A Concept of a Virtual Research Environment for Long-Term Ecological Projects with Free and Open Source Software

Mirko Filetti, Albrecht Gnauck

A Concept of a Virtual Research Environment for Long-Term Ecological Projects with Free and Open Source Software

Mirko Filetti[1], Albrecht Gnauck[1]

[1] Brandenburg University of Technology at Cottbus
Dept. of Ecosystems and Environmental Informatics
Konrad-Wachsmann-Allee 1, D-03046 Cottbus, Germany.
E-mail: Corresponding author: filetti@tu-cottbus.de

Abstract. The management of data and data resources created by different research activities are heavily influenced by various research philosophies and sampling strategies. Within long-term environmental research (LTER) projects data on flows of individuals, chemical substances and other biotic and abiotic materials are collected by different project partners and institutions. This leads not only to different data bases, but also to incomparable data sets. Therefore, a virtual research environment (VRE) for research projects concerning environmental management should be worked out. The facilities of data sharing, interactive data collaboration and data storage as well as the communication within a project team by metadata are in the focus of a VRE which have to be optimised by WEB 2.0 and other collaboration tools. From this background the FOSS application "GeoNetwork – Opensource" (GNOS) is aimed to be used as a central component for data management in a VRE.

Keywords: Long-term ecological research, data management, virtual research environment, software tools

1 Introduction

Annually, around ten billion Euros will be invested for information on the public sector by European countries. Approximately 50% of this information consists on spatial data [1]. But much often, wrong data management strategies concerning quantity, quality, actuality and availability of data leads to high redundancies of collected data sets. Therefore, for a good scientific practice the following aspects of data management should be of interest:

1. A data management concept is needed;
2. Primary data must be sorted with the ability to re-find;
3. Data have to be stored in defined standards.
4. Data security, data privacy, copy rights have to be guaranteed;

5. Open access for the public;
6. All steps of data processing have to be transparent.

According to the Deutsche Forschungsgemeinschaft [2] primary research data should be stored for minimal 10 years on solid storages in the institution that creates the data. Within long-term environmental research (LTER) projects long data sets on flows of individuals, chemical substances and other biotic and abiotic materials are collected by different project partners and institutions. Therefore, an IT-based web 2.0 management concept for the whole project is needed to supply the scientific work with collaboration tools inside and giving representative presentations to the public. The use of free and open source software (FOSS) is an attractive alternative to high cost commercial and often proprietary software solutions because of low cost, short timely software terms, independency and security.

From this background the FOSS application "GeoNetwork – Opensource" [3] is aimed to be used as a central component for data management in a VRE. Based on international standards, GNOS is a data and information management system, which includes interoperability, metadata, data harvesting, and geographical references as well as user groups with different authorisation levels. The development of GNOS has been initiated by the UN in 2001, which has been continued by several partners. In the meanwhile, basic GNOS applications are successfully implemented and modified in many environmental projects of different governmental and non-governmental organisations worldwide.

2 General working steps of long-term research projects

The data resources within a multidisciplinary research project like diagrams, pictures, photographs, measured data curves, visual observations and others should be available for collaboration and permanent long term access. As a first attempt, data sets have to be stored in repositories which should be identified by metadata. The data sets keep their different formats, but they can be searched and collected with their metadata assets by a search engine. For this reason, a modified use of the GNOS application is necessary to guarantee data sharing facilities.

In this context, some questions have to be answered concerning usability of original data, retrieval by metadata and authorised access by specific policies. The following general working steps should be considered:

1. A general database has to be set up to store the data of subprojects and to record them with metadata. The interoperability is warranted with all external and internal data repositories.
2. Technical processes have to be adapted and implemented to reference and to maintain research data. Linkages with and references to different data systems and repositories outside of the research project considered are intended as well.
3. The storage system has to be designed. This will be happen by implementation of interoperability tools and/or interfaces for retrieval and referencing.

4. Long term storage of data and archiving tools have to be set up within the project under consideration or outside the project to save and to store research data over longer time horizons for further usage.

Thereby, it is not necessary to develop a completely new data storage and retrieval system. Adaptations and modifications of an available GNOS application have to be adapted to the requirements of a specific research project.

