Geographically-Aware Information Retrieval for Collections of Digitized Historical Maps

Bruno Martins[1], José Borbinha[1], Gilberto Pedrosa[1], João Gil[1], Nuno Freire[2]
[1]IST - Instituto Superior Técnico. Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[2]BNP - Biblioteca Nacional de Portugal, Campo Grande, Lisboa - Portugal
bruno.martins@tagus.ist.utl.pt, jlb@ist.utl.pt, gilberto.pedrosa@ist.utl.pt,
jgil@ext.bn.pt, nmaf@bn.pt

ABSTRACT

DIGMAP is a project focused on historical digitized maps that will develop a set of Internet services based on reusable open-source software solutions. The main service will provide discovery and access to resources related to historical cartography, based on metadata from European national libraries and other relevant third part providers. These resources will comprise both physical and digitized objects. In the case of digitized maps, available metadata will be enriched by automatic and semi-automatic processes that will try to extract relevant indexing information from the images of the digitized maps, as also from any kind of associated text. This paper presents an early overview on the project, particularly focusing on the aspects related to geographical information retrieval.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: System Issues; User Issues

Keywords

Geographic IR, Digital Libraries, Historical Maps

General Terms

Design

1. INTRODUCTION

DIGMAP is a project co-funded by the European Community program eContentplus. DIGMAP stands for "Discovering our Past World with Digitized Historical Maps", but it could stand also for "digging on maps"! Europe has a very old and heterogeneous political history. Since long, its political and social dynamics have raised the need for a clear perception of the territory. Cartographic materials were also essential when Europeans went outside of their borders, leading mankind in the common discovery of the overall World. As a result, the European legacy in historical maps is enormous, very rich and diversified. This cultural and scientific genre of artifacts represents a special perspective of our history.

Scholars have long been used to exploring these materials, although this was usually only possible on local networks and after special requests. However, at least medium/low resolution copies are nowadays often available in the Web, for the general public.

GIR'07, November 9, 2007, Lisboa, Portugal.
ACM 978-1-59593-828-2/07/0011 ...$5.00.

DIGMAP will address the development of services, to be available on the Internet, for digital libraries of historical digitized maps. This will be made with basis on open standards and open-source software solutions. Existing metadata records, from either national libraries or other third party providers, will be reused in DIGMAP. New techniques will also be studied for enriching the available metadata, through both automated and semi-automated processes. The metadata will be indexed, in order to support advanced retrieval and browsing scenarios. Special attention will be given to the geographic indexing of the maps. From a technical standpoint, the project involves many challenges related to the area of geographical information retrieval (GIR). As such, it will build on previous research efforts such as the Alexandria Digital Library project [7], SPIRIT [9], GREASE [19, 14] and others [17, 10].

At the image level, we can automatically generate metadata by comparing the new maps with a reference knowledge base, as well as by extracting relevant features from the images. Maps can be very rich in decorative and stylistic details, making them very different from, for example, photos. Therefore, this will require new image processing techniques, specific for handling map images.

Considering the existence of metadata information, defined at library catalogues or given informally over Web documents, one can also use automated techniques to enrich the information associated with the maps. Metadata records have often problems of incomplete information. Moreover, geographic metadata in the form of spatial footprints (i.e. coordinates or bounding boxes) is also not commonly available. DIGMAP will explore text mining techniques, specific for handling geographic context information, to extract and disambiguate place references over textual data. The idea is to minimize human intervention in the task of indexing maps.

Using the rich metadata records, DIGMAP will build index structures that combine techniques from information retrieval and geographic information systems, in order to support advanced retrieval scenarios. The DIGMAP service will also offer a browsing environment for humans. In some sense, this will be similar to Google Maps, but with special features for exploring historical maps. This includes timelines (relevant for those maps where it is possible and easy to assess the date) and clustering of related resources.

This paper will proceed by describing the main DIGMAP use cases, followed by a description of the system's architecture. It will then present the main challenges related to the application of techniques from geographical information retrieval to the indexing and retrieval of maps. The paper will end with conclusions and a with a description of the following project activities.

2. DIGMAP USE CASES

The project will result in a set of services according to the main use cases presented in the Figure 1.

