System architecture and the catalogue of data sources

The article describes the system architecture concept applied in the REMSSBOT project, focusing especially on the use of the Catalogue of Data Sources (CDS). The design approach is explained, taking into account the existing situation and the general requirements for environmental information in the REMSSBOT participating regions. The REMSSBOT meta-information system is presented, together with a description of the EEA ETC/CDS and Thesaurus, on which the REMSSBOT CDS datamodel was based. The functionality of the CDS is also shown, as well as the search and navigation approach. Last, the standards and components used for the implementation of the system architecture are given, together with a short description of the implementation, design and process itself.

The design of the system architecture of the REMSSBOT-system was a general task within the REMSSBOT project: to design one system architecture for three computer systems with different front- and back-ends in the Attica, Piemonte and the Scheldt region.

**Design approach**

The main objective is sharing environmental information by keeping the data at their original location and by connecting the local database systems in each region to the REMSSBOT-system. We did consider building a centralised data warehouse as a solution. In that last case, copies have to be made of the original data and these copied data have to be stored in a new database system. A new centralised database system has to be designed and built, all available information needs to be copied and last, but not least, the contents of the new database system have to be maintained. If the original data are changed, the copied data have to be changed in the same way. Within the REMSSBOT project we have chosen the other solution: keeping the data at their original location. The requirements for such a system will be specified below.

**General requirements**

In the three participating regions the REMSSBOT-system will be used to exchange information concerning different environmental topics, such as solid waste, water pollution, air pollution, risk management and ecological management of a river basin. The design of the system architecture had to be useful for all of these different topics; therefore the design had to be very general or, in other words, topic independent.

The technical environment and the communication infrastructure in the three regions are not equal. In the existing environments we find all kinds of different computer platforms, the information can be stored in every possible way, using different kind of file systems and databases. The data communication networks vary from dial-up lines to state-of-the-art data communication networks. In this diversity we have to design a system workable to exchange the available information.

The range of users of the system can be very wide, for example: policy makers, environmental managers, professional users, data-providers, technical specialists, related administrations, universities, general public, etc. In general, we cannot expect this wide range of users to have special knowledge of data communication. Therefore, the data communication part of the new system has to be transparent for the users.

Additionally, in this research and development project the project partners prefer to use new technologies and tools to realise the system, to contribute making innovations in the three regions and to build up experiences and knowledge.

The existing and prospective situation

Currently, the prospective REMSSBOT users (environmental information consumers) generally have direct access to the information of their own organisation only. In many cases they do not know which information, relevant for their work, is available in other organisations and if they do know which information is available, they cannot access that information directly.

Data-providers have comparable problems. They provide information to consumers in their own organisation. Outside their own organisation they do not usually know the prospective consumers. If they do know the external consumers and if they provide information only once, they do not know when they have to provide the updates of the information.

Today, more and more organisations need to supply the public with current information. In the prospective situation the users have direct access to all the available information stored in the databases connected by the REMSSBOT-system (Figures 1 and 2).

The user can be an employee of one of the information providing organisations or of a related administration, but he or she can also be an external person, for example an interested Internet-user.

The user wants to be able to identify what information is available and wants to make a selection of the available information. After that he or she wants to access and retrieve the selected information without the need for all kinds of special technical knowledge to carry out these processes. Finally the user wants to know where the information comes from.

In one selection, information can be provided by one or more data-providers.
This prospective situation defines a lot of technical functions: The system needs a user interface (a client programme or a Web-interface), a meta-information system (a system to store which information is available), search functions to select the meta-information, communication functions (to achieve a connection to one or more data-providers, transparently for the user), access and retrieve functions and functions to present the selected information.

### Meta-information systems

Within the REMSSBOT project we have analysed two main groups of meta-information systems:

- systems based on syntax (text search systems);
- systems based on semantics (abstracts and keywords systems).

Search engines (familiar with Internet users) belong to the first group (examples: Alta Vista, Excite, Infoseek, Lycos, Magellan, Yahoo, etc.). These systems are fast, the maintenance effort is limited, they are non-selective (for example: “this document describes the system architecture” and “this document does not describe the system architecture” are both hits if you search for “system architecture” even though you were not looking for the second one) and they have been specially developed for digital documents available on servers.

The catalogues (well-known of the libraries) belong to the second group, they contain abstracts, indexed by keywords. These systems need a lot of maintenance effort, they are selective and usable for all kind of information (digital and non-digital information, structured and non-structured information).

