Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

IMPLEMENTATION OF A REGIONAL CLEARINGHOUSE FOR SPATIAL INFORMATION

Harri Tolvanen

Department of Geography, 20014 University of Turku, Finland

ABSTRACT

There is a growing amount of spatial data, which is gathered for research and education purposes, and abandoned after the first use. In contrast to publicly managed and distributed information, these data collections do not have a mechanism for storage and delivery. The data are also thematically very diverse and come from different kinds of sources. However, most of the information could be useful for others, but the potential users have no information about these existing data sets. In worst case, the same data is gathered again. This paper describes a regional data archive and discusses the implementation of the service. The most crucial issues in the archive development are legally correct contracts in data transfer, well functioning metadata search, and open policy towards data holders and users. The theme of this project is spatial environmental information, but the archive principles apply to a wider thematic scope in digital data management.

1. INTRODUCTION

A development project with aspects of an Internet-based data storage and delivery mechanism, a spatial data clearinghouse, is described in this paper. The goal of the project is to build an archive service, which enhances the visibility of existing data, provides uniform information (metadata) about the filed data sets, improves the mechanisms of data transfer, and is able to ensure that valuable data sets will remain available after primary use and termination of research projects. The work is based on a regional cooperation initiative in the field of geoinformatics. In this context, “regional” refers to an administrative level between national and local scales.

The project described in this paper has also a thematic content: environment and biodiversity. However, the results of the study apply to all kinds of digital spatial data sets. The specific theme was chosen for the pilot study, because environmental information, and in this case particularly biodiversity information is tied to a location, and thus offers suitable substance for spatial data archive development. In addition, the environmental sector is one of the first and largest users of geographical information applications.

The archive operates on the lowest levels of the spatial network hierarchy: regional and local. It is generally agreed that data should be stored and maintained on the administrative level where it is most feasible [1]. In practice, this is close to the level where the data is originally gathered or constructed. In many cases the data production is decentralized and carried out on the regional level. This means that the data products are mostly designed to serve local needs [2]. Data combination becomes thus difficult due to different standards and practices. Even though ideas and plans to establish a network connected by focal points on each level of the hierarchy are well thought out, the implementation is somewhat slow and troublesome, especially on local level, because this level is actually facing the concrete problems with the concrete data sets.

Although rising from the needs of grass-root level actors in local environmental research and planning scene, the archive concept is connected also to the global context. Local initiatives for information sharing in information networks are crucial elements of the Convention on Biological Diversity and the CHM (clearinghouse mechanism) initiative [3]. The network structure promotes the cumulative combination from local to national, and eventually to global dimension. There are several projects implementing the biodiversity information network on different levels, such as the global GBIF [4 & 5] and ERIN [6 & 7] in Australia. Global Spatial Data Infrastructure initiative presents a variety of theoretical and technical perspectives to spatial data clearinghouses [8].

This paper aims to provide the science community with experiences on building a regional archive system for information storage and delivery. The task harbours several problems, most of which are related to psychological, juridical and economical aspects of data set transfer in the digital realm. The demand for environmental information services is growing, and a solution to a specific case is presented.

2. DATA MANAGEMENT AND PROJECT OUTLINE

2.1 Data types and information levels

The digital datasets concerning environmental issues 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, geological survey information, statistical records and other commercial material, and is not in the scope of a clearinghouse mechanism. However, discussions about pricing and overall availability of these data sets is currently active in Finland, and the trend seems to point towards more open data policy, especially within the public sector.

The second category, on the other hand, is very much in the interest of a clearinghouse. Biodiversity and environmental data, whether originally gathered for basic environmental study or routine 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 to improve the state of the environment, than hiding the data unused.

The 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 [9]. In this case, the archive operates at the lowermost levels of the information hierarchy [10], namely the data and information levels. Thus the archive system is not aiming to offer knowledge or wisdom out of the data, but increase the opportunities to find and combine raw data to achieve a deeper understanding of the phenomena. It is anticipated that the open archive interface enables new and innovative information combinations as people realise how much and how diverse information have been produced.

