José Borbinha[*], Gilberto Pedrosa[**], João Luzio[***], Hugo Manguinhas[****], Bruno Martins[*****]

The DIGMAP Virtual Digital Library

Keywords: Geographic information; cartographic heritage; information systems architectures; metadata; interoperability.

Summary

DIGMAP is a digital library specialized in searching and browsing services for old maps and related resources. The service reuses metadata from national libraries and other relevant third party metadata sources, providing added value services by aggregating all the data in comprehensive collections, browsing indexes and search functions. The services are based in a set of specialized tools, comprising namely a catalogue, an image's feature indexer, a metadata repository, a geographic gazetteer and a geo-parser. The extraction of relevant visual features from images of digitized maps is another focus of the project. The architecture and the technology give it also the ability to easily interoperate with other complementary external services.

Introduction

DIGMAP developed solutions for geo-referenced digital libraries, especially focused on old maps and related resources.

A Resource in DIGMAP is any information object relevant for our scope, such as maps, books or web sites. Resources are described by metadata structures in Dublin Core. For maps, it is possible to register in the metadata its geographic information (geographic boundaries, scales, etc.).

DIGMAP is reusing bibliographic metadata from European national libraries and third party sources, in Dublin Core, UNIMARC[1] or MARC21 [0], provided in generic XML or in specific encodings (e.g. in ISO2709 or in MarcXchange [0]). Old maps are very difficult to describe. But they always represent a limited physical space in the real world. In

* IST – Department of Information Science and Engineering, Instituto Superior Técnico, Lisbon Technical University, Portugal [jlb@ist.utl.pt]

** IST – Department of Information Science and Engineering, Instituto Superior Técnico, Lisbon Technical University, Portugal [gilberto.pedrosa@ist.utl.pt]

*** IST – Department of Information Science and Engineering, Instituto Superior Técnico, Lisbon Technical University, Portugal [gilberto.pedrosa@ist.utl.pt]

**** IST – Department of Information Science and Engineering, Instituto Superior Técnico, Lisbon Technical University, Portugal [gilberto.pedrosa@ist.utl.pt]

***** IST – Department of Information Science and Engineering, Instituto Superior Técnico, Lisbon Technical University, Portugal [bruno.martins@ist.utl.pt]

1 http://www.unimarc.net

DIGMAP, we studied methods for classifying, indexing, searching and browsing maps by their geographic boundaries, with the support of multilingual geographic thesauri (comprising authority files and gazetteers).
The project made a proof of concept reusing and enriching the contents from the National Library of Portugal (BNP), the Royal Library of Belgium (KBR), the National Library of Italy in Florence (BNCF), and the National Library of Estonia (NLE). In a further phase we expect to complement it with contents and references from other libraries, archives and information sources, namely from other European national libraries members of TEL – The European Library (DIGMAP might become an effective service integrated with TEL - 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).

DIGMAP Use Cases

The main DIGMAP use cases are presented in Figure 1.

