Knowledge Extraction by Visual Data Mining of Metadata in Site Planning

Riccardo Albertoni[1], Alessio Bertone[1], Urška Demšar[2],
Monica De Martino[1], Hans Hauska[2]

[1] Istituto di Matematica Applicata e Tecnologie Informatiche,
Consiglio Nazionale delle Richerche, Genova, Italy.
albertoni@ima.ge.cnr.it, bertone@ima.ge.cnr.it, demartino@ima.ge.cnr.it

[2] Department of Infrastructure,
Royal Institute of Technology (KTH), Stockholm, Sweden.
urska.demsar@geomatics.kth.se, hans.hauska@geomatics.kth.se

Abstract. The paper describes a tool designed within the first stage of the European project INVISIP in order to explore geographical metadata in the site planning process. A visual data mining approach is applied to a database of geographical metadata to help the user find an optimal subset of the existing geographical datasets for his particular planning task. It allows the user to perform both confirmative and explorative analysis. The approach is implemented in the Visual Data Mining tool, which integrates different types of visualisations with various interaction functionalities. It includes the interactive communication with the user and the brushing and linking process between different visualisations. The paper also presents an example of an application on a test metadatabase which was created for this purpose.

1 Introduction

This paper addresses issues related to visual data mining of metadata as an aid to the process of site planning. It is part of the research activity of the project INVISIP "Information Visualisation for Site Planning", supported by the European Commission within the Information Society Technology programme (IST 2000-29640). The goal of the project is to create a framework which will support all involved parties in the site planning process: municipal authorities, planning offices, data suppliers and citizens. Different techniques for visualisation of information are used, in order to improve the tasks of search for and analysis of the available information, and to facilitate the decision-making process based on an existing metadata information system for geographic data. The framework created within the project can be used in various fields of application, though its actual implementation concentrates on data for site planning. The project is organised into different phases: the analysis of existing methods for visualisation of information, the demonstrator and the prototype. This paper focuses on the first results of the demonstrator phase and on one of the several tools developed within the project: the Visual Data Mining tool (VDM tool).

The amount of data collected in databases increases from day to day, and databases containing several terabytes of data are no longer uncommon. These data are a potential source of important information. The term *knowledge discovery* refers to the overall process of extracting previously unknown useful knowledge from data, while the term *data mining* refers to the application of algorithms for extracting patterns from data [10]. *Visual data mining* is a novel approach in the knowledge discovery process, which utilizes visualisation as a communication channel between the computer and the user [4]. It integrates the human into the data mining process and combines human flexibility and ability of perception with the huge storage capacity and computational power of the computer. Human ability of perception enables the user to analyse complex events in a short time interval, to recognize important patterns and to make decisions more effectively than any computer can do. To achieve this the user must have data presented in a logical way with a good overview of all information. The development of visualisation techniques that are able to represent large amounts of multidimensional data is therefore especially important [13]. A major advantage of visual data mining techniques over other automatic data mining techniques (such as statistics, machine learning, etc.) is that visualisations allow direct user interaction and dynamic user guidance, which is difficult to achieve in non-visual approaches. They have the advantage of being user-friendly, since they are intuitive and require no understanding of complex mathematical or statistical algorithms and patterns. They provide a high degree of confidence in the findings of the exploration. As a result, visual data mining provides a faster way of data exploration and yields better results, especially in cases where automatic algorithms fail, such as when mining is performed either on extremely large databases, on databases with highly inhomogeneous and noisy data, or when little is known about the data and the exploration goals are vague. [5, 14]. Hoffman et al. [11] give an overview over existing visualisation techniques for single variate and multivariate data. An overview of some other techniques can be found in [14]. Applications of the visual data mining approach include, for example, analysis of telephone calling fraud [6], analysis of atmospheric data [16] and geo-referential statistical information mapping [1–3].

Site planning is the process of arranging structures and shaping the spaces between the structures within a given area. It places the objects in space and time and can concern a small cluster of houses, a single building and its grounds or even a whole community built in a single operation [15]. Site planning is a complicated process that needs large amounts of data in order to find the best placement for a new site. The basic problem in the process of site planning is the search for actual existing data and its analysis. There are several types of data that are needed to analyse and realise planning objectives: geographical data, textual data, images, cadastral data, etc. To analyse such varied types of data is very difficult. However, the analysis is easier to accomplish within a database consisting of metadata information for the planning data [8, 9].

