20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Data and information management for integrated research – requirements, experiences and solutions

F. Zander[a], **S. Kralisch**[a], **W.-A. Flügel**[a]

[a] *Friedrich-Schiller-University, Department of Geoinformatics, Hydrology and Modelling (DGHM) Germany*
Email: franziska.zander@uni-jena.de

Abstract: Environmental management and related interdisciplinary research address a wide range of data and information management demands. The collaborative use of respective data requires providing descriptive meta-information as well as functionality for their storage and access. Further, a detailed and fine-grained permission management is important to preserve intellectual property rights and thus to build-up trust into such systems. A variety of readily available software tools, and standards for data exchange and processing, supports the assembly and deployment of appropriate data management and data sharing platforms that address these requirements and assist researchers to find, access, store, describe, process and disseminate data and results. An example of these systems is the modular-structured, web-based River Basin Information System (RBIS) developed in the Department of Geoinformatics, Hydrology and Modelling at the University of Jena. RBIS focuses on the management of metadata and data (e.g. time series data, geospatial data) and the provisioning of standard compliant data exchange interfaces and services (e.g. WaterML and CSW). RBIS currently has been applied in more than 20 environmental research projects as a data management and sharing platform to support single researchers, bilateral research activities and multi-disciplinary research projects during project runtime and beyond. Environmental data (e.g. Time series and geospatial data) are gathered, preprocessed and made available via interfaces for external applications (e.g. environmental models). Moreover, interim and final results can be shared with project partners as well as local stakeholders. The development of the core system and modules has followed technical, economical and functional requirements raised in integrated environmental research projects during recent years during which RBIS was applied and enhanced. The most important components and their implementation can be summarized as follows:

- **Open source:** To support reuse and extensibility of software, RBIS is based on open source software (e.g. PostgreSQL, PHP, JQuery, UNM MapServer, OpenLayers, pycsw, …).
- **Online / offline access**: RBIS is web-based, but due to the fact that many regions of the world have a unreliable internet connectivity and for an easy distribution, RBIS may also be operated in a virtual server environment to allow for offline accessibility.
- **Flexible and extensible**: To build a system which is flexible, scalable and easy to extend according to actual requirements, RBIS is structured in a modular way using a data description layer to encode the visualization, manipulation and linking of datasets as well as the internal representation of the data in the database.
- **Permission management**: In order to provide a platform for data sharing, a fine-grained user and permission management process was implemented, which includes the ability to apply restrictions to datasets, RBIS modules, and related actions.
- **Interfaces and services**: To expose data on the internet and to exchange data with other applications (e.g. modeling tools), standard compliant interfaces and services are provided (e.g. CSW and WaterML).
- **User acceptance and trust**: In order to raise acceptance, motivate, and build trust, it is not enough to provide a technical solution, but rather to develop training courses for data providers and users, as well as a multilingual web interface, online tutorials and comprehensive support. Additional group communication functions, such as a simple internal calendar, as well as an automatic notification about new or changed events or stored datasets, can also contribute to assist end users.

RBIS is seen as a technical contribution to integrated research projects for bringing different data types and disciplines together, to support different steps in scientific workflows and to assist in the sharing and dissemination of research results.

Keywords: *Environmental information system, research project data management, time series data, geospatial data, data sharing*

1. INTRODUCTION

Data management in research projects, relative to sustainable environmental management, is quite challenging because usually many different disciplines and people are involved. An important role for the collaborative use of respective data involves providing of descriptive metadata and functionality of data for their storage and access, to enable accessibility not only within a project consortium but also for local stakeholders and decision makers. There are a variety of readily available software tools and standards for data exchange and processing, which support the assembly and deployment of appropriate data management and data sharing platforms. These help researchers find, access, store, describe, process, preserve, and disseminate their data and results. Examples are PANGAEA (Data Publisher for Earth and Environmental Science; http://www.pangaea.de)(Diepenbroek et al., 2008), GESIS (Data archive and service provider for social sciences; http://www.gesis.org)(Jensen, 2012) and DRYAD (Repository for biosciences; http://datadryad.org)(Vision, 2010). Beside these repositories there are many other systems on different scales, for certain types of data and purposes. One example is the River Basin Information System (RBIS) which is developed in the Department of Geoinformatics, Hydrology, and Modelling, at the University of Jena, based on gathered experiences and demands in integrated environmental research projects. RBIS focuses upon serving not only as data repository, but rather as an integral part of scientific workflows during project runtime. Therefore, RBIS focuses not only on the management of metadata and data, but also on the analysis and processing of time series data and upon providing standard compliant interfaces and services (e.g. WaterML (Taylor, 2012) and CSW (Nebert et al., 2007)). Besides the support for scientists and construction of a data platform for extended research, RBIS aims at sharing gathered data, interim and final results, not only among project partners, but also with local stakeholders and decision makers.

