Automated Extraction and Retrieval of Metadata by Data Mining

- A Case Study of Mining Engine for National Land Survey Sweden

Author: Zheng Dong
Date: March 2010
Supervisor: Anders Östman
Examinor: Julia Åhlén
Abstract

Metadata is the important information describing geographical data resources and their key elements. It is used to guarantee the availability and accessibility of the data. ISO 19115 is a metadata standard for geographical information, making the geographical metadata shareable, retrievable, and understandable at the global level. In order to cope with the massive, high-dimensional and high-diversity nature of geographical data, data mining is an applicable method to discover the metadata.

This thesis develops and evaluates an automated mining method for extracting metadata from the data environment on the Local Area Network at the National Land Survey of Sweden (NLS). These metadata are prepared and provided across Europe according to the metadata implementing rules for the Infrastructure for Spatial Information in Europe (INSPIRE). The metadata elements are defined according to the numerical formats of four different data entities: document data, time-series data, webpage data, and spatial data. For evaluating the method for further improvement, a few attributes and corresponding metadata of geographical data files are extracted automatically as metadata record in testing, and arranged in database. Based on the extracted metadata schema, a retrieving functionality is used to find the file containing the keyword of metadata user input. In general, the average success rate of metadata extraction and retrieval is 90.0%.

The mining engine is developed in C# programming language on top of the database using SQL Server 2005. Lucene.net is also integrated with Visual Studio 2005 to build an indexing framework for extracting and accessing metadata in database.

Keywords: data mining, geographical metadata, ISO 19115, Lucene.net
Acknowledgement

I would like to thank my supervisor, Anders Östman, for all his wise suggestions on my thesis during the one whole year, as well as giving me this chance to take part in the project of metadata mining computation.

At the same time, I want to thank my examiner, Julia Åhlén, for her evaluation on the result and valuable advices for further improvement.

I would like to thank best friend, Junjun Yin, and all the geomatics classmates of Högskolan in Gävle, for their support and encourage when I suffered the difficulties of computation.

Jesper M. Paasch of National Land Survey provided me the basic information of data environment and great volume of testing data; I also want to thank him for his help.

At last, the acknowledge goes to my parents, daughter, for your patience during my master time in Sweden.
# Table of Contents

Abstract ......................................................................................................................... 2  
Acknowledgement ........................................................................................................ 3  
Table of Contents ......................................................................................................... 4  
List of figures ................................................................................................................ 5  
List of tables .................................................................................................................. 6  

1. Introduction ........................................................................................................... 7  
1.1 Problem and Consequence .................................................................................... 7  
1.2 Historical Evolution of Automatic Extraction of Geographic Metadata ............ 8  
1.3 Existing Methods and Researches on Automatic Extraction of Geographical Metadata .. 10  
1.4 Purpose ................................................................................................................... 14  
1.5 Structure of Thesis ................................................................................................. 14  

2. Literature Review .................................................................................................. 15  
2.1 Data Mining .......................................................................................................... 15  
2.1.1 Motivation and Definition of Data Mining ......................................................... 15  
2.1.2 Classification of Data Mining according to Data Types .................................... 17  
2.2 Geographical Data and Geographical Metadata ................................................... 19  
2.2.1 Geographical Data ............................................................................................. 19  
2.2.2 Definition of Metadata and Geospatial Standard ............................................. 20  
2.3 Lucene.net ............................................................................................................. 22  
2.4 Microsoft SQL Server 2005 and C# language ...................................................... 23  

3. Methodology ........................................................................................................ 23  
3.1 Description of Data Environment and Initial Architecture................................. 24  
3.1.1 Description of Data Environment ...................................................................... 25  
3.1.2 Initial Architecture ............................................................................................. 26  
3.2 Elaboration of Model and Feature Extraction ....................................................... 26  
3.2.1 Data Mining Model based on Data Environment and Data Entities ................. 27  
3.2.2 Identification of Metadata Elements from Different Categories of Data Formats .. 29  
3.3 Construction of Software System ......................................................................... 32  
3.3.1 Creation of SDE .................................................................................................. 32  
3.3.2 Workflow ............................................................................................................ 32  
3.4 Evaluation Phase ................................................................................................... 33  

4. Result of Interface, Efficiency and Success rate .................................................. 37  
4.1 Interface on Webpage ............................................................................................ 37  
4.2 Testing 1: Efficiency of Metadata Indexing ......................................................... 37  
4.3 Testing 2: Success Rate and Influence Facts ....................................................... 38  
4.3.1 Success Rate Restricted in Small Domain ......................................................... 38  
4.3.2 Influence Facts .................................................................................................. 40  

5. Discussion .............................................................................................................. 41  

6. Conclusion .............................................................................................................. 42  

Reference .................................................................................................................... 43
List of figures

Figure 1: Problems of NLS: automated extraction of metadata within LAN cloud ..... 8
Figure 2: Main components of the data mining system ........................................... 17
Figure 3: Metadata packages of ISO 19115 (source: Taussi, 2007) .......................... 21
Figure 4: The developing process of mining method integrated with the Rational Unified Process ........................................................................................................ 24
Figure 5: Basic architecture of the geographic metadata mining method for NLS ... 27
Figure 6: Data mining model based on data environment and data entities .......... 28
Figure 7: Workflow of metadata indexing in database ............................................. 34
Figure 8: Workflow of metadata retrieving ............................................................... 35
Figure 9: The home page of retrieval website ......................................................... 37
Figure 10: Linear change for efficiency of metadata indexing in Testing 1 .......... 38
Figure 11: Success rate for different types of geographical data in NLS ............ 39
List of tables

Table 1: The main metadata contents from vector files basing on file formats (source: Manso et al, 2004) ........................................................................................................ 11
Table 2: The main metadata contents from raster files basing on file formats (source: Manso et al, 2004) ........................................................................................................ 13
Table 3: Core metadata elements and requirements of ISO 19115 (ISO, 2003) .......... 21
Table 4: Optional metadata elements of ISO 19115 supporting automated extraction (Source: Taussi, 2007) ........................................................................................................ 30
Table 5: The efficiency of metadata indexing for Testing 1 ........................................ 37
Table 6: Success rates basing on different metadata elements for Testing 2 ............ 40
1. Introduction

Metadata is used to describe resources of geographical data. It plays a vital role in ensuring the integrity, availability and accessibility of geographical data. The ultimate goal of metadata is to standardize the collection and pre-processing of geographic data to describe what exist in data repositories and which operations can be implemented for possible analysis in the future. From the perspective of resource management, the importance of metadata is self-evident.

1.1 Problem and Consequence

Infrastructure for Spatial Information in Europe (INSPIRE) is a framework with the objectives for data sharing, exchange and re-using. For standardization of metadata in Europe, INSPIRE released a metadata implementing rule based on ISO 19115 and ISO 19119 (INSPIRE, 2007a). All the member states of the EU should provide metadata of their geographic datasets according to this implementing rule before December, 2010. As a member state of EU, the Swedish data providers need to populate a newly-built metadata base with the metadata from geographical data resources. National Land Survey of Sweden (NLS), one of geographical data providers, should create a metadata base and convert the metadata of existing data into this metadata base to comply with the INSPIRE directive.