3 Virtual Research Environment

A virtual research environment (VRE) can be defined as a set of web applications and tools, systems and processes interoperating to facilitate or enhance any research activity within and outside of institutional boundaries. The key issue of a VRE is the development and implementation of an information and data sharing concept [4] where data sharing can be done by different media [5]. Further technological problems and challenges are given by software and hardware aspects. In practice, for a successful use of a VRE it is necessary to have clear ownerships of data, a confirmed research project plan with data policies among the collaborators, clear research objectives and responsibilities, and an adequate personal resource for the IT management. But, the most important point is that VREs need to be more considered as community building projects than as technology projects. VRE's give benefits scientific disciplines at all levels of research. There is neither an 'out of the box-solution' nor a 'one size fits all realizations' approach that will meet the demands of all research activities. Research results will come out very quickly by using a VRE environment. Also new research directions will be supported as well. The range of applications of VRE is very broad. For long-term research, sustainability is required with the same long-term commitment as other parts of the infrastructure of the project lifecycle.

Most VRE's have an international dimension. Therefore it is necessary to attend legal, ethical and other policies and frameworks that govern the sharing of data and other resources. The VRE's offer benefits not only for the project itself but also for improvements of general co operations and best practice solutions. Arguments for national and international funding schemes of VRE's are formulated by Carusi and Reimer [5]. Research funds should be used to organize general networking procedures and interdisciplinary research. They support expensive research infrastructures and the productivity of researchers. It can be expected that an increasing speed of information and communication lead to a faster dissemination of research results, to a better preservation of research outputs, and to a new quality of research outputs.

4 VRE services

Depending from the project goals within a VRE environment several services are available. The external view on a VRE should be a unique monolithic system with fast access. But, neither a total self-development of a VRE is possible (because of time, knowledge and resource conflicts) nor is a software tool known, that fits all the

requirements for such a VRE. The challenge for an IT-developer is to find best practice FOSS-solutions for all services needed and patch them optionally all together to a "stand alone" VRE-application that is according to all (international) standards and to take care for sustainability (support from the FOSS community).

An overview on some services contained in a VRE is presented in table 1. Between definition of services and selection of applications a lot of work is hidden. For testing and evaluating each software part, several extensions, different versions and the complete VRE system including the web server software and hardware several years have to be spent.

Table 1. Services for a Virtual Research Environment (VRE)

VRE service	Characteristic
Access management	Single sign in for different applications
Communication	Web 2.0 elements like messaging, chat, forum, wiki
Data analysis	Data analysis tools, statistical methods
Data visualization	Visualization of information and datasets
Data warehousing	Complex data storage and data analysis
Decision support	Aggregated data for decision makers
E-Learning	Platform for students with E-Learning procedures
Event calendar	Internal and external community events
Group management	Groups- and rights management, organization for teams
Map and spatial data	Map-server, case area maps
Metadata management	Information about data
Mobile access	Optimized layout of webpages, augmented reality, access control
Monitoring	Real time monitoring of sensor data etc.)
Project management	Project organization tools like tasks, milestones, workflow, reports
Project website	Flexible content management system (CMS)
Repository	Data repository and data storage, compression, indexing
Search engine	Global and local comfortable search engines
Social web	Facebook, Twitter integration, etc.
Search engine optimization	(SEO) ranking in top search engines like Google

In table 2 selected State-of-the-Art of FOSS applications are presented for some core services and their relations to other applications. The CMS Joomla [6] will be recommended as a second core component. It is a very popular CMS with a world-

wide developer community and over 7,300 components for all kinds of extensions and hundreds of bridges to third party software.

Table 2. Selected FOSS for VRE Services

Service	FOSS	Relations
Project website, community	Joomla, Typo3, Dupral	LDAP, Shibboleth
Access management	LDAP, Shibolet	
Project collaboration, communication and workflows	BSCW, Joomla components	LDAP, Shibboleth
Repository search with meta data	GNOS, (BExIS, Pangea)	LDAP, Shibboleth
Documentation and best practice	Mediawiki	Bridge for Joomla
Monitoring	Joomla component „Art Data“	LDAP, Shibboleth
Data analysis/Data warehousing	Infobright.org, noSQL, Joomla component „Art Data“	Stand alone on local PC, data Bridge for Joomla
Decision support	Infobright.org, Joomla components,	
E-Learning	Moodle	Bridge for Joomla
Webserver-software	Linux, Apache, MySQL, Tomcat, PHP, XML/XSLT	
Social web	Joomla-Facebook/Twitter integration	Bridge for Joomla
Mobile access	Optimized Homepage	Bridge for Joomla
SEO	Joomla component	

5 GeoNetwork Opensource

The issues of data management, data sharing and data storage are getting more and more important on the different research fields. Only software systems like GeoNetwork Opensource [3] are suitable for an open access onto scientific data within a worldwide net. This software is of low cost, with a worldwide developer scene, independent from commercial dependencies, and follows a unique administrator scheme.

