PROCEEDINGS OF IAMG'98
A. Buccianti, G. Nardi and R. Potenza (Eds.)
Isola d'Ischia, Naples, 1998

INTEGRATED ACCESS TO EARTH SCIENCES DATA: AN EVOLUTIONARY APPROACH

R. Potenza[1], A. Carrara[2], F. Guzzetti[3] and O. Signore[4]

[1]*C. N. R. - CSGAQ,*
Milano, Italy
geomat@icil64.cilea.it

[2]*C. N. R. - CSITE,*
Bologna, Italy
carrara@deis158.deis.unibo.it

[3]*C. N. R. - IRPI,*
Perugia, Italy
F. Guzzetti@irpi.pg.cnr.it

[4]*C. N. R. - CNUCE,*
Pisa, Italy
O.Signore@cnuce.cnr.it

Summary

Four CNR groups cooperated to a joint project for the organisation of geological data sets to be accessed in WWW environment. Data bases on different branches of geosciences were tested in order to develop and refine tools for the conceptual and geographical access to the data. A Thesaurus of Geosciences was prepared to be integrated in the navigation tools. As for the georeferenced data, the current Java based WWW-GIS applications are too elementary to support advanced graphical functions and to process large data sets, by preserving the same accuracy and efficiency of current stand-alone geo-systems. A new collection of Java Applets was therefore developed enabling the remote access, display and query of vector maps and related tabular data. Experiments on the natural catastrophes data base clearly indicate that the technology is still in its infantry, as it is affected by a variety of pitfalls such as software instability and huge virtual memory requirements on the client computer. The Z39.50 standard was kept as the base for the development of a software aimed to cope with both the BIB-1 or GILS profile possibilities.

1. INTRODUCTION

A goal of paramount importance for geological investigation, environmental analysis, land-use planning, geological hazard assessment and risk mitigation is the ability to timely access and retrieve information from distributed geo-spatial data banks [6]. Earth scientists need access to several different kinds of data and information, including scientific papers, maps, photographs, results of calculations and historical data. The relationships among the various sources of information are of different types. If specific software applications can satisfy single user needs, it remains difficult to correlate data that are geographically distributed or that were produced or are available in various hardware and software environments.

From a computer science perspective, hypermedia and networks appear to be the most promising technologies for connecting different sources of data and applications, where information is structured into small *"information nodes"*. In this view, the hypermedia approach is in charge of implementing the complex conceptual relationships (*"interaction paradigms"*) among different information nodes, while the network provides the connectivity among different sites. The implementation of interaction paradigms requires the availability of standards at the user (*i.e.*, metaphors, presentation, browsing), archive (*i.e.*, schema) and protocol (*i.e.*, HTTP, Z39.50) levels, that are yet hard to meet. Moreover, even if efforts are made to consistently enforce such standards, any project requiring a strict conformance to them will face enormous difficulties and will most probably fail.

On the base of these considerations, in 1996 we started the project "IDRISI", a joint research program to enhance the visibility and improve the availability of a large amount of geological information produced in the previous years by various working groups. The goal of the project was a fully integrated access to the available information, aiming to a consistent interface and a coherent architecture. Further project requirements were the integration of various data sources available from different environments, and the search for *"common access points"*, like terminological standards and geographical locations.

Relevant application issues were investigated in detail and tested to access the different data archives, the pattern of which was designed along the project development. A general system architecture based on the use of a conventional Web browser (*i.e.*, Netscape, Explorer, etc.) was outlined. The browser supports different user needs and allows for the definition and refinement of queries, finding concepts to look for and selecting different navigation metaphors, like maps or classification schemes, to access geographically distributed archives. The trade off between imposing a strict conformance to the standards and the flexibility of a less constraining approach, that can increase recall and precision of information retrieval, was also considered. Lastly, a relevant issue was the adoption of an evolutionary approach. This allowed to make available archives and navigation tools that, by itself, was an important achievement.