According to a survey by Östman et.al (2009), 25% of the mandatory metadata elements at the NLS have been deposited in the created-already geographical metadata base which can be converted into INSPIRE metadata base. Additional 67% of mandatory metadata and 23% of optional metadata are stored on Local Area Network (LAN), within hundreds of computers distributed in different observation stations all over the Sweden. This part of metadata elements can be retrieved manually yet have not been extracted and sorted into geographical metadata base. This situation can be treated as a LAN cloud in which the extraction of metadata has not been completed. It is very time-consuming to find these geographical data files and extract metadata elements manually. Therefore, automated extraction of this part
of metadata elements is meaningful and has a very large economic impact. The remaining 8% of mandatory metadata requires a manual investigation and is hard to be extracted automatically. So this part is not involved in the object of this method.

The situation of geographical data and metadata on the NLS involved in this study can be described in Figure 1 as following. The main problem is how to extract the 67% mandatory metadata and 23% optional metadata within LAN cloud automatically.

![LAN cloud of NLS](image)

Figure 1: Problems of NLS: automated extraction of metadata within LAN cloud

1.2 Historical Evolution of Automatic Extraction of Geographic Metadata

During the 1980s, the geographical metadata were often structured as class definitions of geographical attributes and class hierarchy (Sen, 2004). When the metadata are organized in this structure, the corresponding simple metadata schema could be treated as embryonic specification of automated metadata extraction in the follow-up development of database.

During the 1990s, the sharp development of information technology and
tremendous growth of digital information resources motivate the emergence of big data repositories and data warehouse as well as an infrastructure of tools. The simple metadata schema fails in an attempting to effectively manage metadata within the complex data repositories, since they originate from different data resources with different data types and data formats. A specific development environment with integration of metadata management tools, described as a Software Development Environment (SDE), is introduced into the automatic extraction of metadata within complex data repositories for specific applications (Sommerville, 1992). Some robust language are introduced progressively into SDE, especially Extensible Markup Language (XML) and Geography Markup Language (GML), mitigating greatly the barrier to the generation and extraction of geographical metadata.

At this juncture, the concept of data mining was introduced into automatic metadata extraction in SDE. Basic mining method can achieve the functionality of acquiring simple metadata schema by analyzing data structure in database; making extra efforts, the advent mining can dig out data relationship, characteristics and attributes of complicated data (Taussi, 2007). Graph analysis is one of the advanced data mining methods, which is used to find the critical paths and nodes of spatial topological structures based on graph theory (Jungnickel 2005).

From the beginning of 21th century, the point of focus for geographical metadata has shifted to the creation of geographical metadata standards. Due to the high-cost and limited-availability of the geographical metadata, sharing metadata has already been a both acceptable and adoptable way to utilize the original data resource. Moreover, in order to share the metadata not only at the regional level, but also at the national level and even in the global level, making the international geographical metadata standard become a necessary. Among a variety of standards, ISO 19115 (ISO, 2003) standard is an international geographical metadata standard accepted all around the world.

The geographical metadata standard is important for the choice of metadata elements for geo-metadata extraction. Only by depending on these standards, the evaluation of success rate for geo-metadata extraction has practical significance.
Without metadata standards, the usefulness of extraction will be restricted or even been suspicious. Therefore, the geographic data standards must be tightly integrated into the metadata extraction process, to ensure the consistency and accuracy.

1.3 Existing Methods and Researches on Automatic Extraction of Geographical Metadata

So far, there are a large number of related researches and tools developed to deal with generation and extraction of geographical metadata. ArcCatalog is an automated metadata creation tool for shapefiles developed by ESRI, based on CSDGM and ISO 19115 metadata standards. Metadata Validation Service is an online implementation of metadata parser for Federal Geographic Data Committee (FGDC) standard, working on web browser and internet connection.

The University of Zaragoza developed a tool for automated metadata generating and editing, CatMDEdit Metadata Editor (Zarazaga-Soria et al. 2003), for Spanish National Spatial Data Infrastructure. At the same time, a research was carried out also in the University of Zaragoza, aiming at identifying spatial metadata elements from different categories of data format (Manso et al. 2004). The main specification and identification are listed below.

**Vector file format**

The vector files being considered are generated by software like Computer Aided Design (CAD) and GIS. According to the research of Manso et al (2004), possible formats are: ADF and E00 (Coverage file format of ArcInfo ESRI), SHP (Shape format of ArcView ESRI), VEC (vector file of Idrisi), BIN (binary format of Digi: Digi21), DWG and DXF (format of Autocad: Autodesk), MIF and TAB (map formats of MapInfo), and DGN (Design file format of Bentley and Intergraph). For these formats, the main metadata elements which can be acquired are: the coordinate information of bounding box rectangle, the number and name of layers, the number of features, and the number and name of entities for each feature (points, circles, etc). In addition, other metadata properties of spatial reference system in vector files are
horizontal units, projection parameters, datum and Ellipsoid. However, 3 formats of vector files (DXF, DWG and BIN) do not have fields of information of headings or structures of archives to store and present these 4 metadata elements. Different from all other formats, SHP files use a Well-Known Text structured PRJ file (WKT, 2001) to present spatial reference system. The main contents relative to metadata which can be obtained from vector file formats are shown in following Table 2.

Table 1: The main metadata contents from vector files basing on file formats (source: Manso et al, 2004).

<table>
<thead>
<tr>
<th>Format</th>
<th>Title</th>
<th>Bound Box</th>
<th>Number of layers</th>
<th>Name of layers</th>
<th>Numbers of different features</th>
<th>Name and number of features</th>
<th>Horizontal units</th>
<th>projection</th>
<th>Datum</th>
<th>Ellipsoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>E00</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>ADF</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DGN</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DXF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SHP</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>MIF</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TAB</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DWG</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>VEC</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIN</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Raster file format**

Common formats of GIS raster files at NLS are: GIF (Graphic Interchange Format), JPEG (Joint Photographic Experts Group), TIFF (Tagged Image File Format), BMP (bitmap format), PNG (a portable network graphic format), PCX and IFF (Interchange File Format). The metadata elements which can be obtained from these formats of raster files are as following: Width and height (expressed in number of pixel), the bounding box, the bit depth for each pixel, the number of bands, and pixel resolution. Among these formats, the PNG and TIFF files can provide metadata
of author, creation time, abstract as well as a source list. A World File is applied on these raster files to specify a transformation to a world coordinate system, but the disadvantage is lacking the standard information of spatial reference system.

