TR32DB: Implementation of a Scientific Data Management System for the CRC/TR32

Constanze Curdt [(1)], Dirk Hoffmeister [(2)], Christian Jekel [(3)], Guido Waldhoff [(4)], Georg Bareth [(4)]

University of Cologne, Institute of Geography
Albertus-Magnus-Platz, 50923 Cologne, Germany
Email:([(1)]c.curdt, [(2)]dirk.hoffmeister, [(3)]jekelc, [(4)]guido.waldhoff, [(5)]g.bareth)@uni-koeln.de

ABSTRACT

An important duty in every long-term scientific research project is the management of all gained results. This has to include the sustainable storage and backup of all data including accurate description with metadata, as well as the provision and exchange of the data. The realization of such a system raises challenges and problems which need to be resolved.

Therefore, this paper presents the implementation of a scientific data management system for the interdisciplinary, long-term research project Transregional Collaborative Research Center 32 'Patterns in Soil-Vegetation-Atmosphere Systems'. Its design is mainly a combination of file management, database, and self-developed web-interface including web mapping functionalities.

Keywords: scientific data management, interdisciplinary research, metadata, web mapping

INTRODUCTION

In every interdisciplinary, long-term research project, which focuses on environmental field studies and regional modelling, it is essential to collect, store, backup, and exchange data [1]. Therefore, the overall success of such a project depends on a well organized data management. This has to contain the organization of all project data, the implementation, and the maintenance of a system that also encourages intensive data exchange between all involved project sections. Consequently, the sustainable use of all research results during and also after the end of the project is ensured.

In this contribution, we introduce the scientific project data management of the Transregional Collaborative Research Centre 32 (CRC/TR32) "Patterns in Soil-Vegetation-Atmosphere Systems: Monitoring, Modelling, and Data Assimilation" funded by the German Research Foundation (DFG).

Transregional Collaborative Research Centres are applied for up to 12 years, operate on several locations, and also combine cross-disciplinary research fields. They are required by the DFG to establish a so-called 'service project for information management and information infrastructure' which is responsible for the maintenance, management, documentation, and back up of the generated research data. The implementation of the data management system has to be in cooperation with an information provider like a local library or a computing centre [2]. Furthermore, the 'Proposals for Safeguarding Good Scientific Practice' [3] and the 'Recommendation for safeguarded storage and provision of digital scientific primary data' [4] need to be considered.

In this paper, we initially give a short overview of the research project background and the handled data. Then we focus on the implementation of the data management systems and describe its components and functions in detail. The paper ends with a short conclusion.

PROJECT BACKGROUND

The CRC/TR32 is an inter- and multidisciplinary joint research project between the German Universities of Aachen, Bonn, Cologne, and the Research Centre Jülich. The CRC/TR32 started in 2007 and is currently situated in the second funding phase. There are three phases scheduled each running for four

years (Phase 1: 2007-2010; Phase 2: 2011-2014; Phase 3: 2015-2018). The research partners of the CRC/TR32 are from the scientific disciplines of soil and plant science, hydrology, meteorology, geophysics, geography, remote sensing, geophysics, and mathematics. They work in various project sections (SP, Phase 1: 15SPs; Phase 2: 23SPs; Phase 3: ?) on exchange processes between the soil, vegetation, and the adjacent atmospheric boundary layer (SVA). The research goal of the CRC/TR32 is to yield improved numerical SVA models for the prediction of CO2-, water- and energy-transfer by calculating the patterns at the various spatial and temporal scales. The hypothesis of the CRC/TR32 covers the explicit consideration of patterns and structures which lead to a common methodological framework. This will increase the understanding and our capability of describing and predicting the SVA system in a comprehensive manner. The CRC/TR32 study area is the catchment of the river Rur, situated in western Germany, parts of the Netherlands, and Belgium. The field research of the project participants is focused on three sub watersheds that represent the three typical land use forms in of the study area (forest, arable-, and grass land) [5].

