Master thesis

A LOD-based georeferencing tool for historic maps

Alber Sánchez

Director: Dr. Simon Scheider
Advisor: Dr. Carsten Keßler
Institute for Geoinformatics

Abstract

Historic maps represent a snapshot of a past reality no longer available, which makes them an interesting subject of research. Maps are treated usually as regular documents but librarians acknowledge that including additional data concerning space and time and contents would improve search results for users. For achieving this purpose, it is necessary to describe historic maps in such a way that enables search based on spatio-temporal contents.

The inclusion of map metadata such as spatial coverage, publication date and contents in a linked data format provides new opportunities to query maps and to link them with other knowledge sources. This thesis presents a software tool which enables librarians to georeference historic maps, extending the traditional key word search to include space and time criteria. This tool is designed as part of the image referencing process inside libraries and it allows to describe maps as well as their contents by georeferencing and simultaneously linking to contents of the Web of Data. The combination of semantic content descriptions and georeferencing improves the quality and efficiency of both processes. This allows users to discover texts and maps seamlessly using spatial, temporal or content criteria. The relation between maps and its contained geographic features could be calculated using spatial operations or explicitly declared during map georeferencing.

This work as done in the context of the LODUM[1] initiative, the LIFE project[2] and the library of the University of Münster .

[1]Linked Open Data University of Münster `http://lodum.de/`.

[2]Linked Data for eScience Services `http://lodum.de/life/`

Contents

1 Introduction

As part of their foundational mission, libraries keep historic maps. Maps are subject of research with respect to their role as cultural texts, promotional devices, instruments of sovereignty or authoritarian images[26] and they are subject of scientific research. Making these resources available to the public while preserving them is a challenge librarians worldwide are tackling based on modern technologies.

Due to their spatial nature, historic maps deserve special treatment as library resources. Accessing historic maps and relating them to others (modern or historic as well) improves and accelerates research. This is demonstrated by well known historic maps examples like:

- The "*Lienzo de Quauhquechollan*"[3] dates from the 16th century and it is perhaps the oldest map of Guatemala. This map tells a conquest story from the indigenous viewpoint instead of the Spanish Empire's perspective. Unlike traditional snapshot-like maps, the Lienzo represents the history of a geographic region depicting complex spatio-temporal interactions[15]. Handling the Lienzo is difficult because it is composed of 15 cotton pieces but they have been scanned and digitally restored in a project ran by the *Universidad Francisco Marroquin* in Guatemala City.

- The map of Napoleon's march of 1812 over Russia.[4] This Charles Minard's map is a valuable tool for understanding the development and consequences of such a war campaign. This map combines space, time, temperature and the number of soldiers to show "*(...) the catastrophic loss of life in Napoleon's Grand Army*"[21].

- The map of 1854 cholera outbreak in London by John Snow.[5] The story surrounding this map is almost legendary and well-known in the fields of public health, medical geography, history of medicine, geography and cartography. This map displays cholera victims and water pumps suggesting clustering around a specific pump; it is also believed the map showed how an illness is dispersed, to change public health policy and to probe Snow's hypothesis regarding water-transmitted cholera instead of the prevailing miasma the-

[3] *Lienzo de Quauhquechollan* `http://www.lienzo.ufm.edu/`

[4] See the article about Minard under the section on information graphics `http://en.wikipedia.org/wiki/Charles_Joseph_Minard`

[5] Broad Street cholera outbreak `http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak`

ory of that time[41]. Most of these historic assertions are matter of debate today[41][36].

Each of the example maps represents geographic features and people interactions framed as historic events which are difficult to detail using bare text. They not only depict space but also summarizes complex interactions of people with their environment. Details like the Lienzo is emphasizing storytelling sacrificing spatial precision or the fact that Minard preferred data representation accuracy over precise geographical position[21] or the mismatching death count on different version of Snow's map[36] cannot be acquired only by looking at the maps. Similarly, reading documents without observing the maps does not provides a full understanding of the events. For example, linking the historic map of London's cholera outbreak of 1854 to John Snow would help a library user to discover that there are different maps claiming to be the John Snow's map[41].

Academic and scientific research could be improved by including maps as part of result sets on libraries searches. However users searching through library records usually face a gap: Their results consist of written text or images, and their search is based on key word matching, not on the actual geographic contents represented in these images. It would be favourable if search was based on explicit representations of geographic contents instead. This thesis address this gap through the use of Linked Open Data as a means to enable librarians to improve historic map descriptions and allowing them to relate maps to other data sources which improves its discoverability and understanding.

The remaining of this chapter will present general definitions and later it will introduce the research questions guiding this work. The following chapter will explore the concepts presented here in the library and linked data context. The following 2 chapters will present the methodology and introduce the architecture of a software tool implementing some of the ideas exposed here.

1.1 LODUM, LIFE and ULB

The LODUM[6] project is the mechanism by which the University of Münster grants *open acess*[7] and provides *open data*[8] to non-sensitive information[35][33]. LODUM initiative is leaded by the *Institute for Geoinformatics*[9] and the *Universitäts und*

[6]Linked Open Data University of Münster `http://lodum.de/`.

[7]Open access `http://en.wikipedia.org/wiki/Open_access`

[8]Open data `http://en.wikipedia.org/wiki/Open_data`

[9]Institute for Geoinformatics, University of Münster `http://ifgi.uni-muenster.de/`

Landesbibliothek (ULB)[10]; together they launched the LIFE project[11] which aims at improving interdisciplinary collaboration between science and education through spatio-temporal information sharing. This project develops standard-based services and applications to capture, catalogue, manage, find, access and link spatio-temporal data allowing users to improve their spatio-temporal-thematic queries.

The resulting software of this thesis will be integrated with other solutions provided by LODUM and the LIFE project to ULB.

1.2 Historic maps

There are many definitions of map, some function-oriented like "*a particular human way... of looking at the world*"[26] and others concrete and content-oriented:

> *"(Maps) are cartographic products, containing a selection of features for a particular purpose, and created according to the design principles of cartography in terms of abstraction, symbology, colors, labeling, line styles, and legends. Maps are best considered to be a particular type of intellectual object that can be manifested in various physical formats, some fixed and some dynamic".*[30]

The last definition covers the dynamic nature of geographic data nowadays. Besides, it acknowledges that a single map can have many representations. This is interesting for historic maps because the same map could have many printed copies or being part of other library resources like atlases. As other information sources, historic maps are converted to digital formats available on Internet where copies and modifications can be made without control. Here it is an important issue regarding whether or not the map's digital copies must be documented since they are proxies to maps as an intellectual objects.

A definitive map definition is beyond the scope of this thesis but it is acknowledged that a map definition is not complete until roles [26] (cultural text, promotional device, instrument of sovereignty, authoritarian image), functions [19][13](information carrying, explanation of what is in a particular place, enhancing spatial knowledge, spatial knowledge communication, decision support and change of social behaviour due to map use), uses and dimensions [13] are taken into account.

As long as this thesis is concerned, a map becomes historic when its depicted features no longer resemble the present state of those features. This definition is by no means complete but it is enough for the work presented here.

[10]Library of Münster `http://www.ulb.uni-muenster.de/`

[11]Linked Data for eScience Services `http://lodum.de/life/`

The most relevant characteristics of historic maps are[46]:

- They contain information useful to reconstruct past places holding viewpoints of their time. This combined with modern technologies improves historical research.
- Usually, they hold information missing in written sources about contents now forgotten, modified or erased.
- Their accuracy represents the technological state-of-the-art of their time.
- Not only their contents but also the historic maps themselves are research objects .

1.2.1 Historic maps in ULB

The particular case of the *Universitäts und Landesbibliothek*[12] (ULB) has been selected as an example and source of use cases for this thesis. ULB has been realizing challenges regarding its collection of historic maps (4000 approximately):

- Maps do not fit ULB's catalogue. Even though ULB has included maps in its analogue catalogue, there is a plan to move them to a digital one adapted to historic map characteristics.
- There is no way to reference a specific map in a map series (or *Kartenwerk*).
- ULB wants to ease and improve map searching.
- ULB wants to implement map searching using coordinates along with text search.

The library follows this procedure for handling its historic map collection:

1. Cataloguing. ULB produces bibliographic descriptions for new items like historic maps and other documents which are included in the library's and state's catalogue.[13]
2. Scanning. High-resolution digital representations of historic maps are produced. The maps are scanned with calibration devices like rules and color tables which help users get an idea of the original map. Some examples are available on the Internet.[14]

[12]Library of Münster `http://www.ulb.uni-muenster.de/`

[13]North Rhine-Westphalia?s *Hochschulbibliotheken* `http://okeanos-www.hbz-nrw.de/F`

[14]Library of Münster's sample historic maps `http://sammlungen.ulb.uni-muenster.de/nav/classification/116654`

3. Searching tool. A specific piece of software will be develop to ease map finding. ULB users would be able to search using spatial, temporal and content criteria in order to retrieve documents, including historic maps.

4. Publishing. The last step consist on publishing the historic maps, for example using Open Geospatial Consortium Web Map Service (WMS).[15]

Currently, ULB is in step 2. The map's metadata gathered by the software presented here (see chapter 4) will be used by the searching tool mentioned in step 3 for finding the historic maps available in the library's catalogue.

1.3 Georeferencing

Georeferencing is the process of relating information to some explicit geographic location. It can be informal (named-feature or discrete) when places are referenced by name or formal (or continuous or quantitative) when a structured method (like latitude and longitude) is used[30]; the translation between formal and informal georeferencing methods can be done using gazeteers which are dictionaries of placenames[30]. A good georeference must be unique, shared and persistent: Unique implies it assigns a single location to a feature avoiding confusion, shared means it is known by people using it and persistent indicates it should stand through time [40]. In a broader sense, georeferencing not only relates information to space, but also to time and to other information.

In computer science, spatial georeferencing is known as registration and it is mainly applied to digital images. Image registration is "*the process of overlaying images of the same scene taken at different times, from different viewpoints and or by different sensors*"[52]. It has applications in many fields including remote sensing, cartography, computer vision and medicine. It can be useful for[12]: integrating information from different sensors, finding changes at different times or under different conditions, inferring 3D information and modelling based on object recognition.

Image registration can be extended to maps: *Georegistration* is the process of transforming an image (of a historic map) from its local reference system to the coordinate reference system of a map of reference. The map of reference is a map preferably made according to modern standards, using a widely accepted spatial reference system. A reference system is a set of rules for making measurements[14] and "to transform" means changing coordinates between reference systems based

[15]Web Map Service `http://www.opengeospatial.org/standards/wms`

on different datums[31]. Datum is an important concept to reference system and for georeferencing:

> *"A datum defines the position of the origin, the scale and the orientation of the axis or axes of a coordinate system with respect to an object"*[31]

As a matter of fact, the concept of datum also applies to temporal reference systems (i.e the Coordinated Universal Time is a temporal reference system[16]) and to semantic reference systems[37]. As a result, it is possible to reference a piece of geographic information to at least 3 reference systems: spatial, temporal and semantic.