2. SYSTEM APPROACH

In a general sense, accessing Earth Sciences data can be conducted to the more general paradigm known as *"digital libraries"*. A digital library is a collection of real or virtual (*i.e.*, dynamically created) documents made by components stored and managed in heterogeneous environments, like data base management systems (DBMS), information retrieval systems (IRS), or ordinary files. Searching for a document requires to face manifold information sources, adapting to different specific search languages of the hosting systems, like SQL queries for relational databases, free text queries for IRS, and spatial and geometric, properties based queries for GIS environments. In addition, we must remind that some standards are well established and/or emerging, namely: internet browsers as the standard interaction style from a user perspective; Z39.50 as the officially recognised standard for network information retrieval; and SGML and XML as a powerful way of organising complex documents.

From a practical point of view, we must also face the problem of accessing databases described by different schemes, according to the specific needs of the designer, or designed to meet well defined purposes, sometimes ignoring the *"implicit knowledge"* of the administrator. Examples are the frequent omissions of some *"obvious"* geographical data, like country names, or the use of implicit but not standardised information, like units. Moreover, the same data can be stored in one or more *"fields"*, and fields containing the same information can have different names, or vice-versa. However, it is evident that different archives, in spite of these *"lexical"* differences, have semantically equivalent *"access points"* that may be known to a skilled user, but not to the interface software.

In this view, we consider hypermedia the most suitable environment to fulfil the task of accessing heterogeneous sources of information. As pointed out elsewhere [17, 18, 19, 20, 21], the main issues in implementing hypermedia are: the implementation of its cognitive layer; modelling of links; filtering information; and tailoring pages to user needs and interests.

Keeping in mind these considerations, we designed a system architecture based on a

client able to access data archives through recognised standards, where existing and/or suitable, or through well defined interfaces. This architecture is presently widely accepted in the area of Information Systems, and is known as the *"three level"* architecture. Conceptually, a *"fat"* user client includes several components, namely: a Z39.50 client; a *"navigator"* on the thesaurus; and a *"navigator"* on the geographical information. It is worth pointing out that some of the system functions can be performed on an *"access server"*, a machine different from the user or the data-server computers. This will result in a *"lighter"* client, enlarging the range of potential users, since a robust machine at the user end is not needed.

In spite of the intrinsic multimedia nature of the Web technology and of the dramatic advancements in the domain, true geographical data handling requires properly designed services for capturing, retrieving, manipulating and displaying geo-referenced information and maps. Hence, to recreate in an Internet environment the functions and services which traditionally appertain to a GIS is a major, challenging task largely unsolved as yet. A recent survey on the software tools capable to manage geographical data into the world wide web [7, 8], indicates that Web-server based applications are feasible only in very plain circumstances, such as for the search into geographical catalogues and the browsing of pre-defined map collections by generic users whose computers are equipped with a Web browser [23, 12].

The request for powerful graphic tools, including vector data treatment capability, and fast response times, finds an answer in *"Plug-in"* and *"ActiveX"* technologies. In this case, WWW-gateways still support the remote access to GIS data, but software modules, which perform graphical processing locally, enhance Web client platforms. Some vendors are proposing complete suites of integrated tools, including the web server, the authoring software and the client modules (*"Plug-in"*). A drawback of this approach is that users are bound to proprietary platforms and/or specific file formats.

The *"Java Applet"* technology may provide a more general framework. Java classes for graphical and map-oriented functions may supply local intelligence to perform client-side geo-processing. HTTP servers will export base maps and thematic features together with Java Applets, which perform spatial data handling. This approach reduces network traffic and provides a higher level of graphic interaction. Java code runs on the Java virtual machine (JVM) which ensures platform independence and eliminates local modules installation and updating.

Many implementations which exploit these approaches share a common denominator: the client need to acquire huge files (*e.g.*, MWF, Active CGM) properly structured to include multiple layers of graphical objects in raster or vector format, hypermedia links and metadata which describe the file content. Thus, network throughput and geographical data heterogeneity still hamper a wider exploitation of these services on the internet. Moreover, current Java based WWW-GIS applications are too elementary for an exhaustive evaluation. The intrinsic potentialities of Java need to cope with the capability to support advanced graphical functions and to process large data sets, by preserving the same accuracy and efficiency of current stand-alone geo-systems.