The CRC/TR32 participants produce a lot of different data as a result of the interdisciplinary background of the project. They work in various research fields with focus on SVA as well as in divers spatial scales ranges (local, sub-basin and regional/basin scale). Besides a huge amount of data that is collected during several field measurement campaigns, from sensor networks, meteorological and hydrological monitoring or airborne remote sensing, also data resulting from laboratory methods are considered. For example, results of field observations or flight campaigns are: soil water content or soil CO_2 concentration, crop surface/volume models, leaves characteristics, root growth, leave characteristics, as well as land use analysis of the Rur watershed [6-9].

In addition, adequate geodata are purchased within the framework of the CRC/TR32 for modelling purposes from various national or international institutions in different scales. For example topographic, land use, soil, elevation, climate or remote sensing data are purchases from the National Surveying Agency of North-Rhine Westphalia (Germany), Germanys National Meteorological Service, the National Agency of Geology, the European Environment Agency or the United States Geological Service (USGS).

As a result of the research, the project participants also produce a multiplicity of publications, presentations, reports or pictures which need to be handled.

Due to the interdisciplinary background and heterogeneity of the research data, the CRC/TR32 participants provide their data in different data formats. These are for example: ASCII, NetCDF, MS Excel, binary format, PDF, TIF or JPG. Furthermore, the size of a single dataset varies from several gigabytes to few kilobytes.

IMPLEMENTATION OF THE DATA MANAGEMENT SYSTEM

The scientific data management system of the CRC/TR32 (TR32DB) is developed according to the already mentioned guidelines of the DFG as well as to the project background, requirements and needs of the CRC/TR32 participants.

The TR32DB design is a combination of file management, database, and self-developed web-interface including web mapping functionalities (figure 1). The system is physically stored at the Regional Computing Centre of the University of Cologne (RRZK) and is accessible online at www.tr32db.uni-koeln.de. The following paragraphs will describe the components of the TR32DB in detail.

All project data, created by the project participants, are physical stored in the Andrew File System (AFS), a distributed networked file system. The AFS was chosen in cooperation with the RRZK, where it is also located. There are many reasons for this choice. Besides sharing and synchronizing federated file storage, main advantages are aspects like scalability, security, locations independence, cross platform access, easy user administration, as well as simple archive and backup of data. Furthermore, as a result of the cooperation with the RRZK, the availability of all project data is ensured during and beyond the end of the project.

In detail, the AFS data storage is organized in a folder system. This system is divided into a temporary exchange area (Share) as well as temporary and final data storage. They are arranged according to the

hierarchical structure of the CRC/TR32 (project funding phase, research cluster, project section). The temporary and final data storage also distinguishes between five types of data which are (measured/modelled) data, reports, presentations, publications, and pictures. The project participants store their data in the appropriate temporary data storage folder. Within the environment of the AFS a script is operating every night, which moves and sorts all project data from the temporary into the sustainable final data storage/folder. At the same time, an entry for each file is automatically generated into the related MySQL database.
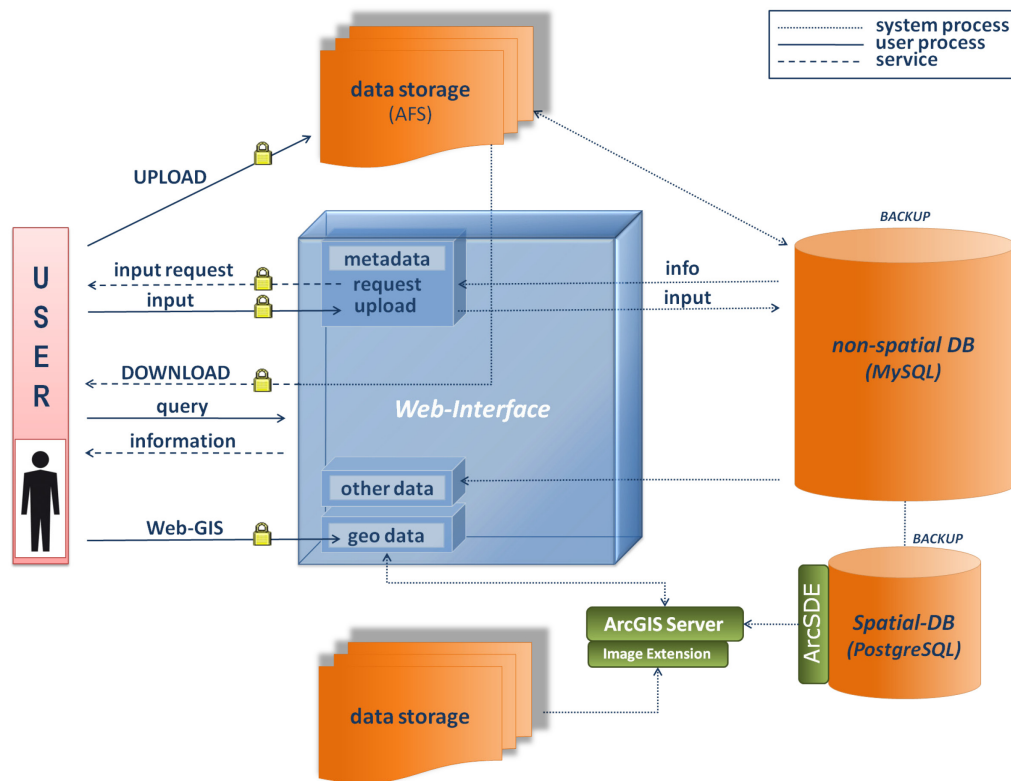
Figure 1: CRC/TR32 data management structure [10]