1.4 Linked Open Data

Linked Data is a set of good practices and principles for publishing and linking structured data in the Web [27][10]. Its principles are: Using Identifiers (URI[17]) for naming things, using HTTP[18] for reaching data, using standards for structuring and querying the data (like RDF[19] and SPARQL[20]) and linking the data to other data[27]. When the data is open available and published using an unrestricted licence then it is Linked *Open* Data (LOD).

LOD inherits Web characteristics by using previous Web standards. Data silos are avoided when URIs and HTTP are used because they enable navigation and data discovery while using RDF enables description of both data and schema, therefore enabling data fusion and expressive querying.

In LOD, data schemata are handled using vocabularies. Vocabularies are collections of classes and properties which use terms of RDF Schema (RDFS)[21] or Web Ontology Language (OWL)[22] or Simple Knowledge Organization System (SKOS).[23] These languages allow to describe domains of interest of any complexity[9][27]. Vocabularies are built using the same language and principles as the entities and relationships they describe; as a result, when some piece of data is unknown, it is possible

[16]Coordinated Universal Time `http://en.wikipedia.org/wiki/Coordinated_Universal_Time`

[17]Uniform Resource Identifier

[18]Hypertext Transfer Protocol

[19]Resource Description Framework `http://www.w3.org/RDF/`

[20]SPARQL Protocol and RDF Query Language `http://www.w3.org/TR/rdf-sparql-query/`

[21]RDF Vocabulary Description Language `http://www.w3.org/TR/rdf-schema/`

[22]Web Ontology Language urlhttp://www.w3.org/TR/owl-features/

[23]SKOS `http://www.w3.org/2004/02/skos/`

to follow its vocabulary's URI to find a definition for it[9]. It is also possible to write a tailor-made vocabulary suiting a specific datasets but good practices encourage first to reuse well-known existing ones [24] [25] [27]. As they are LOD, Vocabularies can be linked to other Vocabularies enabling data translation. In consequence LOD is self-descriptive and an appropriate tool for data heterogeneity handling[27].

The standards, technologies and principles underlying LOD makes possible for anyone to publish and link data just the same way anyone can publish and link Web pages. Historic maps could be annotated by enthusiasts and people native to the historic map's area; they could add valuable information for research. When new links are added, the potential of new links grows exponentially while keeping data distributed. In this environment, historic map' content could be better explained and understood from different viewpoints. This is an interesting point of convergence between LOD, history and Collaboratively Contributed Geospatial Information[8] also known as Volunteered Geographic Information (VGI)[24]: VGI can boost the LOD available while VGI can take advantage of LOD's structuring and linking principles for historic research.

1.5 Research question

Libraries are moving to the Web[2][49][20]. Achieving this successfully is challenging and it requires to move from the traditional keyword searching to search by contents. Realizing that different document-types require different and specific descriptions is a first step towards this goal. At the same time, a homogeneous indexing-describing-searching platform is highly desirable. Web technologies such as Linked Open Data may fit this requirement, however, description standards for non-textual documents remain unknown. The specific case of historic maps brings forward the following challenges:

How to describe and publish historic maps in order to make them queryable by content?

Seamless integration between documents in library's catalogues has the potential to improve historic research. However, there is the additional complexity caused by space, time and contents. Enabling search by any of these 3 dimensions is de-

[24] Best Practice Recipes for Publishing RDF Vocabularies `http://www.w3.org/TR/swbp-vocab-pub/`

[25] Swoogle semantic Web search `http://swoogle.umbc.edu/`

sirable, however, all of them have their own issues raising some subsidiary questions:

What are the specific challenges when georeferencing historic maps?

Georeferencing a historic map demands the identification of at least three underlying reference systems (spatial, temporal and semantic). Particularly, the spatial and semantic reference systems are challenging because of issues related to cartographic projections and precision in the first case and because of heterogeneous contents, purposes and uses of maps in the second.

What metadata should be used for describing historic maps in the library context?

Historic map descriptions in the library context must include information of maps as a documents and information about their cartographic content. The description of the space, time and contents of maps influence their visibility in a library. In order to increase the chances of discovering maps, a careful selection of descriptive vocabularies and standards must be performed.

How can LOD-background knowledge be exploited for improving map search or georeferencing?

LOD is transforming the Web into a global self-describing data graph. LOD provides a large amount of resources but the way they can be used to improve the georeferencing of historic maps or how they can be used to improve the search for resources in a library context is an open question.

1.6 Outline

This master thesis describes a practical approach which answers the aforementioned research questions and tests the proposed solution by means of a Web application which is evaluated in the library context. The application targets librarians and it intends to help them better georeference historic maps in a library catalogue by using Linked Open Data.

This thesis is organized as follows: The next chapter introduces concepts regarding georeferencing along their applicability to historic maps as well as a summary of software tools related to georeferencing historic maps. Chapter 3 presents the approach used to build the tool and it also presents the answers to the research

questions. A detailed description of the software tool is presented in chapter 4 and chapter 5 concludes this document by discussing results and pinpointing future development.

2 Georeferencing historic maps

This chapter presents in more detail the georeference and gazetteer concepts and finishes with a software review of tools used for historic map georefrencing. Concepts are required for answering the research questions presented in this thesis and the software review sets the tool into the context of the state-of-the-art regarding historic map georeferencing.

When a new resource arrives to a library, it is catalogued and an identifier is assigned to it. Afterwards, metadata describing the resource is collected, classified and stored to enable users to discover the new resource. Usually this is enough for analogue catalogues but for a Historic-map-aware digital catalogue the next step is georeferencing.

Georeferencing, in a broader sense, includes finding or defining the spatial, temporal and semantic reference systems[38] of the digital representation of a historic map. A raster image is the result of scanning a historic map and thus, the image can be seen as a representation of that map. The characteristics of a historic map can be described through this image representation however this involves additional issues:

- Accuracy. The scanning process introduces inaccuracies in the resulting digital representation of the map.
- Resolution. Scanning resolution determines the amount of recognizable details captured in the map's raster image. The required scanning resolution for minimizing losing of detail changes from map to map.
- Format. There are many digital formats for storing raster images. Format selection is a trade-off between size and image quality since most formats apply compression algorithms. This decision impacts raster's reading and transmission time.
- Rasters have an image coordinate system of their own. As a result, the process of overlaying a historic map on top of a reference map implies a double transformation of coordinates: From the historic map to the raster image and from there to the reference map.

Finding or defining the spatial, temporal and semantic reference systems of a historic map are influenced by these issues and they must be taken into account by librarians when building digital catalogues.

2.1 Spatio-temporal gazetteers for historic maps

As stated before, gazetteers provide translations between formal and informal means of georeferencing as they are placename dictionaries including geospatial footprints for the named locations[30]. They are made of name-type-location tuples $< N, t, g >$ [30] resembling geographic information tuples $< x, y, z >$[25]. Gazetteers including time and space footprints are more adequate for georeferencing historic maps because matching historic map contents to a gazetteer's tuple is georeferencing actually in the spatial, temporal and content dimensions at the same time. Besides, a gazetteer-linked historic map enables discovery of information such as additional geographic contents or alternative geometric representations.

A glance to the Linked Open Data Cloud[26] reveals that metaphorically "all roads lead to..." DBpedia.[27] The DBpedia project extracts information from Wikipedia[28] for publishing it under Semantic Web standards and Linked Open Data principles[39]. The extracted information includes different language editions of Wikipedia and its internal and external links, images, geographic coordinates and information tables[39]. This raises the question about whether or not DBpedia could be a spatio-temporal gazzeter in the LOD context; the main requirements for the Next Generation Gazeeter (NGG) are[34]:

- Harvesting and integration: The NGGs must include community-maintained features requiring a distributed approach.
- Assessing fitness for purpose: Gazeeter uses differ depending on the application. Deciding if a specific gazetteer fits an application or not can be troublesome specially when including ever-changing user generated content and data accuracy.
- Retrieval, querying and navigation: NGGs must use ontologies for ensuring properties like: to be consistent, to support complex queries, to allow intuitive navigation and to extend the number of available relationships between contents.

DBpedia meets the first requirement through its dependence on Wikipedia and Wikipedia community. Regarding the third one, DBpedia has an ontology of its own described as *(...) a shallow, cross-domain ontology, which has been manually created based on the most commonly used infoboxes within Wikipedia.*[29] This ontol-

[26]The Linking Open Data cloud diagram `http://lod-cloud.net/`
[27]DBpedia `wiki.dbpedia.org`
[28] Wikipedia, the free encyclopedia `http://en.wikipedia.org`
[29]The DBpedia Ontology `http://dbpedia.org/Ontology`

ogy consists of 320 classes described by 1650 properties with a maximal depth of 5[39]. The ontology is generic and its schema depends in Wikipedia contents with no specific focus on historic maps. This constraints its use in the historic research context although it fits well to the broad range of contents available in libraries (requirement number two). A complete evaluation of DBpedia as a gazetteer is beyond the this work's scope, but it is reasonable to think it is a valuable source of space and time footprints where to link historic map features.

Another source of geographic linked data is LinkedGeoData.[30] LinkedGeoData is similar to DBpedia, but instead of using Wikipedia it uses OpenStreetMap (OSM)[31] as data source and by means of a light weight ontology, it publishes OSM data as Linked Open Data[50].

2.2 Map georeferencing

Libraries usually scan historic maps at high resolution (at least 600 pixels per inch) resulting in large files [46]. Some additional elements are scanned along the map like rulers and color scales[32]; those elements keep track of scanner calibration and paper characteristics. When a paper map is scanned, a raster image is produced. Such image resembles the map to some degree. One paramount characteristic of this image is its coordinate system. A raster is an array of pixels; each pixel o "picture element" is squared and usually holds an integer value. The position of a pixel in an image can be represented in terms of x and y coordinates relative to the first pixel in the image. This first pixel is located at the top-left position and it has the index $x = 0$ and $y = 0$. The x position increases when moving rightwards from the origin while the y when moving downwards. The cartographic contents of a historic map can be georeferenced using its raster image coordinate system. In this way, the map contents can be overlaid on top of a map of reference by applying a transformation of coordinates.

In contrast, most historic maps do not state their spatial reference system. Finding the spatial reference system of a historic map is a complex but interesting task since it would provide valuable historic insight[51] [4] [11]. A common image registration process includes 4 steps[52]:

1. Feature detection. It consists in pinpointing (obtaining the coordinates) of

[30] LinkedGeoData. Adding a spatial dimension to the Web of Data `http://linkedgeodata.org/`

[31] OpenStreetMap. The Free Wiki World Map `http://www.openstreetmap.org`

[32] A scanned historic map example is available at `http://sammlungen.ulb.uni-muenster.de/hd/content/pageview/1617755`

outstanding features in both the historic and modern maps. An outstanding feature is also known as Point Of Interest (POI).[33]

2. Feature matching. Features are matched and used for finding a transformation function. A pinpointed POI in bot the Historic and reference maps is known as a Control Point (CP).[34] Unmatched features are rejected or ignored by the transform model.