As a preliminary attempt to cope with such issues, a collection of Java Applets was developed enabling the remote access, display and query of vector maps and related tabular data on natural catastrophes [7, 8]. The experiment clearly indicates that the technology is still in its infantry, as it is affected by a variety of pitfalls such as software instability and huge virtual memory requirements on the client computer.

3. DATA WAREHOUSES

For the project, three main archives containing geological information *s.l.* were available. They refer to: the Italian Thesaurus of Geo-sciences, containing about 8200 entries [16]; information on several thousands geological and hydrological historical catastrophes in Italy [11, 4]; and geological, morphological, mineralogical and bibliographical information for the Valtellina area in northern Italy [14]. A fourth archive, containing information on Paleozoic detrital formations of Northern Apennines was also considered, but its implementation was not completed. The data archives are shortly described below and can be accessed from the project home page at the address http://seal1.cnuce.cnr.it:5000/IDRISI/.

The available archives can be distinguished in two classes: bibliographical and raw data. To access both we rely on the Z39.50 standard, using a software designed to cope with both the BIB-1 or GILS profile possibilities [13, 22]. However, features of high interest to the user like thesaurus browsing, are not provided by the Z39.50 standard and must be implemented aside of the HTTP connection.

3.1 The Italian Thesaurus of Geosciences

At the early steps of the project the need was recognised for a terminological and conceptual support to the structure of the search tools, and the development of a suitable Italian thesaurus was therefore started [2, 15, 16]. Widening the traditional scope of similar older bibliographic aids, the thesaurus now embraces the main terms from geological and related sciences, like physics, chemistry and mathematics. A complex network of hierarchical and relational links between items allows to follow specific knowledge paths, under the control of scientific terminology. At present, the 8200 terms considered are linked by 65,000 hierarchical or conceptual relationships. The consistency of the Italian thesaurus with internationally accepted similar tools, like the IUGS Multilingual Thesaurus [10] and the GeoRef Thesaurus [1], opens the way to the approach in different languages. The thesaurus is currently shared as an independent database and can be accessed clicking the icon CSGAQ in the project home page, for purposes like adding keywords to scientific papers or as an aid to search in on-line bibliographic systems. Its full integration with the other data bases and the extension of the set of relations (now limited mainly to the bibliographic links BT-NT and RT) to geological paradigms [5] will be provided in the near future.

3.2 Information on historical hydrological and geological catastrophes

The Italian Group for Hydrological and Geological Hazard Prevention (CNR–GNDCI) has developed and maintains a set of databases (http://wwwdb.gndci.pg.cnr.it) distributing information on hydrological and geological hazards [11, 4]. The aim of the system is to disseminate a vast amount of data, information and expertise to civil defence authorities, expert users, the media and the general public. Databases presently available include: addresses of research teams currently active within GNDCI; a list of the 1800 GNDCI publications, reports and maps; the IRPI library database; the AVI inventory, containing 17,000 historical

information on landslides and 7000 historical information on flooding events occurred in Italy in this century; and mean daily discharge values for 70 gauging stations in Central Italy.

The system, exploiting the potential of internet and of new software tools, allows both the retrieval of hypertext documents that are dynamically created and to perform remote queries, either pre-defined or free, the latter using the SQL language. Query results provide the opportunity for submitting new queries by pointing and clicking, making information retrieval very efficient. Information is stored into a RDB (Sybase) and is structured into several tables. A specialised software (Sybernet) performs SQL queries and allows for editing, storing and executing command procedures (*i.e.*, Sybase *"stored procedures"*) that can retrieve information from one or more tables and format the output in a variety of different ways, including free text, tables and tab delimited format. Database maintenance is performed from a Pc via ODBC.

3.3 Geological information for the Valtellina area

The Centre for Alpine and Quaternary Geodynamics (CNR–CSGAQ) developed and maintains a system to organise and distribute geological informations on a test area of Valtellina, a region in the central Alps for which geological, morphological, petrological, mineralogical as well as natural hazard data were collected in the past years [14]. Numerical data, descriptions, maps and images are entered in the system as they are validated. Care is taken to provide each data set with suitable references to Thesaurus entries. The system is markedly experimental and is quickly changing along the time. The main upgrade we are providing is the full integration with the Thesaurus structure as well as with the geographic information. The CSGAQ icon in the IDRISI project site gives access to the system.