There are some restrictions for data storage, which need to be considered. These are set by the administrator of the TR32DB for safety and technical aspects. Only authorized users (project participants) are allowed to store or share their project data within the TR32DB structure. Users are just allowed to upload their data in their specific subproject folder of the particular funding phase in which they are working. In addition, the upload of data is limited to 8GB per data type of a project section per day since the moving script operates over night. Therefore, the maximum file size of a single dataset is currently 7GB. However, a maximum size of 4GB is recommended for single files. In addition, the data should be organized, sorted, grouped, entitled and primarily compressed in advanced before submission.

All purchased geodata in raster format, which are needed for visualization purposes in the web mapping application, are currently stored at the Institute of Geography at the University of Cologne. Main reasons for choosing another physical storage are several data access restrictions as well as license regulations.

All descriptive metadata of the submitted data sets are organized in a MySQL database which is also implemented at the RRZK. In addition, administrative data like user information and user rights are stored.

The quantity of the descriptive metadata of a data file depends on its data type because different data types require different attributes. Not every attribute is important for each data type. For example, details about a measuring instrument that should be declared for the data type 'data' are not important for publications or reports. Therefore, a CRC/TR32 metadata profile was designed according to the needs of the project participants, the demands of the data types, as well as to recent metadata standards [11]. To cover all types of data, a multi-level metadata approach was chosen. It is a combination of general, CRC/TR32 specific, as well as data type specific properties. As a 'basic' level, the Dublin Core Metadata

Element Set is chosen. This assures that all types of digital data can be at least described with the 'basic' information. 'Basic' details of a dataset are for example the title, a short description, the creator(s) and the corresponding institution(s), the contributor, publisher, creation and publication date, a citation advice, or the status and the download permission of the dataset. In addition, CRC/TR32 specific properties can be added with focus on the project background and SVA. These are for example CRC/TR32 and SVA-specific related keywords, topics or measurement/modelling regions and locations of the CRC/TR32. To fulfil the requirements of the datasets, additional specific details can be added. For example, a 'picture' requests information about the recorded date, location, event, resolution, size, camera, or copyright. A 'report' just needs further properties like report date and report type. To describe a 'presentation' details like presenter, presentation date, presentation type, presentation event, event title, location, and event period are important, whereas a 'publication' requires information about the publication year, publication place, volume, issue, or page range. While measured or modelled 'data' need to be described with the measurement instrument or modelling method including their resolutions distances, temporal frequency, and parameters, as well as the temporal extend. 'Data' with a spatial relation have to be described for example with the INSPIRE topic category, a lineage, or a geographic bounding box to fulfil the requirements of a geographic metadata standard (e.g. ISO19115, and INSPIRE). In addition, the so called 'metadata on metadata' are stored automatically to complete the metadata properties of a dataset. These are for instance information about the metadata creator, metadata changing date, storage path of the data file as well as details about the project section and funding period the dataset is created in and belongs to. The input of corresponding metadata to a data file into the database is operated by the TR32DB web-interface.

