Visual Analysis of Geographic Metadata in a Spatial Data Infrastructure

R. Albertoni, A. Bertone, M. De Martino
Istituto di Matematica Applicata e Tecnologie Informatiche
Consiglio Nazionale delle Ricerche
{albertoni,bertone,demartino}@ge.imati.cnr.it

Abstract

Nowadays the importance of collecting and sharing geographical data is rapidly increasing. Spatial Data Infrastructures are arising to integrate geographic information services which allow to identify and access geographic information from a wide range of sources. The study of data access addresses different issues such as data delivering, data retrieval, data integration and data selection.

In this paper we focus on the data selection issue: we describe an approach to support data selection activity based on visual analysis of geographic metadata. This facilitates the metadata analysis providing a compact overview of data that are available and a correct interpretation of the result set, allowing to discover properties and relationship among data and to formulate better searching criteria.

1. Introduction

Today geographic data play a crucial role in numerous business and government applications. They are used in many decision-making processes of institutions and organizations from commerce and education to research and healthcare. In particular they represent an indispensable resource to support public administrators during the definition of national policies, to evaluate the environmental impact of political choices. For these purposes, vast collections of heterogeneous geographic data are generated from numerous providers. In addition to publishing the data on CDs and other media, providers are shifting toward making the information available on the Web following the explosive growth of Internet and its users.

The importance of sharing and collecting geographic data is rapidly increasing: usually each country relies on private or public structures to maintain updated geographic information at a regional and at a national level. However, the request for sharing and collecting geographic information crosses the countries border. Spatial Data Infrastructure (SDI) beyond the country borders are arising to integrate geographic information services which allow to identify and access geographic information from a wide range of sources (see [1], [2], [3]).

At European level the importance to access to geographic information has been recognized as essential to ease the definition of coherent European policies [4]. INSPIRE[1] initiative is proposed to make accessible the resources to each European country by defining a framework for the gradual creation of a harmonized spatial information infrastructure. Other initiatives like SPIRIT (Spatially-Aware Information Retrieval on the Internet) [3] propose a worldwide access by getting geographic information directly by surfing in Internet. They usually aim to design and implement a high level of intelligence web search engine to find documents and datasets on the web relating to places or regions referred to in a query.

In a SDI, the quantity and the heterogeneity of data raise the problem to define instruments to manage a large amount of distributed data: the concept of metadata has been introduced to describe geographic data. Digital archives of metadata such as Metadata Information System (MIS) and Catalogue Service (CS) are developed to manage such information. In particular, the following issues need to be addressed:

- Data publishing and delivering: data and metadata entry.
- Data retrieval: accessing to distributed and heterogeneous resources.
- Data integration: integrating information retrieved from the distributed location.
- Data selection: comparison and exploration of data to select appropriate data according with user requirements.

The initiatives that aim to realize a SDI mainly focus on the first three issues. Less attention is given to

IEEE COMPUTER SOCIETY

the comparison and to the exploration of data for data selection activity.

This paper focuses on data selection issue: the importance of defining visual based approaches for metadata analysis to facilitate the users to locate appropriate data is discussed. The principles of an approach to compare and analyse metadata are illustrated. It is out of the aims of this paper to face with the aspects of its integration in a SDI. The proposed approach combines automatic visualization techniques with graphic interaction tools to create a data exploration system. The main purpose is to enable users to uncover and extract hidden relationships in large data sets. The system is the result of a research activity performed inside the EU-funded project INVISIP: Information Visualization in Site Planning, IST-2000-29640 [5], [6].

2. Metadata analysis for data selection

Data selection is the activity that aims to choose the most appropriate data for a specific application. It is strongly affected by the complexity of geographic data. Two main aspects are critical in the selection of geographic data:

- Usually it is not possible to access to the geographic resources to have a look and to realize "at a glance" the information that they contain since geographic data are resources that can be heavy in terms of Kbytes, or that can be not available for free.
- To compare different geographic resources requires strong efforts since data are available in a huge kind of variety. They differ in characteristics like Scale, Reference System, Geographic Extension, Themes, Quality, Fees and so on.