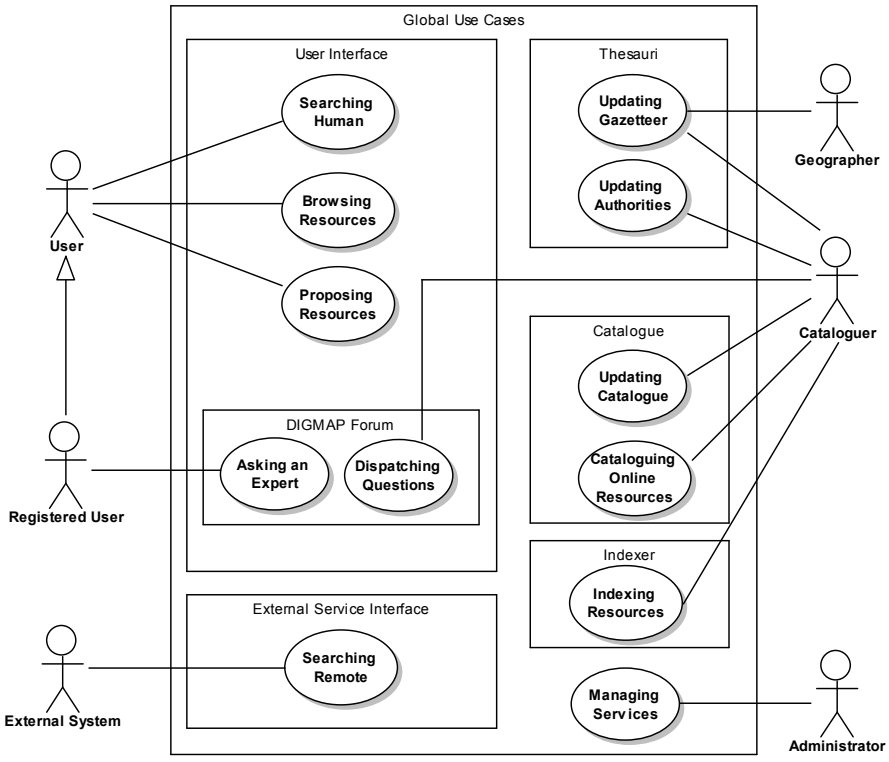

Figure 1: DIGMAP Use Cases.

External users can access resources in two ways: Searching (simple and complex searches in the metadata) or Browsing (browsing in indexes, geographic or temporal spaces). Users also can Propose Resources, or Ask an Expert (ask questions to librarians). The DIGMAP Forum supports interactivity with the users. Through the use case Asking an Expert it is possible to submit questions to be answered by experts registered in the system with specific knowledge in the area. The specialists can answer and manage those questions through the use case Dispatching Questions.

Updating Gazetteer and Updating Authorities support the management of the geographic and authority metadata, which in turn provide support the browsing services.
The collections of bibliographic, authority and geographic metadata records are stored in a catalogue, managed by cataloguers tough the use case Updating Catalogue. Cataloguing Online Resources supports the registration of external resources through, a specific cataloguing interface.
Finally, a back-office component, Managing Services, supports the management of the infrastructure (configuration of services, management of users' accounts, etc.)

DIGMAP Services

The main DIGMAP architecture supporting the use cases described in the previous section is resumed in Figure 2.

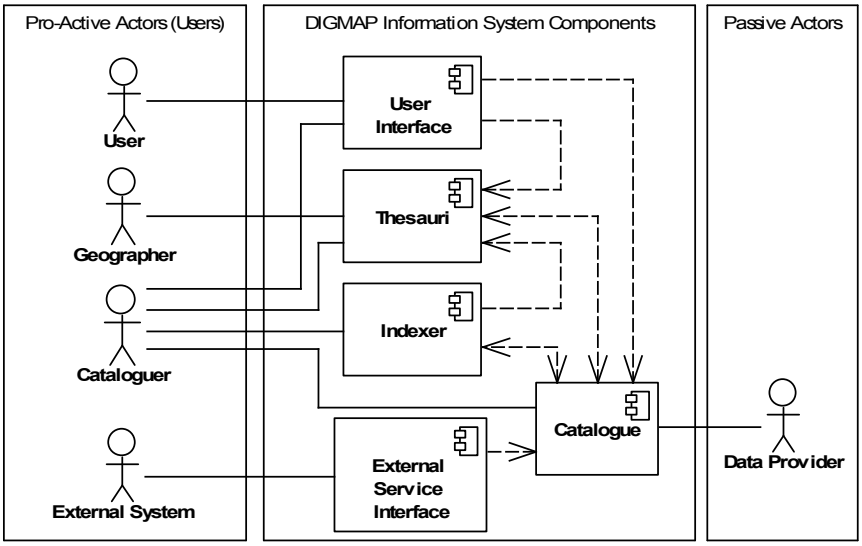

Figure 2: DIGMAP Architecture

The User Interface comprises an OPAC, based in the MITRA search engine [0], and a browsing environment for human users. Browsing can either be made over textual indexes, or using paradigms inspired by Google Maps[2] and Smile Timelines[3].
The Thesauri manages concepts, geographic coordinates, names of places, areas and persons, as also related historical events (with dates or time intervals). The Thesauri is made up of two major sub-systems: the Gazetteer (following the general model proposed for the Alexandria Digital Library Gazetteer[4]) and the Authority File (access to the authors' information, supporting the identification and disambiguation of similar and duplicate names).

