Building a regional clearinghouse for spatial information – views and experiences based on a research project in SW Finland

Harri Tolvanen
Department of Geography, University of Turku

Abstract. The need to improve digital information availability and sharing is currently discussed widely in the geographical information community. Digital data management involves several problems, which are not yet solved in comprehensive and general manner. A research and development project, which approaches the data management issue from the practical perspective, has been launched in south-western Finland. The project aims to establish a functional archive, which gathers and delivers geographical information. The purpose of the service is to offer data owners a place to store data which they are willing to share with others. Also people who need the data are provided with one common source of information. This paper presents experiences of the archive implementation, and some theoretical aspects of the data sharing and networking.

Keywords: spatial data, geographical information, clearinghouse, information archive, data management

Introduction

In this paper, an archive development project with planning and implementation experiences of an Internet-based data storage and delivery mechanism, a spatial data archive, is described. The project is based on a regional cooperation initiative, which aims to improve the overall efficiency of spatial data use in south-western Finland. The cooperation initiative was founded by the University of Turku, South-Western Finland's Regional Environment Centre and the Regional Council of South-West Finland. The archive research project is carried out by the Department of Geography at the University of Turku, and funded by Maj and Tor Nessling Foundation.

The goal of the archive project is to build a service, which enhances the visibility of existing data, improves the mechanisms of data transfer between organisations, and is able to ensure that valuable data will remain available after primary use and termination of research projects. The task harbours several problems, most of which are related to juridical and economical aspects in the digital realm. While the goal of the project is to produce mechanisms for spatial data management and administration, there is a thematic orientation in the first phase of the development: the pilot archive hosts digital data sets which consider environment and biodiversity issues. Environmental

theme was chosen, because it is practically always spatial data, and there is a strong international trend towards creating a global network of data archives, *clearinghouses*.

The demand for environmental information services is existing and growing. There are several different types of questions and needs, which need to be answered by creating appropriate information systems. This paper aims to provide experiences on building an archive system for environmental information storage and delivery in a regional level.

Data management perspectives

The digital spatial datasets can be classified into two major categories: data that are produced and marketed commercially, and data that are gathered by researchers and research institutions for scientific or environmental management purposes. The first category comprises of national cartographical agencies' base maps and other commercially available spatial data, and is not in the scope of a clearinghouse mechanism. However, discussions about pricing and overall availability of these data sets is currently active, and the trend seems to point towards more open data policy, especially within the public sector.

The second category is very much in interest of a clearinghouse. Biodiversity and environmental data, whether originally gathered for basic environmental study or routine management survey, is an asset in creating a better understanding of the environment, and thereby for improving conservation and resource management. The worst failure of the information sharing concept would be a case, where available data is left unused only because there is no awareness of its existence. The cases where the data producer does not allow use by others must be accepted as such. However, it is assumed, that majority of scientists would rather see their abandoned data being used by someone else, than hiding it indefinitely without any purpose to use it further themselves.

The general development of the information systems, in this case in the field of biodiversity and environmental research, has led to different kinds of solutions, which are obviously meeting different kinds of needs. Some information systems are smart databases, computer-assisted identification tools or document retrieving systems, which serve defined content-oriented purposes of information management (Schalk 1998). In this case the project operates in the lowermost levels of the information hierarchy presented by Laihonen et al. (2003), namely the data and information levels. Thus the system is not aiming to offer knowledge out of the data at hand, but to increase the opportunities to find and combine data to achieve deeper understanding of the phenomena. Bearing in mind the request for seamless combination of multi-source spatial data in European context (INSPIRE 2002), the archive project can be seen as one of the first steps on the way. There are several projects which are meeting the technical challenges for multi-source spatial data integration in international scale, such as GIMODIG (*http://gimodig.fgi.fi/*).