In the following sections, the general structure and selected functions of RBIS will be explained, based on selected key requirements and gathered experiences during the application of the system for research projects over the past few years. Furthermore, some example applications of RBIS will be used to illustrate the results.

2. REQUIREMENTS, EXPERIENCES AND SOLUTIONS

The development of the core system, and of all RBIS modules, followed technical, economical and functional requirements which arose, based on experiences and demands gathered from integrated environmental research projects, in which RBIS was applied and enhanced, during the past 10 years. Selected requirements and problems and their implemented solution within RBIS are described in the following sections.

2.1. Open source

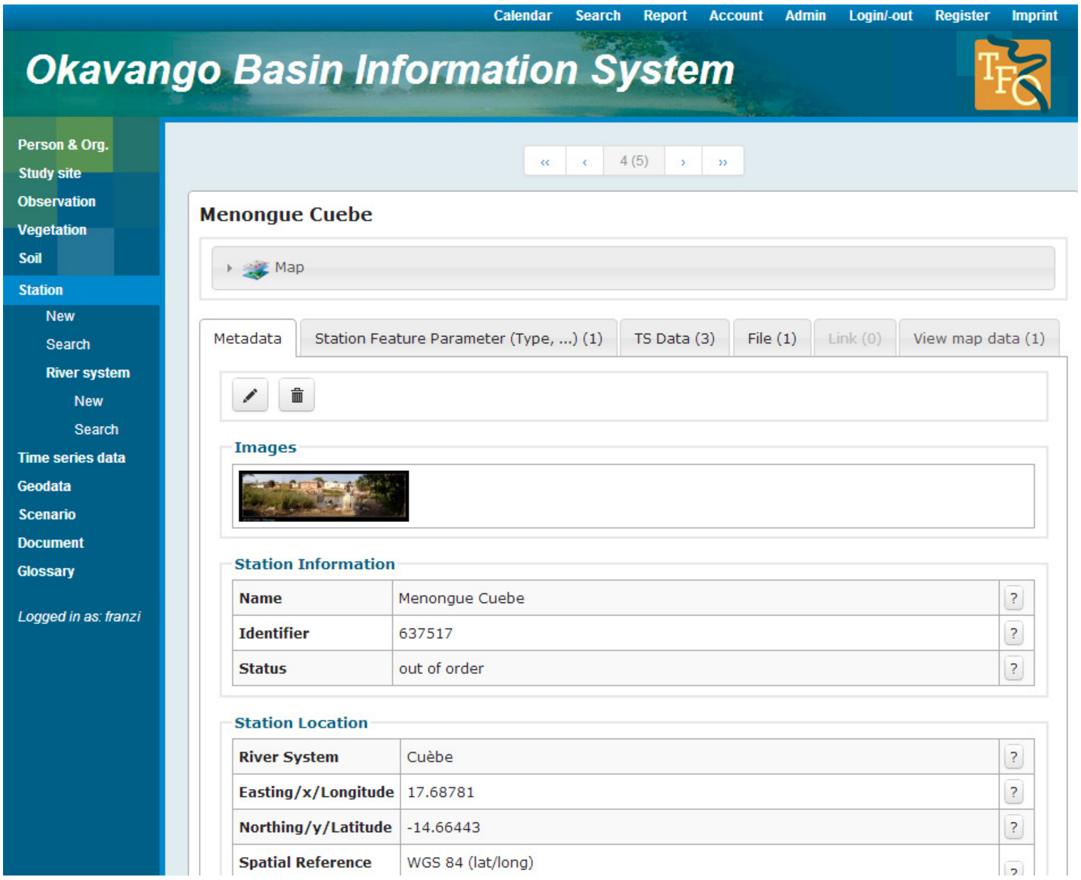
RBIS is based upon freely available open source software in order to ensure low costs for deployment and operation and to support reuse and extensibility of software. The common layout of RBIS follows a 3-tier architecture. On the server side, the system is implemented using a standard Linux web stack with Apache web server, PHP programming language, PostgreSQL database management system (http://www.postgresql.org) and PostGIS extension for spatial data support. To manage and visualize spatial data, the UMN MapServer is used to create maps and the OpenLayers library is for a user friendly display of map data in a web browser. The web interface for the management of metadata and associated data has been restructured using the JavaScript library JQuery.