In order to eliminate this limitation of specifying geographic reference in separate files, some new formats are developed which includes spatial reference system. They are GTIFF (Tagged Image File Format focusing on Geographical reference system), JPEG2000 (format of standardized compress file according to ISO 15444), GeoJP2 (JPEG2000 format related to the spatial reference system stored as GTIFF by Mapping Science), MrSID (raster format based on wavelets by LizarTech), ECW (raster format based on wavelets by Ermapper), INGR (Intergraph raster format), and NITF (North American standard of transfer of images of NIMA). In addition to the general metadata elements mentioned for the raster file formats, these formats also have useful information about horizontal units of pixel and projection information (datum, projection parameters, ellipsoid).

The formats of digital terrain models in raster grid can be: ADF (ArcInfo Data Format GRD of ESRI), GRD (Ascii Grid format of ESRI and surfer grid file format of GoldenSoft), DEM (Digital Terrain Elevation), DTE (Digital Terrain Elevation), DT0 (Digital Terrain Model of American), HGT (Digital Elevation Models of NASA), BII, BIP and BSQ (raster interchange format in MapInfo). The metadata elements they can provide are similar to aerial photographs, orthogonal photographs and scanned cartography, but also new elements of max and min statistics are added.

For the high-resolution satellite image, the formats can be: IMG (Erdas Image format), PIX (PCI and Geomatics image database format), ERS (image format of Ermapper), NOAA1B (raster format for low resolution images from NOAA satellite), EOAST (raster format for Earth Observatory Satellite), F-EOAST (format for LandSat 7 and IRS satellite). The metadata elements that can be obtained are:
Wide and height (expressed in number of pixel), the bounding box, the bit depth for each pixel, the number of bands, pixel resolution, horizontal units, Projection information, data type for store pixel, acquisition date, satellite platform, statistical parameters and other parameters which are difficult to extract as metadata (Manso et
Table 2: The main metadata contents from raster files basing on file formats (source: Manso et al, 2004).

<table>
<thead>
<tr>
<th>Format</th>
<th>W/H</th>
<th>Bound Box</th>
<th>Pixel resolution</th>
<th>Bits/Pixel</th>
<th>Bands/dimension</th>
<th>Max, min statistical</th>
<th>Other statistics</th>
<th>Horizontal units</th>
<th>Projection</th>
<th>Datum</th>
<th>Ellipsoid</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIF</td>
<td>X</td>
<td>X1</td>
<td>X1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JPEG</td>
<td>X</td>
<td>X1</td>
<td>X1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIFF</td>
<td>X</td>
<td>X1</td>
<td>X1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMP</td>
<td>X</td>
<td>X1</td>
<td>X1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFF</td>
<td>X</td>
<td>X1</td>
<td>X1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PNG</td>
<td>X</td>
<td>X1</td>
<td>X1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCX</td>
<td>X</td>
<td>X1</td>
<td>X1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTIFF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JPEG2000</td>
<td>X</td>
<td>X1</td>
<td>X1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeoJP2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MrSID</td>
<td>X</td>
<td>X12</td>
<td>X12</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X2</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ECW</td>
<td>X</td>
<td>X12</td>
<td>X12</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X2</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ADF</td>
<td>X</td>
<td>X</td>
<td>X4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X4</td>
</tr>
<tr>
<td>GRD ESRI</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X3</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>GRD surfer</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEMUSGS</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>MicroDEM</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>DTE</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOQ2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>DTED</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>IMG Erdas</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>E00 grd</td>
<td>X</td>
<td>X</td>
<td>X4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X4</td>
</tr>
</tbody>
</table>
1 In case of a world file existed
2 If version file contains this information.
3 If all the data file has been read and its values may be computed.
4 If all the file sections are presented.

### 1.4 Purpose

The aim of this project is to develop and evaluate a suitable method of data mining to extract geographical metadata from the geo-referenced materials of NLS on the LAN. This method should be able to index basic attributes of geographical files and corresponding metadata into metadata schema automatically according to the categories of data formats existed in NLS. After acquiring the metadata schema, it needs support searching metadata according to the input keyword (in English or Swedish). As a first study for further improvement in future, four data entities existed on the LAN of NLS are considered: web data, spatial data, document data and time-series data. The web data will not be dwelled on deeply but given an appropriate development environment. For the left 3 entities, all the metadata elements supporting automatic extraction will be specified. However for simplifying the actual encoding and testing, the extracted metadata elements will be restricted to several ones according to a few typical data formats.

Overall, the ultimate goal is to complete an extraction and retrieving functionality for geographical metadata on the LAN, and achieve a success rate of at least 50%.

### 1.5 Structure of Thesis

The remainder of this thesis is organized as following:
Chapter 2 gives the theoretical background of geographical metadata mining. Definition and classification of data mining methods are introduced as the first part. After that, metadata are interpreted simply with focusing on ISO 19115 geographical metadata standards. Finally, a simple description of related software platform and tools (Lucene.net, c# language and SQL Server 2005) is given out.

Chapter 3 is concentrating on the methodology of this thesis. The whole process of data mining is divided into 4 phases based on Rational Unified Process (RUP).

In chapter 4, the final results of this method are exhibited respectively on the interface, efficiency of indexing and success rate. The facts influencing the success rate are also described.

The discussion and collusion of the methods developed are shown in chapter 5 and chapter 6, where further thinking for subsequent development and optimization are presented.

2. Literature Review

2.1 Data Mining

2.1.1 Motivation and Definition of Data Mining

Along with the development of information technology and computer hardware technology, especially since mid-1980s, abundant amount of digital data has been collected and stored into various kinds of information repositories. However, if lacking of powerful data analysis tools, the human ability cannot afford to acquire an overall and deep comprehension to the growing amount of data and extract valuable information embedded in the data repositories. Han and Kamber (2006) described this situation as “data rich but information poor”. In order to promote efficient extraction and turn these rich data into valuable information even knowledge, the data analysis tools are therefore calling for a systematic progress, which is vividly described as data mining.
Han and Kamber (2006) give the following definition of data mining:

**Data mining** is the process of uncovering interesting knowledge and patterns from large amounts of observational data stored in databases, data warehouses, World Wide Web, or other information repositories.

It is a promisingly interdisciplinary field that has been evolved from the intersection of vast fields such as machine learning, data management, statistics, pattern recognition, artificial intelligence, spatial data analysis, business and economics (Hand et al., 2001). To accommodate multiple kinds of numerical topics and materials, the repository types include not only the major databases such as relational databases, data warehouse and transactional databases, but also advanced database system and World Wide Web, which can handle special data types.

Broadly speaking, data mining is regarded as a synonym in terms of Knowledge Discovery from Data (KDD) (Demšar, 2004). The process of KDD can be divided into several sequential steps: data selection, data pre-processing (data cleaning, data integration and data transformation if necessary), pattern extraction and evaluation by performing data mining with presenting the discovered patterns and knowledge as the last step (Hand et al., 2001). Standing on this point of view, 6 components are involved in a data mining system, coupling with the steps mentioned (Han and Kamber, 2006):

1. Numerical data repositories, including database, data warehouse, World Wide Web and others where data selection, data cleaning, data transformation and data integration could be applied.
2. Database servers and data warehousing server are used to store the cleaned data after pre-processing.
3. Knowledge base which is treated as the basis for search and evaluate the novelty, understandability and usefulness of the patterns discovered in the data.
4. Data mining engine that takes responsibility of major functionalities and manipulations of data mining.
5. Pattern evaluation which closely cooperated with the data mining engine, even sometimes integrated with it directly, so as to help identify and find the valuable
and useful patterns more easily by applying some interestingness operations.

