BauDenkMalNetz – Creating a Semantically Annotated Web Resource for Historical Buildings

Anca Dumitrache

Computer Science
Class of 2011
Jacobs University Bremen

Guided Research Proposal
Date: January 15, 2011
Supervisor: Prof. Dr. Michael Kohlhase

Abstract

The purpose of this project is creating a semantically annotated web resource of urban historical landmarks. The annotations will highlight the most relevant information about the landmarks (e.g. the buildings' architect, architectural style or construction details), for the purpose of extended accessibility and smart querying. The project will be implemented starting from a series of touristic books on architectural landscape. These books will be analyzed using text mining techniques and automated tools, which will lead to coming up with an ontology, together with a lexicon, that will represent the necessary vocabulary to express all the relevant architectural and historical information. After gathering all the necessary metadata, a web site will be built around it, providing advanced browsing functionality based on the semantic content, and such features as special queries for generating custom touristic leaflets.

Contents

1 Introduction

1.1 Motivation

The architectural landscape of a city is generally made up not just of well-established landmarks, but historical buildings with a rich cultural background that lie outside the mainstream touristic circuit. People wanting to explore the more personal and less-known places of a city have little access to information about these hidden architectural gems and the stories behind them, even though all required data on historical buildings in Germany has been meticulously collected by the offices for historical monuments. However, this data is not easily accessible, and often tedious to browse through.

In Bremen, an effort to collect all of this information and present it to the general public way was made by the publisher Nils Aschenbeck, who released a series of guide books [7] about the city. However, for the moment, these books are only accessible in printed format. The BauDenkMalNetz (German for "listed buildings web") project proposes a way of discovering Bremen's architectural landscape that is suited for the tech-savvy tourist.

1.2 Structure of the Proposal

In the beginning, the details of the preliminary work done for this project will be described and an initial prototype of BauDenkMalNetz will be presented. Afterwards, the proposed work plan for the project will be described.

Beginning from some light semantic annotations on the series of books that was made available to us, we intend to define the necessary metadata for building a web portal where these texts could be easily accessible to the general public. Relevant information like the architect, architectural style and time during which it was built will be clearly delineated for each building, thus enabling browsing based on these keywords. For extracting the metadata, we plan to process the text using automated tools and natural language processing techniques. The first step in setting up the web resource will be creating a model of the relevant data and the relations it comprises (i.e. an ontology), together with a lexicon that will contain the keywords from the preprocessed texts.

The web resource will then be built using the Drupal framework, as it facilitates the incorporation of RDF semantic annotations. A possible enhancement to be studied will be an extended querying functionality, that will enable people visiting Bremen to create their own custom tourist guides (i.e. a guide of all buildings belonging to the Bauhaus period).

2 Planned Investigation

2.1 Preliminary Work

We have previously explored different ways of approaching the BauDenkMalNetz project. A prototype version exists [9], built using **Semantic MediaWiki** (SMW [3]).

Semantic MediaWiki was built as an extension of MediaWiki, the well known wiki application which powers Wikipedia. SMW provides enhanced features for browsing and organizing its contents as a result of the added semantic annotations to its text.

Our motivation for using SMW in deploying the web portal was its suitability for rapidly creating a working prototype of our resource. SMW allows for easily adding and editing of the necessary data and metadata available on historical buildings. New information could be easily incorporated and linked to the already existing data by making use of SMW page creation and editing tools. At the same time, the structure of the metadata (i.e. the ontology representing the data) could be easily modified, simply by adding in-text annotations. However, after a careful analysis, we have encountered some drawbacks that led us to reconsider our approach.

The text that we need to add semantic annotations to is already known, and not liable to substantial changes in the future. The same goes for the structure of the metadata (e.g. details on various architectural styles, as well as the buildings that adhere to it, are more or less set). Because of this, we have concluded that there is no real benefit coming from the feature of flexibly changing the contents, as provided by SMW.

Because the ontology to which the texts adhere is never explicitly specified, but rather implied from the annotations done directly on the text, the conceptual model of our metadata less obvious and never explicitly formalized.