2.2. Online and offline access

RBIS is web-based to enable the collaborative use of stored data and information. As long as the internet connectivity is good it works fine. Unfortunately most of the project regions where RBIS has been applied until now are located in countries with unreliable internet connectivity (such as Africa, Latin America or Asia). To enable offline access and for ease of distribution RBIS may also be operated in a virtual server environment using the VirtualBox (http://www.virtualbox.org) software package. A handover of the system with all collected and processed data to local stakeholders should be part of the dissemination process in the last phase of the project or, in additional as a snapshot during the course of the project. This procedure allows for collected data to stay in the region of their origin and also gives additional incentives to provide data, although RBIS might not always be accessible locally during the project due to connectivity problems or other limiting.

2.3. Flexibility, scalability and extendibility

Initially a primary aim of RBIS was to build a system that can be readily applied, and partly reused, for other research projects with similar or additional demands. Therefore, the underlying software structure is very flexible and adaptable with XML-files used as description layers to encode the visualization, manipulation and linking of datasets (Figure 1), and the internal representation of the data in a relational database (for more detail see Kralisch et al., 2009). Furthermore all datasets can be linked with each other or used as part of other datasets (e.g. a person as responsible party). Datasets with a spatial relation (e.g. from monitoring sites) are automatically linked to a map, based on predefined rules in a corresponding XML-file. Furthermore, it is possible to associate files to each RBIS dataset. These files might contain the data described by the metadata, but also images (e.g. of a measurement station) or other files with additional information.



Menongue Cuebe

Station Information

Name	Menongue Cuebe	?
Identifier	637517	?
Status	out of order	?

Station Location

River System	Cuèbe	?
Easting/x/Longitude	17.68781	?
Northing/y/Latitude	-14.66443	?
Spatial Reference	WGS 84 (lat/long)	?

Figure 1: Station metadata information (partial) and linkage (3 time series data, 1 file, 0 links to other datasets, 1 geometry object (point) in a map)

2.4. Modules

RBIS is built in a modular way. There are several mandatory and other optional modules (mainly related to certain data types) and associated functions. The description of responsible parties, persons and organizations based on the ISO 19115 standard (ISO, 2003), and all administrative core functions, belong to basic modules and functions. Overarching modules are optional, such as the map component (RBISmap) for the management and visualization of raster and vector files, or modules like the management and processing of station and time series data (RBISts), or the management of geodata metadata according to ISO 19115 (as part of RBISmap). The 'observation' module (RBISobserv) represents a generalized type of dataset with a spatial relation (can be either a point or polygon). The location information is stored as study site or area and field observations, such as measurements, field trip descriptions, soil and water samples or survey data, can be easily linked to study sites or stored in other RBIS modules, adapted to the respective data types, and linked to additional study sites (e.g. to group datasets related to sub catchments defined as study areas). An overview of the main components is shown in Figure 2.