The development of GNOS has been initiated by the UN in 2001, which has been continued by several partners as Food and Agriculture Organisation (FAO), UN Office for the Coordination of Humanitarian Affairs (UNOCHA), Consultative Group on International Agricultural Research (CSICGIAR), The UN Environmental Programme (UNEP) and the European Space Agency (ESA). In the meanwhile, basic GNOS applications are successfully implemented and modified in many environmental projects of national and international governmental and non-governmental

organisations worldwide. Figure 1 shows the start-up page of a fresh GNOS installation.

Fig. 1. Start-up page of GeoNetwork Opensource with map-viewer and hit list

For the facilities of data management, which should be realised in such a project (collaborative processing, aggregation, sharing, publication and long tern archiving) there is obviously no need to initiate a completely new development of GNOS or to create a new application. But it is necessary to modify the available GNOS application, which it fits to the specific requirements of the different sub-projects and workflows. For this reason, a communication tool should be developed which fulfils the actual data requirements, and connects all long-term data partners within a research project.

5.1 Consideration of standards

The implementation of a research project on information and storage systems has to be in congruence with national and international standards. Based on GNOS the some international standards for metadata and harvesting which should be introduced in a project work are presented in table 3:

Table 3: Considered metadata standards for GeoNetwork Opensource

Standard	Description

Dublin Core (DC) Metadata (International Organization for Standardization, ISO)	Basic standard for a (minimal) metadata description (not special for spatial data).
Content Standard for Digital Geospatial Metadata (CSDGM) (Federal Geographic Data Committee (FGDC; ESRI FGDC);	Standard for digital geospatial metadata from the leading GIS manufacturer ESRI and the Federal Geographic Data Committee
ISO 19115	Geographic Information Metadata, common base with FGGC, but more detailed
ISO19139	ISO 19139 provides the XML implementation schema for ISO 19115
INSPIRE (Infrastructure for Spatial Information in Europe)	INSPIRE is based on the infrastructures for spatial information established and operated by the 27 Member States of the European Union. The Directive addresses 34 spatial data themes needed for environmental applications, with key components specified through technical implementing rules. This makes INSPIRE a unique example of a legislative "regional" approach [7].
OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).	The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP' [8].

5.2 Implementation of GNOS

From the statements given above some requirements of an appropriate hardware environment will be derived. Such an environment enables the implementation of the project objectives. GNOS is mainly built on XML/XSLT and offers excellent opportunities for a structured and platform independent programming. Fast processors, large storage devices as well as redundant data provision and back-up services are needed. The system can run on Windows or LINUX including the implementation of an Apache Web-Server, a MySQL database, Tomcat and Java.

Generally a fast multicore server is recommended for GeoNetwork. The running database application and the integrated map-server require sufficient RAM and processing power. LINUX or Windows can be used as operational systems, according to other basic conditions. The hard drive capacity should be calculated by the expected data volume. If only the metadata will be stored, no big capacities are needed. But typically the GNOS-Server is the central storage unit even for the big raw data. The

GNOS database can be built with Postgress, MySQL or with a GNOS integrated third party database, which is not so powerful. For different web 2.0 components JAVA and Tomcat needs to be installed on the webserver. As an alternative to the provided third party open source map-server it is possible to use the enterprise ESRI ArcIMS (Map-Server).

For system scaling and reliability of the service cloud computing- and virtualisation technologies can be used. Depending on the last mentioned parameters the following system requirements will be recommended:

- Fast (web)server with multicore processor;
- > 500 GB Hard drive;
- > 4 GB RAM;
- Operation System: Windows Server 2003/2008, Linux;
- Add on: MySQL, PHP, Tomcat, Java SDK;
- ArcIMS (GIS webserver) (optional);
- cloud-server (optional).

5.3 Functions

The core function of GeoNetwork is the metadata referenced search for local and distributed spatial data in a repository by keywords and time- and location filters (what, when where?). The raw data for the repertory can be up-/downloaded in different data types and formats (like maps, PDF, Excel, Word, JPG, etc.) by using of a web interface. An interactive map server provides a fast intuitive search for spatial data with the ability to use layer technology an own or third party (web) maps like Google Maps (see figure 2). Own online composed maps from GNOS can be downloaded as PDF directly.