Geographical data are highly multidimensional: they can have up to four dimensions of spatial and temporal information, which provide the measurement framework for all other dimensions. Because of this spatio-temporal dependency, the objects and relationships in geographical databases tend to be more complex than in non-geographical databases. Today geographical databases tend to contain other types of data than the traditional raster and vector formats, such as imagery and geo-referenced multimedia. Data mining in such heterogeneous complex data is very difficult, but mining in the metadata instead of the actual data can make the task of data exploration easier [7, 18]. In the last years the geodata market has been structured into metadata information systems and infrastructures on regional, national or international level, which enable the providers to describe and the users to find appropriate data [8, 9].

Metadata are usually defined as data about data. They describe the attributes and contents of an original document or work. The library card catalogue is a standard example of metadata: each card represents and leads the user to a much larger body of information, the book or other item catalogued. When applied to electronic resources, it refers to data in the broadest sense: datasets, textual information, geospatial information, images, graphics, music and anything else that is likely to appear in digital form [17].

Using automatic exploration techniques for interpretation of complex metadata is often not very effective. By using visualisation techniques instead of or alongside the automatic techniques, the understanding and interpretation of data behind the metadata can be improved. This gives the user a deeper insight into the data and its structures [18]. Göbel et al. [9] present an overview of applications for exploration of metadata. These applications usually provide some kind of keyword search, geographic search and occasionally also temporal search mechanisms. Search results are presented as textual result lists, which makes the comparison and interpretation of results difficult. The applications offer access to a large amount of geodata, but provide the user with only a very basic level of interactivity and control of the search.

In the next sections we present how the visual data mining approach was applied to site planning metadata within the INVISIP project.

2 Visual Data Mining within INVISIP

The project INVISIP will provide a technical platform based on the exploration of metadata as an aid to facilitate information access and data handling for the site planning process. One of several systems within INVISIP is the Visual Data Mining tool (VDM tool), which is the implementation of the visual data mining approach on the metadata of site planning data. The aim of the VDM tool is to improve the information retrieval process providing visual presentations and intuitive selection mechanisms.

The process of visual data mining in the metadata of geographical data used in the site planning process is shown in Figure 1. The problem can be defined as finding the optimal subset of all the existing geographical datasets for the particular site. The optimisation criteria for this subset are initially set by the user, e.g. finding the most up-to-date datasets, finding the datasets with the appro-

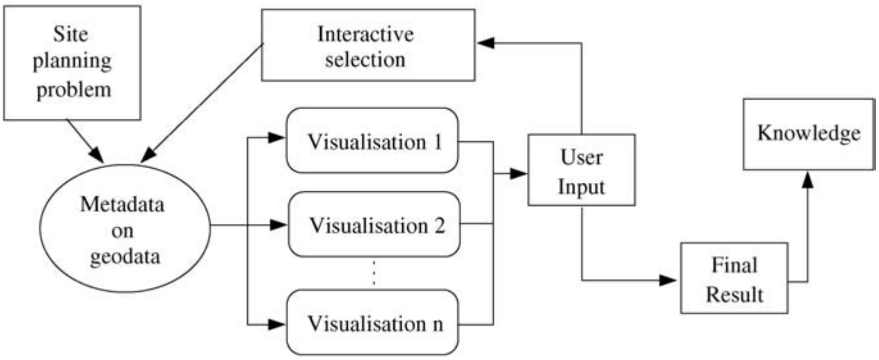


Fig. 1. Visual data mining process on metadata

priate resolution, etc. The user formulates these suitability criteria according to his specific discipline (traffic analysis, environmental aspects, etc.). In automatic data mining these criteria are usually expressed as logical statements, connected by either conjunctions or disjunctions. In the case of visual data mining, however, the user opens visualisations based on these criteria. Using different types of visualisations, the user can display one or more metadata attributes. He recognises patterns in open visualisations and bases his choice on a subset of items that he is interested in. His input is the desired selection of subsets. When this selection is performed, it is performed simultaneously in all open visualisations. The result of this selection is a restriction of the search space which may show new patterns to the user, some of which he might not have been aware of before. He can then repeat the whole process immediately on the selected subset of metadata records or add new visualisations based on additional criteria and repeat the interactive selection on all these old and new visualisations. By interacting with and operating on the visualisations the user has full control over the search activity. The process continues until the user is satisfied with the result, which represents a solution to his initial problem. The final result is a subset of metadata that contains the most suitable geographical datasets for the user's purpose.