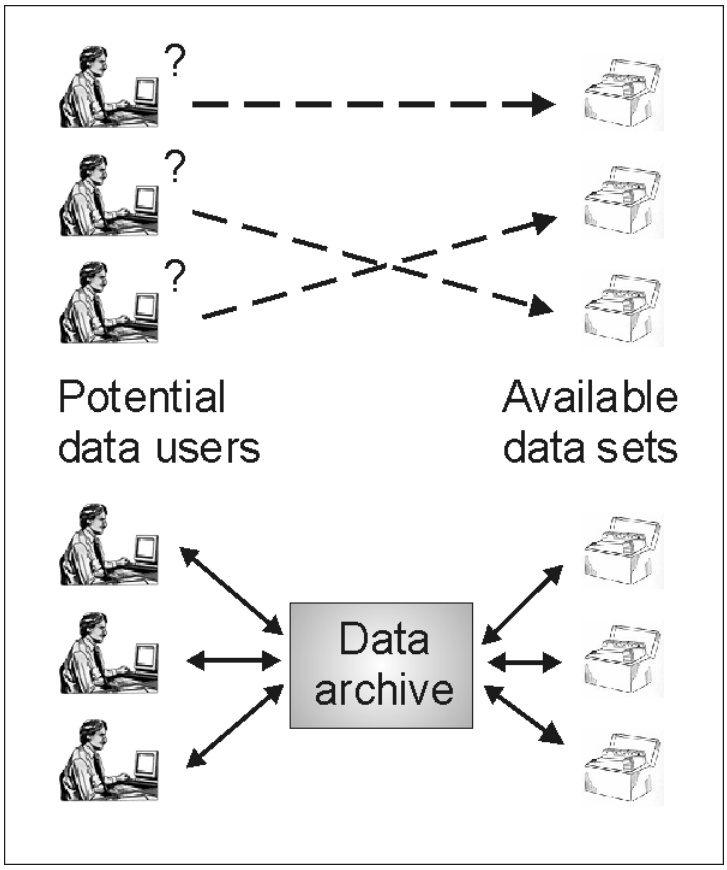

Fig. 1 In the current situation knowledge of existing data is poor (above). The archive aims to offer one common contact surface for data users and data holders (below).

2.2 The project outline

The SW-Finnish environmental information archive development is a part of an ongoing GI cooperation initiative [11] of a regional environmental authority, a regional council responsible for land use planning, and a university in Southwest Finland. Regional network partners include other universities, municipalities, enterprises and governmental and non-governmental organisations. The initiative was launched in late 1999, and it covers all sides of spatial 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 GIS (Geographical Information System). The project has launched a map-based bibliographic information service demonstration on the Internet [12].

Although the geographical area in regional integration measures is often defined by purposes of coastal zone, river basin, urban area and rural area management [1], in this case the area is defined by the founding partners' administrative boundaries. The area is quite diverse in landscape properties – it ranges from the open Baltic Sea, through an archipelago, to agricultural and forested landscape on SW-Finnish mainland. The area has quite a good register of environmental information already produced and held by different organisations, and therefore provides a good ground for data management development.

A common will to enhance data sharing has been the motivation for the cooperation initiative. Therefore, it is emphasized that a local level clearinghouse requires local level connections and partnerships between the actors in the field of operation, in this case the

environmental scene. Existing connections and a history of bilateral data exchange between regional organizations form a good building ground for regional cooperation. Constant communication and networking guarantees awareness of the current situation and projects in the area, as well as discussion about ideas and suggestions within the community. The possibility to meet other people in the regional community in frequent seminars and meetings promotes successful cooperation atmosphere among professionals in the field.

3. THE ARCHIVE STRUCTURE AND PROCESS

3.1 Data flow

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

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 remain with the author, only the right to use the data set for a specified purpose is granted. The data producer can restrict the use of the data by defining terms of use in the contract.

Upon data acquisition, the data set 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 the data producer. The user states the purpose of data use in the contract, and accepts the responsibility to use the data properly according to copyright legislation. This includes a ban of copying and delivering the material further.