(6) User interfaces, which allow the users to interact with the data mining system, visualize and present the interesting patterns discovered.

2.1.2 Classification of Data Mining according to Data Types

With respect to the fast developments of data mining, there are a lot of new fields emerging based on traditional data mining. In relation to this complicated situation, it is a controversial task to specify an accurate classification. By referring to different viewpoints from researchers, data mining about geography can be classified roughly into: spatial data mining, document mining, time-series data mining, web mining, visual data mining and distributed data mining (Han and Kamber, 2006; Hsu, 2003; Wang et al., 2006).

![Figure 2: Main components of the data mining system](Source: Han and Kamber, 2006)
**Spatial Data Mining**

Spatial data mining is the integration of data mining and spatial database analysis, aiming to understand spatial data, discover spatial relationship, identify relationship between spatial data and non-spatial data, and manipulate the spatial databases (Han and Kamber, 2006). It contains special data objects such as points, lines and polygons, which are used to present the spatial shape and location (Shekhar, Zhang and Huang, 2005).

**Document Mining**

Document mining is interpreted as a progress of extracting meaningful and novel knowledge and patterns from unstructured or semi-structured text documents (Wang et al., 2006). As described, text mining is similar to the technology of Information Retrieval (IR), which is developed to help query valuable information. Three major methods are used: keyword-based association approach, the classification approach and document clustering approach.

**Web Mining**

Web mining technology is used to access to the information resource of World Wide Web (WWW) effectively, and extract interesting and implicit patterns. However, web information distinguishes itself from other information recourses because of its high dynamics, complexity and imperceptibility, huge volume, wide distribution and usefulness confined with small part of the whole pages. These particularities present significant challenges to the flexibility and functionality of web mining technologies. According to Madria, et al. (1999), web mining technologies can be categorized into: web content mining, web structure mining and web usage mining.

**Visual Data Mining**

Visual data mining is an integration of data mining technology and visualization technology. It makes use of human ability of visual sense to discover hidden patterns which was presented and influenced by dynamic parameters (Thearling et al., 2001). 2-D, 3-D graphics and even table forms of data are employed in the visualization technologies to simplify the process of data mining.

**Distributed Data Mining**

Distributed data mining (DDM) refers to the mining of data distributed at
heterogeneous sites, which means different places or different physical locations (Wang et al., 2006). The common solutions for distributed data mining are applying algorithms on a database or data warehouse. However from a global sight, this solution often seems be not efficient enough. In order to gain more feasible solutions, the current development of distributed data mining has turned to develop partial data model by local analysis as the first step, then combine the partial data models to a global data model by creating and utilizing some new algorithms (Hsu, 2003). It is widely adopted in many fields, such as credit card management and Wi-Fi technologies (Battiti et al., 2003). The data mining based on the local area network can be regarded as one form of distributed data mining.

*Time-series data mining*

Time-series data are collected and measured along with time changing and aiming at discovering the frequently-occurred patterns. For modeling time series, 4 trends of movements are characterized: Trends or long-term movement, cyclical variations, seasonal variations and random movements. Similarity search is one of the typical methods for time-series analysis to uncover the patterns by comparing similarity to the given one (Hsu, 2003).

### 2.2 Geographical Data and Geographical Metadata

#### 2.2.1 Geographical Data

Geographical data, also called geospatial data, usually refers to high-resolution satellite image, spatial-temporal data, maps as well as other geo-referenced materials. In a generic sense, the geographical data are traditionally presented in raster and/or vector format. In recent years, some new types of geographic data have emerged with obvious characteristics of semi-structured or unstructured. For instance, multimedia data is a typical unstructured data as well as the web data while plain text data are semi-structured.

Due to the immense amount of digital geo-data and the complexity of data formats, it is hard to guarantee the data integration and accessibility when coming to
manage and manipulate the data. In this case, metadata has been proposed to help improve the operation on geographical data.

2.2.2 Definition of Metadata and Geospatial Standard

Taylor (2003) gives the following definition of metadata:

“Metadata is the data associated with objects, which relieves their potential users of having full advance knowledge of their existence or characteristics.”

In general, metadata is described as data about data, which could be understood as a structured data series created for describing the data resource and facilitating the resource management. To improve the resource discovery and retrieval, in the recent several years, metadata also has a rapid development in the fields of administrative management, information management, content rating, security management, and data preservation (Taylor, 2003). In any case, a typical metadata record must contain a series of elements, which was pre-defined specifically to describe certain attributes, in order to facilitate some possible operations for the data sets and metadata itself. For example, the information on a student card can be regarded as a catalogue of metadata, which comprise of student name, student number, nationality, specialty, as well as all the other necessary information.

Metadata is especially useful in the geographical field. It contains the identification of the resource, specific information of the geographical location, accessibility to the geographic datasets, the suitability of the datasets for a pre-defined purposes and necessary information about how to manage and process the data. Because the geographical information is often unreachable and expensive, extracting and sharing the metadata seems to be a good method to understand and manage the original data (Demšar, 2004).

For sharing the metadata globally, several geographical metadata standards are developed. The ISO 19115 standard is published by International Standard Organization (ISO), which comprehensively defines the element schema to describe the geographical resource and services. The 14 metadata element packages of ISO
19115 are shown in Figure 3 with the core elements in Table 3 below. What deserved
to be mentioned here is that the Infrastructure for Spatial Information in the European
Community (INSPIRE) is using the ISO 19115 geographical metadata standard to
foster the coordination of spatial metadata and the interoperability of services for all
the Europe communities (INSPIRE, 2007b).

![Figure 3: Metadata packages of ISO 19115 (source: Taussi, 2007)](image)

Table 3: Core metadata elements and requirements of ISO 19115 (ISO, 2003).