Fig. 2. Search with keyword by location and time

The metadata model is implemented by a template machine with predefined meta-templates according to the international norms and standards. Templates can be new generated or customised to the own requirements. Internal and external thesauri catalogues provides sensitive search functions and suggestions for categorisation while the data input process (controlled vocabulary). A fine granulated user- and group management grants controlled access to each data set. The worldwide exchange of data and the synchronisation of the metadata on distributed (GeoNetwork) server architectures is realised by a standardised harvesting interface. Further core functions of GeoNetwork are given by up- and downloading of data and documents (for instance maps, PDF, Excel-Sheets), by an interactive map-viewer with own maps and layer technology, by map- and layer export procedures as PDF-files, by exchange and synchronization of metadata with standard harvesting technologies, by privacy and access control on each dataset, by an internal/external Thesauri catalogues for keywords, and by international (multilingual) translation procedures.

6 Long-term data management

Water quality management deals with diverse tasks as river basin management, nature conservation, and pollution control. All of these tasks are covered by experts from different scientific, engineering and social-economic disciplines. The resulting statements for water management are based on data sets with different origins and sampling intervals. The management of freshwater ecosystems requires long-term observations of water quantity and water quality indicators like water flow, DO, BOD, algal biomass, nutrients, water plants and others. In table 4 some groups of data for a sustainable water management are given which have to be combined by GNOS to establish a unified research and management information system.

Table 4. Data types of different origin for water management

Origin of data	Types of data
Hydrology	Morphology, water level, water flow
Ecology	Biodiversity, biomass, individuals
Land use	Land cover, patches, farming
Socio economy	Anthropogenic uses, tourism, industry
Energy	Energy production,
Administration	Planning, statistics
Politics	Environmental law, decision support

7 Conclusions

The development of optimal management strategies can only be achieved by using powerful informatic tools like a virtual research environments (VRE), that have a collaborative focus also. The most important point is that VREs need to be more considered as community building projects than as technology projects. To combine data sets from different project partners within a research project is often not only a difficult, but also a complicated task store, to handle and to analyse these data sets. Long-term storage and retrieval of such data is mostly impossible because of the inconsistent spatio-temporal structure of the data sets. The objectives of most research projects don't aim a data warehouse oriented design to evaluate the data sets or the collection of data in a single database and cause the usage of GeoNetwork Opensource. On the base of GNOS an interoperable metadata database will be implemented to aggregate, store, and share datasets from the different projects. In addition a high flexible content management system like Joomla, Dupral or Typo3 or collaboration tool like BSCW is needed to include several enhancements and a single sign in system like LDAP and Shibboleth to build a monolithic VRE in one piece. The exchange of data with harvesting technologies becomes very popular in the last years. Each VRE might have a standard interface for harvesting like the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). The main challenges and problems for the development and success are not only from technical nature but also effected by anthropogenic factors: space, time, funding, isolation, procrastination, poor motivation, trust, commitment, working style, ownership, data access, difficulty of learning software and technology, lack of appropriate skills and ready access to technical support and extensive training needs, rapid advantage of technology. Research activities in the future can be established in the socioeconomic field, in visualisation of data and structures and network structures with harvesting strategies. New useful fields for VREs from the IT-branch are also cloud computing and virtualisation techniques and exploring the semantic web technologies with ontologies.

8 References

1. Interministerielle Ausschuss für Geoinformationswesen (IMAGI): Geoinformation und moderner Staat. http://www.imagi.de (accessed 10.02.2011).
2. Deutsche Forschungsgemeinschaft (DFG): Sicherung guter wissenschaftlicher Praxis. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf (accessed 10.02.2011).
3. GeoNetwork Opensource: Homepage - GeoNetwork Opensource http://geonetwork-opensource.org (accessed 10.02.2011).
4. Rueppel, U., Gutzke, T., Petersen, M., Seewald, G.: An internet-based spatial decision Support system for Environmerntal data. In: CERN, Sharing. Proc. of EnviroInfo 2004, pp. 331-338, Geneva, Switzerland, (2004).
5. Carusi, A., Reimer, T.: Virtual Research Environment Collaborative Landscape Study; JISC the UK's Joint Information Systems Committee. www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf (accessed 13.04.2011).

6. Joomla: Homepage. www.joomla.org. (accessed 13.04.2011).
7. European Commission INSPIRE: Homepage, Inspire Directive on start-up page http://inspire.jrc.ec.europa.eu/ (accessed 13.04.2011).
8. Open Archives Initiative: Interoperability through Metadata Exchange. http://www.openarchives.org/pmh/ (accessed 13.04.2011).