This approach to metadata exploration can be used for confirmative analysis, where the user already has some idea what he is looking for and only needs to confirm a prior hypothesis, as well as for explorative analysis, where no prior hypothesis about the search goal exists. In the latter case, the user starts from scratch and forms the hypothesis along with the selection. He visualises some arbitrary attributes, recognises a pattern, adopts this pattern as one of his initial criteria and repeats the process from the beginning. The interaction between the user and the computer and the simultaneous selection of metadata in all visualisations form a dynamic process that can be repeated a number of times, until a satisfactory outcome has been found.

One of the tasks within INVISIP is to create a database of metadata. This database will represent the site planning data obtained in case studies, performed in the initial stage in all participant countries of the project (Sweden, Italy, Poland and Germany). At present this database is not complete, therefore a special (and smaller) test metadatabase has been created to evaluate the VDM tool. The items in this test metadatabase are multivariate records, whose attributes were created according to the ISO 19115 [12] standard for metadata for geographic information. The attributes of the metadata in the test metadatabase are described in Table 1. They represent a subset of the attributes specified by the ISO standard. In the final metadatabase of INVISIP other attributes from the standard will be used as well.

A practical example of an application of the visual data mining approach to this metadatabase is described at the end of the next section.

Table 1. Attributes of the test metadatabase

Metadata attribute	Description with some examples
Keyword	keyword describing the resource dataset (air pollution, climate, topology, ...)
Metadata file ID	humanly understandable identifier of the resource dataset
ID_main	unique numerical identifier of the resource dataset
Language	language of the resource dataset (English, German, Italian, Polish, Swedish, other, ...)
Reference date	reference date of the resource dataset
Theme	overall topic of the resource dataset (environment, agriculture, climatology, ...)
Update frequency	frequency of updating of data (biennially, daily, monthly, continually, never, ...)
Update level	level of updating (feature, dataset, ...)
Format	file format of resource dataset (ASCII, MapInfo, ARC/Info, ...)
Spatial representation type	spatial type of the resource dataset (vector, raster, image, ...)
Resolution	resolution of the resource dataset
Metadata level	does an entry represent a dataset or a collection of datasets or something else
Metadata parent ID	which other collection includes the resource dataset
Progress	development stage of dataset (planned, ongoing, completed, ...)
Fees	cost of the resource dataset
Order instructions	how to order the resource dataset
URL	the URL location of the resource dataset
Distribution media	how the resource dataset is distributed (via FTP, CD-rom, ...)
Online resource function	what is available online (order, download, request for additional information, ...)

3 The Visual Data Mining Tool

This section describes the main functionality of the VDM tool, its architecture and an example of its application.

3.1 The Functionality of the VDM Tool

The VDM tool is characterised by the visualisation techniques and the interaction functionalities.

The *visualisation techniques* at present include two different types of visualisations: visualisations of one attribute (a pie chart and a histogram) and visualisations of multiple attributes (a table and a parallel diagram). Other visualisations can be included in the VDM tool in the future.

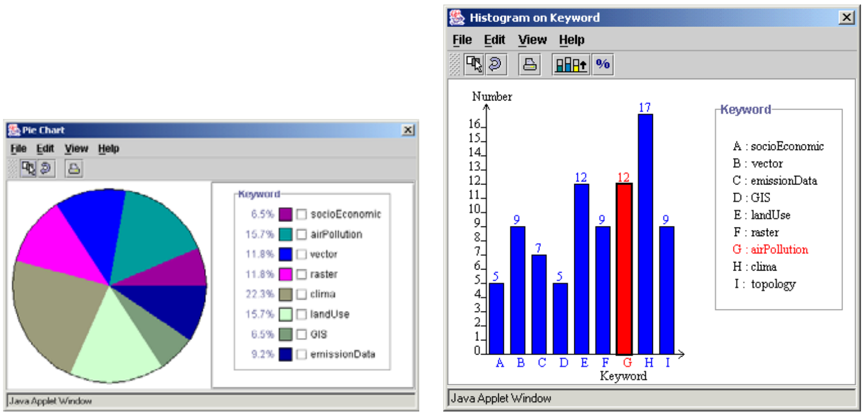
A brief description of each of the visualisations follows:

A *pie chart* (Figure 2(a)) shows the proportional size of values of one chosen attribute. It is useful when the user wants to recognise a significant element within the attribute.