Figure 2: Request on data according to data type 'measured/modeled data' with metadata view

The connective element between the data storages and database is the TR32DB web-interface. It is located at www.tr32db.de. The main tasks of the web-interface are the management and visualisation of metadata, other data (e.g. climate data), and geodata. Therefore, a couple of diverse basic functionalities

and also web mapping functionalities are provided. They will be described in detail below. With regard to licences restrictions of purchases basic (geo-)data, unpublished research data of the project participants, as well as management, download or visualisation of large data files, it is essential to develop a safe and stable web-interface according to recent standards and principles.

Every visitor of the TR32DB web-interface can use the basic functionalities. These are for example a 'quick' search for data files according to the topic (e.g. soil, vegetation, atmosphere, land use, remote sensing) or type (e.g. data, presentation, publication) of data, funding phases (phase 1 or phase 2), clusters (A, B, C, D, Z), or project sections (e.g. A1, A2, B1, B3, C1), as well as to CRC/TR32 modelling/measurement regions and sites. Furthermore, a search form is available where it is possible to combine various queries with each other. These are for example full-text search, data type, funding phase, temporal extend, CRC/TR32 keywords, creator, or measurement/modelling regions and sites. Subsequent to the data query, every visitor is allowed to access the data file details/metadata (figure 2). The representation of the metadata follows the CRC/TR32 metadata profile multi-level approach. Therefore, the metadata pop-up window initially presents the 'basic' information followed by the specific metadata of each data type. The download of the data file is only allowed for authorized users (depending on the chosen download permission of the dataset). Users who download a data file are encouraged to obey the data policy of the CRC/TR32 for example concerning its use and adequate citation.

Only authorized users (project participants) are allowed to use the internal functions of the web-interface that will be describes now. One of the main important functions provided by the TR32DB web-interface is the input of corresponding metadata to a data file (figure 3). Users who are logged in are requested to deliver their metadata to a specific data file which they have just stored in the AFS before. The input of metadata is provided by a simple structured entering form which is adjusted according to the type of the data file. The form is arranged according to the CRC/TR32 multi-level metadata profile. At first, the user has to enter the general properties, following by the CRC/TR32 specific properties and the data type specific properties. As a special function within the metadata input form, the user has the possibility to use the already inserted metadata of another data file as a template. Furthermore, the user has the opportunity to edit the inserted metadata of his data files.

Figure 3: Metadata input form for data type 'measured/modeled data'

Besides, authorized users can apply CRC/TR32 Digital Object Identifiers (DOI, http://www.doi.org) for their data files. Generally one can say that the aim of a DOI is to cite and link to electronic resources. This is widely uses in the scientific community for scientific articles. The usage of DOIs for scientific primary data is getting more and more popular. In this case the DOI is used to make primary scientific data worldwide citable as publications. Thus, a DOI offers a persistent and stable reference to scientific data. In addition, it is an easy way to link an article with the underlying scientific data. [12]

For example, the CRC/TR 32 scientific dataset

- Waldhoff, G. (2010): Land use classification of 2009 for the Rur catchment. doi: 10.1594/GFZ.TR32.1. Retrieved September 20, 2011, from http://www.tr32db.uni-koeln.de/DOI/doi.php?doiID=1

is the underlying scientific data to describe the study of the article

- Bolten, A., Waldhoff, G. (2010): Error Estimation of ASTER GDEM for regional applications - Comparison to ASTER DEM and ALS elevation models. Proceedings of the DIGITAL EARTH SUMMIT, Nessebar, Bulgaria, June, 12-14 2010. Retrieved September 20, 2011, from http://www.digitalearth-isde.org/Bulgaria/pdf/14_Bolten-Waldhoff_Germany_paper.pdf

There are specific guidelines to apply a DOI for scientific data which need to be considered [13-14].

To apply a CRC/TR32 DOI and make the primary scientific dataset citable as publication, the TR32DB users have to follow and accept some conditions. These contain for example the user (author) responsibilities for the quality control of the primary dataset. Furthermore, the user has to agree not to modify or change the primary dataset after the application of the DOI and also the open access to the dataset. The user and the TR32DB are responsible for the quality control of the descriptive metadata. In addition, the TR32DB has to guarantee the long-term availably of the published dataset.