Metadata concept, data about data, is adopted to overcome these drawbacks: metadata give a detailed description of the geographic data characteristics according with a specific standard. They provide a first level of data integration and allow to compare sources provided by different organizations. Moreover they represent a mean to choose geographic data without resource download.

The vast collections of geographic data determine the generation of a large set of metadata. Furthermore the complexity of geographic data forces metadata to be characterized by many attributes and to be represented in a multidimensional information space. Instruments able to manage this large set of metadata and their multidimensionality are needed. A lot of Metadata Information System (MIS) and Catalogue Systems (CS) are generated to organise and manage metadata. [7] gives an overview of MIS and CS for geographic data and provides more details about the metadata concept and the related initiatives.

In particular, different initiatives have been carrying out to define metadata standard (ISO 19115[8], FGDC[9], CEN/TC 287 ENV 12657[10]) and to facilitate the searching of metadata (UDK[11], [12]). They propose browsing tools for metadata that provide the results as a list of textual information. Considering the multidimensionality and the quantity of metadata, such list of information overwhelms user abilities of comprehension. Analysis tools are needed to support the user in the comprehension of the searching results.

Visual analysis of metadata appears as a first step towards the solution of such problems [13]. Visualizations enable the user to have a compact overview of data that are available, a correct interpretation of the result, to mine properties and relationship among data and to formulate better searching criteria. Different visual analysis approaches are proposed. They mainly differ in accordance with the types of attributes they consider or the type of representation used for the spatial extent. Focusing on ISO 19115 Metadata standard, the attributes can be represented by categorical values or full text values and bounding box is used to express spatial extent attributes. In [14] a visualization approach is proposed to solve the problem of the exploration of metadata working mainly on attributes whose value is expressed as full text. This approach combines different visualizations into a so called SuperTable. In [15] GeoCrystal system is proposed: it focuses on the spatial extent and lets the user compose complex queries and visualize search results in a 3D space for geographic data.

In this paper we propose an approach applied to the categorical attributes of ISO 19115 metadata standard. Categorical attributes play an important role both since they are numerically relevant (more than twenty metadata attributes are defined as categorical) and they represent important information such as maintenance attribute, progress; type of spatial representation, resolution, theme classification, etc. The complexity of the ISO 19115 standard and the number of attributes that characterize it, may lead to the following main problems in the data selection:

- *Unfamiliarity with attributes*: the user usually has only a partial knowledge of the available attributes and must perform his selection in an unfamiliar information space. The searching criteria that he is able to perform might not be enough to successfully end its selection activity. Therefore he needs to refine his criteria using some attributes he is not familiar with.

- *Data missing*: the repository might not contain the data the user is looking for; hence he is forced to define new criteria in order to find similar data.

We propose a solution to such problems applying an exploration approach to metadata. It combines the functionalities of automatic visualization and graphical interaction to enable users to uncover and extract hidden relationships in large data sets.

3. An approach for visual metadata analysis

In this paragraph we describe a visualization-based approach to analyse metadata. The main idea is to simultaneously use visualization techniques, graphic interaction and a dynamic link among the visualization themselves using Brushing and Linking techniques[16].

The approach is characterised by three iterative phases: a visualization phase, an exploration phase, and a query-building phase.

During the first phase (*visualization phase*) different representations of metadata attributes and values are provided in order to give the user a compact and human understandable view of the available data. Different visualisation techniques are provided and can be applied at the same time. They are classified according to the number of attributes they can display: single attribute and multi attribute visualization.