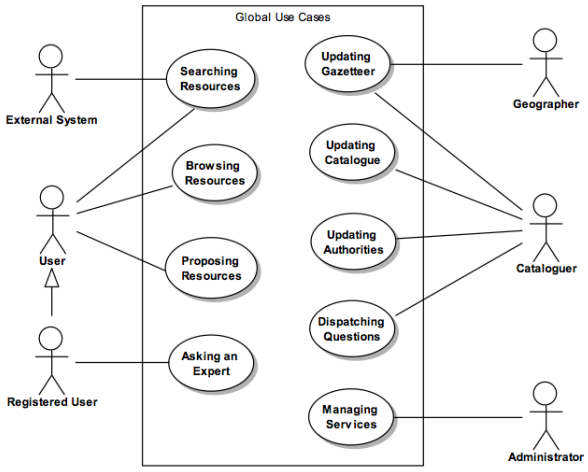

Figure 1: DIGMAP Use Cases.

Concerning the actors, a **User** is an anonymous person that interacts with the system. A **Registered User** is a user that has been recognized by the system successfully. **External System** is a machine which accesses to the system for searching resources. **Cataloguer** is a person that will be updating the metadata or answering questions posted by Registered Users. **Geographer** is a specialist who manages the gazetteers (i.e. place name thesauri). **Administrator** is a DIGMAP responsible who will manage all intervening actors.

The users of the system will be able to access resources in two ways: **Searching Resources** (simple and complex searches in the metadata) or **Browsing Resources** (browsing in indexes built from the metadata). Particular attention will be given to the support of scenarios were users search and browse according to geographical and temporal criteria. Users will also be **Proposing Resources** for inclusion in the system. Through **Asking an Expert**, users can submit questions to be answered by experts, registered in the system and with specific knowledge in the area. This will be supported by a software module that allows access to previous questions and answers, so that repeated and frequent questions can be easily dispatched. The specialists will answer and manage those questions through the use case **Dispatching Questions**.

Finally, there will be three different data management services in DIGMAP. **Updating Thesaurus** will maintain multiple thesauri and gazetteers. **Updating Catalogue** will feed the Catalogue with metadata from remote data providers or through a local cataloguing interface. This local interface will also support the management of the authority metadata (**Updating Authorities**) and the indexing of the maps. **Managing Services** is a back-office component where DIGMAP will be managed by several distinct actors with appropriate security permissions.

3. DIGMAP ARCHITECTURE

The software solutions produced in DIGMAP will be based on open standards and released as open-source. The results of the project will be useful for local digital libraries of maps, as either a standalone system or as interoperable components for wider and distributed systems. The generic architecture will follow the design shown in Figure 2, which represents a "Deployment Diagram" as defined by the Unified Modeling Language (UML).

A **Resource** is a relevant information object that is registered in the System by records holding descriptive and indexing metadata. Examples of resources are maps, atlases, books or Web sites about historical cartography. Special focus will be given to map

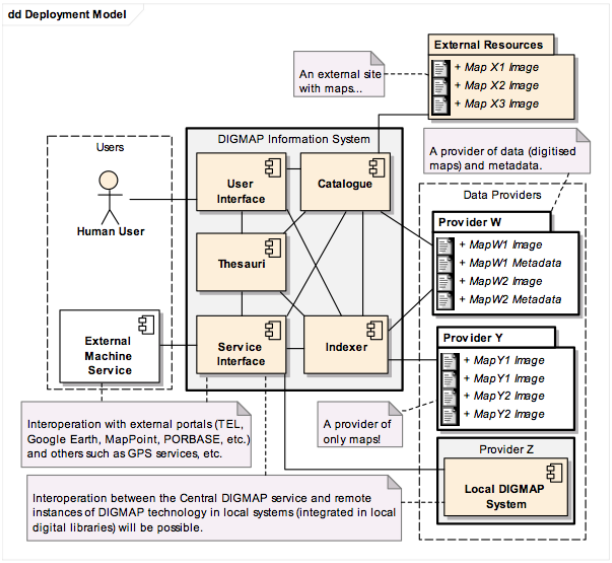

Figure 2: DIGMAP Architecture.

resources. When a Resource is a map, it must be possible to associate it to metadata structures describing its geographic details (geographic boundaries, scales, details of sub-maps inside a wider generic map, etc.). On-line Resources are recognized by the existence of a URL in one field of the record.

A central component of the **DIGMAP Information System** will handle service configuration and interconnection of the different modules. This central component will also be responsible for storing **User** data, most likely through the use of an LDAP server.

The **Catalogue** is the service responsible for the submission, editing and storage of the records describing the Resources. The Catalogue must support the definition of collections, i.e. groupings of the Resources according to some criteria. Examples of criteria for collections can be the type or genre of the resource, the source or provenance of the record, etc. It will be possible to register Resources in the Catalogue through a local user interface, or importing records (e.g. through Z39.50 and SRW/SRU, although the preferred interface for metadata interoperability will be OAI-PMH). The local user interface of the Catalogue must make it possible to edit any existing record. The Catalogue must be able to maintain the descriptions of authorities and of the maps in multiple metadata formats, especially in UNIMARC, MARC 21 and Dublin Core.