The access to the purchased climate data (from German Weather Service, Meteomedia Group, Rur Water Res. Assn, and Erft Water Res Assn.) is also available via the TR32DB web-interface. Users are able to quick and easy search for climate data from the German Weather Service via an internal search form. They can query data by multiple stations in combination with the time extend and finally export the result to a file. Finally, the TR32DB web-interface provides a tool to quickly generate diagrams of a climate data query by selection of multiple stations, climate parameters, and temporal extend (figure 4).
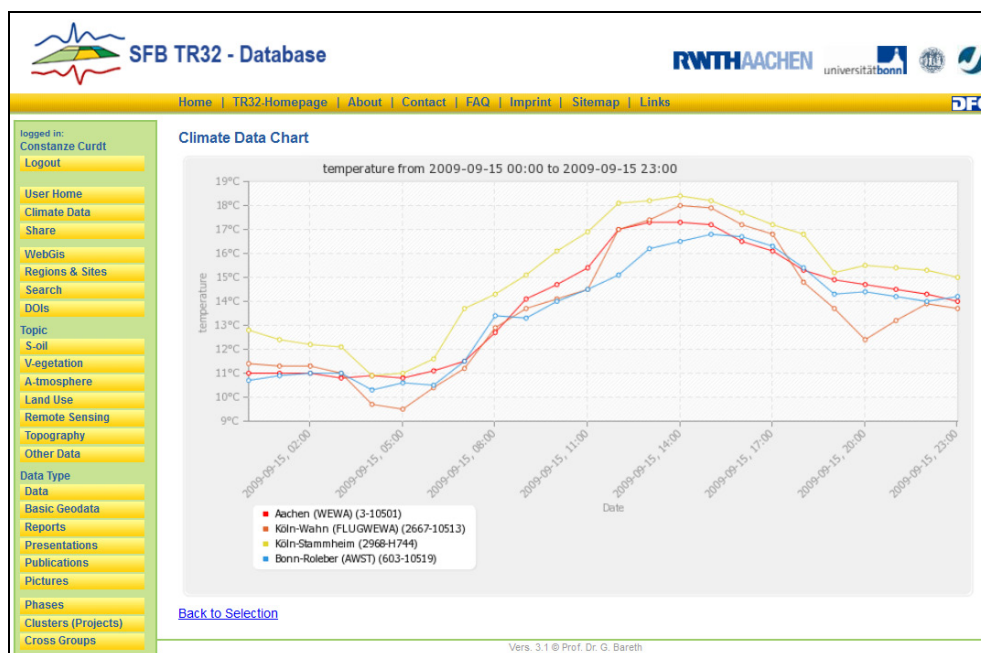

Figure 4: Request and visualisation of climate data

An important part of the TR32DB is the visualisation and supply of purchased CRC/TR32 specific geodata. For this reason an integrated WebGIS (TR32 Viewer) is realized. This WebGIS (figure 5) is only accessible for authorized users because of access restrictions and licenses restrictions concerning the purchased geodata issued by the responsible organisations. The TR32 Viewer is implemented using ESRI's ArcGIS Server API for Flex. Flex provides a fast and pleasant visualisation of all purchased spatial data for the Rur catchment. Currently, the main functionality of the TR32 Viewer is the visualisation of purchased geodata. Within the 'Map' functions (globe symbol) it is possible to display the 'Table of contents'. Here, the user is able to switch on the various layers. These are for example the