3. Transform model. It includes the selection of a type of transformation and its parameters are estimated from the coordinates of the matching features.

4. Image resampling generates a new map-image where the pixels of the scanned historic map are relocated in their corresponding locations in the map of reference. Since origin and destination rasters seldom overlap exactly, some pixels in the new map-image are unmatched. Resampling gives those pixels a value too, most of the time by interpolating nearby pixels.

In the library context, feature detection and matching (steps 1 and 2) are done manually due to the large set of different cartographic symbols used in maps, however the use of a gazetteer can speed the process. Regarding the third step, the transformation models used in cartography are[31]:

- Similarity transformation: It preserves the internal geometry of the transformed coordinate reference system, that is, the orthogonality and scale along the X and Y axes. This transformation applies 3 operations: scale, rotation and translation. It requires at least 4 parameters or in other words, at least 2 matching features (represented by points).

- Affine transformation: It is used when the scale or the orthogonality along X and Y axes in the involved spatial reference systems is not kept. It requires at least six parameters (3 matching features) and for this reason its able to correct effects such as paper shrinkage in X and Y directions independently.

- Polynomial transformation: It is a 12 parameter transform method. It can correct distortions specific to satellite image . The equations for this transformation can become unstable when datasets of very different geometry are being linked.

[33]Point of interest `http://en.wikipedia.org/wiki/Point_of_interest`

[34]Control point `http://support.esri.com/en/knowledgebase/GISDictionary/term/control_point`

There are two ways for handling spatial relations between geographic features in LOD: One is by explicitly stating the spatial relationship and the second one is by testing the relationship using each feature's geometry. In the first case, a RDF triple is built using two features and a property defined as a spatial relationship, for example, the Open Time and Space Core Vocabulary[35] includes as properties the region connection calculus relationships. In the second case, already existing GIS algoritmhs can be used for testing Egenhofer spatial relatonships [16][17]. In the first case plain SPARQL can be used to discover a relationship, however it is constrained to the stated spatial relation. The second case requires processing the geometries involved and testing the relation. This second approach is actually implemented in GeoSPARQL.

GeoSPARQL[36] is an Open Geospatial Consortium (OGC)[37] standard establishing a vocabulary for describing geographic data using RDF, a set of spatial functions and a set of query transformation rules[5]. GeoSPARQL can use two encodings for describing a feature's geometry: The first one is Geographic Markup Language (GML)[38] which is XML[39] based and the second one is Well Known Text (WKT) which has been slightly modified to mention the spatial reference system in each feature encoding[5].

2.3 Temporal georeferencing

Historic maps register a set of features in their content but they say little about the temporal characteristics of their features. It could be assumed a feature was observed when the map was published but map making is a process spanned over time. Besides, it is known it was customary to create maps from other maps to which no reference was kept. In other words, a feature in a map doesn't mean it was observed in the map's publication year or even worse, it doesn't mean the feature existed at all.[40][41][42]

The temporal referencing of contents is constrained because maps do not provide

[35] Open Time and Space Core Vocabulary Specification `http://observedchange.com/tisc/ns/`

[36] GeoSPARQL - A Geographic Query Language for RDF Data `http://www.opengeospatial.org/standards/geosparql`

[37] Open Geospatial Consortium `http://www.opengeospatial.org/`

[38] Geography Markup Language `http://www.opengeospatial.org/standards/gml`

[39] XML. Extensible Markup Language `http://en.wikipedia.org/wiki/XML`

[40] Phantom island `http://en.wikipedia.org/wiki/Phantom_island`

[41] Sandy Island: The Island That Never Was `http://newswatch.nationalgeographic.com/2012/11/29/sandy-island-ile-de-sable-or-ile-de-sables-the-island-that-never-was/`

[42] South Pacific Sandy Island 'proven not to exist' `http://www.bbc.co.uk/news/world-asia-20442487`

time footprints for them and the map's publication date cannot be extended to its contents as an observation time. Unfortunately, historic maps tell more about features' spatial rather than temporal properties. As a consequence, the only map content with some temporal certainty is the map itself as a published document.

From the LOD perspective, some initiatives for a time SPARQL have been proposed but so far none of them have become an standard. An implementation of Allen's interval algebra[1] would be useful in the historic map and library context because it would enable temporal reasoning and simplify SPARQL syntax the way GeoSPARQL did it for the spatial dimension.

2.4 Content georeferencing

Once the historic map's spatial reference system is found or defined, the map's content can be georeferenced. The proper exploration and use of a historic map requires two kinds of information: Information related to the map as artifact and information regarding its cartographic contents[23]. Map layout elements like the legend, the title, the scale and the map index are source of the first kind of information which is usually stored as metadata following bibliographic standards. The cartographic contents are those displayed in the map area. The map area is a holder for cartographic features and its importance lies in the fact it can be used for querying spatial relationships based on contents coming from a different source.

As mentioned earlier, LOD can use classes and properties from OWL ontologies to describe features. An approach from the museum community which address the ontological description of historic maps is portrayed in [22]. It employed the CIDOC Conceptual Reference Model (CIDOC CRM)[43] to describe a historic map. CIDOC CRM is a top-level ontology which states definitions for describing concepts and relationships used for documenting cultural artefacts[22]. The authors had trouble to unambiguously describing map scale, spatial reference system, orientation, legend, heights, accuracy and the link between a map and a collection. As an answer, they suggested to introduce additional classes and properties to CIDOC CRM and to determine a geographic ontology for describing historic maps' geographic features and its relations. In [23], unlike [22], an OWL ontology was build from scratch but it was not aligned to a top-level ontology.

Other approach consists in assuming that semantic reference systems for historic map contents can be based on ontologies. These ontologies need to be found (or created) and liked to historic maps. However, it must be kept in mind that librarians

[43]The CIDOC Conceptual Reference Model `http://www.cidoc-crm.org`

cannot be asked to create ontologies on the fly while georeferencing historic maps. Instead, they may declare the map' contents as instances of a predefined ontology which must be generic in order to cover the wide range of topics handled in a library.

2.5 Software tools

Applications regarding historic maps have two main functions: Spatial georeferencing and feature annotation. Here are some examples:

- Georeferencer[44] is a Web application for georegistering online images of historic maps. The users can georegister a scanned historic image available on internet using control points and Google Maps[45] as background. The application can generate KML[46] files or alternatively it can host the map as a Web Map Service (WMS)[47]; both alternatives can be explored with a standard compliant software with Internet access. An interesting additional feature allows displaying distortion grids enabling the visual assessment of map's precision, this is built on top of MapAnalyst. Georeferencer is not an open source project and it charges companies for hosting historic maps and offer them as WMS.

- MapAnalyst[48] is a desktop application for analysing old maps' accuracy. It calculates distortion grids, displacement vectors, inaccuracy circles and scale isolines.

- ACME mapper[49] is a general purpose mapping application. It allows users to pinpoint and link places.

- EuropeanaConnect is a project for enabling users get involved in the Europeana digital cultural heritage Web portal.[50][48] It extends europeana with annotating capabilities for portal's contents (images, maps, hypertext, audio and video). It provides 3 main functionalities: Map browsing, map image georegistration and annotating capabilities for points, lines and polygons with server-side storing. A feature's footprint is georegistered using SVG respect to the map image's spatial reference system. This implies that a change in

[44] Georeferencer. Online georeferencing tool for scanned maps. `http://www.georeferencer.org/`

[45] Google maps `http://maps.google.com`

[46] KML `http://www.opengeospatial.org/standards/kml`

[47] Web Map Service `http://www.opengeospatial.org/standards/wms`

[48] MapAnalyst. The map historian's tool for the analysis of old maps `http://mapanalyst.org/`

[49] ACME Mapper `http://mapper.acme.com`

[50] Europeana. Think culture `http://europeana.eu/`

the historic map's SRS (i.e by adding a new control point) has not impact feature's location, however, footprints need to be transformed each time they are requested for displaying on a map.

- YUMA map annotation tool provides scholars with a social annotation tool for studying historic maps[47]. It consists on a digital map in top of google maps which allows users to create, edit and reply annotations. The annotating process has a semi-automatic linking approach: The user draws a polygon and then the application suggest the countries and relevant cities nearby. Then the user types text which is analysed using Named Entity Recognition NER (OpenCalais webservice) and the recognized ones are linked to a LOD dataset URI. Annotation are themselves LOD.

- *maphub* is a georeferencing and annotation tool for historic maps[51] built by reusing YUMA's code. It implements the same functionality as YUMA.

From this thesis approach, not only spatial, temporal and content georeferencing is important but also access. These criteria can be used to compare software (see table 1 on page 19). Additionaly, some insight can be extracted from reviewing these applications:

- Using a well known map of reference. Google maps[53] and OpenStreetMap[54] are common choices. Furthermore, the more popular the spatial reference system, the more transformation available to local reference systems.

- Enabling users to explore historic maps in different applications.[55] A *georeference once, see anywhere* policy can help library's users overlay historic maps with other geographic information in their GIS software of choice. Publishing historic maps as WMS or downloadable KML files are common options.

- User-generated content sharing. Users want to share data which is fundamental for VGI initiatives and it has the potential to improve data quality.

From the LOD perspective, the aforementioned characteristics could be provided:

[51]maphub `http://maphub.github.io/`

[53] Google maps `https://maps.google.com/`

[54]OpenStreetMap. The free wiki world map `http://www.openstreetmap.org/`

[55]David Rumsey Map Collection. View collection `http://www.davidrumsey.com/view/view`

- GeoSPARQL vocabulary can be used in combination with geographic information (e.g. OpenStreetMaps as map of reference) and a map API (e.g. OpenLayers[56] or LeafLet[57]) to create a ligh-weight georeferencing application for historic maps.
- GeoSPARQL vocabulary supports two geometric encodings (GML and WKT, see 2.2) which are widely used by GIS software and both of them support a large quantity of spatial reference systems.
- LOD principles like using URIs and make them dereferenceable[27] enhance data sharing by allowing unambiguously resource identification and providing resource descriptions when browsing URI.

Additionally, LOD applications follow a pattern in which users can employ a tool to find resources or directly access the data repository through an SPARQL endpoint. This enables an unsatisfied user to build his own application. For example LODUM[58] and DBpedia[59] follow this pattern.