The REMSSBOT-system uses a catalogue. The catalogue is based on the guidelines of the Catalogue of Data Sources (CDS) and Thesaurus developed by order of the European Environment Agency (EEA) in Copenhagen by the European Topic Centre on Catalogue of Data Sources and Thesaurus (ETC/CDS and Thesaurus).

The Catalogue of Data Sources and Thesaurus

The Catalogue of Data Sources is developed by the CDS Topic Centre of the EEA. In 1996 they produced the first version of the CDS. To produce this version they used the experiences of other datamodels of meta-information systems or standards. For example: the CORINE CDS pilot product, UDK (Umweltdatenkatalog), DIF (Directory Interchange Format), IFEN, Nordic, etc.). In the last version of the CDS (September 1997, version 2.0) the standard element set GELOS (Global Environmental Information Locator Services), proposed by the MITWG (Meta-Information Topic Working Group) of the G7 ENRM project, is an integrated part of the CDS.

In the REMSSBOT-system we use the ETC/CDS version 1.0, September 1996. The CDS contains the meta-information of the environmental data-objects (abstracts, addresses of institutions and persons and other additional specifications of the data-objects).

In order to index, retrieve and select the abstracts of the data-objects a thesaurus is used. A thesaurus is a structured list of terms. The ETC/CDS uses the GEMET (General European Multilingual Environment Thesaurus). The first version contained about 4,000 terms in English, with classification, upper hierarchy and partial equivalence in Dutch, French, German, Italian and Spanish.
It was the result of merging the following multilingual documents:
• Umwelt Thesaurus of Umweltbundesamt, Berlin;
• Thesaurus quadrilingue per l’Ambiente of CNR, Rome;
• CNR/EEA Classification Scheme;
• Multilingual Environment Thesaurus of NBOI, Amsterdam;
• EnVoc UNEP Infoterra Thesaurus, Nairobi;
• Thesaurus of Medio Ambiente, Madrid;
• Lexique Environnement Planète, Paris.

The last version of the GEMET (version 1.0, September 1997) has two arrangements: a classification scheme (four super groups, 34 groups: top-terms, significant descriptors, terms) and a thematic order (40 themes).

The GEMET has a tree structure of descriptors: broader terms, terms and narrower terms.

In the REMSSBOT-system we do not need the complete contents of the GEMET. We use a so-called micro-thesaurus containing a number of selected terms of the GEMET, supplied with special terms, concerning the environmental topic of the REMSSBOT-system. We use the tree structure of terms, not limited to three levels.

This tree structure is very efficient for searching, therefore we use more micro-theauris: one for thematic search (what-descriptors), another for spatial search (where-descriptors) and the third for search for data-owners (who-descriptors).

Additional information for the EEA and the ETC/CDS and Thesaurus can be found from the addresses shown in Figure 3.

The search and selection approach
The REMSSBOT approach for selecting data-objects is based on four main keywords: wwww (what, where, who and when).

With the what-keyword the user specifies the thematic coverage of the required data-object. He/ she is searching the (thematic) thesaurus and selects one or more descriptors.

With the where-keyword the user specifies the spatial coverage. In the REMSSBOT-system he or she is searching a geographical micro-thesaurus and selects one or more geographical descriptors.

With the who-keyword the user specifies the owner (author, data-provider, etc.). He is searching the address information of institutions and persons (or an administration micro-thesaurus) and selects one or more items.

With the when-keyword the user specifies the time coverage of the required data-object. This input is a selection criterion and is related to the time specification of the data-object.

After this input the user gets an overview of all abstracts complying with the required wwww-specifications.

The functionality of the CDS
The main functionality of the ETC/CDS is to provide meta-information of environmental data-objects and to provide the current information available by the URL (Universal Resource Locator; for example: http://www.netor.gr/remssbot).

To realise this functionality, address information, indexed by the GEMET, and meta-information of data-objects, also indexed by the GEMET, are stored in the CDS-database.

The REMSSBOT-system has the same functionality as the ETC/CDS supplied with the functionality to access and to retrieve actual information and structured data available on other servers (not by URL).