The second phase (*exploration phase*) is based on the analysis of the visualizations previously displayed to extract knowledge about metadata. In particular, single attribute visualizations provide the knowledge of the available values and quantitative information of metadata attributes, whereas multi attribute visualizations provide the knowledge on metadata attributes and the existing relationships. This task is performed using both interaction functionalities with the element displayed in a visualization, and brushing and linking to combine different visualization methods. The result of the exploration phase assists the user in the choice of both attribute and its values to define new query criteria.

In the third phase the criteria are completed and the query is generated (*query building phase*). Finally the attribute values are graphically selected to express the query and the starting subset is reduced. As soon as the query is performed, all displayed visualizations are updated showing the new (sub)set of data. Furthermore, if necessary, a new step of the process can be performed starting from the previous visualizations or activating new visualizations. Otherwise, if the results obtained does not satisfy user requirements, it is possible to delete some selections previously performed and return to an "old" set (a so called *Undo*).

Illustrative example

We illustrate an application of our approach to study the site planning of a shopping centre. We suppose that the user needs to search for data in conformity with the following requirements: data about specific themes (environment, planning, infrastructure), written in English and in MapInfo format. Performing a query on the repository, he does not obtain any result. Then he needs to perform an explorative analysis in the repository to search for data that are the most suitable to his requirements.

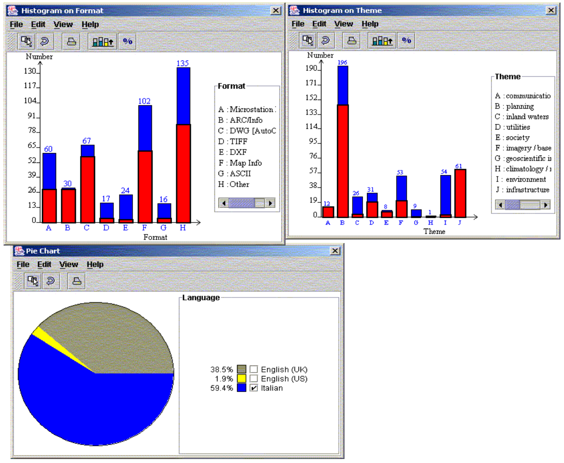

Figure 1: Pie chart on language, a histogram on theme, and a histogram on format.

The explorative activity is performed in different steps.

- The user opens a Pie Chart to represent Language, a Histogram to represent Theme, and a Histogram to represent Format.
- The user interacts with the visualizations and performs some reasoning on the available data to understand which requirement could be disregarded.
- If he clicks on "English" in the Pie Chart, he can easily realize that there are many records dealing with "infrastructure" and "planning" in the Histogram of Theme, but there is no one related to "environment"; moreover many records have the format MapInfo. Similarly, when he clicks on "Italian" in Pie Chart (Figure 1), he realizes that there are many datasets dealing with "environment", and some to "infrastructure" and "planning". Thus, supposing the user is also able to understand Italian, he can realize that changing his requirements on language he obtains data

related to the required themes. This solves the problem of *data missing*.

After some steps of the process, that is, once the user has selected MapInfo and all of the three required themes, the resulting set of data satisfies all the requirements but is still too large. A further exploration needs to be performed on other attributes. Probably he has to deal with the problem of *unfamiliar attributes*.

- To proceed in the analysis, he opens a new visualization, i.e. a Parallel Coordinates Plot (PCP) representing Resolution, Progress, and Update Frequency attributes.