2 http://maps.google.com

3 http://simile.mit.edu/timeline/

4 http://www.alexandria.ucsb.edu/gazetteer/

The Catalogue is the subsystem responsible by the management of the records describing the Resources. The Catalogue supports the creation, editing and removal of records, as also the definition of collections, which are ways to group the Resources according to any desired and technically possible criteria. Examples of criteria for collections can be the period of the resource (century, etc.), the type or genre of the resource, the source or provenance of the record, etc.

It is possible to edit or register a new Resource in the Catalogue through a local user interface, or importing sets of records (Z39.50[5] or SRW/SRU[6] or OAI-PMH[7]). Recognized metadata formats are UNIMARC[8], MARC 21[9], and Dublin Core[10], but any other format can be easily integrated. The Catalogue maintains the descriptions of authorities and of the maps in multiple metadata formats, especially MARC21 and UNIMARC (the format provided by the libraries) and Dublin Core. Anyway, once harvested by DIGMAP, all the bibliographic records are converted in the central service to Dublin Core, for a profile that extends the Dublin Core TEL profile [0].

In the following sections we'll stress some of the most relevant detailed components or services.

Adding geographic coordinates

Usually, libraries don't include structured geographic metadata in their bibliographic records. Since we need that information for our specialized search and browsing services, we developed a specific application to make it possible to add that information, as represented in the Figure 3.

It is important to stress that DIGMAP is a virtual digital library, in the sense that it holds only the metadata that describes the resources, but not the resources themselves, which remain in the local libraries or web sites. The resources also can be a) digital-born, b) digitized, or even c) physical resources existing only in the shelves of the libraries. When the resources are digitized maps, it is possible to index them by their geographic boundaries.

In this sense, the workflow to add the geographic coordinates to the maps starts with the harvesting of the metadata from the local servers, and only after that the librarians can perform the indexing. For that purpose they can use the on-line service provided by DIGMAP, which makes it possible to extend the metadata with the new elements.

---

5 http://www.loc.gov/z3950/

6 http://www.loc.gov/standards/sru/

7 http://www.openarchives.org

8 http://www.unimarc.net/

9 http://www.loc.gov/marc/

10 http://dublincore.org/

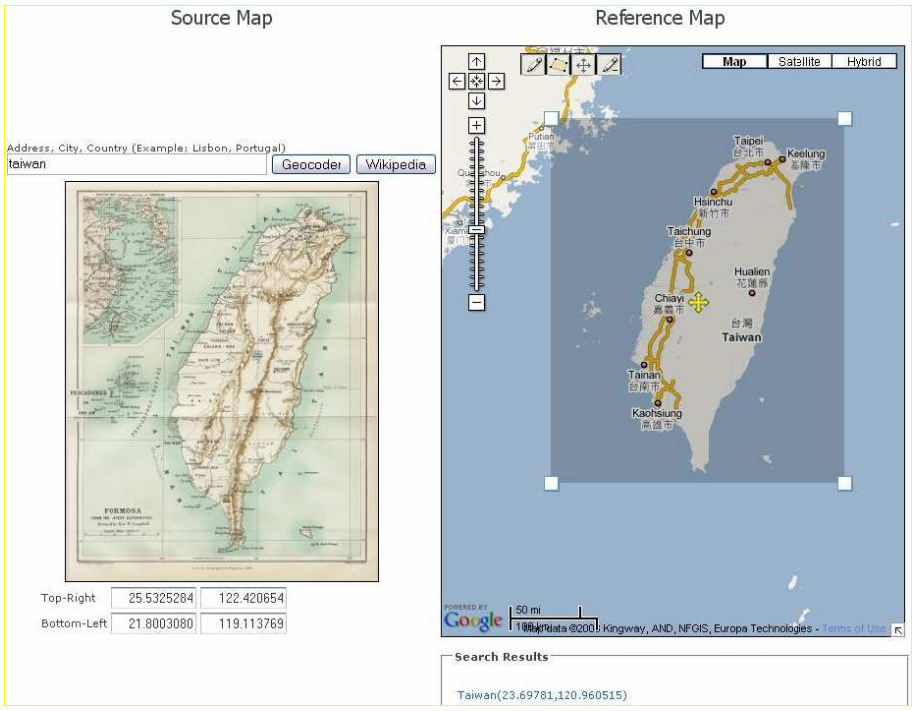

Figure 3: Geographic Indexing of Resources

By default, those extensions are stored only in the DIGMAP central service, but it is also possible to send back to the libraries new versions of their original records (in Dublin Core, MARC21 or UNIMARC) enriched with these new elements.