A *histogram* (Figure 2(b)) shows the number of objects for each value of the chosen attribute. It is useful to recognise the distribution of data objects and can help to identify potentially suspicious objects which can be removed from further analysis by appropriate selection.

A *table visualisation* allows the user to choose one or several attributes and visualise them in a table of values. The columns represent the metadata attributes and the rows the data objects. It is not a graphical visualisation but is nevertheless useful to display a large number of attributes when data reduction has already been performed by previously using some other visualisation.

A *parallel diagram* or *a parallel coordinates plot* (Figure 3) maps the attributes of the dataset onto vertical axes. Each data object in the dataset is



(a) A pie chart

(b) A histogram

Fig. 2. Two visualisations of one attribute

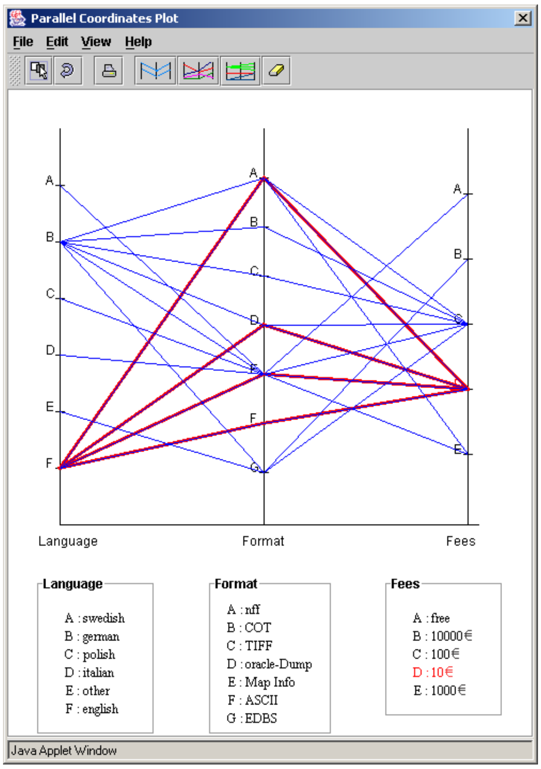


Fig. 3. A parallel diagram

represented as a continuous piecewise linear line connecting the axes. The line intersects the vertical axes at the points that correspond to its attribute values. Since the line representing an object connects different attributes, it is necessary to select at least two attributes for a non-trivial plot.

The VDM tool has two different kinds of *interaction functionality*: the interaction between a single visualisation and the user and the interaction among different visualisations.

The interaction between a visualisation and the user enables the user to explore the content of the visualisation and to extract graphically the selected subset of metadata. The possibility of a graphical selection of metadata exists in all the visualisation techniques, but varies according to the type of each technique. It is based on the selection of graphical entities in the visualisation. Each graphic entity is linked to the value of the attribute which it represents. Graphical entities in question can be angular segments of a pie chart, bars of a histogram, rows of a table or lines in a parallel diagram. The values of the attributes are shown in the legend next to the visualisation. The user can select a desired subset of objects by clicking on the graphical entities that represent their attribute values or by choosing one or several values in the legend. A special type of graphical selection is implemented in the parallel diagram. Unlike in the other visualisations the polygonal lines represent the correlation among different attributes rather than one attribute value only. Figure 3 shows an example: the selected polygonal lines in red show the correlation between the specified cost of the datasets (10€) with their format and their language. The datasets with this cost can be obtained in four different data formats, but they are all in English.

The second kind of interaction helps to discover correlation among graphic entities represented in different visualisations. All the visualisations are interconnected according to the concept of *brushing and linking*. Brushing is an interactive selection process, while linking connects the selected data from the current visualisation to other open visualisations. If the user has several different visualisations open (of one type or of more different types) and decides to perform a selection of objects in one of them, the graphical entities that represent this selection and correspond to the same subset of selected data objects are highlighted in each visualisation, providing a better visual impression. When the selection is performed, all other graphical entities disappear from each open visualisation.

3.2 The Architecture of the VDM Tool

The VDM tool is designed as a Java applet in order to easily handle web-based explorations. It consists of three main components: the data manager connecting the VDM tool to different resources of metadata, the control panel which integrates all components and the visualisation wrapper which provides a common template for the different visualisations.