<table>
<thead>
<tr>
<th>No</th>
<th>Core Elements</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dataset title</td>
<td>Mandatory</td>
</tr>
<tr>
<td>2</td>
<td>Dataset reference date</td>
<td>Mandatory</td>
</tr>
<tr>
<td>3</td>
<td>Dataset responsible party</td>
<td>Optional</td>
</tr>
<tr>
<td>4</td>
<td>Geographical location of dataset</td>
<td>Mandatory under certain conditions</td>
</tr>
<tr>
<td>5</td>
<td>Dataset language</td>
<td>Mandatory</td>
</tr>
<tr>
<td>6</td>
<td>Dataset character set</td>
<td>Mandatory under certain conditions</td>
</tr>
<tr>
<td>7</td>
<td>Dataset topic category</td>
<td>Mandatory</td>
</tr>
<tr>
<td>8</td>
<td>Spatial resolution of dataset</td>
<td>Optional</td>
</tr>
<tr>
<td>9</td>
<td>Abstract of dataset</td>
<td>Mandatory</td>
</tr>
<tr>
<td>10</td>
<td>Distribution format</td>
<td>Optional</td>
</tr>
<tr>
<td>11</td>
<td>Additional information of dataset</td>
<td>Optional</td>
</tr>
<tr>
<td>12</td>
<td>Spatial representation type</td>
<td>Optional</td>
</tr>
<tr>
<td>13</td>
<td>Reference system</td>
<td>Optional</td>
</tr>
<tr>
<td>14</td>
<td>Lineage statement</td>
<td>Optional</td>
</tr>
<tr>
<td>15</td>
<td>On-line resource</td>
<td>Optional</td>
</tr>
<tr>
<td>16</td>
<td>Metadata file identifier</td>
<td>Optional</td>
</tr>
<tr>
<td>17</td>
<td>Name of metadata standard</td>
<td>Optional</td>
</tr>
</tbody>
</table>
Based on ISO 19115, some selected parts are translated from English into Swedish language and published in a technical report, named as “SIS TR 14: metadata på svenska” (metadata in Swedish). It gives translation and explanation of some metadata elements as well as some descriptions and terminologies which are helpful to use some software tools more conveniently. With respect to these, it is actually a supplement of ISO 19115 metadata standard in Sweden.

2.3 Lucene.net

Lucene.net is an algorithmatic port of Java Lucene which runs on the dot net platform, aiming at greatly speeding up the development of search engines. It can be understood as an Information Retrieval (IR) library, as well as a series of APIs which encapsulates both the functionalities of indexing and searching from the viewpoint of application. The metadata scheme can be obtained by indexing data attributes according to detailed requirements, and the input keyword of metadata be retrieved in the schema to find the original data file. However, Lucene.net is an abstract framework rather than a practical procedure; the logical function must be achieved by the developers when implementing specific applications.

It owns the capability to deal with many formats of data when they can be converted into text format. It can achieve a functionality of content searching by word segmentation method. Word segmentation in computation refers to dividing written language into meaningful units of words. For English, the space is a conventional delimiter of word segmentation. Directly, Lucene.net provides word segmentation functionality for English language. In the scope of word segmentation, it is not consummate but has a high flexibility to support expansive development. Developers could choose the most appropriate method from the provided word segmentation methods, or create a new method by combining and summarizing these provided
methods to meet their actual demands.

Lucene.net can be applied to develop big search engine, but also small desktop search engine. Even if the amount of data is massive, the time cost is still affordable.

2.4 Microsoft SQL Server 2005 and C# language

Microsoft SQL Server 2005

Microsoft SQL Server 2005 is a database platform, which not only improves querying and storing capability for relational data and structured data over all the SQL versions before, but also advances to handle semi-structured data by developing an XML data type. This advantage makes it appropriate to be applied as the main database management system for geographical metadata mining method (Rizzo et al., 2006). Because it integrates with Common Language Runtime (CLR), it can employ .net framework and a lot of programming language such as C#.

C# programming language

C# (C sharp) is an enhanced version of C++ programming language suitable for .net environment developed by Microsoft. Component-orientation as a new feature of C#, it could be treated as a kind of object-orientation owned by JAVA and C++. In addition, C# is adding capability of type security and firmly coupling with web development. According to the official declaration of Microsoft, the C# language approaches the development of XML more easily. For these characteristics of C#, it could be regarded as a suitable programming language to mine the data categories of distinctive formats of geographic data and facilitate the transformation to XML for further development.

3. Methodology

The aim of this project is to develop and evaluate a method of data mining in order to extract and retrieve metadata from the geographical materials of NLS on the LAN. To achieve this purpose, the design of this method is integrated with the Rational Unified Process (RUP), an object-oriented framework of software development process (Kruchten, 1998). It is presented in four lifecycle phases with
key objectives for each phase. Firstly going through the basic step, the data environment is analyzed and an initial architecture is acquired. In the next elaboration phase, data entities are divided into four types and then a theoretical model basing on data entities is set up. Afterwards, the geographical metadata from different categories of data formats are identified and then conformed to geographical metadata standard ISO 19115. After solving these problems, the design of this method falls into the implementation phase that the construction of a suitable Software Development Environment and choice of suitable metadata mining algorithm and workflow. For further refinement, the validation of this method should be checked to ensure the success rate higher than 50% in the evaluation phase. The basic process is shown in Figure 4 step by step.

<table>
<thead>
<tr>
<th>Embryonic phase: Basic processing step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of data environment</td>
</tr>
<tr>
<td>Initial architecture design</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elaboration phase: Model and Feature extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model on data entities</td>
</tr>
<tr>
<td>Metadata from Categories of data formats</td>
</tr>
<tr>
<td>Geographical metadata standard</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Construction phase:</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDE</td>
</tr>
<tr>
<td>Workflow</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluation phase:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface</td>
</tr>
<tr>
<td>Efficiency of indexing</td>
</tr>
<tr>
<td>Success rate</td>
</tr>
</tbody>
</table>

Figure 4: The developing process of mining method integrated with the Rational Unified Process

3.1 Description of Data Environment and Initial Architecture

In this phase, a general outline about data environment of NLS should be described, including data description and a basic architecture. The description includes requirement analysis and data specification.
3.1.1 Description of Data Environment

The geographical data and corresponding metadata are distributed over hundreds of computers on the LAN of NLS. However, the data files lack normalized operational formats, and at the same time, are not formalized in a database environment.

In NLS, there are four kinds of data files existed: Time-series data, spatial data, documents data and webpage data.

Document file mainly refers to massive files of PDF, Microsoft Word file and plain text files. Due to the semi-structure of these files, mining should focus both on the contents and the fundamental attributes. In this study, the mining method must support both Swedish and English languages.

In GIS, one type of time-series data is the real-time satellite reports and GPS-collecting data. A large number of these reports are in formats like Excel file. Therefore, the mining functionality for the contents and basic attributes of Excel file is required.

There is a variety of spatial data types, mainly vector files and raster images. Feature extraction for these spatial data is the hot topic of graphical analysis and processing. For example, the metadata from satellite image content can be information of bounding box, the depth of each pixel, the number of bands, and so on. Some information cannot be obtained by general data mining method. Because spatial data often are operated and presented in specific software, it requires some specific APIs for each file format.

Web data is distinctive from all the data structure mentioned due to its dynamic changes and high complexity. However for this case, the source files of NLS are provided as specific webpage in the NLS cloud. The web mining is not applied to the whole World Wide Web (WWW). This characteristic makes it possible to mine the webpage content by using simple methods for feature extraction, without paying too much consideration to the capacity and time-efficiency of the application.
### 3.1.2 Initial Architecture

After analyzing the data environment and requirement, an initial architecture can be developed in Figure 5.

The first two-operating steps are data selection and data positioning, where data positioning means specification of IP numbers, volume, directory path and filename. In the latter process, the data attributes are extracted and entered into a knowledge base. The disks and folders which contain these valuable resources are located by IP address and accessible from any computer. Afterwards, these data are recorded in the Knowledge Base (KB) on the metadata mining server, and the metadata schema would be generated.