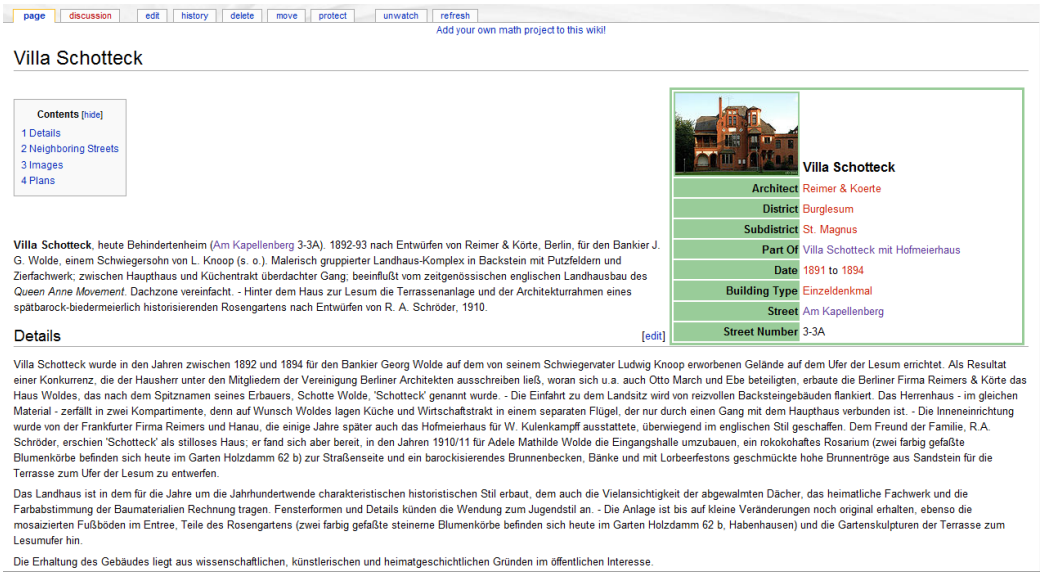

Figure 1: Screenshot of the SMW prototype

In this case, alignment to other similar ontologies (in keeping with the linked-data philosophy of reuse) is still possible, yet it is rendered more difficult by the lack of an explicit formal definition of the ontology.

The issue of having to annotate large blocks of text was not really addressed by our prototype. SMW provides some tools suited for database import, however the texts we want to analyze are stored in simple HTML files. There is a file for each individual building, with picture attachments, and information like the name of the architect being highlighted. The volume of data that needs to be processed (four books have been published thus far, with more than one hundred buildings being described) makes it almost impossible to have the texts annotated manually, like we did for building the prototype, while also making the project rather suited for the employment of natural language processing techniques in order to get the needed semantical annotations. Therefore, a much more intuitive way of approaching the project is to preprocess the texts and determine the relevant metadata before actually building the web portal.

2.2 Processing the Text

From the original touristic guides on which the web resource will be based, we want to identify the vocabulary that relates to historical building. This

vocabulary will then be used to build an **n-gram model** [14] of our guides, but also to refine our ontology, as detailed in the next section.

The guidebooks were made available to us in an HTML format. Therefore, the first step in order to be able to process the text would be stripping it down to a plain-text format. We will make use of the **LaMaPUn** [11] Perl library for processing the text. Now the text is ready to run through an automated tool that will run a set of n-gram collections over it.

An n-gram model refers to a probabilistic model that, given the first $n-1$ words in a sentence, will predict the n^{th} word. By creating such a model, we will get a formalization of the way the text is structured (e.g. which words refer to architectural styles), which will then enable us to understand what concepts will be mapped to resources and properties in our RDF model of the metadata.

2.3 Building an Ontology

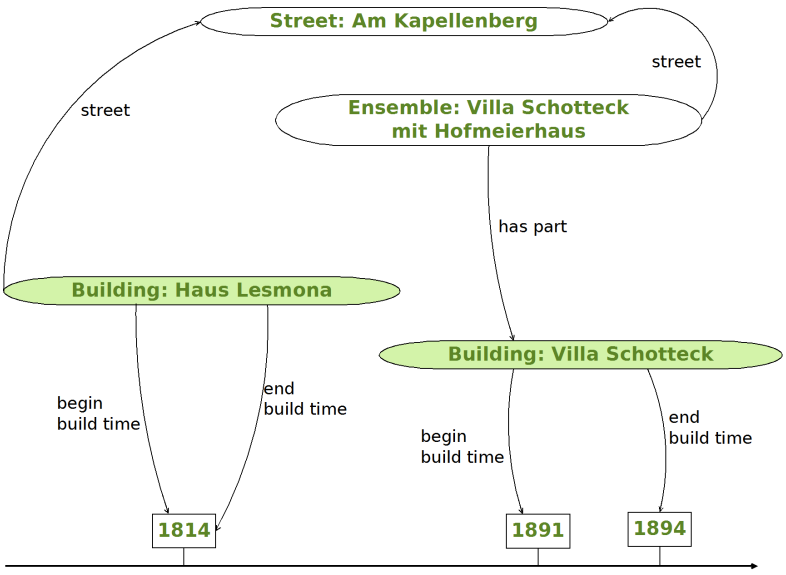

Figure 2: A tentative fragment of the BauDenkMalNetz ontology

At the heart of the web resource we wish to build lies the BauDenkMalNetz ontology, a formal representation of the metadata on historical buildings, together with the relations formed inside this data. The ontology will be specified in the Web Ontology Language (OWL), based on the RDF model for serialization, for easier integration in the Drupal framework.