[56]OpenLayers: Free maps for the Web `http://openlayers.org/`

[57]leaflet. An open-source javascript library for mobile-friendly interactive maps `http://leafletjs.com/`

[58]LODUM SPARQL endpoint `http://data.uni-muenster.de/php/sparql/`

[59]DBpedia SPARQL endpoint `http://dbpedia.org/sparql`

Tool	**Spatial & temporal georef.**	**Content georef.**	**Map access**
Georeferencer (Web app.)	Spatial georeferencing using Google maps as reference. Accuracy analysis. It uses map's publication year.	Unavailable	KML files, WMS hosting (organizations are charged for WMS hosting).
MapAnalyst (Desktop appl.)	Spatial georeferencing and accuracy analysis.	Unavailable.	Files.
ACME mapper (Web app.)	Spatial georef. not provided. Google maps as map of reference.	Content pinpointing.	URL with encoded key-value pairs.
Europeana Connect[48] (Web app.)	Spatial georef, unknown map of reference.	Users draw a geometry and type a description. The application suggest tags using tag clouds based on Web services (Geonames)	Europeana Website.[52]
YUMA[47] (Web app.)	Same as Europeana Connect.	Same as Europeana Connect except for the tag clouds.	Unknown.
maphub (Web app.)	Same as YUMA.	Same as YUMA.	Unknown.

Table 1: Software comparison.

3 Methodology

This chapter presents details on what principles the application was built and how it answers the research questions introduced in section 1.5.

This thesis' approach consist of georeferencing historic maps in the spatial, temporal and content dimensions. Correspondingly, the application introduced here uses as datums the World Geodetic System of 1984, the Gregorian calendar and DBpedia ontology and it assumes a scanned map is a proxy for the purpose of describing the source historic map.

Regarding georeferencing, the application avoids the issue of finding the historic map's spatial reference system and instead assumes the historic and reference maps have the same one. The reference map employed by the application introduced here is provided by OpenStreetMap[60] using the World Geodetic System of 1984 (WGS84) as spatial reference system.[61][62] Finding reference features for positioning control points (or feature detection, see section 2.2) is manually done by users as well as the feature matching. Since the more Control Points (CP) the better the georeferencing accuracy, LOD is used here to accelerate the process by suggesting Points Of Interest (POI) as possible reference features. These POIs are retrieved from DBpedia after the users has matched 3 features (the minimum required for a similarity transformation from the raster to WGS 84 coordinates) by querying the most populated cities in the map area. When the user clicks on a suggestion, both Historic and reference maps are panned to the location for the user to generate a new control point.

The temporal georeferencing is handled by asking the user to type the map's publication year. Then the application is able to suggest links to modern places and also to historic events happening during the map's publication year. The former by retrieving DBpedia resources covered by the map area and the latter by retrieving events from DBpedia using the map area and the map's publication year. This can enhance the historic map with links to regions on it or to non-existing placenames in the map's time. For example, in the first case, a historic map from 1630 depicting an small part of west Germany can be linked to the Berg State (from the 12th to the 19th century)[63]; in the second case, the same map can be linked to the modern Federal Republic of Germany.

Users can also type a free text description of the map. In this case, the application

[60] OpenStreetMap. The free wiki world map `http://www.openstreetmap.org`

[61] EPSG:4326 `http://spatialreference.org/ref/epsg/4326/`

[62] World Geodetic System `http://en.wikipedia.org/wiki/World_Geodetic_System`

[63] Berg (state) `http://en.wikipedia.org/wiki/Berg_(state)`

suggests links by means of DBpedia Spotlight web services[42] which finds DBpedia's resources matching user's map description. As with the other suggestions, it is the librarian using his expertise who chooses to link or not the map to a specific DBpedia's resource. The approach presented here is convenient for users (librarians) because they do not have to define map's contents but link the map to them.

3.1 Why LOD?

In the World Wide Web search engines have prevailed because key word tagging is the fastest way to index large sets of heterogeneous Web resources. However, the key word approach is limited when users have complex queries for specific data and no time to browse over large result sets. Search engines return a list of Web pages instead of straight answers to questions such as *what countries are depicted in Minard's map of Napoleon's campaign of 1812?* or *what are the modern names of those territories?*; *what are the names of the streets in John Snow's map?* or *do they still exist?*; *what is the temporal extent of the story told by the the "Lienzo de Quauhquechollan"?*.

This is a potential encounter point between libraries and LOD: Libraries are experienced in organizing, structuring and cataloguing knowledge, which is needed to overcome key word searching limitations and on the other hand, LOD allows to link data resources to libraries in a structured fashion.

The role of gazetteers in libraries is enabling spatial and temporal georeferencing while disambiguating resources[30]. For this reason, they are being taken to the Web[29][30][34] as web services.[64] Gazetteers are important to georeferencing because they *(...) provide translation between formal and informal means of georeferencing*[30], that is, they allow to translate place names into coordinates.

However, it is better to use LOD-enabled gazetteers for georeferencing historic maps instead of web services in order to use a single technology. A next generation of LOD-enabled gazetteers will increase data discovery, blur silos and provide structure to Web contents. In this way, the same map features could be discovered through LOD-gazetteers in different historic maps displaying its changes over time.

It could be argued that an XML-based language could be used to describe historic maps instead of RDF. This is possible and in fact, there is the "Historical Event Mark-up and Linking" (HEML)[65] language which *(...) contains useful elements to annotate digital maps with historical information*[28]. The advantages of using RDF

[64]WFS Gazetteer Profile `https://portal.opengeospatial.org/files/?artifact_id=46964`
[65]Heml `http://www.heml.org/`

lie in data interoperability and re-use, but both characteristics cannot be appreciated when looking at an individual application but the whole Web.

XML-based languages focus on data representation and interchange instead of semantics and linking, which is the case of RDF. Besides, XML is tree-oriented which means it is suited to represent hierarchies whereas RDF is graph-oriented and this resembles better the structure of the hyperlink-based Web. In addition, XML can represent the same information in many different ways and of course flexibility is good for data creation, but it can lead to ambiguities in data consumption. Not to mention that combining Schemas describing data from different sources is more difficult in XML than when RDF is used[7]. Additionally, the LOD principle of identifying things with URIs does not reduce information duplicity but provides the mechanisms to realize when two identifiers refer the same thing (e.g the Web Ontology Language's *sameAs* property[66]). This improves information re-use[7][6].

One final argument in favour of using RDF and LOD is discoverability. This can be achieved using LOD to relate resources inside a library; self descriptive links going from one resource to other help users narrow data exploration to relevant data instead of the blind hyperlink-based exploration offered by HTML.

3.2 How to describe and publish historic maps in order to make them queryable by content?

A historic map can be published on the Web as a hyperlinked image in an HTML page or as a spatially georeferenced image on top of a map of reference. An example of the former is the Library of the US congress[67] and of the latter is the David Rumsey Map Collection.[68] In order to make their maps discoverable, Websites depend on Web search engines' ability to index the map's description or they develop search tools relying in the maps' extension, publication year, title or key words. The metadata used to find maps is hidden behind the search application and users need to employ different applications across different Websites to find the information needed. On the contrary, this thesis proposes the use of LOD principles to describe historic maps. This will counteract data silos by removing the software layer that separates users from data.

The main datasource in a historic map is its contents, however traditional search

[66] owl:sameAs `http://www.w3.org/TR/owl-ref/#sameAs-def`

[67] The Library of Congress. American memory `http://www.loc.gov/rr/hispanic/frontiers/gutierrz.html`

[68] David Rumsey Map Collection. Map of India 1804 `http://rumsey.geogarage.com/maps/g2310061.html`

applications can only say if a map covers the same space and time requested by the user but they cannot guarantee the map actually contains what the user is looking for. However, listing all the map's contents is too much work. This problem can be attenuated by georeferencing the map area and linking the map to external contents. The map area is a property which can be spatially described using GeoSPARQL vocabulary. A polygon describing the map area can be used to test and discover spatial relationships with other spatially described contents available on the Web. Additionally, the map as a whole can be linked to external LOD contents, avoiding digitalizing spatial footprints or creating detailed descriptions for each of the map contents. Nevertheless this is still a lot of effort, an application could use the map's spatial and temporal reference systems to retrieve suggestions from triple stores reducing the georeferencing time substantially.

3.3 What metadata should be used for describing contents?

The metadata required for describing a historic map includes:

- Document (historic map) identification and standard (or regulatory) requirements. These metadata are gathered by librarians during cataloguing (see section 1.2.1). In a LOD-enabled library catalog, the map identification is an URI linked to the map's metadata and adding new data consists on creating new links to the map's URI.

- Metadata linking the original (paper-based) and the scanned historic map. This information is collected by librarians during historic map scanning.

- Map properties. Georeferencing historic maps in the library context only requires to georeference the map area as a map property. The map area can be described using the GeoSPARQL vocabulary and the Geographic Markup Language or Well Known Text encoding (see section 2.2).

- Map contents. This thesis' approach consists on retrieving content based on the map area without specifying the geometries for each contained feature. This avoids users to deal with geographic feature's vagueness[43]. However, contents' spatial footprints remain an open issue. For example, such spatial footprints could be provided by DBpedia although they are only served as points.[69]

[69] vague-places is a software project for extracting polygon geometries from DBpedia points `https://github.com/kxtells/vague-places`

The Dublin Core Metadata Initiative (DCMI) Metadata Terms vocabulary[70] is a set of standard terms used to describe web resources. This vocabulary can be used to encode the information related to historic maps as artifacts[23]. The Friend-of-a-Friend vocabulary (FOAF)[71] is meant to provide properties and classes for linking people and information using the Web. Combined, DCMI and FOAF can be used to describe a historic map and the people related to it, for example, the map's author. This allows to answer queries for documents including historic maps as well.

On the other hand, the information regarding historic maps' cartographic contents[23] can be described using the GeoSPARQL vocabulary and the *maps ontology*. The GeoSPARQL vocabulary correspond to some modules of the GeoSPARQL specification: The core module contains the top-level classes for spatial objects, the topology vocabulary module defines properties for encoding topological relations and the geometry module which establishes the data types for serializing geometry data[44].

Currently, the Muenster Semantic Interoperability Lab (MUSIL)[72] is developing the *maps ontology*[73] for specific purpose of describing historic maps. It is still an ongoing effort which takes place during the regular meetings of MUSIL. However, some of its properties can be used to link data to historic maps. Complex queries can be answered using this information, for example, *what are the historic maps of the 16th century depicting the surroundings of the Berg State in central Europe?*.

Table 2 presents an overview table of some classes, properties and literals that can be used to describe historic maps.

3.4 What are the specific challenges when georeferencing historic maps?

As mentioned in section 2.2, historic maps do not state their spatial reference systems. The issue of historic maps with unknown spatial reference systems is challenging as seen in [51] [4] [11]. Moreover, uncertainty about historic maps would prevail even if their spatial reference system were known because of the available technology of that time. However, under the assumption that topological relations between map features are kept in time, it is possible to overcome this problem by *Rubbersheeting*[74] the historic map to make it fit a modern reference map. This is equivalent to assume both maps have the same reference system at a cost of introducing inaccuracies or

[70] Dublin Core Metadata Initiative `http://dublincore.org/documents/dcmi-terms/`
[71] FOAF Vocabulary Specification `http://xmlns.com/foaf/spec/`
[72] Muenster Semantic Interoperability Lab `http://musil.uni-muenster.de/`
[73] Maps ontology `http://www.geographicknowledge.de/vocab/maps.rdf`
[74] Rubbersheeting `http://en.wikipedia.org/wiki/Rubbersheeting`

Source	**Class, Property or Literal**	**Description**	**Use in historic maps**
DCMI	Agent (C)	A resource that acts or has the power to act.	To represent a historic maps' author.
	description (P)	An account of the resource.	To link a historic map to a description given by a librarian.
FOAF	name (P)	A name for some thing.	To link the map's author to its name.
GeoSPARQL vocabulary	wktLiteral (L)	A Well-known Text serialization of a geometry object.	Datatype given to the map of reference's coordinates of the map area.
maps ontology	hasScale (P)	To link to a map scale.	
	mapsArea (P)	To link a to map's area coordinates.	
	mapsPlace (P)	To link to places.	
	mapsTime (P)	To link to a year.	
	title (P)	To link to a title.	