Cataloguing external resources

The registration of online resources is an important feature of DIGMAP. It allows adding content to DIGMAP from the largest available source: the Internet. For that we developed a specific cataloguing application, cat.on.map, which makes it possible to create records for on-line resources. During that process it is also possible to create automatically thumbnails of the page described.

The on-line resources catalogued in DIGMAP belong to the following collections: maps, bibliography, collections, collectors/dealers, journals, libraries, societies, research/scholar and also a generic collection for “others”

When the resources to be described are on-line maps, it is also possible index them with their geographic boundaries, using the service described in the previous point.

Figure 4 gives an overview of the cat.on.map interface.

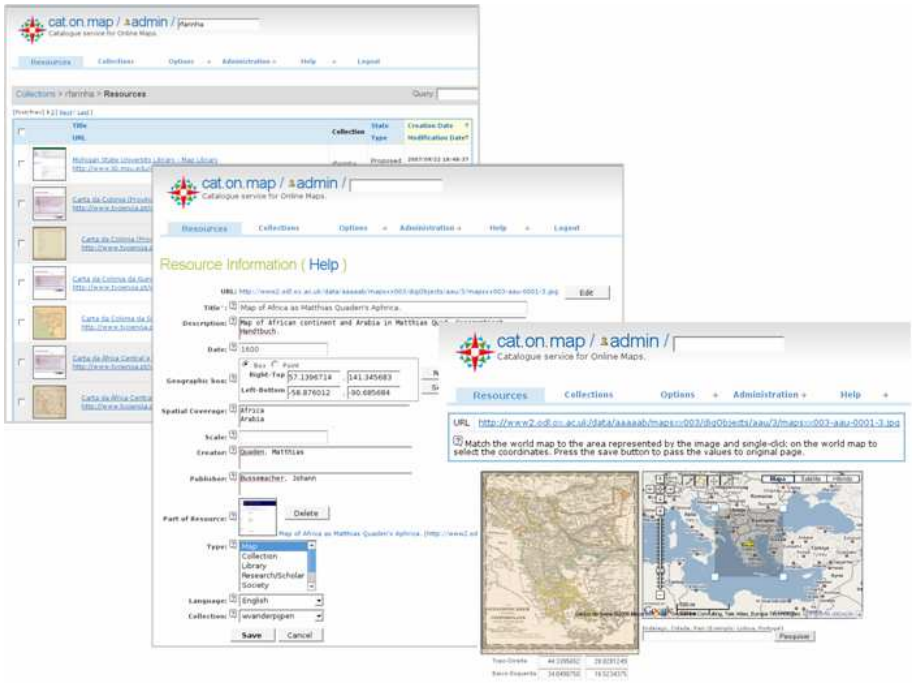

Figure 4: The interface of the cataloguing tool.

Extracting extra features from the maps

Old maps can be very rich in decorative and stylistic details (galleons, mermaids, unicorns, monsters, etc.). For that, we also developed solutions to segment the images and make it possible to index those features.

This is an application that is provided to the libraries, preferably to be used locally, in a network with easy access to images of the digitized images. The application can be also used from the central DIGMAP system, but since the preferable scenario is to use images at 300 dpi (dots per inch), which libraries not always made available through the Internet, this can be an issue.

This tool processes the images and, using only automated image processing techniques, identifies potential features. After that the librarian has to select the relevant one, removing those not relevant.

As a result of the usage of this tool, it is created an object made of a collection of images (one for each feature) and an XML file describing the features.

Figure 5 shows an example of a map, where all the potential features are marked. Figure 6 shows in more detail the features that were considered relevant.

Examples of classes of features that can be considered for this purpose are cartouches, maps' orientations (compass roses, etc.), relief representations, vegetation, graduated borders, scales, boats, sea monsters, etc.