The *data manager* handles input and output of data. It is based on a table that contains all metadata that can be visualized. The data connection implemented is based on ODBC. The VDM tool is therefore open to integrate other database managers such as Oracle, SQL server and so on.

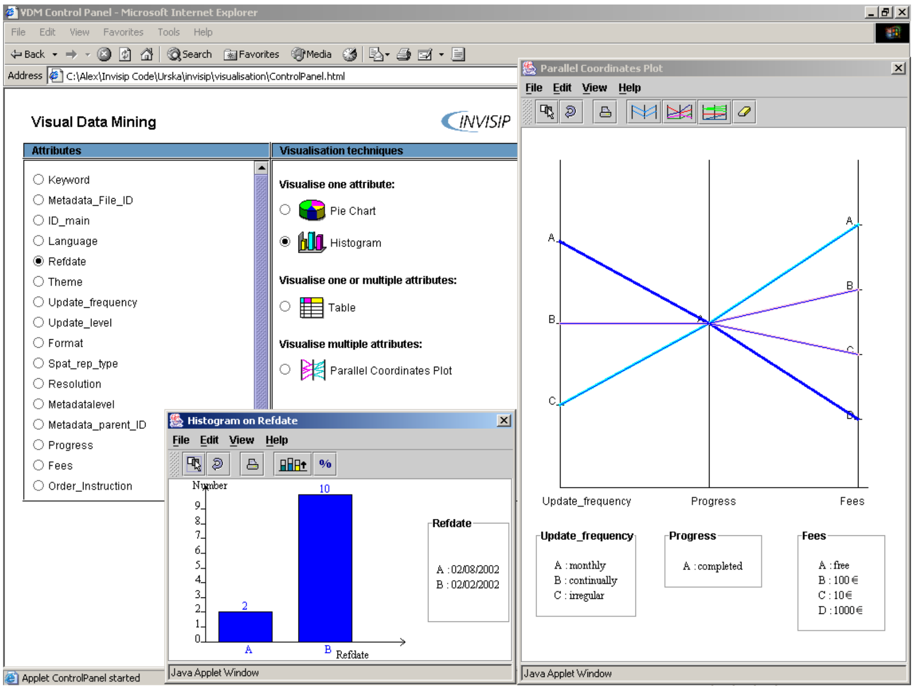
The *control panel* is the main component of the VDM tool, providing a Graphical User Interface (GUI) as shown in the background of Figure 4. The left side of the control panel shows the list of metadata attributes, while the right side shows available visualisations. The control panel activates the visualisations of selected attributes and manages the general layout of the different visualisations.

All visualisation techniques are based on the *visualisation wrapper* Java class. It is an abstract class that all visualisations extend and provides the interface between the visualisations and the control panel, as well as functionalities to draw and to update the graphs contained in the wrapper. It is also responsible for the look & feel (colours, character fonts, etc.) of all visualisations and for the common characteristics such as toolbar and menu.

3.3 An Application Example

This section describes an application example of the VDM tool during the data acquisition phase.

Let us suppose that the goal of the user is to look for datasets that are cheap, complete and continually updated. Using the VDM tool, the scenario shown in Figure 4 can be generated. It shows different visualisations activated by the user on several metadata variables: a parallel diagram of the update frequency, the progress and the cost, and a histogram of the reference date. From these visualisations the user can extract useful information about the available geographical datasets. The parallel diagram suggests that all datasets are complete, but that the datasets available for free are not updated frequently. The histogram shows that some of the datasets are updated in February and others in August. Assuming the user would like to purchase the most up-to-date datasets, he selects those produced in August. All active visualisations are updated according to this selection as shown in Figure 5. From the resulting parallel diagram in Figure 5 it becomes obvious that the most current datasets are updated continually and cost 100€. Based on this result the user can now decide if these datasets are suitable for him, e.g. if their cost is still low enough for him to afford. If the resulting datasets do not fulfil his criteria, he can start the analysis from the beginning with different initial conditions.