The **Indexer** is the component that supports to the automatic indexing of resources. It will be composed of three major modules: the **Image Analysis** module, addressing the automatic generation of metadata records through image analysis techniques; the **Text Analysis** module, addressing the automatic enriching of metadata records through text mining techniques; the **Retrieval** module, addressing the generation of appropriate indexing structures for supporting the required retrieval operations.

The **Thesauri** is the component of the system that registers auxiliary information for the purpose of indexing, searching and browsing the Resources. An initial survey specifically focusing on historical and geographical thesaurus is available at [21]. In DIGMAP, the thesauri component will be composed of two major sub-systems: the **Gazetteer**, to register auxiliary geographic information; the **Authority File** to maintain and provide easy access to the author's information and identification and disambiguating from similar and duplicate authors. The Gazetteer will build on previous efforts in Alexandria Digital Library project [7], and on a proposal by

the Open Geospatial Consortium (OGC) for a Web gazetteer service [3]. It will essentially provide access to place names, geographic concepts, relationships among geographic concepts, spatial footprints, and other associated geographic metadata. As for the Authority File, it will mainly deal with duplicate metadata, aggregating the different references for the same data record.

The **User Interface**, besides a traditional OPAC service, will offer a browsing environment for humans. It will explore paradigms inspired by Google Maps, Virtual Earth, TimeMap [8] and other similar systems and previous research efforts. Some of the specific functions that will be provided by the User Interface include timeline visualizations, clustering of related resources, and advanced geographical information retrieval mechanisms. The user interface will also include a module for handling user questions, which will likely be developed by extending an existing forum system.

The **Service Interface** will provide access for external services. DIGMAP will explore different solutions for interoperability with other systems and services (e.g. Web portals, Mapping clients, and applications such as Google Earth). For this purpose, the applications and services to be developed will be based on open standards (e.g. XML Web services), sharing features with other applications.

4. GIR METHODS FOR COLLECTIONS OF DIGITIZED HISTORICAL MAPS

This section surveys GIR methods that can be used in DIGMAP to process digitized maps and associated metadata. These techniques will, in one hand, support the Updating Catalogue use case through the extraction of relevant features for indexing (either from the textual metadata or from the images themselves), and, in the other hand, support the Searching and Browsing use cases through the offering of advanced retrieval mechanisms for exploring the geographic domain. Geographic information retrieval is a core aspect of the system, as the Indexer, Catalogue, Gazetteer and User Interface all require techniques researched within this field.

4.1 Handling textual descriptions of maps

Maps can have textual descriptions, given either as metadata explicitly associated by librarians, or given as text in the context of a Web page describing the map. The main challenge in handling these textual descriptions relates to finding instances of interesting words and phrases, such as a placename or a time period. For doing this, we will build on previous research efforts [10, 14], as well as on an OGC proposal for a Web Geoparsing Service [11]. The algorithm will essentially consist of a two-step process of first identifying interesting entities in the text (e.g. looking up placenames in the gazetteer) and, in the case of place references, using heuristics to disambiguate these entities into spatial footprints defined in the gazetteer. Since each of the discovered placenames will be linked to a gazetteer entry, other DIGMAP modules will have a clear way of knowing where to look for more information about the places associated to a given map. The gazetteer will in turn follow another OGC proposal for a Web gazetteer service [3], reusing information from existing gazetteers and geographic ontologies [2, 22].

Ideally, we would like to be able to associate each map with an exact spatial footprint (i.e. the geographical area that it covers). DIGMAP will therefore also study techniques for combining the extracted place references into a single encompassing geographic scope. This will again build on previous research efforts [1, 15].

Finally, it should be noted that political borders and geographical names change over time. Ancient maps that were accurate when they were made may no longer correspond to the actual reality. Having historical placenames defined in the gazetteer is of foremost importance in DIGMAP. Besides placenames, the DIGMAP gazetteer will give particular attention to the definition of temporal features (i.e. historical periods) and to the linking of placenames to the corresponding temporal periods.