This knowledge base and data mining method achieves the functionalities of metadata extraction. When the users enter the keyword they are interesting in, the data mining server extract the corresponding metadata record stored in Knowledge Base, and then retrieve the corresponding source data file.

The results of the searching are presented in an html-based webpage, similar to the interface of Google searching engine.

### 3.2 Elaboration of Model and Feature Extraction

The objective of the elaboration phase is to refine the initial design by expanding the detailed analysis to specific problem domains. For reification of basic architecture, this phase is to develop a mining model and extract features of geographical data in NLS. Generally, feature extraction refers to extract characteristics about a large data set, for instance a data file. In this application, feature extraction refers to extract important attributes of geographical data which can be regarded as metadata elements.

The types of data entities are verified in the data environment of NLS, and the geographic metadata elements are identified with respect to different categories of data formats. Then, the elements of ISO19115 metadata standard are compared with the elements obtained from categories of data formats, with the aim to get a clear
impress of the consistency between ISO 19115 and the metadata available. Going through this phase, the theoretical model in detail has been refined as a logical guidance to the construction of software system followed.

3.2.1 Data Mining Model based on Data Environment and Data Entities

One design goal is to have a distributed data mining application, because all the

Figure 5: Basic architecture of the geographic metadata mining method for NLS

3.2.1 Data Mining Model based on Data Environment and Data Entities

One design goal is to have a distributed data mining application, because all the
data are distributed not only in local computer but also other computers on the LAN. First of all, a knowledge base is established in the metadata mining server to handle all the valuable data files on the LAN. Some optimization algorithms are also added to improve the performance. This design makes it possible to extract and mine all the data on the LAN, only knowing the IP address and path on the data resources. The users could use any computer as a metadata mining server, instead of a necessary large-scale public server. This design greatly improves the convenience of data management on the LAN.

Figure 6: Data mining model based on data environment and data entities

Based on the LAN distributed design, the system is further broken into 4 sub-models work-parallel for different entities respectively: document data mining, web data mining, spatial data mining, and time-series data mining.

At last, the mining results are presented in a series of hyperlinks on a webpage, with the directory and descriptions for each hyperlink. Visual data mining is a rather new concept in which 2-D and 3-D graphics are used to present the results. Broadly
speaking, visual methods for adjusting the parameters can also be considered as visual data mining. When the user examines these hyperlinks, they could easily acquire an approximate impression if this result is un-useful in order to adjust the input metadata keyword.

3.2.2 Identification of Metadata Elements from Different Categories of Data Formats

For different categories of data formats from different data entities, the importance of different metadata elements varies distinctively. Furthermore, the choice of mining method and software must be adapted to the categories of data format. Hence, identifying the metadata elements according to the data formats is an important processing step before this method reaches the mature step of encoding.

According to the report of manso et al. (2004), the spatial data are separated into vector data and spatial data, and the metadata elements are introduced respectively.

For the document data at NLS, the formats enumerated below are involved: Word, PDF, and plain text. Meanwhile with respect to time-series data, the main format is referred to Excel. For these four formats, the valuable information for identifying metadata are mainly file name, file extension, title, author, creation time, version, key words, key content, language, abstract.

The main representation of webpage is in Hypertext Markup Language (HTML) format. The metadata is embedded in the header of webpage or stored in an individual file, which is not visible for web users but support machine-read easily. The described metadata ranges from descriptive text and keywords, even to the standard information such as Dublin Core. The mining of webpage supports the metadata marked in XML format, and the extraction on web is different from usual databases. Hence, there is not much practical significance to roughly classify the web format for this unique method. In the scope of this thesis, the web data has not been dwell on too much, just a suitable software environment is created with series of APIs for development of web mining in future.
3.2.3 Metadata elements of ISO 19115 supporting automatic extraction

For verifying the consistency with ISO 19115 geographical metadata standard, the metadata from different file formats are compared with the elements in ISO 19115. In ISO 19115, the mandatory elements in Table 3 can be considered as subjective ones and are necessary to be included. According to the paper of Taussi (2007), it is worth mentioning that the optional elements which are supported by simple automated extractions are located in four ISO 19115 packages: spatial representation, the reference system, the extent of the dataset, and the application schema. The elements within these packages are related to spatial data tightly and listed in table 4. It could be regarded as guidance to the choice of metadata elements from table 1 and table 2 to ignore the unnecessary elements. As to the document data and time-series data, the mandatory elements are necessary, and the emphasis is on content in this method.

In the package of spatial representation, the elements of dimension size (number of elements along the axis) and resolution can be extracted out of raster data by simple automatic method. In contrast, the Geometric Object Type and Geometric Object Count, which are used to describe the name and number of vector objects, can only be extracted from vector data.

For the package of reference system, these four elements can only be obtained when a reference system is specified. They are name of reference system, projection, ellipsoid, and datum.

As to the packages of dataset extent and application schema, the elements can be acquired automatically only when the software and application system allows specifically.

Table 4: Optional metadata elements of ISO 19115 supporting automated extraction
(Source: Taussi, 2007).

R – Raster data, V – Vector data, X – both for raster data and vector data

<table>
<thead>
<tr>
<th>Metadata package</th>
<th>Element</th>
<th>Yes</th>
<th>Maybe</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial representation</td>
<td>Number of dimensions</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cell Geometry</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Corner Points</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Center Points</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point in Pixel</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimension Size</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometric Object Type</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometric Object Count</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Reference system**

| ReferenceSystemIdentifier | X |
| Projection | X |
| Ellipsoid | X |
| Datum | X |

**Extent of dataset**

| Polygon | X |
| West Bound Longitude | X |
| East Bound Longitude | X |
| South Bound Longitude | X |
| North Bound Longitude | X |
| Extent (Temporal) | X |
| Minimum Value (Vertical) | X |
| Maximum Value (Vertical) | X |
| Unit of Measure | X |
| Vertical Datum | X |

**Application schema**

| Schema Language | X | X |
| Constraint Language | X | X |
| Schema Ascii | X | X |
| Graphics File | X | X |
| Software Development File | X | X |
| Software Development File Format | X | X |
3.3 Construction of Software System

3.3.1 Creation of SDE

When a Software Development Environment (SDE) is proposed for the realization of the method, this method is changed from design documents into practical encoding.

The implementation of this metadata mining method was realized by making use of the Microsoft development tool, Microsoft Visual Studio 2005, as the main developing platform on top of the operation system Window XP. Microsoft SQL Server 2005 is adopted as the database management system to manipulate and preprocess the raw geographical data and the extracted metadata on the LAN. Lucene.net is the main tool for metadata indexing and extraction of document files and time-series files, using the C# as the programming language.

A large part of files in NLS are in document format and Excel format directly which can be converted into text files. There are also a great number of time-series data collected by position-aware instruments. Although they are not explicit document files, for example with .k as the file extension, they can be converted into plain text file. Therefore, the metadata of these files are appropriate to be extracted within Lucene.net. It can achieve a functionality of mining both basic attributes and content for these file formats which can changed into text format.