Table 2: Classes and properties used for describing historic maps.

even hiding feature changes over time because of map overfitting. However, bringing back Linda Hill's argument, the value of historical information on paper maps more than compensates for the residual error in their georeferenced versions[30]. In this thesis, rubbersheeting is achieved by estimating an affine transformation from the historic map to the map of reference (see section 2.2). Polynomial transformations were discarded because they can become unstable in some circumstances and because of the complexity associated to represent curve lines using the Well Known Text geometry encoding in GeoSPARQL.

As pointed in [18], the use of digital copies introduces problems related to the readability of the original analogical maps. One of them is losing track of the paper size in favour of the "real world" proportions represented by the map. This problem can be mitigated by taking advantage of the fact that historic maps are scanned along with rulers (see section 2.2): A measurement over the scanned ruler

can enable the estimation of the paper size of the historic map.

Historic maps are commonly decorated with figures not related to their cartographic content. Sometimes this conceals the map area changing its rectangular shape. This requires flexibility for representing the shape of the map area and GeoSPARQL can handle this non-regular polygons for both representations and querying.

Historic maps have a large amount of contents making georeferencing a tiresome work. An application can mitigate this by using the spatial and temporal georeferencing of the map to suggest potential contents from LOD data sources. Human intervention would be required to choose the proper suggestions among the contents retrieved form the LOD data cloud.

Regarding contents which no longer exist (e.g the Roman Empire), they can be linked from a historic map to a spatio-temporal gazetteer. If the contents are not registered in the gazetteer, new records must be first created. For example, DBpedia partially meets the requirements of a gazetteer (section 2.1) and it has the advantage of including point footprints for some of its resources. It includes a large amount of historical resources and new ones can be included by the creation of a new article in Wikipedia.

3.5 How can LOD-background knowledge be exploited for improving map search or georeferencing?

Historic maps help reconstructing past places and they are potential source of new data about the technological, artistic and social conditions of former times. Linking historic maps to places or events occurring when they were published improves their discoverability and usage, for instance , historic maps can be used to reconstruct the change of land use through time[32] or to understand the decision-making process at certain points in history.

In the specific case of libraries, LOD data sources can be used to improve:

- Historic map georeferencing. In the registration process (see section 2.2), , spatially-enabled LOD can be used to suggest additional matching features. The use of suggestions diminishes the time required to add control points during feature detection and matching steps. This potentially improves georeferencing because control points can be added faster and the larger the amount of matched features the better the georeferencing accuracy.

- Map search. In the library's processing of historic maps (see section 1.2.1),

maps are georeferenced after scanning. In this point, maps can be linked to its contents using external LOD-datasources. This is possible by using the map's georeferencing information to retrieve resources from a LOD-enabled spatio-temporal gazetteer (e.g DBpedia) and present them as suggested links. A librarian can include the new links in a historic maps description enhancing its discoverability and extending the global data graph of LOD at the same time.

As a result of the above mentioned uses of LOD in the library context, the scientific work is improved by reducing the time required to gather historic maps pertinent to a research.

3.6 Software development method

The general principle guiding the development of this tool was to tightly interlink georeferencing with semantic descriptions and external knowledge (DBpedia) in order to improve both. In this way, georeferencing becomes part of the semantic descriptions allowing enrichment of historic map descriptions. Discussions regarding this were held during the LIFE project's weekly meetings and the MUSIL meetings as well.

The software development methodology employed was to take good practices from agile software methodologies[75] and project management[45] and keeping in mind that an open source solution is preferable because libraries are subject to regulations (e.g. *RAK* the German rules for alphabetical cataloguing[76]).

The product of this thesis is the specification and implementation of a software tool which demonstrates how the research questions can be addressed (See section 1.5). The software allows to describe maps as well as their contents by georeferencing and simultaneously linking to contents of the Web of Data. The combination of semantic content descriptions and georeferencing improves the quality and efficiency of both processes.

The main stakeholders identified in the project were the ULB (as the tool client), the LIFE project (as the project's sponsor), the tool users (as themselves) and the developer (see section 1.1).

The project scope was obtained from the sponsor's needs that were detailed in meetings with the client. Additional features came from reviewing applications identified by the client and the sponsor which were said to resemble the needs of

[75] Agile software development `http://en.wikipedia.org/wiki/Agile_software_development`

[76] *Regeln für die alphabetische Katalogisierung* `http://files.d-nb.de/pdf/rak_wb_netz.pdf`

both. The project deadline was given by the schedule and rule book of the University of Münster regarding master thesis. The project cost is considered irrelevant since it is developed as a thesis.

Emphasis was given to mitigate top risks[3] during the project:

1. Misunderstanding of requirements.
2. Lack of management commitment and support.
3. Lack of adequate user involvement.
4. Failure to gain user commitment.
5. Failure to manage end user expectation.
6. Changes to requirements.
7. Lack of an effective project management methodology.

The measures taken to answer the risks were:

- Weekly meetings with the sponsor. The weekly LIFE project staff meeting were source of feedback and a communication channel useful for mitigating risks number 1 and 2. The meetings also help identify key technologies employed in LIFE project and by using them, the product's chances of incompatibility were reduced.
- Risk number 3 and 4 were accepted since access to users (librarians) was scarce due to time limitations.
- Risk number 5 was mitigated by hosting the product in Internet making it available for the stakeholders as well as scheduling kick-off, intermediate and final meetings with them.
- In accordance to agile methodologies, writing programming code was preferred over documentation. This mitigates risk number 6 since it avoids documenting rejected or changed features.
- Adopting good practices from agile software development and project management mitigated risk number 7.

The tool presented here has a place in ULB's historic map process (see section 1.2.1). This tool collects historic map's metadata required by library's LOD-enabled query tools. The map's metadata is fundamental to find historic maps through spatial, temporal or feature queries.

4 Implementation

This chapter describes an implementation of the suggested approach. The application tests the feasibility of the ideas presented in this document and it allows to describe maps as well as their contents by georeferencing and simultaneously linking to contents of the Web of Data.

The application presented here is meant to be used on the Web and it is completely developed using HTML and Javascript.[77] Javascript was chosen not only for its qualities (it is a standard, light-weighted, object oriented and it's part of most Web browsers) but also because other developments by LODUM initiative and LIFE project were built using the same language. This eases integration with other software and it also reduces its time to market.[78] Furthermore, the application's inception came from ULB's historic map process (see section 1.2.1) in order to be sure the application fits the library's process.

This tool is designed for librarians and its purpose is to georeference historic maps by means of Linked Open Data. The tool is openly available [79] and it uses the following Javascript APIs: JQuery[80] and JQueryUI[81] were used for the graphical user interface and event handling, DataTables [82] was used for additional table events and Sylvester[83] for matrix operations. In addition, the application uses the services provided by the DBpedia SPARQL endpoint[84] and DBpedia Spotlight[85]: The former to retrieve resources and the latter to find matches from DBpedia articles to user's descriptions of historic maps.

The application is able to suggest links from the map to DBpedia; this aims to accelerate the georeferencing process. Additionally, linking historic map's data to other LOD data sources grants the 5th star in the Tim Berners-Lee's rating system. The 5th star system's purpose it to encourage the publication of new LOD by assigning points (stars) to data. For achieving the top rating (the 5th star), the data must be available on the Web, accessible as machine readable data, published using non-proprietary formats and following open standards and finally, the data must be linked to other people's data.[86]

[77] Javascript is a light-weight scripting language available on most Web browsers.
[78] Time to market `http://en.wikipedia.org/wiki/Time_to_market`
[79] georef `https://github.com/albhasan/georef`
[80] JQuery `http://jquery.com/`
[81] JQuery user interface `http://jqueryui.com/`
[82] DataTables `https://datatables.net/`
[83] Sylvester. Vector and matrix math for Javascript `http://sylvester.jcoglan.com/`
[84] DBpedia SPARQL endpoint `http://dbpedia.org/sparql`
[85] DBpedia Spotlight `https://github.com/dbpedia-spotlight/dbpedia-spotlight`
[86] Linked Data `http://www.w3.org/DesignIssues/LinkedData.html`

Regarding the user interface, the application layout consist of 3 main vertical panels: The left panel contains tabs depicting the tool's workflow, the central panel is a container for the historic map and the right panel contains the map of reference and tabs for metadata.

4.1 Research question evaluation

"What metadata should be used for describing contents?". The metadata tags required by libraries to catalogue historic maps should encode the map's contents and additionally the mandatory metadata of local standards or regulations. An answer to this question is the usage of well-known ontologies and vocabularies like Dublin Core, Friend of a Friend, GeoSPARQL and for the remaining metadata, properties and classes were taken from the *maps ontology*. However, the LOD representation of mandatory metadata is still an open issue.

"What are the specific challenges when georeferencing historic maps?". Georeferencing historic maps of unknown spatial reference system while enabling content search is the biggest challenge addressed in this thesis. The answers to these challenges are rubbersheeting historic maps to match a reference map, describe the historic map's area and linking it to DBpedia's spatio-temporal resources. This is acceptable in the library context but specific knowledge fields would require detailed description using additional vocabularies or ontologies.

How can LOD-background knowledge be exploited for improving map search or georeferencing?. LOD can be used to suggest features, places, or events which can be linked to historic maps. Since linking is faster than typing descriptions, it can be said LOD-background knowledge speeds the georeferencing process. In a posterior stage, the links from historic maps to external resources increase their chances of being discovered because they enable library users to reach them by following the links.

Answering the question *"How to describe and publish historic maps in order to make them queryable by content?* requires the concurrence of the answers to the other questions. The response to the challenges, the metadata used and the confluence of LOD-background knowledge makes historic maps queryable by their contents which is an advantage in comparison with the traditional search based on key words.

4.2 Application usage

A typical user interaction with the tool includes 3 stages: Georeferencing the historic map image, annotating the map contents and obtaining the map description. The first stage is georeferencing where the spatial transformation parameters are obtained. The second stage consists of creating links from the map to DBpedia resources. In the third stage, the map description is stored or given to the user. These links can be of two kinds: Links from a map to modern places and links from the map to historic places (see example in section 3.2).