VDM Control Panel - Microsoft Internet Explorer

Address C:\Alex\Invisip Code\Urska\invisip\visualisation\ControlPanel.html

Visual Data Mining

Attributes

○ Keyword
○ Metadata_File_ID
○ ID_main
○ Language
● Refdate
○ Theme
○ Update_frequency
○ Update_level
○ Format
○ Spat_rep_type
○ Resolution
○ Metadatalevel
○ Metadata_parent_ID
○ Progress
○ Fees
○ Order_Instruction

Visualisation techniques

Visualise one attribute:
○ Pie Chart
● Histogram

Visualise one or multiple attributes:
○ Table

Visualise multiple attributes:
○ Parallel Coordinates Plot

Histogram on Refdate

Refdate
A : 02/08/2002
B : 02/02/2002

Parallel Coordinates Plot

Update_frequency
A : monthly
B : continually
C : irregular

Progress
A : completed

Fees
A : free
B : 100 €
C : 10€
D : 1000 €

Applet ControlPanel started

Fig. 4. A screen shot of what the tool looks like

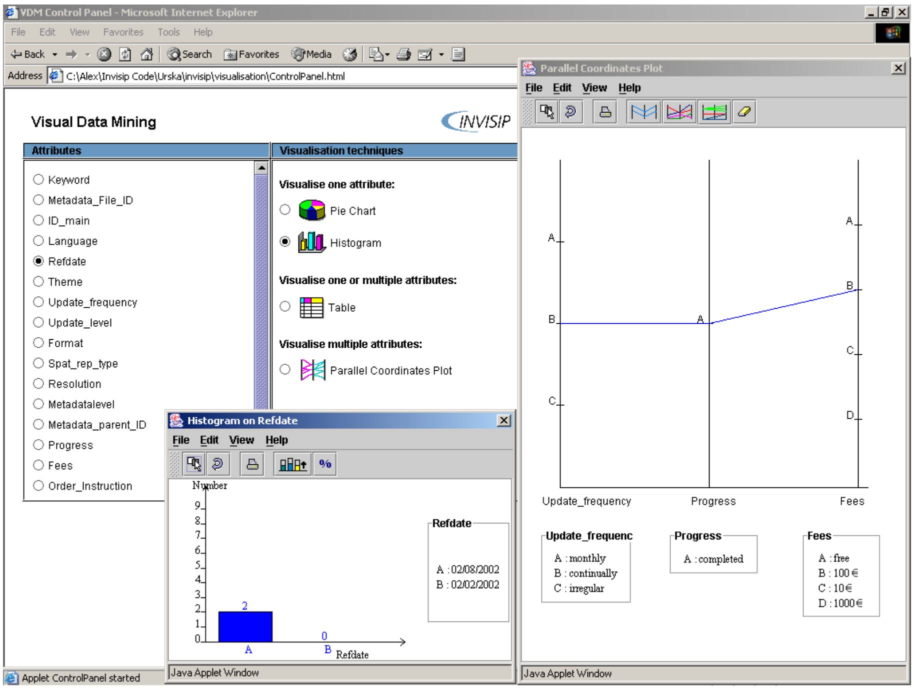


Fig. 5. The screen shot of the tool after a selection task

4 Conclusion

This paper presents the research activity performed in INVISIP, a project funded by European community that aims to assist users with different skill level during the site planning process. The first part of the project focuses on problems related to the task of data acquisition. It is a critical task, because the choice of wrong data can seriously affect the success of the whole process.

An approach based on visual data mining is proposed for exploration of geographical metadata in order to increase the user awareness about which geographical data is available. It allows the user to perform both the confirmative and the explorative analysis, helping him to find a compromise between data that he needs and data that is available. By using visualisation as a communication channel, this approach provides a friendly way to analyse metadata even for a not very skilled user.

A first demonstrator tool has been developed. It is general enough to be used in a different context than INVISIP. One of possible other applications is to configure it as a tool to explore the web-catalogue of geo-data vendors instead of the INVISIP metadata repository.

In the next phase of the project we plan to include new visualisation techniques and selected automatic data mining techniques into the tool. Other different types of visualisations could be developed, such as those that are linked to either spatial or temporal distribution of geographical data. Automatic data mining techniques suitable for application on geographical data and metadata (such as for example statistical methods, clustering, mining using association rules, methods for analysing incomplete data, etc.) will be analysed and possibly integrated into the tool. The final aim is to provide an instrument that will support the analysis of the geographical data involved in the site planning process as well as its metadata. In this way the tool can help the user throughout the whole site planning process.