The linked-data community [12] advocates the reuse of knowledge models and vocabularies, in order for achieving interoperability across the web. Indeed, there already exist various ontologies that model some of the relevant data about historical buildings, out of which the following were found relevant for aligning with the BauDenkMalNetz ontology:

- The **GeoNames** [2] ontology models geospatial semantic information. In particular, it maps unique locations on the globe, determined by their geoname, to a unique URI with a corresponding RDF web service. For our project, it can be used to uniquely identify each historical building based on its coordinates. Reusing this ontology brings the added advantage of explicitly specifying the geolocation of a building, which allows for easier integration with web mapping services.
- The **CIDOC CRM** [1] ontology represents the detailed scientific documentation of cultural heritage objects, which include historical monuments. By aligning our ontology to CIDOC CRM, we can formulate a full description of the historical information related to a building (e.g. the architectural style of the monument, the official sources which document the monument etc.).

However, a complete and functional ontology for our web resource cannot be created simply by reusing content from other ontologies. GeoNames only refers to the physical location of a building, and CIDOC CRM refers to museum curated knowledge on a general level, encompassing everything from small artifacts to monuments, whereas our ontology only deals with buildings described in tour guides. Therefore just those resources and properties related to buildings need to be reused, while the documentation of the landmarks will be less strict. Based on the vocabulary obtained by processing the text in the previous step, the BauDenkMalNetz ontology will be manually tweaked in order to encompass all this information.

2.3.1 Related Application: MANTIC

MANTIC [15] is a project similar to BauDenkMalNetz, that deals with cultural heritage sites of the city of Milan. It has the CIDOC CRM ontology at its core, which it uses in order to store information about the archeology of the city. This information is then incorporated into the Google Maps API, making for an easy to use application for browsing Milan's historical landmarks,

that is quite similar in scope to our project. However, unlike BauDenkMalNetz, MANTIC deals with public archeological sites, which only refer to part of the buildings that we are interested in. For the buildings that are listed as touristic landmarks, MANTIC provides a good example of how CIDOC CRM can be reused for our project, however, the buildings that are privately owned will need to be represented differently.

2.4 Ideas for Deployment

Once we have an ontology and a comprehensive vocabulary, we can start building our web resource. The website will be deployed making use of the **Drupal** [4] framework, a content management system that provides an RDF API [8], which will enable us to incorporate the metadata represented by our OWL ontology, and make semantic queries on it.

The text of the tourist guides will be then uploaded to the web resource, with individual entries for each building, and cross-links to pages about types of locations (e.g. streets, neighborhoods etc.), but also architectural styles and architects. A search feature based on these criteria will be made available. Also, for increased functionality, navigation based on the Google Maps API [5] will be made available.

For even more advanced querying features, we can make use of the **XSPARQL** [6] query language. XSPARQL combines XQuery, an XML-based query language, with SPARQL, the query language for the semantic web, which allows for getting XML results for queries over the semantic metadata of our web resource. By selecting from a list of available queries, tourists will be able to create personalized guides of historical buildings.

2.5 Evaluation

For the purpose of evaluating our project, we consider BauDenkMalNetz in the context of semantic digital libraries. The concept of semantic digital libraries refers to classic digital resources for storing knowledge which have been enriched with semantic metadata. The BauDenkMalNetz web resource fits this description, and therefore is suitable for evaluation according to existing standards for digital libraries [10] and semantic digital libraries [13]. Out of the three concepts discussed by Fuhr: performance, **usability** and **usefulness**, we will focus on the last two for our evaluation. A group of

test-users will navigate through the resource, providing feedback based on the points to be discussed below.

Usability refers to the ease with which users navigate the content. The test users will provide feedback on how easy/difficult it is to find a particular building, by querying the system based on a criteria of their own choosing (e.g. location, architectural style etc.), and also about how they managed to find their way from one particular building to another, based on a common characteristic.

Usefulness refers to the quality of the content. The users will be asked to provide their input on how accurate the query results are in relation to what they were expecting to find, and also about the informative character of individual buildings' pages.