A typical user interaction would go over the following steps:

1. An user opens an Internet browser and then he types the application URL. [87]

   Georeferencing the map image

2. The default tab in the left panel is called `Image`. The user types the URL of a scanned historic map and then he presses the `Load image` button. See figure 1.

3. The user starts georeferencing the map by choosing the `control point tool` and pinpoints a feature in both the historic and reference map (figure 2). When the user has pinpointed and matched certain number of POIs, the map of reference displays the borders of the historic map (figure 3). The user can go to `Control Points` tab where a table displays the coordinates of the pinpointed features; by selecting a row, the icons in the maps change accordingly. The user can also press the button `Suggest POIs` which retrieves POIs from DBpedia, when the user click on a suggestion both maps pan to the suggestion's location.

4. The user selects the tool `Draw a map area on image` located in the historic map panel. With this tool the users can draw the map area polygon in the historic map. Then the map area is displayed in the map of reference. figure 4.

5. Now the user can draw a 1 cm line in the historic map using the tool `Draw an 1 cm line`. These data is used to estimate the historic map's size and scale. An example drawn line is showed in figure 5.

6. The user can now type metadata in the tabs under the map of reference. The second tab is for general information (`Metadata`). Here the user can type descriptive information such as the identifier (URI) of the paper map. See figure 6.

[87] For example `http://giv-siidemo.uni-muenster.de:81/code/georef.html`

Annotating the map contents

7. The next tab under the map of reference is `Places`. This is for linking historic map contents to modern places. Here the user can type a list comma-separated names of modern places and when the `Find matches` button is pressed, the application retrieves matches from DBpedia. The user can check the ones matching his entries. See figure 7.

8. The `Links` tab retrieves time-enabled places form DBpedia using the map's area and publication date. The user must press the button `Suggested tags` and then he checks the places matching the map contents. See figure 8.

9. The `Description` tab allows the user to type a description for the map. By pressing the `Find matches` button the application retrieves potential matches to DBpedia pages. The users can check those which apply to the map.

Obtaining the map description

10. Now the user can go back to the left panel where he will find the `Results` with the `KML` button. This button opens a new browser window with KML code (figure 10). The user can save the KML code in a text file which can be used later for displaying the historic map in a KML-enabled software as shown in figure 11. This is just an approximation of the georeferencing results since this application does not resample the historic map in order to get a transformed image.

11. Likewise, the `Get RDF` button in the `Results` tab opens a new browser window with a SPARQL query for inserting the collected metadata in a triple store (See figure 12).

Figure 1: A historic map loaded in the application.

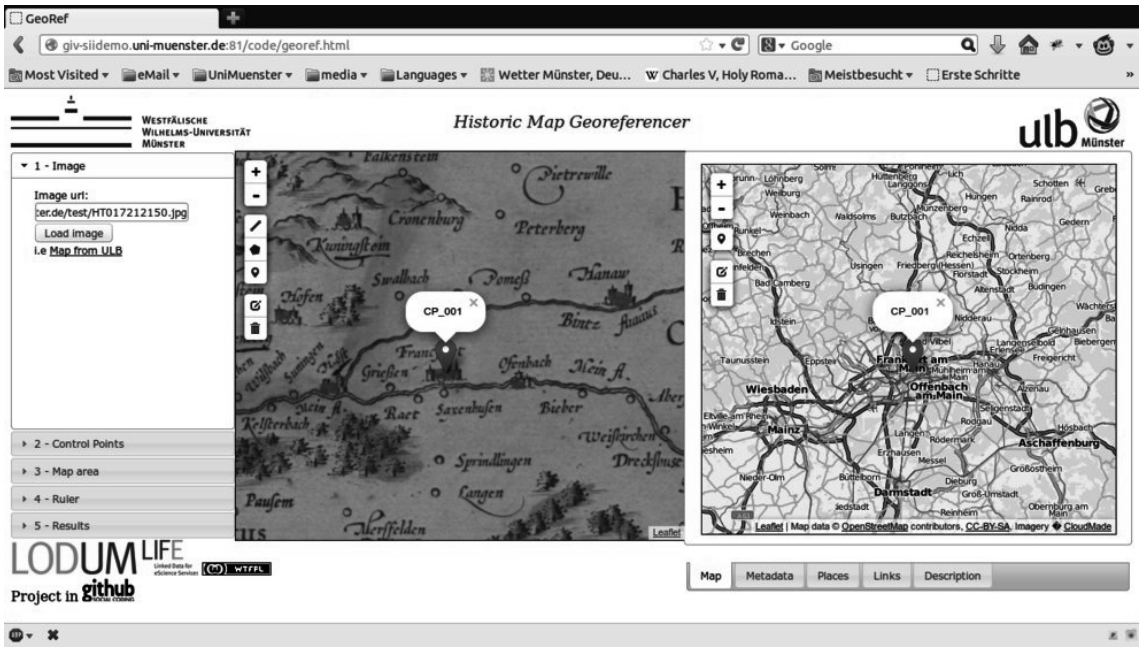

Figure 2: A pinpointed Point of Interest.

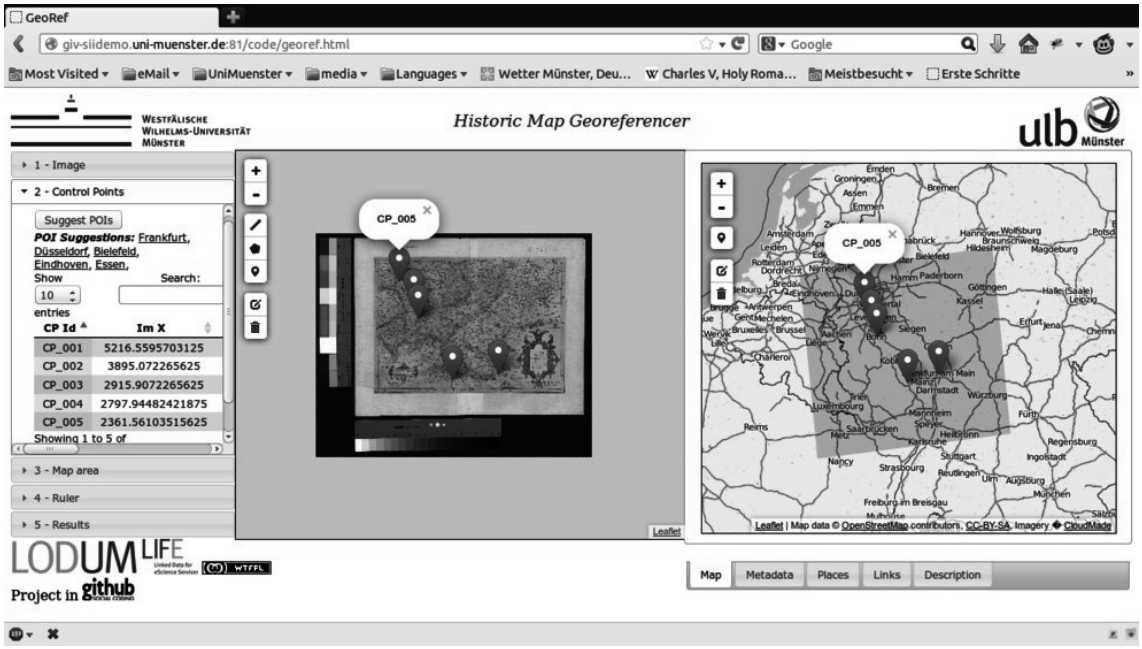

Figure 3: Image of historic map depicted over the map of reference.

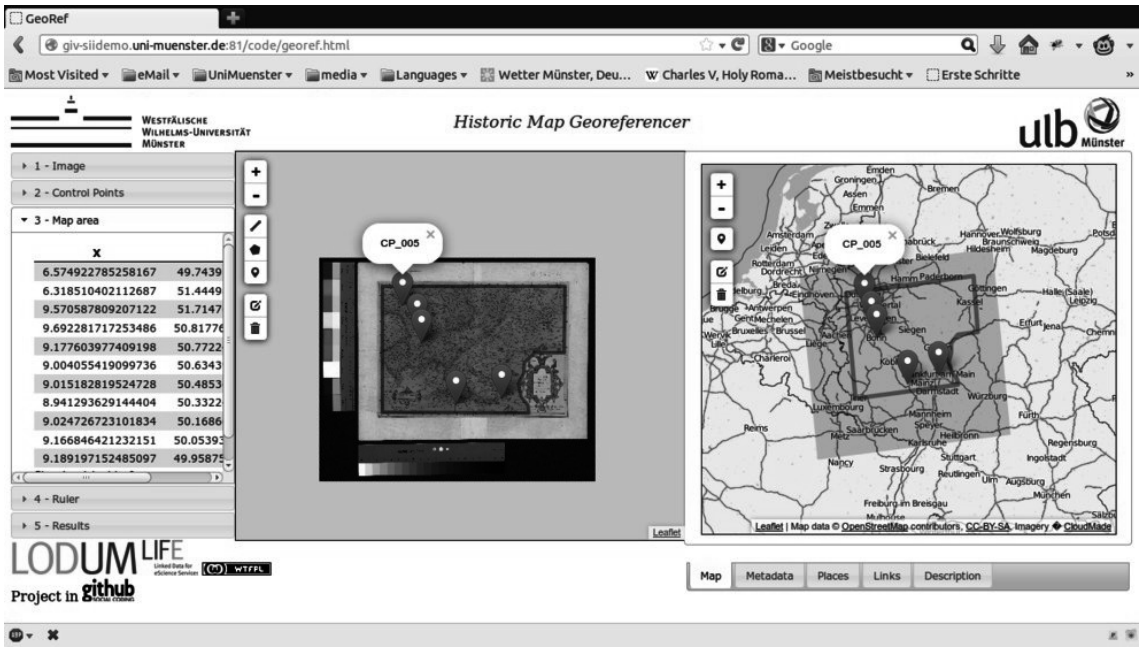

Figure 4: Map area depicted over the map of reference.