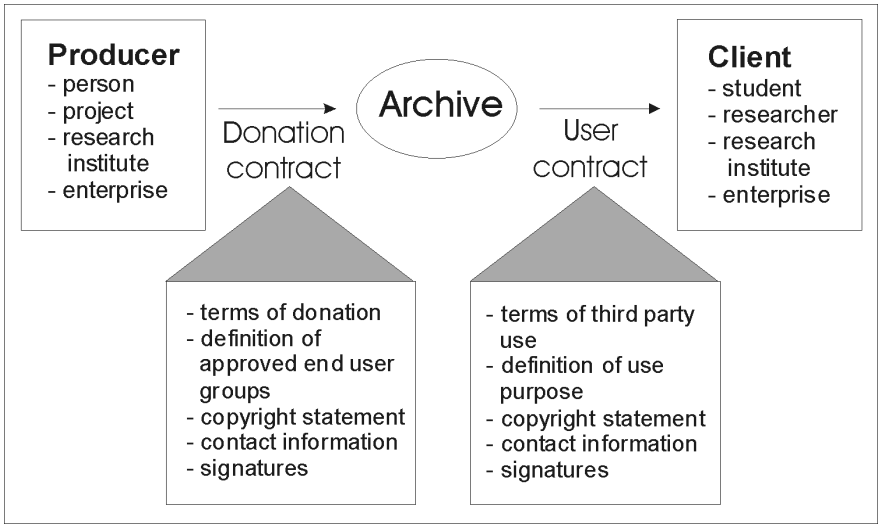

Fig. 2 Data set flow and contractual phases in the archive system.

3.2 Archive services

The database includes metadata not only for the archived material, but also for data that are available directly from the producers. This is necessary to fulfil the goal to produce a service, which would include as many environmental data sets from the region as possible and thus serve the local users better. It is anticipated that in the implementation phase it is important to gather a critical mass of material to the searchable metadatabase. Once the service is established, the aim is to obtain more datasets into the archive through negotiations and contracts, which is 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 a thematic classification, author name, index word or year of publication. The aim was to make the database simple enough to enable efficient operation regardless of the amount of content. On the other hand, the database should be flexible for further development: as the tasks may grow more complex in the future, the database should not need a complete redesign, but additional components and increasing functionality.

The spatial coverage of a given data entry is only a text field property in the metadatabase, but not a search term. At the moment the project does not have resources to create a system which could derive the coordinate information directly from the data set itself, since the input format is undefined and there is also plain metadata in the database. As the map server does not cover the metadatabase material comprehensively, it can not be used as a search method, but only to browse a selection of data sets visually on a map. Having spatial search possibility in the metadatabase would mean a major geocoding effort to create the spatial information for the database entries.

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, view a sample map, and open map browser (if available for the data set in question) for the particular data set entry. After the metadata review, the user can find the contact information of the archive to proceed to a material request.

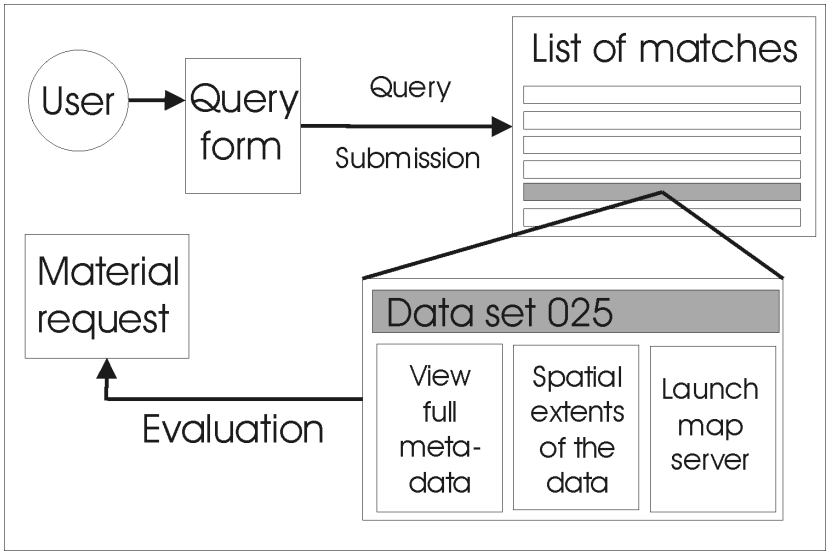

Fig. 3 The data search flow in the user interface.

4. FROM IMPLEMENTATION TO INTEGRATION

4.1 General settings

First, it is emphasized that it is important to build a social network among regional key actors in the field of interest, in this case environmental research and geoinformatics, because the data issues are very sensitive by nature. Previous familiarity with the process and the people behind it, as well as the transparency of the whole structure from planning to implementation and finally operation, give the people involved a sense of security and trust. These are important, because the research data archiving is based on voluntary data donations.