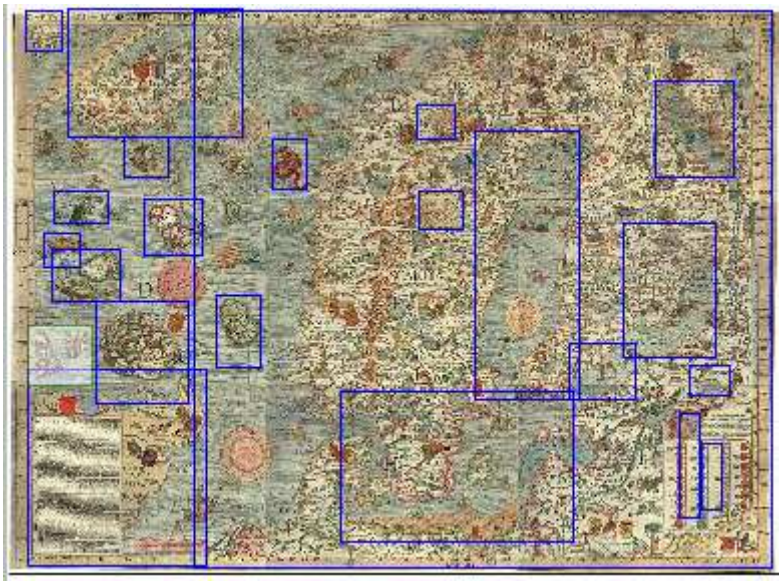

Figure 5: Proposed features from a digitised map

Figure 6: Features extracted from a digitized map

The DIGMAP Gazetteer

The DIGMAP gazetteer is a service integrating data from multiple sources (e.g. the geonames[11] website, the ECAI time period directory [0], etc.). The DIGMAP gazetteer follows the XML Web interface and the data model proposed for the Alexandria Digital

11 http://www.geonames.org

Library (ADL) gazetteer [0], with changes related to the handling of temporal information.
This gazetteer offers the support for another relevant component, the geo-parser, which can identify occurrences of geographic terms in the metadata records. This made it possible to offer geographic browsing, based on textual indexes, the same technique used for the Authority File, as illustrated in the Figure 7.

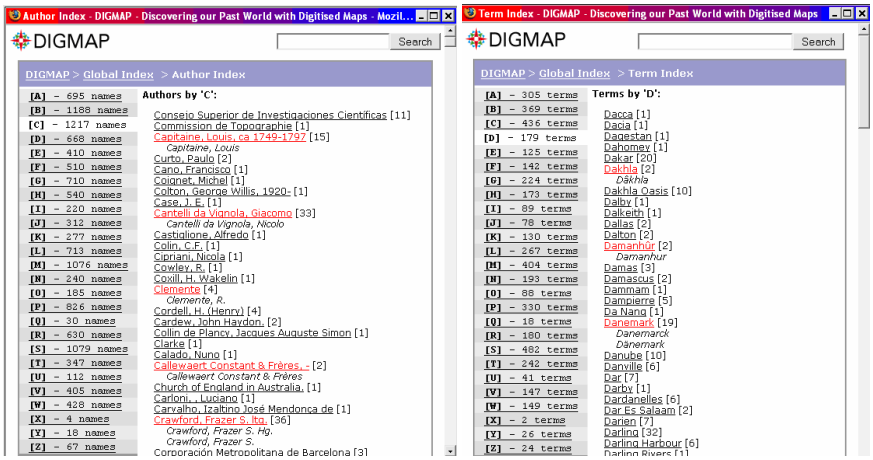

Figure 7: Browsing supported by the Authority File and the Geographic Index

Metadata Repository

REPOX is an XML infrastructure to store, preserve and manage metadata sets. It can play the role of a broker or other specific service in a Service Oriented Architecture, to manage, transparently, data sets of information entities in digital libraries, independently of their schemas or formats. The main default functions of this service are submission, storage, schema transformations and retrieval.
The architecture of REPOX is presented in the Figure 8. The main component is the REPOX Manager, which harvests the records from the data sources via the data source interfaces. A Record Signature is created to assure the integrity of the records, the record and its signature are wrapped in a Record Package and archived in the file system. After that, the Access Point Manager creates and stores the indexes of the access points in a relational database.