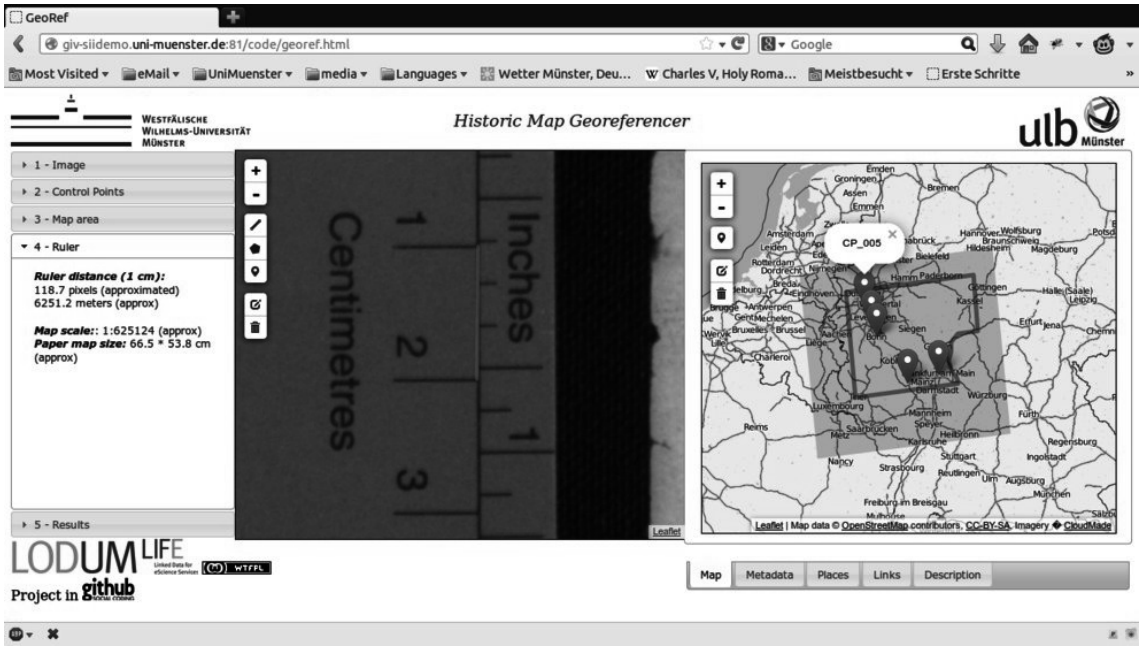

Figure 5: Map area depicted over the map of reference.

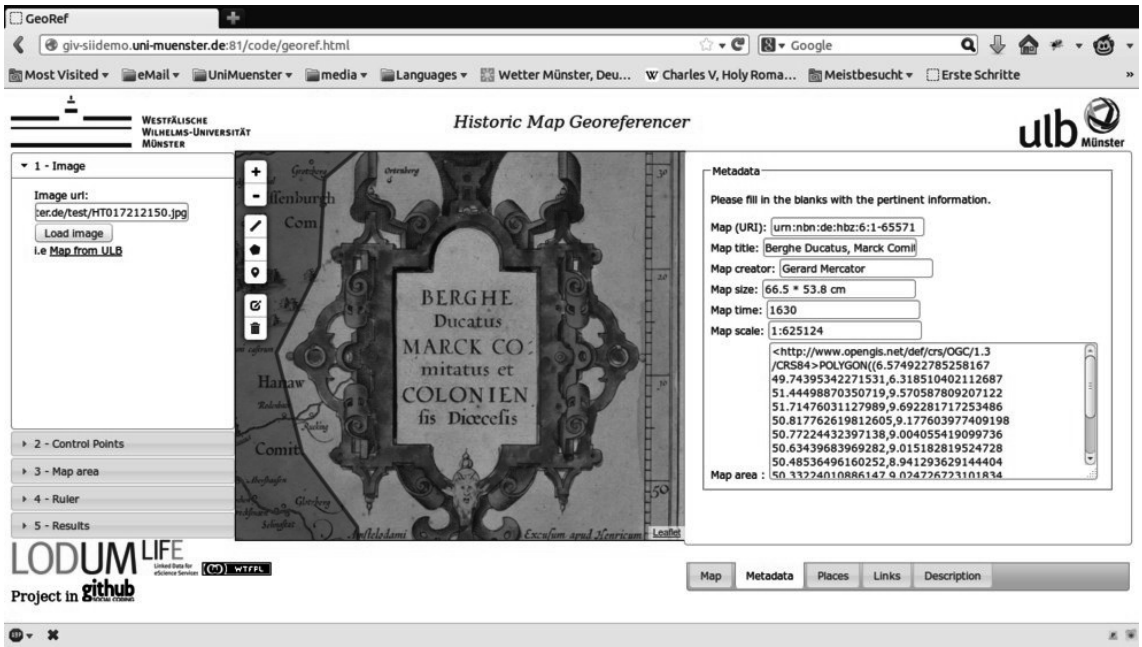

Figure 6: Historic map information in the Metadata tab.

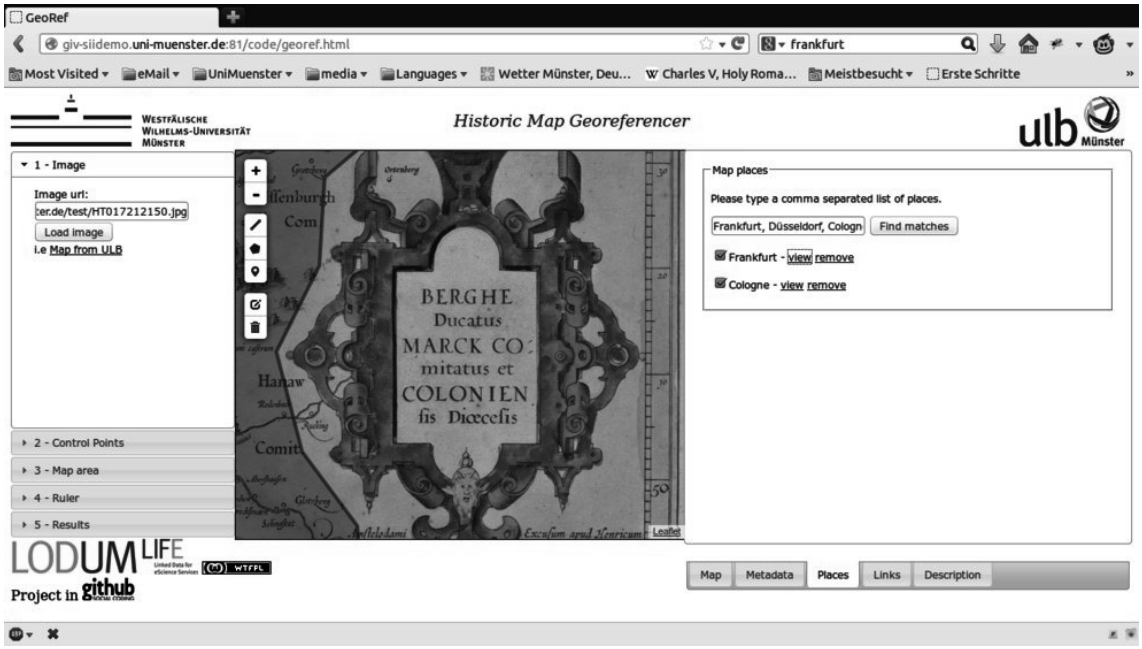

Figure 7: Suggested links derived from user's place list.

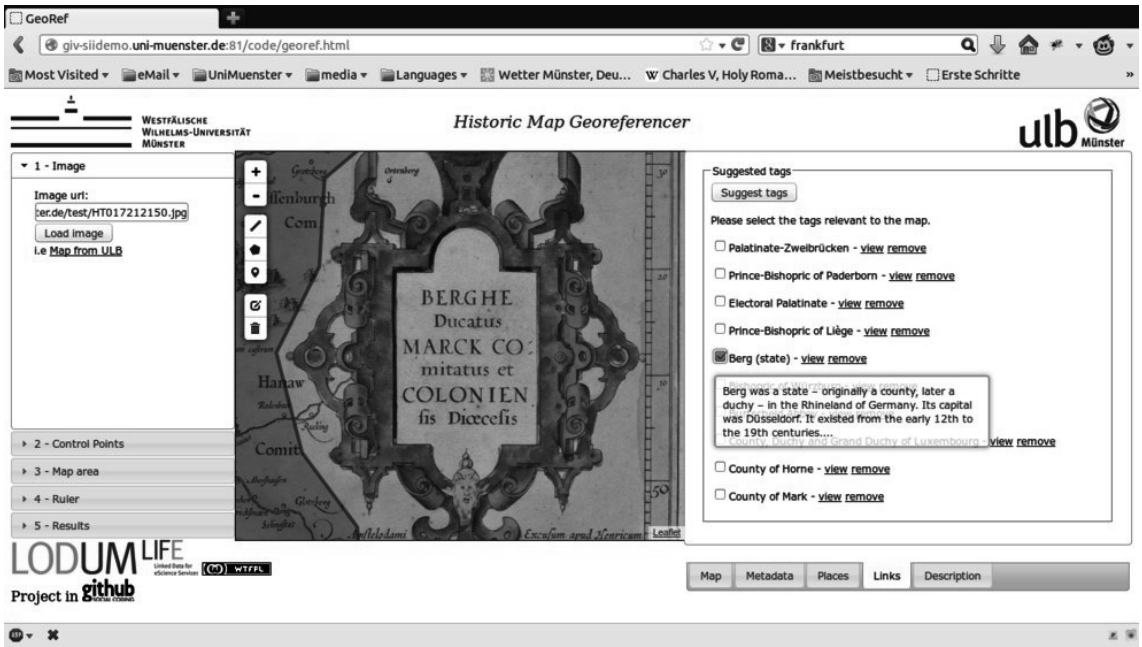

Figure 8: Suggested links derived from map's spatio-temporal georeferencing.

Figure 9: Suggested links derived from map's spatio-temporal georeferencing.

Figure 10: Browser window with map's spatial georeferece (KML).

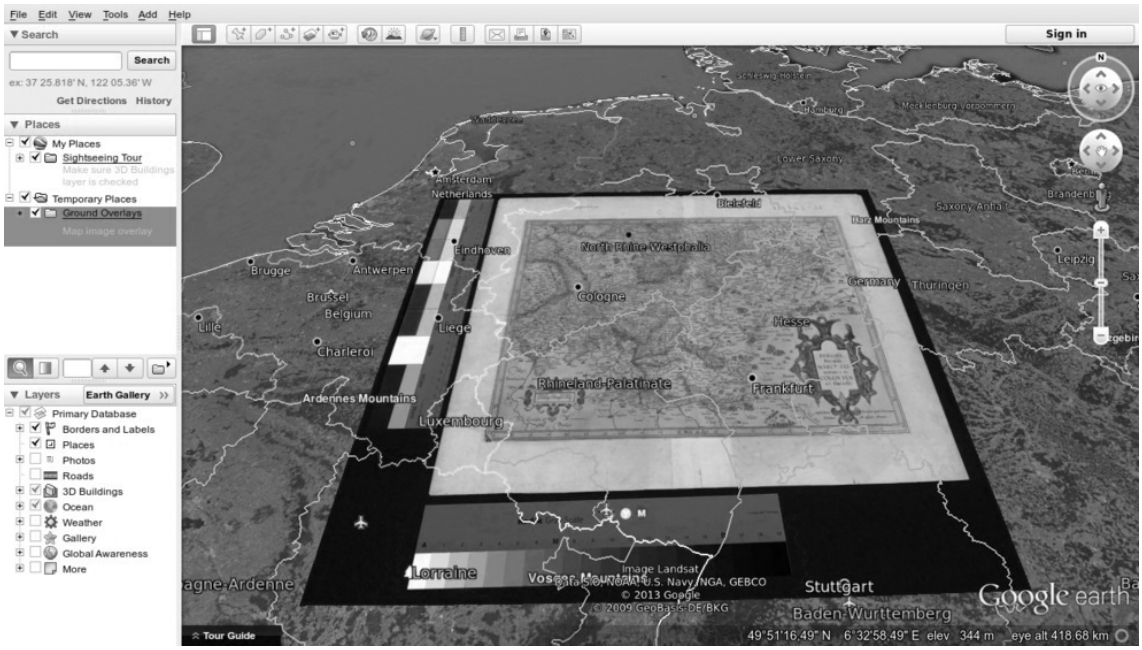

Figure 11: Georeferenced historic map in a KML-enabled viewer.