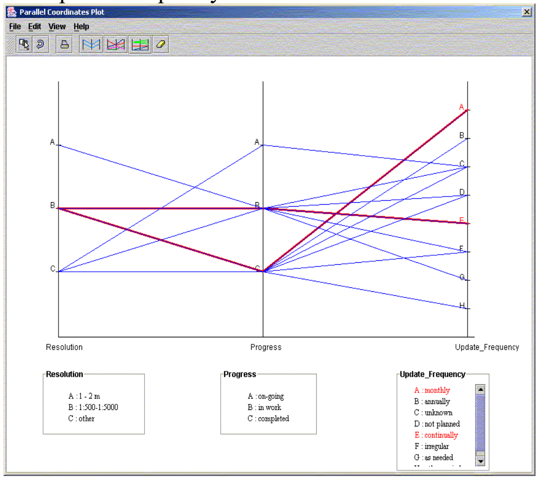

Figure 2: PCP on Resolution, Progress and update frequency

- The user interacts with the visualization and performs reasoning on the new attributes. For example, he can focus on the Update Frequency attribute. When he clicks on "continually", all the data sharing this value for Update Frequency attribute are highlighted in red: he can easily observe that their Progress value is "in work" and their Resolution is "1:500-1:5000". Similarly when he clicks on "monthly", he realizes that such data have Resolution "1:500-1:5000" and their Progress value is "completed" (Figure 2).
- The knowledge provided by such interactions allows the user to discover which criteria are the most suitable taking into account the new attributes.
- Supposing that he is more interested into data that are "completed" rather than data that are "continually" updated, he reduces the datasets according to "Update Frequency = monthly".

If the data set is still too large, it can be reduced by considering other attributes that the user previously did not take into account. Using reasoning similar to the previous ones, the user may reduce the datasets according to new criteria that he dynamically builds. At the end of the process, if the results are not satisfying, he may perform some steps back or the whole process from the beginning.

4. Discussion and future development

The proposed approach has been designed and a system has been developed within the European project INVISIP to analyze geographic metadata in a real case study. The system provides components for visual feedback and graphic interaction and work on categorical metadata attributes. In particular it has been tested in the different phases of the process of planning a commercial center [6]. The results of the tests outline the system capabilities to discover relationships among metadata and to increase user awareness about available geographic data.

The problems of unfamiliarity with attributes and data missing are solved:

- The visual approach provides at a glance quantitative information and qualitative information and increase user knowledge about the values that the attributes may assume. This knowledge is essential to refine the searching criteria using unfamiliar attributes.
- The visual approach provides a compact view of data. This allows the user to easily understand which data are available, which attributes might replace the missing ones and to find the data that are the most similar to those he is looking for.

The approach is satisfactory when dealing with a huge amount of data. However user comprehension becomes difficult if data amount increases i.e. over 100000. That is independent of the applied visualization, therefore different strategies need to be adopted. A possible solution is to organise the available data according to the attributes the user is more familiar with and to the relevance he gives to them. In our future work we are going to define an approach [17] based on hierarchical clustering with proper visualizations such as Magic-Eye View, [18], or Cone Trees [19]. Other problems may occur when applying clustering techniques to geographic data. Clustering techniques are based on similarity criteria hidden in the semantics of the words (and the sentences) that are used to express the value of the metadata attributes. To make explicit the semantic neighbourhood of the attribute value and to define a similarity criterion among metadata items, an ontology specification of value has to be adopted. In our future work we also are planning to study the application of ontology concepts to geographic context.

5. Conclusion

This paper addresses the issue of data selection that characterises the process of data access in a SDI. An overview of the main drawbacks that affect the search activity of geographic data are discussed and a visual approach is illustrated. An approach for metadata analysis is introduced to support users during the data selection phase. It is based on well-known visualizations and powerful graphic interaction techniques. The approach facilitates the user in the comprehension of the results of a browsing search as well as to discover relationship among data.
Future work will be investigated to extend this approach.

Acknowledgement

The authors acknowledge the support of the European Commission, IST Program. They thank all partners of the INVISIP project who contributed to this study: a special thank goes to KTH Team in Stockholm, UoC Team in Krakow, D’Appolonia Team in Genoa and Municipality of Genoa for the important discussions.

10. References

[1] P.C. Smith, U. Düren, O. Ostensen, L. Murre, M. Gould, U. Sandgren, M. Marinelli, K. Murray, E. Pross, A. Wirthmann, F. Salgé, and M. Konecky, *INSPIRE Architecture and Standards Position Paper*, JRC-Institute for Environment and Sustainability, European Commission, JRC, 2002.