4.2 Handling images from digitized maps

An initial DIGMAP survey, available at [6], describes image analysis methods that can be applied to maps. Although there is a vast literature on this subject, older historical maps present specific challenges and are, in general, harder to deal with than modern maps. DIGMAP requires methods flexible enough to deal with the following challenges:

- The quality of the digitized images is often poor, either due to restrictions in the digitalization process, or due to degradation from ageing or other factors.
- The geographical standards used in ancient maps, if any, are not as consistent as modern ones. Scale, position and meaning of symbols and colors may be unknown or inaccurate.
- Modern cartography is able to represent detailed and accurate information about the location and shape of geographic features, whereas ancient maps are based on incomplete or incorrect data.
- Text on ancient maps is likely not typeset in a modern font and may even be handwritten, making it harder to recognize through automated OCR processes.

Given these challenges, DIGMAP will not be over-ambitious in its image processing purposes. Unlike many algorithms and systems developed to support geographic information services, we do not require full feature extraction and recognition methods. Instead, we are looking for the most effective ways to add any kind of useful geographic metadata to a given image of an old map. Concerning the applicability to old maps, the most effective automatic methods appear to be geometry-based vectorization [12] and text extraction [5, 13, 20]. These are the main approaches that we will be considering for image processing in the DIGMAP project. The complete solution for geographic feature extraction is also not likely to be composed of just one algorithm. Instead, a pipeline of techniques should be employed. For DIGMAP, we are considering the following set of processing operations:

1. Map restoration and pre-processing, correcting flaws introduced by document degradation and preparing the image for the more advanced processing stages.
2. Map segmentation, grouping pixels according to color and spatial homogeneity and also preparing the image for the more advanced processing stages.
3. Map vectorization and comparison, obtaining vector-based representations of the map images so that we can compare them with a knowledge base containing well-known geographical footprints.
4. Feature extraction, recognizing useful features (i.e. text and symbols) in the pixel domain.
5. Feature disambiguation, associating meaning to the extracted features (i.e. discovering the spatial footprint from the placenames in the map)

Note that the feature disambiguation operation can also use similar techniques to the ones described in Section 4.1. Besides placenames occurring in the map, other symbols (e.g. cartouches) can be particularly useful to infer metadata such as the date of creation.

4.3 Map indexing and retrieval

A requirement for the DIGMAP user interface is that it should support the combination of geographic search parameters with parameters like text content and time periods, allowing users to search for resources that are about a designated geo-temporal context and also contain other specific metadata associated. The underlying retrieval component should therefore support all these mechanisms in an integrated approach.

On what concerns text retrieval, DIGMAP will use standard techniques such an inverted index and the BM25 ranking scheme [18]. Although DIGMAP proposes to explore temporal periods associated with the resources, it should be noted that we are not considering versioned resources changing over time. To support searching by time periods or other numeric metadata elements, the index will have built-in support for payload indexes. Payload is the name given to metadata that can be stored in the index together with each occurrence of a term. Support for geographical retrieval will be a special case, for which we will use separate index structures, afterwords merging the results.

Regarding the types of supported geographical queries, we are initially only planning to support the region containment operator. Result ranking will be based on a linear combination of the BM25 metric with specific heuristics for measuring geographical similarity (e.g. the area of overlap) [14].

Many indexing schemes have been proposed for spatial footprints, including grid indexes, quad-trees, R-trees, k-d-trees, and space filling curves such as Z-order [4]. The most popular method is the R-Tree, a balanced tree derived from the B-tree which splits space in hierarchically nested, possibly overlapping rectangles. R-tree indexes are also commonly supported in the open-source world, for instance through extensions to relational database systems. They are therefore good candidates for usage in DIGMAP.

Another possibility, potentially advantageous in terms of simplicity, is the use of the C-Squares indexing system [16]. Having the surface of the earth divided in a grid of labeled squares, at one of a range of scales, the spatial footprints can be represented as a list of C-Square codes, given as a textual string comprising numbers and a separator character. Indexed resources can be matched to a designated query footprint (itself expressed as one or more C-Square codes) using standard text search methods.

5. CONCLUSIONS

DIGMAP started in October 2006, and will run for 24 months. It is not a pure research project, but a tentative to develop innovative services reusing existing methodologies, software and data. An initial version of the system is already undergoing tests, although the GIR functionalities are still at an early stage of development. In the next months, we will concentrate our efforts on developing a fully-functioning prototype.

From a technical standpoint, the project will address methods related to the area of geographic information retrieval. Some of the main challenges have been surveyed in this paper, particularly on what concerns the development of an initial version of the system. In case of success, the project will also direct some attention to the development of advanced functionalities related to the interactive exploration of digitized maps, for instance through interfaces similar to that of Google Maps, built on top of the Web Map Service specification from OGC [3].