Although rising from the needs of grassroot level actors in local environmental research and planning scene, the data archive is connected also to the global context. Local initiatives for information sharing in information networks are crucial elements in the convention on biological diversity (Rio) and the CHM (clearinghouse mecha-

nism) initiative (UNEP 1997). The network structure promotes the cumulative combination from local to national, and eventually to global dimension (fig. 1). There are several projects implementing the biodiversity information network in different levels, such as the global GBIF (Global Biodiversity Information Facility *http://www.gbif.org/*)(see Edwards et al. 2000) and Australia's ERIN (Environmental Resources Information Network *http://www.ea.gov.au/sdd/erin/*)(see Bisby 2000). Global Spatial Data Infrastructure initiative presents a variety of theoretical and technical perspectives to spatial data clearinghouses (GSDI 2002).

Project background

The archive development work originates from an ongoing cooperation initiative of a regional environmental authority, a regional council responsible for land use planning, and a university. Regional network partners include universities, cities, enterprises and governmental and non-governmental organisations in south-western Finland. The project covers all sides of geospatial information handling; remotely sensed information and vector data, applied use of data in land-use planning, as well as data production from field measurements into a GIS (Geographical Information System). A common will to enhance data sharing has been the motivation for the project.

Constant communication among the participants of the network guarantees awareness of the current situations and projects in the area, and discussion about ideas and suggestions within the community. The possibility to meet other people in the regional community in frequent seminars and meetings is necessary for successful cooperation atmosphere. The network, consisting of professionals in many fields, is also an important source of ideas and

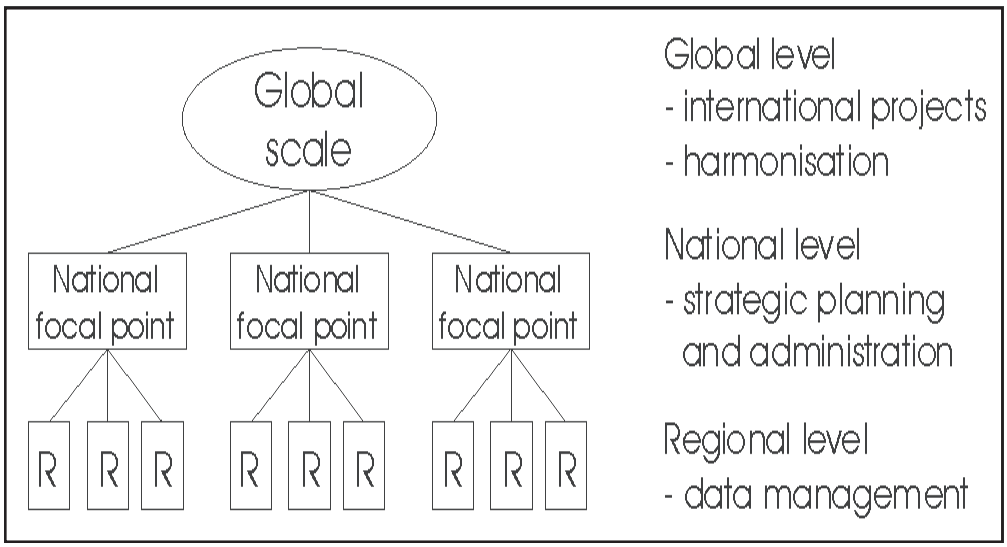

Figure 1. The networking hierarchy levels and project trends.

knowledge in the development process.

Although purposes of coastal zone, river basin, urban area and rural area management often define the geographical area in regional integration measures (INSPIRE 2002), in this case the area is defined by the founding partners' administrative area. The area is quite diverse in landscape properties – it ranges from the open Baltic Sea areas through a mosaic of sea and islands, to agricultural landscape on SW-Finnish mainland. The area has a quite wide selection of environmental information held by different organisations, and therefore provides a good ground for data management development.

The archive deals with the lower level units of the network hierarchy: regional and even local. It is generally accepted, that data should be stored and maintained on the administrative level where it is most feasible (INSPIRE 2002). In practice, this is close to the level, where the data is originally gathered or constructed. Even though ideas and plans to establish focal points on each level of the hierarchy are well thought out, the implementation especially in local level is somewhat slow and troublesome, because this level is actually facing the concrete problems with the concrete data sets.

The archive is handling the data sets as units, and does not operate with the data set contents. Thus, the aim is not to create an intelligent expert system, which would answer questions regarding biodiversity, environment or any other phenomenon. Although the archive is not offering any content services, it can be seen as a social and administrative facility, more than a technical or content-driven service.