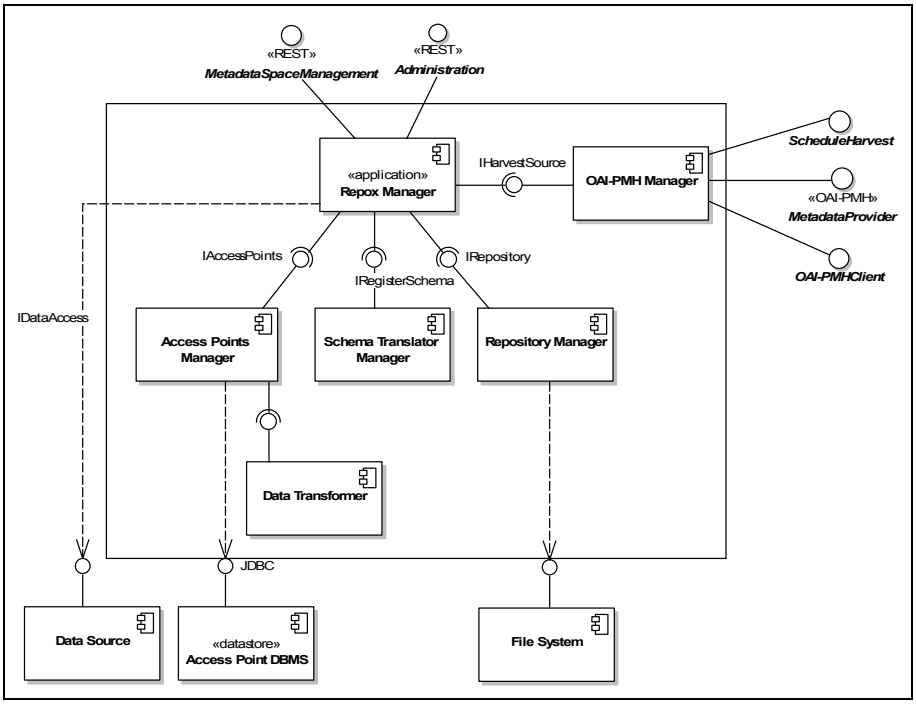

Figure 8: REPOX architecture.

The Schema Translator Manager is responsible for registration of translation schemas and the process of translation between to schemas with an available translation.
An administrative REST interface (Administration «REST») allows a system administrator to create, update or eliminate access points; to export records; to manually trigger the execution of a data source harvest, etc. The scheduling of harvests, OAI-PMH Client and Server requests will be handled by the OAI-PMH Manager.
REPOX plays a fundamental role in DIGMAP, since it is the component that makes it possible to manage the multiple metadata interfaces with the data providers, as also the transformations.
REPOX in itself is an independent valuable service for metadata management. For that reason it is being made available as open-source.

Conclusions and future work

DIGMAP is a project aligned with the TEL – The European Library[12]. An important next step will comprise the interoperability between the DIGMAP service and the TEL portal, making it possible to motivate users searching old maps in TEL to be redirected to DIGMAP.
Currently ongoing work is also addressing the optimization of some of the software components (e.g. adding more data into the gazetteer or improving the automated extraction

12 http://www.theeuropeanlibrary.org/

methods). Finally, the biggest challenge is going to be the integration of this service with the TEL portal, and in the future with the Europeana portal[13].

Acknowledgments

We would like to express our gratitude to our partners of the DIGMAP project for their contribution to this work. This research was funded by the DIGMAP project (ECP-2005-CULT-038042) of the European Community eContentplus programme.

References

LOC - MARC Standards, MARC Format Documentation Overview, December 2006 http://www.loc.gov/marc/status.html

ISO. ISO/DIS 25577: Information and documentation – MarcXchange. 30 November 2005. http://www.niso.org/international/SC4/n577.pdf

Hugo Manguinhas and José Borbinha, Quality Control of Metadata: A Case with UNIMARC. Proc. 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), pages 244-255. Lecture Notes in Computer Science (LNCS) 4172, Springer, Heidelberg, Germany.

Jorge Machado, José Borbinha (2006). MITRA: A Metadata Aware Web Search Engine for Digital Libraries, XATA 2006 – XML: Aplicações e Tecnologias Associadas, Fevereiro 2006, Portalegre, Portugal, pp. 392-393, XATA.

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

M. Buckland and L. Lancaster 2004 Combining Place, Time, and Topic: The Electronic Cultural Atlas Initiative. D-Lib Magazine, 10(5)

Theo van Veen, Robina Clayphan. Metadata in the Context of the European Library Project. Proceeding of the International Conference on Dublin Core and Metadata for e-Communities 2002: 19-26 - Florence, October, 13-17, 2002. Firenze University Press

13 http://www.europeana.eu/