Some general problems, which are the same in every clearinghouse, arose 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 principle: 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 path from the producer to the user, and there is no point in harmonizing all data to one format. This is simply because there are as many user systems as there are producer systems, and the conversion would benefit only a small group of users, if any. The same principle applies to coordinate systems. This may be a subject to change in the near future, as the software packages enable more versatile use of different formats, and possibly one common exchange format could be introduced.

The basic problem in the field is the lack of knowledge. While many different individuals or organisations produce spatial data for their work, it is impossible to be aware of all existing data. Holders of data, in this case environmental data, can be basic monitoring operatives, academic researchers, NGOs or businesses, who do not generally share their data, or even information about their data with others. The archive aims to create a common marketplace for the data: there would be a place to search for metadata, the information about information (fig. 1), and even obtain data sets quickly and easily.

The data sets in environment and biodiversity field are diverse and also potentially beneficial in several ways. It is not only the academic community that requires environmental information for research, but also regional land use planners, environmental administration and environmental education need accurate and up-to-date information for their individual tasks. Location is seen as a key factor in the future communication technology, as GPS and other positioning devices become connected to mobile Internet browsers. Environmental information has a strong potential to become a significant content group in the new location based services. The archive does not offer technical solutions to mobile services, but may act as a source of data.

4.2 Main challenges in the archive development

One of the most crucial questions that have emerged during the establishment of the pilot service has been the definition of data ownership and copyrights. During the project a diverse group of people and organisations, which produce digital spatial data, were involved. Whether produced by governmental organisations or undergraduate students, each data set has an individual history. The most complex ownership structures may involve a project, which is managed by an organisation, has funding from an outside source, and cooperates with other organisations and projects. When people working in a project like this, maybe even on a personal research grant, produce digital data sets, it is very difficult to determine who owns the copyrights and who can actually donate the material further. When such data sets are combined with each other, the issue gets even more complex.

These questions make writing of the archive contracts difficult. On one hand, the contracts should be of general nature, readily applicable to any kind of data set and any kind of producer background. On the other hand, each case presents unique settings, which require unique contract paragraphs. This contradiction can, however, be overcome by

careful planning of the contract forms, including the general fields, but leaving enough room for individual contract terms.

One of the most frequently asked questions concerns the data producers' willingness to donate their data to common use. The motivation of data holders is a very important task in the process, and the experiences so far are positive. The people who were contacted in the pilot phase were very interested in the idea and ready to give out the metadata of their data sets. When asked for data donation, some hesitation occurred. This was mainly due to the fact, that the people were not familiar with the archive. In the vast majority of cases, people thought it is reasonable to give out their data after their own targets have been accomplished.

To function properly, the interface of the archive system should be designed for easy access and comprehensive services to satisfy the needs of data producers and data users. A twofold approach to the data search was taken: there is a form based database query and a map browser interface. These two will be merged as soon as the resources allow.

4.3 The metadata question

The decision upon metadata format was taken after reviewing many spatial data standards and standard proofs. The situation between the global ISO and European CEN standards seems unfinished, although some organisations and countries in Europe see that the ISO/TC 211 would be the way to go [13]. In this case, it seemed feasible to use a format designed for this specific purpose. The archive metadatabase consists of general information fields only. This proved to be sufficient at this point, since all metadata forms received were filled comprehensively and correctly. The adoption of a standard metadata format is not urgent, but it may appear necessary in the future. However, the system is ready to be modified into that.

It must be stressed, that even though no national or international standards were followed as such, the general format used is easily extended to include more detailed information and eventually meet the standard. The approach of selected general information on the metadata was chosen because the standardised metadata formats are quite heavy to implement on such low-intensity regional-scale project and diverse group of data producers. The possibility of having a long list of unfilled fields in the metadata is evident and unwanted.

In the future, it could be possible to manage the metadata in a more effective way. There could be a centralised catalogue, which provides a common interface for each producer's metadata registers [13]. The point is, however, that the metadata is produced and managed by the individual data producers. This requires a well harmonised metadata template. The need to create an interface simple enough for metadata collection is recognised [13], as different data producers have different kinds of practices in their own data description. As the exchange formats, metadata standards and common interface languages evolve, the archive will most likely develop in this direction.