Structure of the archive

Actors, interactions and contracts

The data archive operations can be presented in three phases: data acquisition, archiving and delivery (fig. 2). Acquisition includes negotiations with data producers, metadata creation and contractual agreements for data transfer. Archiving consists of metadatabase maintenance, web service maintenance and service unit operations (customer service). User contacts and contracts for data use are included in the delivery phase.

Data acquisition requires two documents: a contract to transfer the data to the archive, and a metadata form. The contract is the actual document enforcing the transaction, while the metadata form describes the data ownership and technical issues. By assumption the ownership and copyright of the material remains with the author, only the right to use the data set for a specified and separately defined purpose is granted. The data producer can restrict the use of the data by setting terms of use in the contract.

As the archive receives a data set, it is validated through test use and metadata evaluation. If the metadata is sufficient, it is recorded to the database. The database itself is optimised for data search through an Internet user interface. The user interface is a query form, which hosts the search terms as drop-down menus or free text fields. One service of the user interface is a map server, which is able to present the datasets on a map background for browsing and examining. Finally, the delivery is made by a contract, which defines the terms of data use. These terms include the general terms, and the specific terms set by

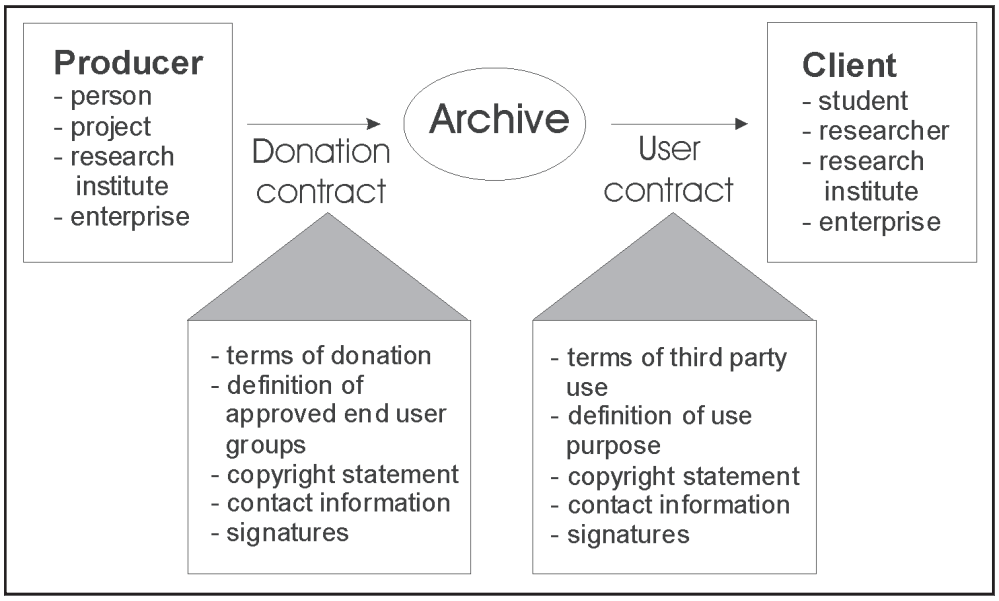

Figure 2. Data set flow and contractual phases in the archive system.

the data producer.

Metadata and database

The decision upon metadata format was taken after reviewing many spatial data standards and standard proofs. At this point, it seemed feasible to use a format designed for this purpose. Is must be stressed, that even though national or international standards were not followed as such, the general format used can be easily extended to include more detailed information and eventually meet the standard. The approach of selected general information on the metadata was selected mainly because the standardised metadata formats are quite heavy to implement on such low-intensity project and diverse group of data producers. The possibility of having a long list of unfilled fields in the metadata is evident and unwanted.

The database includes metadata not only for the archived material, but also for the data that are available directly from the producers. This proved necessary to fulfil the goal to produce a service, which would include as many environmental data sets from the region as possible. It is anticipated, that in the implementation phase it is important to gather a critical mass of material to the searchable metadatabase to provide a better service for the environmental professionals in the region. Once the service is established, the aim is to obtain more datasets into the archive through negotiations and contracts, which operations are time-consuming.