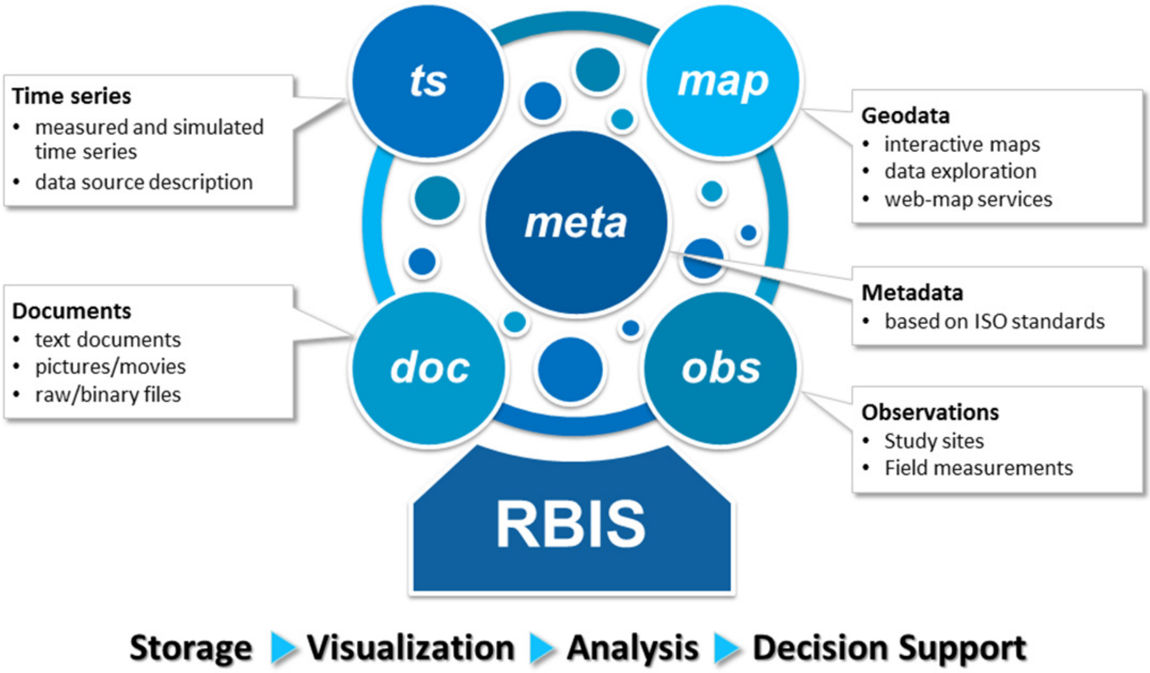


Figure 2: RBIS primary components.

2.5. Support for scientific workflows and environmental modeling

Data should not be only managed and stored within RBIS, but also continually used. To allow for this function, data and data provenance need to be described in detail within the metadata. Functions to process and export the data (e.g. time series data) also should be included. Furthermore, data management should be somehow integrated with a scientific workflow. One of these workflows is covered by the Jena University's Integrated Land Management System (ILMS), which provides an integrated modular software platform and covers different steps in environmental systems analysis and planning with a flexible and user-friendly workflow (http://ilms.uni-jena.de). One integrated part of ILMS uses ILMSinfo (or RBIS) for the centralized management, analysis, visualization and presentation of different types of data. Other software components in ILMS are as follows: ILMSimage (a software tool for the identification and classification of real-world objects from satellite imagery using methods of object based image analysis), ILMSgis (a software for the derivation of modelling entities using a Web Processing Service based on GRASS GIS and following the Hydrological Response Unit (HRU) approach) and ILMSmodel (an environmental modeling framework (Jena Adaptable Modelling System - JAMS) for building, running and analyzing environmental simulation models (http://jams.uni-jena.de), e.g. hydrological models (Kralisch et al., 2012) .

In order to provide time series data (e.g. meteorological and hydrological data) ready to use for modeling (e.g. in ILMSmodel), the data need to be checked and corrected in advanced (e.g. date/time and data format consistency and/or data gaps). This is done during the uploading of time series data, and detected data gaps can be filled with the rule-based gap filling toolbox in RBIS (Zander et al., 2011). After visualization and analysis, time series data can then be exported or directly accessed by models (see more details in Kralisch et al., 2009).