5 Acknowledgments

The authors would like to acknowledge the support of the European Commission, IST Program. They thank the INVISIP partners who contributed to this study: D'Appolonia and Agricultural University of Krakow. A special thanks goes to the other members of the programming team who have contributed to the development of the VDM tool: Kajetan Hemzaczek, Igor Podolak (Agricultural University of Krakow) and Isabelle Colin (Istituto di Matematica Applicata e Tecnologie Informatiche).

6 Contributors List

Riccardo Albertoni, Alessio Bertone and Urška Demšar (authors' names are given in alphabetical order) have contributed to the development of the VDM tool under the supervision from Monica De Martino and Hans Hauska. This paper was written as a joint contribution.

References

1. Andrienko G. and Andrienko N., Intelligent Visualisation and Dynamic Manipulation: Two Complementary Instruments to Support Data Exploration with GIS, *Proceedings of AVI98: Advanced Visual Interface Int. Working Conference*, pages 66-75, ACM Press, 1998.
2. Andrienko G. and Andrienko N., Knowledge-Based Visualisation To Support Spatial Data Mining, *3rd International Symposium, IDA-99 Proceedings*, Lecture Notes in Computer Science 1642:149-160, Springer-Verlag, Berlin, 1999.
3. Andrienko G. and Andrienko N., Interactive Maps for Visual Data Exploration, *International Journal of Geographic Information Science*, 13(4):355-374, 1999.
4. Ankerst M., *Visual Data Mining*, PhD Thesis, Ludwig-Maximilians-Universität, München, 2001.
5. Buono P., Costabile M. F. and Lisi F. A., Supporting Data Analysis Through Visualisations, *Proceedings of the International Workshop on Visual Data Mining, in conjunction with 2nd European Conference on Machine Learning (ECML'01) and 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, pages 67-78, Freiburg, 2001.

6. Cox K. C., Eick S. G., Will G. J. and Brachman R. J., Brief Application Description, Visual Data Mining: Recognising Telephone Calling Fraud, *Data Mining and Knowledge Discovery*, 1(2):225-231, Kluwer Academic Publisher, 1997.
7. Gahegan M., Wachowicz M., Harrower M. and Rhyne T. M., The Integration of Geographic Visualisation with Knowledge Discovery in Databases and Geocomputation, *Cartography and Geographic Information Science*, 28(1):29-44, 2001.
8. Göbel S., Ranking Mechanisms in Metadata Information Systems for Geospatial Data, *Proceedings of the EOGEO 2002 workshop for developers of Geospatial data services over the Web*, Ispra, 2002.
9. Göbel S. and Jasnoch U., Visualisation techniques in metadata information systems for geospatial data, *Advances in Environmental Research*, 5(4):415-424, Elsevier Science, 2001.
10. Hand D., Mannilla H. and Smyth P., *Principles of Data Mining*, MIT Press, Cambridge, Massachusetts, 2001.
11. Hoffman P.E. and Grinstein G.G., A Survey of Visualizations for High-Dimensional Data Mining. In: Fayyad U., Grinstein G. G. and Wierse A., editors, *Information Visualisation in Data Mining and Knowledge Discovery*, pages 47-82, Morgan Kaufmann Publishers, San Francisco, 2002.
12. ISO 19115 Final Draft: International Standard on Metadata for Geographic Information, Status: approval stage, http://www.iso.org, 2002.
13. Keim D. A. and Kriegel H. P., Visualisation Techniques for Mining Large Databases, A Comparison, *IEEE Transactions on Knowledge and Data Engineering*, 8(6):923-938, 1996.
14. Keim D. A., Müller W. and Schumann H., Visual Data Mining, State of the art report, *Eurographics 2002*, European Association for Computer Graphics, Saarbrücken, 2002.
15. Lynch K. and Hack G., *Site planning*, MIT Press, Cambridge, Massachusetts, 1984.
16. Macedo M., Cook D. and Brown T. J., Visual Data Mining in Atmospheric Science Data, *Data Mining and Knowledge Discovery*, 4(1):69-80, Kluwer Academic Publisher, 2000.
17. Milstead J. and Feldman S., Metadata: Cataloguing by any other name . . ., *Online*, http://www.onlinemag.net/, 23(1), Information Today Inc., 1999.
18. Nocke T. and Schumann H., Meta Data for Visual Data Mining, *Proceedings Computer Graphics and Imaging, CGIM 2002*, Kauai, Hawaii, USA, 2002.