The database for metadata entries was designed to host the information gathered with the metadata form, and optimised to serve the searches by pre-arranged thematic classification, author name, index word or year of publication. The aim was to make the database simple enough to enable effi-

cient operation regardless of the amount of content. On the other hand, the database should be flexible to further development: as the tasks may grow more complex in the future, the database should not need a complete redesign, but additional components and functionality.

Archive operation

The pilot service launched in the Internet provides a flow of search (fig. 3), beginning with the search criteria. Once a query to the database is completed, the system returns a list of matching documents to the user. From the list, the user can browse the full metadata description, sample map, and web map browser (if available for the data set in question) for the particular data set entry. After the user has reviewed the metadata and found the material interesting, he is given the contact information to proceed to a material request – either with the archive, or the producer, in the case that the material is not stored in the archive.

The spatial coverage of a given data entry is at the moment only a text field property in the metadatabase, but not a search term. Having spatial search possibility would mean a major geocoding effort to create the information for the database. However, a map interface was created to visually present a selection of the data. At the moment there are no resources to make major conversions to all the incoming material, which means that the map service includes only datasets that are already in a suitable format and coordinate system as

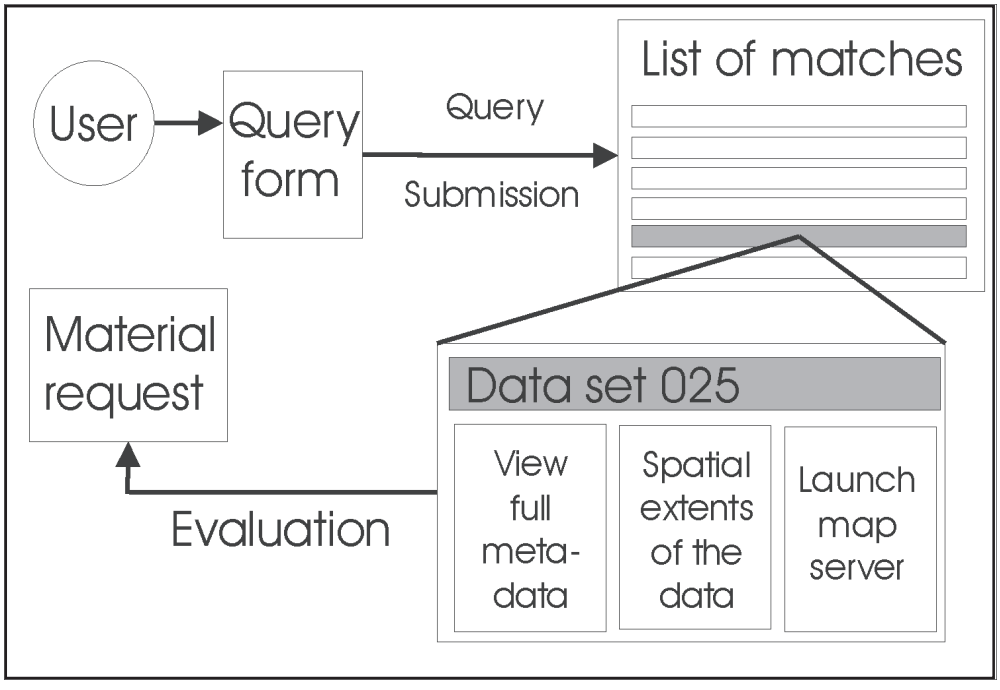

Figure 3. The search process in the user interface.

they are deposited.

As the map server does not comprehensively cover the database material, it can not be used as a search method, but only to browse some data sets visually on a map base. As soon as every data set can be included in the web-based map server, the location can be used as a search term, for example by giving the user a possibility to delineate an area of interest on a map.