2.6. Interfaces and services

In addition to the web-based user-interface for data and metadata import, manipulation, visualization, analysis, and export, RBIS provides several import and export interfaces and services for data as well as metadata. To expose a catalogue of selected metadata records (raster, vector, and time series data) on the internet, an OGC standard-compliant CSW-Service (Catalogue Service for the Web) (Nebert et al., 2007) based on the CSW server implementation 'pycsw' (www.pycsw.org) has been set up. Currently the RBIS CSW-service is used in an overarching global search for raster, vector and time series data on selected RBIS installations (http://leutra.geogr.uni-jena.de/RBISsearch) and in the GLUES Geodata infrastructure (http://geoportal.glues.geo.tu-dresden.de), which is a common data and service platform for the international research program 'Sustainable Land Management' sponsored by the German Ministry of Education and Research (BMBF)(Bernard et al., 2013).

RBIS also provides several interfaces for the automatic import and update of time series data. The data sources can be local or online repositories as well as web services. An example of an online repository is the Global Surface Summary of the Day (GSOD) product (global data from weather stations), calculated by the National Climatic Data Center (NCDC) in Asheville, NC USA and based upon the database of the Integrated Surface Database (ISD)(DSI-3505)(Lott, 1998; Smith et al., 2011). GSOD data can be imported and updated in RBIS very readily, including a unit conversion from English to SI-units.

2.7. Data sharing, access and protection of Intellectual Property

In a collaborative project data should be shared and accessible to all project members without any restrictions. In reality, however not all data can be shared, due to specific license restrictions (e.g. political or institutional). Sometimes intellectual property has to be protected, because it may be necessary in some disciplines to get authorization or approval before publication of the corresponding work. In addition, certain data, especially interim results or purchased data, should not necessarily be accessible to all stakeholders or other external users. Therefore, it is necessary to have a very fine-grained user and permissions management system. The user and permissions management in RBIS is based on permission groups, actions (view, download and edit) and data types (e.g. time series data), with users being assigned to permission groups. This allows the granting of permissions on all datasets related to one module. If this indirect permission management is not enough, additional restrictions can be set on each dataset by the dataset owner (e.g. invisible or download of data only on request) and explicit access permissions can be granted to specific individuals.

2.8. User acceptance, trust and motivation

Some of the most important requirements are to motivate, raise acceptance, and build trust. It is not enough to simply provide a technical solution with a ‘nice’ interface, but rather, the requirements might be finally achieved with the help of training courses for data providers and users, a multilingual web, online tutorials, and comprehensive support. In the case of RBIS, training courses are offered and conducted for project partners and local stakeholders. In order to reduce problems with language barriers, the RBIS frontend currently is available in English, Vietnamese, Portuguese and German and can be translated to other languages if needed. Furthermore, an online tutorial is available, and there is a comprehensive support provided to assist data providers and users of RBIS.

Additional functions are available to support research projects during a project and to keep the existence of the system in the users mind. For example, a simple internal calendar is available with the possibility of linking events and files as well as automatic notifications on new or changed calendar events or stored datasets.

3. APPLICATIONS

RBIS has been applied in more than 20 environmental research projects to serve as a data management and information system. The size of associated projects ranges from single PhD projects (small group of users; at least one) up to multi-disciplinary research projects (many registered user accounts).

RBIS installations, related to single PhD projects, focus mainly on meteorological and hydrological data management, pre-processing, analysis, access and preservation of time series data, and processing (Zander et al., 2012), as one part of the applied workflow covered by ILMS (see section 2.5). One example is the Kosi RBIS (http://leutra.geogr.uni-jena.de/kosiRBIS), which is related to recently a completed PhD research project (Nepal, 2012). Furthermore, the built data collection can be or is used for further research.

In larger research projects, the focus is more on the management of basic and result data, sharing, exchange and presentation, for all project members and local stakeholders within an inter-/multi-/trans-disciplinary environmental research project. Examples of on-going projects are the BMBF-funded project ‘The Future Okavango’ (http://www.future-okavango.org), with OBIS (Okavango Basin Information System) to store environmental information from the Okavango basin, and the BMBF-funded ‘LUCCI’ project (Land Use and Climate Change interactions in the Vu Gia Thu Bon River Basin/Central Vietnam - http://www.lucci-vietnam.info), where environmental data from the Vu Gia Thu Bon River Basin are stored within the Vu Gia Thu Bon RBIS.