The REMSSBOT-system distinguishes in principle three types of data provision (three types of services) to the users:
1. To provide meta-information and the address where to find the real information. The information is not available in digital format.
2. To provide meta-information and the matching real information, available more or less in a fixed format, copied from a server (for example: by URL).
3. To provide meta-information and the real information by carrying out of selections on databases. This service is an extension to the ETC/CDS.

These three types of services are called REMSSBOT-services (R-services): R-info, R-product and R-process.
To realise the extension to the ETC/CDS-system we need more technical information about data-objects and we have to store this technical information in an extension to the ETC/CDS-database.

The datamodel of the CDS and Thesaurus

The ETC/CDS and Thesaurus consists of three main components: a collection of keywords (terms or descriptors), a collection of addresses of institutions and persons and a collection of abstracts of data-objects and specifications (Figure 4).

Environmental information (and thus meta-information) can be classified in a number of categories: activities, datasets, data, stations and sites, maps, tools and applications and documents.

An activity or a dataset consists of a collection of miscellaneous environmental information. An activity describes a project or a programme; within a project different kinds of data are generated or collected. The data collection and documentation are described by a dataset. The data collection consists of different objects dealing with a certain issue or related to a common spatial reference. Objects may take the form of maps, measurement data (e.g. in a file or database), reports, etc.

In general, this means we have different classes of data-objects and one data-object can have one or more related data-objects. In that case we speak of parent and child data-objects.

Addresses from institutions and persons play a major role in the ETC/CDS, because any meta-information is affiliated to an institution or to one or more persons and can be localised by these addresses. Furthermore, persons are related to an institution.

This means one data-object is related to one or more addresses. Of course one person or institution can also be related to more data-objects; and one address can have one or more related addresses. In that case we speak of parent and child addresses.

The multilingual thesaurus (GEMET) is a structured list of authorised terms, grouped in logical categories and sub-categories: four super groups, groups, top-terms, significant descriptors, terms. The top-terms, significant descriptors and terms are used as keywords for indexing addresses and data-objects.

The REMSSBOT datamodel has two extensions to the ETC/CDS-datamodel (Figure 5):

1. To specify users and user groups and to specify user privileges some tables are added to the ETC/CDS-datamodel, related to addresses. This functionality is optional. The security level and the access control can be different in the Attica, Piemonte or Scheldt region.

2. To carry out the REMSSBOT-services (R-product and R-process) some tables to store technical information are added to the ETC/CDS-datamodel, related to data-objects.

Within the REMSSBOT-system the R-process is carried by the E-service (Elementary-service). The E-service sends a query and returns single or multiple values, structured data, a map or a picture from a REDP (regional environmental data-provider).
Standards and components of the system architecture

The system architecture is realised with the following standards and components:
• open industry standards;
• object technology;
• Web technology;
• CORBA (Common Object Request Broker Architecture);
• VisualWave/VisualWorks and Distributed Smalltalk (Attica/Scheldt);
• Forté environment (Piemonte);
• A meta-information system based on the EEA ETC/CDS and Thesaurus datamodel;
• Client: PC (Windows 3.11 + Windows 95), workstation (Windows NT + Unix);
• Server-platform: Windows NT, Unix, Aix.

Design and implementation process

The design and development team for the system architecture, composed of experts of the subcontractor EDS-Belgium (Electronic Data Systems), produced three documents:
1 “Survey and impact of CDS situation and research of the technical situation.” This document helps to select the general purpose server-independent technology for use in developing distributed applications and gives examples of technical architectures.
2 “Guidelines for the System Architecture” documents the technical building blocks and the guidelines for the architecture, based on a research of the technical alternatives and a survey of the infrastructure in the three regions (Figure 6).
3 “Technical Guidelines for the System Architecture” describes the implementation of the system architecture in more detail.

At the same time a prototype of the architecture was built to prove that the design and the chosen development environment is a real working solution. The prototype was produced in close co-operation with the three development teams in the Attica, Piemonte and Scheldt regions.

The technical guidelines and the prototype of the system architecture (including a CDS-database, a functional front-end client application and a back-end server) functional for a limited number of datatypes, was the starting point for the “implementation-phase” of the demonstrators in the three regions.

Further reading


ETC/CDS project (1996d), Third ETC/CDS Workshop on CDS and Thesaurus, Copenhagen, 16-17 September 1996, Proceedings, ETC/CDS project.


Lievens, E. and Gallo, P. (1996), REMSSBOT Project D03.02 Guidelines for System Architecture REMSSBOT project.