Discussion

The approach taken in the archive development is predominantly contributing to the needs of regional level actors. However, as the regional data archive is interlinked to the national focal point, and thereby to the global scene, it is anticipated that the archive model could at least partly be enforced in a larger context. All projects build their own databases to fulfil their specific needs, thus there is a need for an agreed exchange format (Allkin 1998). Even though several technical solutions, such as GML (Geographical Markup Language), are being developed to improve data interoperability, there is no possibility to implement these methods immediately in a small-scale archiving project.

Some general problems, which are the same in every clearinghouse, arose also in this project. Particularly the question of multiple data formats was faced from the beginning. In this case, since at this point there is no intention to create on-line usability, the problem was solved with an "as is" answer: the archive accepts files in the format that the producer chooses, and delivers them unchanged. The archive is only a step on the data set's way from the producer to the user, and there is no interest to cause extra work to the producer. On the other hand, there is no point in harmonizing all data to one format, because there are as many user systems as there are producer systems, and the conversion would benefit only a small group of users. The same principle applies to coordinate systems.

The development of the archive will continue in the future. Main targets are to improve the Internet map service, to create more advanced search facility and increase the amount of material stored in the database. Through the pilot service test use, the correctness of the contracts is evaluated, and user feedback is collected. In general, the target is to achieve a status as a trusted, well organised and smoothly operating service, which benefits the data producers as well as the data users, thus improving the quality of environmental research and management in south-western Finland.

It is emphasized, that building a social network among regional key actors in the field of interest, in this case environmental research and geoinformatics, is important. This is especially true in the sense that data ownership issues are very sensitive in nature: nobody wants to be part of something that appears from nowhere. The transparency and possibility to participate in the whole process, from planning to implementation and finally operation, gives the people involved a sense of security and trust.

The biodiversity CHM initiative (UNEP 1997) is one example of thematically oriented global network. It can be seen as a horizontal connector between regional databases, but with a limited thematic content. The structure can be built also the other way around: regional databases with no thematic limitations. This way the region would be the main category of interest, instead of, in this case, environmental information. Inter-

linking these regional databases would lead into massive structures, but in regional use it may be more efficient in some thematic branches, for example urban area planning or nature conservation projects.

In the future, it is expected that digital information archives become a feasible way to store and organise the vast amount of information gathered in different branches of the society. The new technologies, especially mobile networks enable more efficient services based on spatial data, and there will be an increasing demand for spatially referenced environmental and biodiversity information. Whether this demand is directed to raw data material or knowledge derived from it, remains unknown. However, the original digital data sets, which are in scope of this archive project, form the basis for the future services. Therefore it is important to secure the data gathered to date, and manage the material in a way it may be integrated to new developing information systems.

References

Allkin, R. (1998). Effective Management and Delivery of Biodiversity Information. *In* Bridge, P., P. Jeffries, D. Morse & P. Scott (eds.): *Information Technology, Plant Pathology and Biodiversity*, 87-102. Cab International, Wallingford.

Bisby, F. (2000). The Quiet Revolution: Biodiversity Informatics and the Internet. *Science* 289: 2309-2312.

Edwards, J., M. Lane & E. Nielsen (2000). Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science* 289: 2312-2314.

GSDI (2002). Developing spatial data infrastructures. Version 1.1. Global Spatial Data Infrastructure, Technical Working Group.

INSPIRE (2002). Environmental thematic user needs – Position Paper, version 2. INSPIRE Environmental Thematic Coordination Group. European Environmental Agency.

Laihonen, P., M. Rönkä, H. Tolvanen & R. Kalliola (2003). Geospatially structured biodiversity information as a component of a regional biodiversity clearinghouse. *Biodiversity and Conservation* 12:1, 103-120.

Schalk, P. (1998). Archiving Biodiversity: Information Technology Applied to Biodiversity Information Management. *In* Bridge, P., P. Jeffries, D. Morse and P. Scott (eds.): *Information Technology, Plant Pathology and Biodiversity*, 213-220. Cab International, Wallingford.

Tolvanen, H. & R. Kalliola (2002). Digitaalisten tietoaineistojen arvo ja saatavuus. *Terra* 114: 4, 253-256.

UNEP (1997). Introduction to the clearing-house mechanism of the Convention on Biological Diversity to facilitate and promote technical and scientific co-operation.