4. FUTURE WORK

Geologists for their investigations need to access and retrieve heterogeneous information from a variety of Earth Sciences Digital Libraries (ESDLs). The request for timely access and efficient retrieval of sound scientific and technical information from heterogeneous, geographically distributed ESDLs is growing and it can be foreseen that the demand will increase in the future. Due to the calls for international efforts for geological investigations, and to the multidisciplinary character of geological studies, the search for Earth Sciences information is not limited to geological data, but spans many other disciplines and fields. In this view, the integration of a properly structured thesaurus in an ESDL represents an important step.

Despite the dramatic advancements in the availability of multimedia authoring and navigation tools, and a certain improvement in networks performance, designing, maintaining and accessing ESDLs remains a complex task, that requires technological skills and advanced software tools. Even more difficult appears the design and the implementation of *"software agents"* capable of interconnecting geographically distributed, heterogeneous *"information nodes"*, looking for *"common access points"* in the different nodes, and linking archives containing data of different types (*i.e.*, free text, tables, vector and raster graphics, images, maps, etc.), despite lexical and structural differences.

Future developments of our project will include: a) the development and integration of other data systems, possibly spanning wider geoscience branches, and the in depth integration of the Italian Thesaurus of Geosciences with the databases, with priority to the archive on

historical hydrological and geological catastrophes; b) improved capabilities to generate and deliver virtual (*i.e.*, generated on demand) hypertext documents containing heterogeneous information types from different databases, and; c) an attempt to recreate in an internet environment GIS functions and services, that is, retrieving, manipulating and displaying geo-referenced information and maps, exploiting a WWW Java server gateway, geo-processing systems compliant to OGIS interface specifications [3, 9] and a collection of Java *"servlets"* and *"applets"* exchanging geo-referenced objects.

ACKNOWLEDGEMENTS

We are grateful to R. Bartoli, G. Fresta and G. Tonelli for their invaluable help in the design and implementation of the browsers and the RDBMS query interfaces. G. Bardazza designed the home page of the project IDRISI.

REFERENCES

1. AMERICAN GEOLOGICAL INSTITUTE (GOODMAN, A.B. ed.), 1994. *GeoRef Thesaurus, Seventh edition*. American Geological Institute, Alexandria, VA, USA, 10+842 pp.
2. BARTOLI, R., POTENZA, R., SIGNORE, O., 1998. *Thesaurus Italiano di Scienze della Terra*. Edizione ipermediale, Consiglio Nazionale delle Ricerche, Milano – Pisa
3. BUEHLER, K., MCKEE, L., (eds.), 1996. *The OpenGIS Guide*. OGIS TC Document 96-001, 1996. http://ogis.org/guide/.
4. CARDINALI, M., GUZZETTI, F., REICHENBACH, P., TONELLI, G., 1998. *Conveying scientific information to the users: the experience of the GNDCI information delivery system*. Annales Geophysicae, Vol. 16, Suppl. IV, Part IV, Nonlinear Geophysics & Natural Hazards, C 1218.
5. CARIMATI, R., MARINI, A., POTENZA, R., 1982. *The mathematical formalisation of the geological relations identifying the basic structure of a geological data bank*. In: Cubbitt, J., and Reyment, R.A., (eds.), Quantitative stratigraphic correlation. Wiley, Chichester, 13-18.
6. CARRARA, A., GUZZETTI, F., (eds.), 1995. *Geographical Information Systems in Assessing Natural Hazards*, Kluwer Academic Publishers, Dordrecht, 354 pp., ISBN 0-7923-3502-3.
7. CARRARA, A., PASQUI, V., 1997. *Interfaccia WWW ai GIS nella gestione dei rischi naturali. Rapporto Tecnico*, Consiglio Nazionale delle Ricerche, CSITE, Bologna.
8. CARRARA, A., PASQUI, V., 1998. *WWW-GIS gateways and natural hazards*. (in preparation).
9. FEDERAL GEOGRAPHIC DATA COMMITTEE, 1994. *Content standard for digital geospatial metadata (June 8)*, FGDC, Reston, VA, USA.
10. GRAVESTEIJN, J., KORTMAN, C., POTENZA, R., RASSAM, G.N., 1995. *Multilingual thesaurus of geosciences*. Second edition. Information Today inc., Medford NJ, USA. L+645 pp.
11. GUZZETTI, F., CARDINALI, M., REICHENBACH, P., 1994. *The AVI project. A bibliographical and archive inventory of landslides and floods in Italy*. Environmental Management, Vol. 18: 623-633. http://avi.gndci.pg.cnr.it/wwwavi/PaperEn/avien.html.
12. HUSE, S.M., 1995. *GRASSLinks: A New Model for Spatial Information Access in Environmental Planning*. PhD dissertation, University of California at Berkeley, 1995. http://www.regis.berkeley.edu/sue/phd/.