regional soil type map of Germany, the digital landscape model, the digital cadastral map, or digital orthophotos. Furthermore, the locations of gauging stations, Eddy Flux Towers, EC Stations, and weather radar coverage are provided in several layers. The user is able to change the transparency of all layer on the fly. In addition, the TR32 Viewer provides 'Tools' (dice symbol). For example the user is able to use the 'DWD stations'-tool. Here, all DWD stations of the catchment area can be queried concerning their location or station attributes. For example, attributes like name, measuring period, parameters, or measurement gaps are displayed for each climate station on request. Furthermore, the user can use the 'Identify'-tool to receive further information about a map layer. The TR32 Viewer also contains functions for 'Navigation' (compass symbol) or 'Help' (question mark symbol).
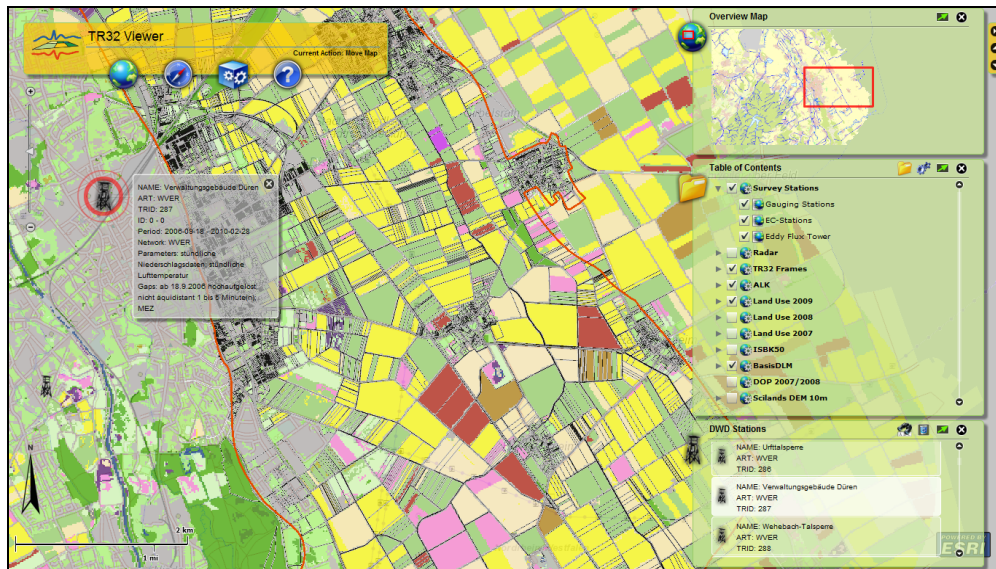

Figure 5: CRC/TR32 Viewer

CONCLUSION

The centralised organisation of scientific research data within interdisciplinary research projects involves several problems. These arise because of various demands on the system. [15] point out, that for example heterogeneous data should be stored using an explicit structure including metadata. Unfortunately, many research projects have experienced that the participation of the users and the time for the development of the system are not appropriate [16]. In addition, [17] notices that the willingness to share research data depends on the scientific community.

In this paper, we have discussed the centralized scientific data management of the DFG funded CRC/TR32, a large, interdisciplinary, long-term research project with focus on regional monitoring, measurement, and modelling in the field of SVA. To support the overall communication and data exchange within the interdisciplinary research project, the implementation of a stable and sustainable system is very important. To face some of the mentioned problems within scientific data management, the design of the TR32DB is arranged to be as simple and user friendly as possible and expandable according to the requirements of the project participants.

REFERENCES

[1] – Mückschel, C.& Nieschulze, J. (2004). Editorial zum Schwerpunktthema dieser Ausgabe: Datenmanagement in interdisziplinären Umwelt-Forschungsprojekten. Zeitschrift für Agrarinformatik 4. Retrieved September 15, 2011, http://www.gil.de/publications/zai/archiv/R8_20040014.pdf.

[2] – German Research Foundation (2009). Service-projects for information management and information infrastructure in CRC – INF. Retrieved September 15, 2011, from http://www.dfg.de/foerderung/programme/koordinierte_programme/sfb/module/modul_inf/index.html.

[3] – German Research Foundation (1998). Proposals for Safeguarding Good Scientific Practice. Recommendations of the Commission on Professional Self Regulation in Science. Retrieved September 15, 2011, from http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf.

[4] – German Research Foundation (2009). Recommendations for 'Safeguarded storage and provision of digital scientific primary data. Retrieved September 15, 2011, from http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901_en.pdf.

[5] – CRC-TR32 (2011). Funding proposal, 2nd phase. Retrieved September 15, 2011, from http://www.tr32.uni-koeln.de.