3.3.2 Workflow

In order to gain a better global control from an overall situation, the workflows of metadata extraction are separated into two small modules. Indexing module is used to record and arrange metadata into schema, and retrieving module to acquire and represent the retrieval result. It is quite time-consuming to index metadata for a large amount of geographical data, so it is ineffective and impractical to index metadata for each time of retrieval. When indexing and retrieving are separated into two modules, the indexing could be processed in the background. It avoids to invoking the useless
files of the database in the process of retrieving, thus greatly reduces the cost time and searching complexity. As long as the metadata generated by indexing has not changed, the retrieval could be carried out at any time without the need to re-indexing. Generally speaking, the indexing is the precondition of retrieving, and retrieving is the inspection for the result of indexing.

**Metadata Indexing Module**

First, the indexing module traverses all the files within the targeted file folders or disks. Afterwards, all the necessary attributes of the file will be extracted and saved into the indexing-file in database. These can be characterized as metadata schema. The workflow of indexing is shown in Figure 7.

**Metadata Retrieving Workflow**

After entering keywords for characterization of files by the user, the system searches the records containing these keywords automatically in the index metadata file of database. Since spatial concepts are treated like any other keyword, the emphasis of content mining is mainly on the document file and time-series file. For these two data entities, the keyword is contained either in the basic attribute records such as title, or embedded in the key content. This data mining system searches the keyword in the basic attribute fields firstly, and then searches in the key content of file. When having finished traversing all the records in indexing file, the search results will be displayed in a prescribed interface on the webpage. The workflow is shown in Figure 8.

**3.4 Evaluation Phase**

The user’s satisfaction mainly lies on three conditions: operability on interface, efficiency of indexing and success rate of retrieving.

For the operability, an explicit interface can convenience and deepen the understanding of users. The interface of this method is based on webpage, similar to Google search webpage. The titles of the resulted files are enumerated in a series of
Figure 7: Workflow of metadata indexing in database

1. Begin of indexing
2. Input the disk names and IP of targeted computer
3. Inspect a folder in target disks
4. If have subfolder:
   - Yes: Go to subfolder
   - No: Traverse all files in this subfolder
5. Extract attributes of all files and corresponding metadata
6. If last folder in disk:
   - No: Go to next file folder of this disk
   - Yes: Go to next disk
7. If the last disk:
   - No: Go to next disk
   - Yes: Update of index file in database is completed
8. Indexing file (metadata schema) in database

End of indexing
Figure 8: Workflow of metadata retrieving
options which are in hyperlink form, maximum in 50 options. The better the searching result satisfies the evaluating purpose, the more front the option ordered. Below the title, maximum 200 bytes are used to describe this file, with the key word for searching are highlighted both in title and/or description. The user could analyze the validity and accuracy of the results roughly according to the description and titles, lying on their expertise of this subject, in order to choose out the best options. This interface supports and improves the capacity of human-computer interaction effectively.

When evaluating the extraction of metadata in this method, an appropriate way used is to acquire the efficiency of indexing relatively to cost-time. By testing efficiency of indexing for several times and showing them in the histogram, the change is acquired obviously that if linear or threshold in a scope of data files.

For a given number of geographical data files from NLS, the success rate here refers to the percentage of files in which the input metadata keywords indeed existing can be retrieved successfully. For different data entities and formats, the success rate takes effect on different data attributes and entering keywords. The success rate is the most powerful indicator for evaluating the effect of this method, and indeed the visualization of success rate makes it clear at a glance. By the visualization of success rate, users can get insight into data and gain an intuitive impression and hypothesis of the patterns and result the data attempting to reveal (Keim, 2002).
4. Result of Interface, Efficiency and Success rate

4.1 Interface on Webpage

![Home page of retrieval website](image)

**Figure 9:** The home page of retrieval website

4.2 Testing 1: Efficiency of Metadata Indexing

The efficiency of indexing has been tested for five times, using different volumes of files for each time. The specific data are listed in Table 5. Generally speaking, the total volume of files for testing was 34323, and the sum of time-costing was 89 minutes. In average, the efficiency of indexing is 385 files per minutes. Form the Figure 10, the efficiency is almost linear attenuation, but the velocity of attenuation is slow.

When the user in NLS wants to index files which are restricted within volume of 20,000 in several disks or folders, the indexing is quick and efficient.

<table>
<thead>
<tr>
<th>Number</th>
<th>Volume of files</th>
<th>Time (Minutes)</th>
<th>Efficiency of indexing (volume/minute)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>395</td>
<td>1</td>
<td>395</td>
</tr>
</tbody>
</table>
4.3 Testing 2: Success Rate and Influence Facts

4.3.1 Success Rate Restricted in Small Domain

For testing the success rate of this method (without giving full consideration on web mining), 148 spatial raster images, 203 document files, and 44 Excel files were loaded on a computer. In order to simplify the testing process, only a few data formats and metadata categories for extraction were tested. The detailed information is given below.

**Image**

The image mining supports the searching of resolution, title and file extension. Of the 148 image files, there were 122 JPEG files, 9 Gif files and 17 bmp files. When retrieving by file extension (JPG, GIF and BMP) and titles, all raster files were found
easily. The image resolution of 148 files was extracted with a success rate of 91.2%.

**Document file**

In the document set, there were 105 PDF files, 49 plain text file and 49 Word files.

When searching the file extensions (PDF, txt, and doc), the success rate for each format are 92.4%, 95.9% and 81.6%. Searching the keywords in title, the success rate changed to 81.9%, 95.9% and 81.6%. When searching in the document content, the success rate has a sudden drop for PDF files and Word files (till 72.4% and 75.5%); lower than text files (still 95.9%).

**Time-series data**

Time-series mining supports the functionality of searching for Excel file (Title, format and content). For the 44 Excel files, according to both format and title, the success rate is 95.5%. However in the process of content searching, the success rate has decreased to 79.5%.

![Success rate graph](image)

**Figure 11:** Success rate for different types of geographical data in NLS.

When mining the file resolution, 148 spatial images are involved and 135 are
found. For mining of title and file extension, all the files are tested and the success rates are respectively 91.9% and 94.7%. When retrieving according to file content, document file and Time-series file (PDF, Word, text and Excel) are tested and the success rate is 78.9%. As a conclusion for retrieving different metadata elements (resolution, title, file extension and content), 1185 files are used and 1067 ones can be found successfully. The final success rate in average is 90.0%.