Despite the effort on developing automated approaches, DIGMAP is also considering the integration of human operators into the processing workflow. The key is to provide intuitive user interfaces that supply maximum information from minimum user effort, leveraging differently the strengths of humans and machines. Human librarians can, for instance, correct the metadata records generated through automated mechanisms.

The project will make a proof of concept by reusing and enriching the contents from the National Library of Portugal (BNP), the Royal Library of Belgium (KBR/BRB), the National Library of Italy in Florence (BNCF), and the National Library of Estonia (NLE). In a second phase, this will be complemented with contents and references from other libraries, archives and information sources, namely from other European national libraries members of TEL – The European Library. In case of success, the ultimate goal of DIGMAP is to became a service fully integrated with TEL, and in this sense the project is fully aligned with the vision "European Digital Library" as expressed in the "i2010 digital libraries" initiative of the European Commission.

6. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: geotagging web content. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, 2004.

[2] M. Chaves, M. J. Silva, and B. Martins. A geographic knowledge base for semantic web applications. In *Proceedings of the 20th Brazilian Symposium on Databases*, 2005.

[3] J. Fritzke and R. Atkinson. Gazetteer service profile of the web feature service implementation, 2006. OCG Draft Implementation Specification 05-035rl.

[4] V. Gaede and O. Gunther. Multidimensional access methods. *ACM Computing Surveys*, 1997.

[5] A. Gelbukh, S. Levachkine, and S. Y. Han. Resolving ambiguities in toponym recognition in cartographic maps. In *Proceedings of the 5th IAPR International Workshop on Graphics Recognition*, 2004.

[6] J. Gil, P. Castro, R. Pestana, P. Teixeira, C. Ribeiro, A. Wyttenbach, and J. Borbinha. State of the art in image processing for digitised maps, 2007. DIGMAP Deliverable D4.1.

[7] L. Hill and Q. Zheng. Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. In *Proceedings of the 1999 Annual Meeting of the American Society for Information Science*, 1999.

[8] I. Johnson. Putting time on the map: Using TimeMap for map animation and web delivery. *GeoInformatics*, July/August 2004.

[9] C. Jones, A. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In *Proceedings of the 3rd International Conference on Geographic Information Science*, 2004.

[10] A. Kornai. Proceedings of the HLT-NAACL 2003 workshop on analysis of geographic references, 2003.

[11] J. Lansing. Geoparser service draft candidate implementation specification, 2001. OGC Discussion Paper 01-035.

[12] S. Levachkine. Raster to vector conversion of color cartographic maps for analytical GIS. In *Proceedings of the 5th IAPR International Workshop on Graphics Recognition*, 2003.

[13] H. Luo and R. Kasturi. Improved directional morphological operations for separation of characters from maps/graphics. In *Selected Papers From the Second international Workshop on Graphics Recognition, Algorithms and Systems*, 1998.

[14] B. Martins, N. Cardoso, M. Chaves, L. Andrade, and M. J. Silva. The university of Lisbon at GeoCLEF 2006. In *Proceedings of the 7th Workshop on Cross Language Information Retrieval*, 2006.

[15] B. Martins and M. J. Silva. A graph-ranking algorithm for geo-referencing documents. In *Proceedings of the 5th IEEE International Conference on Data Mining*, 2005.

[16] T. Rees. C-squares, a new spatial indexing system and its applicability to the description of oceanographic datasets. *Oceanography*, 16(1), 2003.

[17] J. Reid. geoXwalk - a gazetteer server and service for uk academia. In *Proceedings of the 7th International Conference on GeoComputation*, 2003.

[18] S. E. Robertson, S. Walker, and M. Hancock-Beaulieu. Okapi at TREC-7. In *Proceedings of the 7th Text REtrieval Conference*, 1998.

[19] M. J. Silva, B. Martins, M. Chaves, N. Cardoso, and A. P. Afonso. Adding geographic scopes to web resources. *CEUS - Computers, Environment and Urban Systems*, 30(4), July 2006.

[20] A. Velázquez and S. Levachkine. Text/graphics separation and recognition in raster-scanned color cartographic maps. In *Proceedings of the 5th IAPR International Workshop on Graphics Recognition*, 2003.

[21] L. Vilches-Blázquez, B. Martins, A. Wyttenbach, M. Poveda, M. Álvarez, J. Luzio, and J. Borbinha. Geographical and historical thesauri : The state of the art, 2007. DIGMAP Deliverable D2.1.

[22] M. Wick. Geonames website. http://www.geonames.org/.