung und mobile Geodienste*, Wichmann, Heidelberg, Germany.

Bartelme, N., 2005, *Geoinformatik, Modelle Strukturen Funktionen*, Springer, Berlin Heidelberg, Germany.

Bernhard, L., Fitzke, J. and Wagner, R. M., 2005, *Geodateninfrastruktur, Grundlagen und Anwendungen*, Wichmann, Heidelberg, Germany.

Brinkhoff, T., 2008, *Geodatenbanksysteme in Theorie und Praxis, Einführung in objektrelationale Geodatenbanken unter besonderer Berücksichtigung von Oracle Spatial*, Wichmann,

BSI, 2008, IT-Grundschutz-Kataloge: M 1 Maßnahmenkatalog Infrastruktur Strukturierte Datenhaltung, http://www.bsi.bund.de/gshb/deutsch/m/m02138.htm (accessed 14. April 2008).

DFG, 1998, Proposals for Safeguarding Good Scientific Practice - Recommendations of the Commission on Professional Self Regulation in Science, Weinheim, Germany. http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf (accessed 14. April 2008).

Elmasri, R. and Navathe, S. B., 2005, *Grundlagen von Datenbanksystemen*, Addison-Wesley, München, Germany.

ESRI, 2008, ArcGIS Server, Comprehensive Server-based GIS, http://www.esri.com/software/arcgis/arcgisserver/index.html (accessed 14. April 2008).

Fluxnet Canada, 2004, The Fluxnet-Canada Data Management Plan, http://www.fluxnet-canada.ca/pages/protocols_en/Data_Management_Plan_23_jan_2004.pdf (accessed 14. April 2008).

Heidelberg, Germany.

ISO 19115, 2003, Geographic Information – Metadata, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020 (accessed 14. April 2008).

Kalle, C., 2005, Kolloquium Ausgewählte Themen der Datenverarbeitung 14.12.2005 „Datenhaltung am ZAIK/RRZK“, Cologne, Germany. http://www.uni-koeln.de/rrzk/multimedia/kolloquium/ws0506/_pdf/Datenhaltung-am-ZAIK.pdf (accessed 14. April 2008).

Mückschel, C. and Nieschulze, J., 2004, Editorial zum Schwerpunktthema dieser Ausgabe: Datenmanagement in interdisziplinären Umwelt-Forschungsprojekten. Zeitschrift für Agrarinformatik, Heft 4, p. 68.

Mückschel, C., Nieschulze, J., Weist, C. Sloboda, B. und Köhler, W., 2007, Herausforderungen, Probleme und Lösungsansätze im Datenmanagement von Sonderforschungsbereichen, Schwerpunkt "Daten- und Informationsmanagement - aktuelle Ansätze aus interdisziplinären Forschungsprojekten", http://www.ezai.org/index.php/eZAI/issue/view/7 (accessed 14. April 2008).

MySQL, 2008, http://www.mysql.com (accessed 14. April 2008).

Nauss, T. Göttlicher, D, Dobbermann, M. und Bendix, J., 2007, Central Data Services in Multidisciplinary Environmental Research Projects, Schwerpunkt "Daten- und Informationsmanagement - aktuelle Ansätze aus interdisziplinären Forschungsprojekten", http://www.ezai.org/index.php/eZAI/issue/view/7 (accessed 14. April 2008).

Nogueras-Iso, J., Zarazaga-Soria, F. J. and Muro-Medrano, P. R., 2005, *Geographic Information Metadata for Spatial Data Infrastructures, Resources, Interoperability and Information Retrieval*, Springer, Berlin Heidelberg, Germany.

Open SSL, 2008, http://www.openssl.org (accessed 14. April 2008).

TR32-Wiki, 2007, http://www.meteo.uni-bonn.de/projekte/tr32-wiki (accessed 14. April 2008).

Türker, C. and Saake, G., 2005, *Objektrelationale Datenbanken. Ein Lehrbuch*, dpunkt.verlag, Heidelberg, Germany.

Yeung, A. K. W. and Hall, G. B., 2007, *Spatial Database Systems, Design, Implementation and Project Management. The GeoJournal Library 87*, Springer, Dordrecht, The Netherlands.