Table 6: Success rates basing on different metadata elements for Testing 2

<table>
<thead>
<tr>
<th>Number</th>
<th>Metadata elements</th>
<th>Formats of files</th>
<th>Volume of all files</th>
<th>Volume of success files</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Resolution</td>
<td>JPEG, Gif, Bmp</td>
<td>148</td>
<td>135</td>
<td>91.2%</td>
</tr>
<tr>
<td>2</td>
<td>Title</td>
<td>JPEG, Gif, Bmp, PDF, Word, Excel, Text</td>
<td>395</td>
<td>363</td>
<td>91.9%</td>
</tr>
<tr>
<td>3</td>
<td>File extension</td>
<td>JPEG, Gif, Bmp, PDF, Word, Excel, Text</td>
<td>395</td>
<td>374</td>
<td>94.7%</td>
</tr>
<tr>
<td>4</td>
<td>Content</td>
<td>PDF, Word, Excel, Text</td>
<td>247</td>
<td>195</td>
<td>78.9%</td>
</tr>
<tr>
<td></td>
<td>Sum of all files and average success rate</td>
<td></td>
<td>1185</td>
<td>1067</td>
<td>90.0%</td>
</tr>
</tbody>
</table>

4.3.2 Influence Facts

For this small Testing 2, the inherent restriction of Lucene.net and expertise of developers when detailed encoding have played great effects on the success rate for explaining why it cannot reach 100%. However, there are two reasons that should take great part of responsibilities of the sudden sliding of success rate for content mining of time-series files and document files.

The choice of word-segmentation algorithms is especially important for the final success rate of searching. Lucene.net has offered several established word-segmentation algorithms, but they seem too simple and general in nature to complete content mining for all document and time-series file. This method requires support for diglossia (Sweden and English), which puts forward a higher demand on optimization of word segmentation algorithms. However, this application chooses the basic word-segmentation algorithm for Latin language which the Lucene.net owns.
Inherently, and without doing some optimizations.

In the current method of reading content of PDF, using the third-party software is the most common way. In C#.net, PDFBox, one of the third-party software, are used to achieve reading PDF content. However, the reliability of PDFBox is not best since a high percentage of PDF can’t be read successfully. For example, a large part of the existing PDF document is originating from the Word files, which are transformed by variety of software. Some such PDF file can’t be read by PDFBox when testing, even not be accessed in the database. Another typical PDF file it could not read is the one changed from the scanning documents by Optical Character Recognition (OCR). In general, there are some limitations of third-party software to read PDF content. Of course, it is also the limitation of the method designed in this thesis.

5. Discussion

For gaining an explicit progress of the metadata mining method, the Rational Unified Process (RUP) is proposed to control the process and sub-blocks of the overall design. Firstly, a basic architecture is built up on the detailed data environment and requirements. For a more efficient process, the geographical data are categories into four types of data entities (spatial data, webpage data, time-series data and document data) to adapt to the different preformed mining methods. For each data entity, the metadata elements for extraction are indentified according to the data format. When encoding to realize the design, a Software Development Environment (SDE) has been built up. The open source Lucene.net is integrated into this SDE to provide a framework of indexing and retrieving. The indexing and retrieving modules take the responsibilities of metadata extraction and evaluation in this method respectively.

However in the actual encoding, the development and evaluation of this method is not complete and thorough (for which it is a massive project and a large number of tests would be needed). The mining of metadata on website has not been given full
consideration, but just a suitable software development environment given out. At the same time, the metadata elements for the other types of data entities (spatial data, time-series data and document data) are restricted to a small domain. The file title, file extension and content are mined for document file and time-series data (typically in format of PDF, Microsoft Word, plain text and Microsoft Excel). For the spatial data, another metadata, resolution, is added to test for formats of JPEG, Bmp and Gif. However, it is hard to mine the document information embedded in spatial files, so it has not been mined in this testing work. Moreover, although the test data is typical for geographical material from NLS, the amount of data is not abundant enough.

Despite all these limitations, the result of testing is still a meaningful indication for its potential and usefulness. Nevertheless as a referenced result of first study, the average success rate has achieved 90%. The barrier to higher success rate is mainly laying on three influence facts. The choice of word-segmentation algorithm in Lucene.net is the most important fact. The third-party software for reading content of documents is not efficient for all the text-based files. Both of them play great impact on the success rate, especially for content mining. In addition, the expertise of developers affects the success rate of this method, both from the whole architecture building and the detailed refinements.

6. Conclusion

At the National Land Survey of Sweden, 67% of all the mandatory geographical metadata and 23% of optional ones are embedded in the Local Area Network without being extracted. This thesis develops a data mining method to extract these metadata automatically with the aim of providing geographical metadata for Europe according to the Metadata Implementing Rules of Infrastructure for Spatial Information in Europe.

The attributes of the geographical data in NLS and corresponding metadata are proposed to be extracted and indexed into a knowledge base, and then a metadata
schema will be generated. When the implementation of retrieving, the keywords which the users are interested in are inspected in the metadata schema and the information of the related files are extracted and presented on the interface on webpage.

For evaluating the efficiency of extracting metadata, the indexing work has operated 5 times in Testing 1. There are abundant files are involved ranging from volume of 395 to 12350. Within this scope, the efficiency is high with average of 385 files per minutes, and the linear attenuation is slow.

In the simple Testing 2 for success rate, four metadata elements in a few formats of 395 typical files are extracted and retrieved:
(1) Resolution, title and file extension for spatial files in format of JPEG, Gif and Bmp.
(2) Title, file extension and content for document files in format of PDF, Microsoft Word and plain text.
(3) Title, file extension and content for time-series files in format of Microsoft Excel.

Generally speaking, the success rate is lowest on content mining in average of 78.9%, and highest on file extension mining of 94.7%. Overall, the average success rate of all files reaches 90.0%, higher than the original demand of 50%.

Reference


Hsu J. (2003), Critical and Future Trends in Data Mining: A Review of Key Data
Mining Technologies / Applications, in: Wang J. (ed.), Data Mining, Opportunities

INSPIRE (2007a), INSPIRE Metadata Implementing Rules: Technical Guidelines
based on EN ISO 19115 and EN ISO 19119, Infrastructure for Spatial Information in
Europe, European Commission Joint Research Centre.

Council of 14 March 2007, Establishing an Infrastructure for Spatial Information in
the European Community, Infrastructure for Spatial Information in Europe, Official
Journal of the European Union.

ISO (2003), ISO 19115-Geographic information-metadata, International Standard
Organisation.

Jungnickel D. (2005), Graphs, Networks and Algorithms, Springer-Verlag, Berlin

Keim D. A. (2002), Information Visualization and Visual Data Mining, IEEE

Technology Series: Addison-Wesley.

Madria, S. K., S. S. Rhonwmich, W. K. Ng, and F. P. Lim, (1999), Research Issues in
Web Data Mining, Proceedings of Data Warehousing and Knowledge Discovery,
First International Conference. DaWak’, pp. 303-312.


Rizzo T., Machanic, A., Skinner, J., Davidson, L., Dewson, R., Narkiewicz, J., Sack,


Shekhar S., Zhang P. and Huang Y. (2005), Spatial Data Mining, in: Maimon O. and
Rokach L. (eds.), Data Mining and Knowledge Discovery Handbook, Springer

York.

Taussi, M. (2007), Automatic Production of Metadata out of Geographic Datasets,
Helsinki University of Technology, Department of Surveying, master’s thesis (in