Based on this evaluation, an assessment about the user-friendliness of the digital library will be made, and the possible improvements to the web resource will be considered.

3 Proposed Time Line

This section will discuss the time line and deadlines for the various steps in completing the project:

- *January* – An initial processing of the text will be done, and a tentative version of the BauDenkMalNetz ontology will be sketched.
- *February* – After analyzing the initial results, the ontology and its affiliated vocabulary will be refined. Towards the end of the month, the deployment phase will begin.
- *March* – The building of the web resource interface with Drupal will continue. After the structure of the website is done, and most of the data is in place, we can start focusing on the querying functionality.
- *April* – The beginning of the month will be reserved for the web resource to be evaluate by test users. Their input will be collected, and minor tweaks that can improve functionality will be implemented before the project deadline. Then, once the evaluation is done, the thesis write-up can commence.

4 Conclusion

Once finished, the BauDenkMalNetz project will provide a comprehensive and easy-to-use guide to the city of Bremen, and possibly even help boost the touristic appeal of Bremen. A possible enhancement to the resource will be creating a mobile version of the website, so that tourists can create virtual itineraries that they can access on the go.

However, the scope of the project is not limited to Bremen. Both the ontology and the vocabulary are general enough to adapt to any resource of historical landmarks. Therefore, the project could be further extended to deal with more cities, provided that relevant touristic guides about them are made available.

References

[1] The CIDOC Conceptual Reference Model, 03 2010. URL `http://cidoc.ics.forth.gr`.

[2] GeoNames, 04 2010. URL `http://www.geonames.org`.

[3] Semantic MediaWiki, 03 2010. URL `http://semantic-mediawiki.org`.

[4] Drupal – Open Source CMS, 01 2011. URL `http://drupal.org`.

[5] Google Maps, 01 2011. URL `http://maps.google.com`.

[6] Waseem Akhtar, Jacek Kopecký, Thomas Krennwallner, and Axel Polleres. XSPARQL: Traveling between the XML and RDF worlds – and avoiding the XSLT pilgrimage. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications*, number 5021 in Lecture Notes in Computer Science. Springer Verlag, 2008.

[7] Nils Aschenbeck and Ilse Windhoff. *Landhäuser und Villen in Bremen*. Aschenbeck Verlag, Bremen, 2009.

[8] Stéphane Corlosquet, Renaud Delbru, Tim Clark, Axel Polleres, and Stefan Decker. Produce and Consume Linked Data with Drupal! In Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum,

Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web – ISWC 2009*, number 5823 in LNCS. Springer Verlag, October 2009.

[9] Anca Dumitrache, Christoph Lange, Michael Kohlhase, and Nils Aschenbeck. Prototyping a browser for a listed buildings database with Semantic MediaWiki. In Christoph Lange, Jochen Reutelshöfer, Sebastian Schaffert, and Hala Skaf-Molli, editors, *5th Workshop on Semantic Wikis*, number 632 in CEUR Workshop Proceedings, 2010. URL `http://kwarc.info/clange/pubs/semwiki2010-baudenkmalnetz.pdf`.

[10] Norbert Fuhr, Giannis Tsakonas, Trond Aalberg, Maristella Agosti, PrebenHansen, Sarantos Kapidakis, Claus-Peter Klas, László Kovács, Monica Landoni, András, Micsik, Christos Papatheodorou, Carol Peters, and Ingeborg Sølvberg. Evaluation of digital libraries. In *International Journal of Digital Libraries 8*, 2007.

[11] Deyan Ginev, Constantin Jucovschi, Stefan Anca, Mihai Grigore, Catalin David, and Michael Kohlhase. An architecture for linguistic and semantic analysis on the arXMLiv corpus. In *Applications of Semantic Technologies (AST) Workshop at Informatik 2009*, 2009. URL `http://www.kwarc.info/projects/lamapun/pubs/AST09_LaMaPUn+appendix.pdf`.

[12] Tom Heath et al. Linked data – connect distributed data across the web. URL `http://linkeddata.org`.

[13] Sebastian Ryszard Kruk. *Semantic Digital Libraries*. PhD thesis, National University of Ireland, Galway, 2009.

[14] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*, chapter 6. MIT Press, Cambridge, Massachusetts, 1999.

[15] Glauco Mantegari, Matteo Palmonari, and Giuseppe Vizzari. Rapid prototyping a semantic web application for cultural heritage: The Case of MANTIC. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *The Semantic Web: Research and Applications (Part II)*, number 6089 in Lecture Notes in Computer Science. Springer Verlag, 2010.