[6] – Graf, A., Prolingheuer, N., Schickling, A., Schmidt, M., Schneider, K., Schüttemeyer, D., Herbst, M., Huisman, J.A., Weihermüller, L., Scharnagl, B., Stennpass, C., Harms, R., & Verecken, H. (2011). Temporal Downscaling of Soil Carbon Dioxide Efflux Measurements Based on Time-Stable Spatial Patterns. Vadose Zone Journal 10 (1). doi: 10.2136/vzj2009.0152. Retrieved September 15, 2011, from http://vzj.geoscienceworld.org/cgi/content/abstract/10/1/239.

[7] – Schickling, A., Graf, A., Pieruschka, R., Plückers, C., Geiß, H., Lai, I-L., Schween, J., Erentok, K., Schmidt, M., Wahner, A., Crewell, S. & Rascher, U. (2010). The influence of leaf photosynthetic efficiency and stomatal closure on canopy carbon uptake and evapotranspiration - A model study in wheat and sugar beet. Biogeosciences. 7. doi: 10.5194/bgd-7-7131-2010. Retrieved September 15, 2011, from http://www.biogeosciences-discuss.net/7/7131/2010/bgd-7-7131-2010.html.

[8] – Hoffmeister, D., Bolten, A., Curdt, C., Waldhoff, G. & Bareth, G. (2009). High-resolution Crop Surface Models (CSM) and Crop Volume Models (CVM) on field level by terrestrial laser scanning. Proc. SPIE 7840. doi: 10.1117/12.872315. Retrieved September 15, 2011, from http://link.aip.org/link/?PSI/7840/78400E/1.

[9] – Waldhoff G. & Bareth, G. (2009). GIS- and RS-based land use and land cover analysis - case study Rur-Watershed, Germany. Proc. SPIE 7146 (26). doi: 10.1117/12.813171. Retrieved September 15, 2011, from http://spie.org/x648.html?product_id=813171

[10] – Curdt, C., Hoffmeister, D., Jekel, C., Brocks, S., Waldhoff, G., & Bareth G. (2011). TR32DB – Management and Visualization of Heterogeneous Scientific Data. Proceedings of the 19th International Conference on GeoInformatics, Shanghai, China. doi: 10.1109/GeoInformatics.2011.5981092.

[11] – Curdt, C., Hoffmeister, D., Waldhoff, G., & Bareth G. (2009). Implementation of a metadatabase for a regional SDI for soil-vegetation-atmosphere modelling. Proc. GSDI 11 World Conference: Spatial Data Infrastructure Convergence: Building SDI Bridges to address Global Challenges, Rotterdam, The Netherlands. Retrieved September 15, 2011, from http://www.gsdi.org/gsdiconf/gsdi11/papers/pdf/189.pdf

[12] –Brase, J. and A. Farquhar (2011). Access to Research Data. D-Lib Magazine 17(1/2). Retrieved September 15, 2011, http://www.dlib.org/dlib/january11/brase/01brase.html

[13] – Brase, J. (2004). Using Digital Library Techniques – Registration of Scientific Primary Data. Lecture Notes in Computer Science. 3232.

[14] – Brase, J. & Klump, J. (2007). Zitierfähige Datensätze: Primärdaten-Management durch DOI. WissKom 2007 : Wissenschaftskommunikation der Zukunft. R. Ball. Jülich, Forschungszentrum, Zentralbibliothek. Schriften des Forschungszentrums Jülich 18.

[15] – Mückschel, C., Nieschulze, J., Weist, C., Sloboda, B. & Köhler, W. (2007). Herausforderungen, Probleme und Lösungsansätze im Datenmanagement von Sonderforschungsbereichen. eZAI, 2: 1-16. Retrieved September 15, 2011, http://www.preagro.de/ezai/index.php/eZAI/article/viewPDFInterstitial/21/22

[16] – Nelson, B. (2009). Empty archives. Nature 461, (7261). Retrieved September 15, 2011, http://www.nature.com/news/2009/090909/pdf/461160a.pdf

[17] – Borgman, C.L. (2010). Research Data: Who will share what, with whom, when, and why? China-North America Library Conference. Retrieved September 15, 2011, from http://works.bepress.com/borgman/238/.