4.4 Expanding the archive

The technological challenges in data interoperability are diminishing, and the harmonization of contents becomes emphasized. In this case, the archive is handling the data sets as units, but it does not manage 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, but to create an effective way to disseminate the data as such. 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.

The approach taken in the project is predominantly contributing to the needs of regional level actors. However, as the regional data archive can be linked 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. Bearing in mind the request for seamless combination of multi-source spatial data in European context [1], the archive project can be seen as one of the first steps on the way. There are several top-down projects, which are

meeting the technical challenges for multi-source spatial data integration in international scale, such as GiMoDig [14]. While several technical solutions are being developed to improve data interoperability, there is no possibility to implement these methods immediately in a regional small-scale archiving project.

The biodiversity CHM initiative [3] is one example of a 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 sector. This way the region would be the main category of interest, instead of, in this case, biodiversity information. Interlinking 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 [15]. Whether this demand is directed to raw data 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 data so that it may be integrated to new information systems, which emerge in the future.

5. ACKNOWLEDGEMENTS

The author wishes to thank project leader professor Risto Kalliola and coordinator Lasse Nurmi for their cooperation during the project, Mia Rönkä and Pasi Laihonen for comments on the manuscript, and Marja Vieno for language revision. The study is financed by Maj and Tor Nessling Foundation and the Finnish Geography Graduate School.

6. REFERENCES

[1] INSPIRE, *Environmental thematic user needs – Position Paper*, version 2. INSPIRE Environmental Thematic Coordination Group. European Environmental Agency. 2002.

[2] Riecken, J., «The improvement of the access to public geospatial data of cadastral and surveying and mapping as a part of the development od a NSDI in Northrhine-Westfalia, Germany», *Proceedings of the 4th AGILE Conference*, 215-221, Brno, 2001.

[3] UNEP, *Introduction to the clearing-house mechanism of the Convention on Biological Diversity to facilitate and promote technical and scientific co-operation*, UNEP/CBD/CHM/RW/3/2, 1997.

[4] GBIF, *Global Biodiversity Information Facility,* http://www.gbif.org/, 20.2.2003.

[5] Edwards, J.L., Lane, M.A. & Nielsen, E.S., «Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop», *Science*, vol. 289, 2312-2314, 2000.

[6] ERIN, *Environmental Resources Information Network*, *http://www.ea.gov.au/sdd/erin/*, 20.2.2003.

[7] Bisby, F.A., «The Quiet Revolution: Biodiversity Informatics and the Internet», *Science,* vol. 289, 2309-2312, 2000.

[8] GSDI, *Developing spatial data infrastructures. Global Spatial Data Infrastructure*. Technical Working Group, version 1.1., 2002.

[9] Schalk, P.H., «Archiving Biodiversity: Information Technology Applied to Biodiversity Information Management», *in*: Bridge, P., Jeffries, P., Morse, D.R. & Scott, P.R., (eds.), *Information Technology, Plant Pathology and Biodiversity*, 213-220, Cab International, Wallingford, 1998.

[10] Laihonen, P., Rönkä, M., Tolvanen, H. & Kalliola, R., «Geospatially structured biodiversity information as a component of a regional biodiversity clearinghouse», *Biodiversity and Conservation,* vol. 12, 103-120, 2003.

[11] Regional Council of Southwest Finland, *Lounaispaikka, a regional GI service,* http://www.varsinais-suomi.fi/lounaispaikka/, 20.2.2003.

[12] Tolvanen, H., «Lounais-Suomen paikkatietoyhteistyön kehittämishanke – 2. vaiheen raportti», *UTU-LCC Publications 2,* 39 p., 2001.

[13] Gouveia, C., Henriques, P., Nicolau, R., Rocha, J. and Santos, M., «Moving from CEN TC 287 to ISO/TC 211 – The approach of the Portuguese National Geographic Information Infrastructure». *Proceedings of the 4th AGILE Conference*, 260-269, Brno, 2001.

[14] Finnish Geodetic Institute, *GiMoDig - Geospatial info-mobility service by real-time data-integration and generalisation*, http://gimodig.fgi.fi/, 20.2.2003.

[15] Burnett, C. & Kalliola, R., «Maps in the information society», *Fennia*, vol. 178, 81-96, 2000.