[2] H. Scholten, A. LoCashio and B. Bonn, “ESMI, towards a European Spatial Metadata Infrastructure”, *Proceedings of EOGEO'98*, Salzburg, 1998.

[3] C.B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel, “Spatial Information Retrieval and Geographical Ontologies An Overview of the SPIRIT Project”, *SIGIR’02*, Tampere, Finland , August 11-15, 2002.

[4] EIONET: European Environment Information and Observation Network, http://www.eionet.eu.int/

[5] S. Göbel, J. Haist, H. Reiterer, and F. Müller, “INVISIP: Usage of Information Visualization Techniques to Access Geospatial Data Archives”, *DEXA 2002* , Aix-en-Provence, France, September 2-6, 2002, pp. 371-380.

[6] R. Albertoni, A. Bertone, M. De Martino, U. Demšar, and H. Hauska, “Knowledge Extraction by Visual Data Mining of Metadata in Site Planning”, *Proceedings of SCANGIS03*, Espoo, Finland, June 4-6, 2003.

[7] S. Göbel and K. Lutze “Development of metadata databases for geospatial data in www” *ACM GIS ’98*, Washington, D.C.,USA, 1998.

[8] ISO 19115, Geographic Information Metadata, International Standard Organization, http://www.isotc211.org/, 2003

[9] FGDC, Document FGDC-STD-001-1998, Content Standard for Digital Geospatial Metadata, Meta-Data Ad Hoc Working Group, Federal Geographic Data Committee, USA, 1998.

[10] CEN/TC 287 ENV 12657, ENV:“Euro-norme Voluntaire for Geographicinformation – Data description-Metadata”. CEN:European Committee for Standardization, CEN/TC 287: Geographic Information European Prestandards, http://www.cenorm.be, 1998.

[11] W. Swoboda, F. Kruse, R. Nikolai, W. Kazakos, D. Nyhuis, H. Rousselle, “The UDK Approach: the 4th Generation of an Environmental Data Catalogue Introduced in Austria and Germany”, *Meta-Data'99*, Third IEEE Meta-Data Conference, Bethesda, Maryland April 6-7, 1999.

[12]D. Stein, Geospatial Data Sharing through the Exploitation of Metadata, *ESRI International User Conference*, San Diego, California, July 8-11, 1997.

[13] N. Alper, and C. Stein, “Geospatial Metadata Querying and Visualization on the WWW Using Java Applets”, IEEE, 1996.

[14] P. Klein, F. Müller, H. Reiterer, and M. Eibl, “Visual Information Retrieval with the SuperTable + Scatterplot”, 6th *Conference Information Visualization*, London, England, 2002.

[15] S. Göbel, J. Haist, J. Uwe, “GeoCrystal: Graphic-Interactive Access to Geodata Archives”, *Visualization and Data Analysis 2002 Proceedings*, USA, 2002, pp. 391-402.

[16]Keim, D., W. Müller, and H. Schumann, Visual Data Mining.in Eurographic *STAR Proceedings*., Saarbrücken, 2002.

[17] R. Albertoni, A. Bertone, M. De Martino, U. Demšar, and H. Hauska, “Visual and automatic data mining for exploration of geographical metadata”, *Proceedings of the 6th AGILE*, Lyon, France, April 24-26, 2003.

[18] M. Kreuseler, N. Lopez; and H. Schumann, “A Scalable Framework for information Visualization”, *Proceedings InfoVis’2000*, Salt Lake City, 2000, pp.27-36.

[19] J.D. Mackinlay, G. Robertson, and S.K. Card, “The Perspective Wall: Detail and Context Smoothly Integrated”, *ACM Conference on Human Factors in Computing Systems (CHI '91)*, 1991, pp.173-179.