Figure 12: Description of the map in RDF..

5 Discussion and further developments

This thesis explores how Linked Open Data can be used for georefencing historic maps in the spatial, temporal and content dimensions in the context of libraries. As an illustration, a software tool is introduced;[88] this tool allows the description of historic maps and their contents while at the same time it uses and adds new Linked Open Data. The tool employs the World Geodetic System 1984 , the Julian calendar and the DBpedia ontology as spatial, temporal and semantic reference systems respectively. However, this thesis' tool does not confront issues such as drawing the spatial footprints of map features, performance issues related to keeping large images in web browsers' memory, georeferencing historic maps belonging to a index map[89] or overview maps.

The importance of georefencing historic maps in the aforementioned dimensions lies on the increment in discoverability; in this fashion, library users obtain result sets including historic maps when querying library catalogues. This potentially improves scientific research by increasing the amount of relevant information available, and by narrowing searches to pertinent resources.

The contributions of this thesis are two: First, it enhances historic map discoverability through a software tool and second, it helps leaving behind keyword-based search in favour of content-based search by improving libraries' historic map georeferencing process. It is important to realize that descriptions of map contents allow to answer more complex spatio-temporal questions regarding maps than the traditional key words. In particular, historic map processing is improved when using linked open data by suggesting features to act as control points and suggesting links from the historic map to places or events contained in it.

Regarding the future work, the next step in the ULB's processing of historic maps consist on developing a searching tool. That tool will enable library's users to include historic maps as results in their searches. This could be achieved in two ways: First, by enabling users to type the year and draw an interest area on a map and second,by enhancing searches by matching terms in a spatio-temporal gazetteer for obtaining footprints and using them to test spatio-temporal relationships with historic maps.

[88] georef `https://github.com/albhasan/georef`

[89] Index map `http://en.wikipedia.org/wiki/Index_map`

References

[1] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November 1983.

[2] D. Andresen, L. Carver, R. Dolin, C. Fischer, J. Frew, M. Goodchild, O. Ibarra, R. Kothuri, M.Larsgaard, B.Manjunath, D.Nebert, J.Simpson, T.Smith, T.Yang, and Q.Zheng. The www prototype of the alexandria digital library. In *In Proceedings of ISDL'95: International Symposium on Digital Libraries*, pages 17–27, 1995.

[3] Tharwon Arnuphaptrairong. Top ten lists of software project risks: Evidence from the literature survey. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume Volume 1, pages 732–737, March 2011.

[4] Caterina Balletti. Georeference in the analysis of the geometric content of early maps. *e-Perimetron*, Volume 1(No 1):32–42, 2006.

[5] Robert Battle and Dave Kolas. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web*, 3(4):355–370, 2012.

[6] Tim Berners-Lee. Using xml for data. http://www.w3.org/DesignIssues/XML-Semantics.html.

[7] Tim Berners-Lee. Why rdf model is different from the xml model. http://www.w3.org/DesignIssues/RDF-XML.html, 1998.

[8] Mohamed Bishr and Werner Kuhn. Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In SaraIrina Fabrikant and Monica Wachowicz, editors, *The European Information Society*, Lecture Notes in Geoinformation and Cartography, pages 365–387. Springer Berlin Heidelberg, 2007.

[9] C. Bizer. The emerging web of linked data. *Intelligent Systems, IEEE*, Volume 24(Issue 5):87–92, 2009.

[10] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, Mar 2009.

[11] Chryssoula Boutoura. Assigning map projections to portolan maps. *e-Perimetron*, Volume 1(No 1):40–50, 2006.

[12] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24(4):325–376, December 1992.

[13] James R. Carter. The many dimensions of map use. In *Proceedings, International Cartographic Conference*, 2005.

[14] N.R. Chrisman. *Exploring geographic information systems*. Wiley, 2nd edition edition, 2002.

[15] Noel Cressie and Christopher K Wikle. *Statistics for Spatio-Temporal Data*. Wiley, 2011.

[16] M. J. Egenhofer. A formal definition of binary topological relationships. In *3rd International Conference, FODO 1989 on Foundations of Data Organization and Algorithms*, pages 457–472, New York, NY, USA, 1989. Springer-Verlag New York, Inc.

[17] M. J. Egenhofer and J. Herring. Categorizing binary topological relations between regions, lines, and points in geographic databases. *The*, 9:1–28, 1994.

[18] Piero Falchetta. Perception, cognition and technology in the reading of digital cartography. *e-Perimetron*, Volume 1(No 1):77–80, 2006.

[19] Ulrich Freitag. *The Selected Main Theoretical Issues Facing Cartography: Report of the ICA-Working Group to Define the Main Theoretical Issues on Cartography for the 16th ICA Conference, Cologne 1993*, chapter Chapter 1: Map Functions. University of Toronto Press, 1993.

[20] James Frew, Michael Freeston, Nathan Freitas, Linda L. Hill, Greg Janee, Kevin Lovette, Robert Nideffer, Terence R. Smith, and Qi Zheng. The alexandria digital library architecture. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98, pages 61–73, London, UK, UK, 1998. Springer-Verlag.

[21] Michael Friendly. Re-visions of minard. *Statistical Computing and Statistical Graphics Newsletter*, 11:13–19, 2000.

[22] Eleni Gkadolou and Emmanuel Stefanakis. A formal ontology for historical maps. In *26th International Cartographic Conference*, August 2013.

[23] Eleni Gkadolou, Eleni Tomai, Emmanuel Stefanakis, and Georgios Kritikos. Ontological standardization for historical map collections: Studying the greek

borderlines of 1881. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences - XXII ISPRS Congress*, volume Volume I-2, pages 203–208. International Society for Photogrammetry and Remote Sensing, August 2012.

[24] Michael Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, August 2007.

[25] Michael Frank Goodchild. Keynote address: spatial information science. In *Fourth International Symposium on Spatial Data Handling*, July 1990.

[26] John Brian Harley. Deconstructing the map. *Cartographica: The International Journal for Geographic Information and Geovisualization*, Volume 26(2):1–20, 10 1989.

[27] Tom Heath and Christian Bizer. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, February 2011.

[28] Charles van den Heuvel. Modeling historical evidence in digital maps: a preliminary sketch. *e-Perimetron*, Volume 1(No 2):113–126, 2006.

[29] Linda L. Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. In *In J. Borbinha and T. Baker (Eds.), Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000*, pages 280–290. Springer, 2000.

[30] Linda L. Hill. *Georeferencing: The Geographic Associations of Information*. MIT Press, 2009.

[31] Jonathan Iliffe and Roger Lott. *Datums and Map Projections: For Remote Sensing, Gis, and Surveying*. Whittles Publishing, 2nd edition edition, 2008.

[32] Louis R. Iverson. Land-use changes in illinois, usa: The influence of landscape attributes on current and historic land use. *Landscape Ecology*, 2(1):45–61, 1988.

[33] Tomi Kauppinen, Alkyoni Baglatzi, and Carsten Keßler. Linked Science: Interconnecting Scientific Assets. In Terence Critchlow and Kerstin Kleese-Van Dam, editors, *Data Intensive Science*. CRC Press, USA, 2013.

[34] Carsten Kessler, Krzysztof Janowicz, and Mohamed Bishr. An agenda for the next generation gazetteer: geographic information contribution and retrieval. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 91–100, New York, NY, USA, 2009. ACM.

[35] Carsten Kessler and Tomi Kauppinen. Linked open data university of muenster-infrastructure and applications. In *Demos of the Extended Semantic Web Conference 2012*, May 2012.

[36] Tom Koch. The map as intent: Variations on the theme of john snow. *Cartographica: The International Journal for Geographic Information and Geovisualization*, Volume 39(4):1–14, January 2004.

[37] Werner Kuhn and Martin Raubal. *AGILE 2003: 6th AGILE Conference on Geographic Information Science*, chapter Implementing semantic reference systems, pages 63–72. Collection des sciences appliquées de l'INSA de Lyon. Presses polytechniques et universitaires romandes, 2003.

[38] Werner Kuhn and Martin Raubal. Semantic reference systems. *International Journal of Geographical Information Science*, 17:405–409, 2003.

[39] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2013. Under review.

[40] Paul A. Longley. *Geographic Information Systems and Science*, chapter Chapter 5: Georeferencing. Wiley, 2 edition edition, 2007.

[41] Kari S. McLeod. Our sense of Snow: the myth of John Snow in medical geography. *Social Science and Medicine*, 50:923–935, 2000.

[42] Pablo N. Mendes, Max Jakob, AndrÃ©s GarcÃa-silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics*, 2011.

[43] Daniel R. Montello, Michael F. Goodchild, Jonathon Gottsegen, and Peter Fohl. Whereś downtown? behavioral methods for determining referents of vague spatial queries. *Spatial Cognition and Computation*, 2003.

[44] OGC GeoSPARQL - A Geographic Query Language for RDF data, 2012.

[45] Project Management Institute PMI. *A guide to the project management body of knowledge (PMBOK guide)*. Project Management Institute, Newtown Square, PA, 2008.

[46] David Rumsey and Meredith Williams. *Past Time, Past Place: GIS for History*, chapter Chapter 1: Historical Maps in GIS. ESRI Press, 1st edition edition, 2002.

[47] Rainer Simon, Bernhard Haslhofer, Werner Robitza, and Elaheh Momeni. Semantically augmented annotations in digitized map collections. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 199–202, New York, NY, USA, 2011. ACM.

[48] Rainer Simon, Christian Sadilek, Joachim Korb, Matthias Baldauf, and Bernhard Haslhofer. Tag clouds and old maps: Annotations as linked spatiotemporal data in the cultural heritage domain. In *Workshop On Linked Spatiotemporal Data 2010, held in conjunction with the 6th International Conference on Geographic Information Science (GIScience 2010)*, Zurich, Switzerland, September 2010.

[49] Terence R. Smith. A digital library for geographically referenced materials. *Computer*, 29(5):54–60, May 1996.

[50] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. LinkedGeoData: A core for a web of spatial open data. *Semantic Web*, 2012.

[51] W. R. Tobler. Medieval distortions: The projections of ancient maps. *Annals of the Association of American Geographers*, 56(2):351–360, 1966.

[52] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.

I hereby assert that this master thesis with the title "*A LOD-based georeferencing tool for historic maps*" is written by myself and that I did not use any other than the declared resources. All parts which are literally and logically taken from external sources within this work are marked as being external.

Münster, NRW, Germany 27-09-2013

Alber Sánchez