4. OUTLOOK AND CONCLUSION

Currently, there are new modules under development, for example, a web-based execution of ILMSmodel model runs based on time series data stored in RBIS and a module to manage models, model runs, parameter sets, and model results. Results in the form of indicators will be stored in the existing module, these then can be adjusted and enhanced with regard to the produced data. Furthermore, all other ILMS components will be executable and web-based as well. This results in a closer integration with using RBIS as central data storage.

It has been shown that RBIS is a useful tool for different sizes of environmental research projects. Certain challenges and experiences have influenced the RBIS developments during the past few years. Therefore, RBIS has become a powerful tool to support, on the one hand, scientists in time series data and geodata management, and presentation, and on the other hand, data sharing and management within environmental research projects. The maintenance, deployment and enhancement effort is easily managed due to its modular, flexible structure, and the application of open source software. The integration of data from different disciplines and the web-based execution of model simulation runs is based upon aspects, such as land use change and climate change scenarios. The presentation (e.g. visualization) of result data and findings of RBIS will result in it being more capable of supporting decision makers deliberations while serving as an information platform for local stakeholders.

ACKNOWLEDGMENTS

The authors acknowledge the support of the German Ministry of Education and Research which has funded the RBIS development within various research programs (grant numbers 03IP514, 01LL0912, 01LB0801, 01LL0908D).

REFERENCES

Bernard, L., Mäs, S., Müller, M., Henzen, C., Brauner, J., 2013. Scientific geodata infrastructures: challenges, approaches and directions. Int. J. Digit. Earth 0, 1–21.

Diepenbroek, M., Schindler, U., Grobe, H., 2008. PANGAEA ® - platform for an ICSU World Data Center as a networked publication and library system for geoscientific data Network of ICSU WDCs. epicawide.

ISO, 2003. International Standard ISO 19115 Geographic information – Metadata.

Jensen, U., 2012. Dienstleistung für eine internationale Community? Das Datenarchiv für Sozialwissenschaften der GESIS.

Kralisch, S., Böhm, B., Böhm, C., Busch, C., Fink, M., Fischer, C., Schwartze, C., Selsam, P., Zander, F., Flügel, W.-A., 2012. ILMS – a Software Platform for Integrated Water Resources Management, in: Seppelt, R., Voinov, A.A., Lange, S., D. Bankamp (Eds.), Leipzig, Germany.

Kralisch, S., Zander, F., Krause, P., 2009. Coupling the RBIS Environmental Information System and the JAMS Modelling Framework, in: Proc. 18th World IMACS/and MODSIM09 International Congress on Modelling and Simulation, Edited by: Anderssen, R., Braddock, R., and Newham, L., Cairns, Australia. pp. 902–908.

Lott, N., 1998. Global surface summary of day. Natl. Clim. Data Cent. Asheville NC Httpwww Ncdc Noaa Govcgibinres40 Pl.

Nebert, D., Whiteside, A., Vretanos, P. (Eds.), 2007. OpenGIS Catalogue Services Specification 2.0.2.

Nepal, S., 2012. Evaluating upstream downstream linkages of Hydrological Dynamics in the Himalayan Region (Dissertation). Friedrich-Schiller-University, Jena.

Smith, A., Lott, N., Vose, R., 2011. The Integrated Surface Database: Recent Developments and Partnerships. Bull. Am. Meteorol. Soc. 92, 704–708.

Taylor, P. (Ed.), 2012. OGC WaterML 2.0: Part 1 -Timeseries. Open Geospatial Consort.

Vision, T., 2010. The Dryad Digital Repository: Published evolutionary data as part of the greater data ecosystem. Nat. Preced.

Zander, F., Kralisch, S., Busch, C., Flügel, W.-A., 2011. RBIS - An Environmental Information System for Integrated Landscape Management, in: Hřebíček, J., Schimak, G., Denzer, R. (Eds.), Environmental Software Systems. Frameworks of eEnvironment. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 349–356.

Zander, F., Kralisch, S., Busch, C., Flügel, W.-A., 2012. Data management in multidisciplinary research projects with the River Basin information System, in: Hans-Knud Arndt, Gerlinde Knetsch, W.P. (Eds. . (Ed.), Shaker Verlag. Shaker Verlag, pp. 137–143.