13. LOFFREDO, M., BALDACCI, M.B., SIGNORE, O., 1997. *Tailoring Z39.50 on Existing Databases: the ARCA Project*, Proceedings of DEXA'97 - 8th International Workshop on Database and Expert Systems Applications, September 1-2, 1997, Toulouse, France, (Edited by Roland R. Wagner) IEEE Computer Society, ISBN 0-8186-8147-0, Library of Congress Number 97-80040, p. 332-338.

14. MORANDI, S., POTENZA, R., (ed.), 1997. *Progetto pilota alta Valtellina - A multimedia system of geoscience informations*. http://csgaq.terra.unimi.it, Consiglio Nazionale delle Ricerche, Milano.

15. POTENZA, R., BARTOLI, R., SIGNORE, O., 1994. *Tools for accessing large geological data systems*. Proceedings of IAMG '97 - CIMNE, Barcelona, 450-452.

16. POTENZA, R., (ed.), 1997. *Thesaurus Italiano di Scienze della Terra*. Quaderni di Geodinamica Alpina e Quaternaria , Consiglio Nazionale delle Ricerche, CSGA, Milano.

17. SIGNORE, O., 1995. *Issues on Hypertext Design*. DEXA'95 - Database and Expert Systems Application, Proceedings of the International Conference in London, United Kingdom 4-8 September 1995, Lecture Notes in Computer Science, N. 978, Springer Verlag, 283-292, ISBN 3-540-60303-4,

18. SIGNORE, O., 1996. *Exploiting Navigation Capabilities in Hypertext/Hypermedia*. HICSS-29 Annual Hawaii International Conference on System Science, Maui, Hawaii - January 3-6, 1996, 165-175, ISBN 0-8186-7327-3, ISSN 1060-3425,

19. SIGNORE, O., 1995. *Modelling Links in Hypertext/Hypermedia*. In: Multimedia Computing and Museums, Selected papers from the Third International Conference on Hypermedia and Interactivity in Museums (ICHIM'95 - MCN'95), October 9-13, San Diego, California (USA), 198-216, ISBN 1-88-5626-11-8,

20. SIGNORE, O., BARTOLI, R., FRESTA, G., 1997. *Tailoring Web Pages to Users' Needs*. Proceedings of Workshop Adaptive Systems and User Modeling on the World Wide Web, at the Sixth International Conference on User Modeling, Chia Laguna, Sardinia, 2-5 June 1997, 85-90.

21. SIGNORE, O., BARTOLI, R., FRESTA, G., LOFFREDO, M., 1997. *Implementing the Cognitive Layer of a Hypermedia*. Museum Interactive Multimedia 1997: Cultural Heritage Systems Design and Interfaces, Selected papers from ICHIM 97 the Fourth International Conference on Hypermedia and InterActivity in Museums, Paris, France, 3-5 September, 1997, (Bearman, D. and Trant, J., eds.), Archives & Museum Informatics, 15-22, ISBN 1-885626-14-2

22. SIGNORE, O., LOFFREDO, M., SABINA, S., 1997. *Z39.50-SQL gateways: technical description*. Deliverable D5.1, Telematics Application Programme, Information Engineering Sector, Project IE-2005, Aquarelle: Sharing Cultural Heritage through Multimedia Telematics (July 1997).

23. SMITH, T.R., 1996. *A Digital Library for Geographically Referenced Materials*. *IEEE Computer*, Vol. 29: 5, May 1996, 54 - 60.