BauDenkMalNetz – Creating a Semantically Annotated Web Resource for Historical Buildings

by

Anca Dumitrache

Supervised by Prof. Dr. Michael Kohlhase

Computer Science

Jacobs University Bremen

May 2011

Executive Summary

BauDenkMalNetz (listed buildings web) deals with creating a semantically annotated website of urban historical landmarks. The annotations cover the most relevant information about the landmarks (e.g. the buildings' architects, architectural style or construction details), for the purpose of extended accessibility and smart querying. BauDenkMalNetz is based on a series of touristic books on architectural landscape. After a thorough analysis on the requirements that our website should provide, we processed these books using automated tools for text mining, which led to an ontology that allows for expressing all relevant architectural and historical information. In preparation of publishing the books on a website powered by this ontology, we analyzed how well Semantic MediaWiki and the RDF-aware Drupal 7 content management system satisfy our requirements. Finally, the website was deployed by using Drupal 7 together with a custom module implemented for our ontology, and various external modules used for working with RDF.

Acknowledgements

I would like to thank Prof. Dr. Michael Kohlhase and Christoph Lange for offering guidance and support throughout this guided research. Furthermore, I would like to thank Lin Clark for the invaluable advice on how to configure and use Drupal 7. Finally, I would like to thank Deyan Ginev for helping with the LaMaPUn library, and the KWARC group for their suggestions and support.

Contents

1 Introduction

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 [Lan11]. 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 [AW09] about the city. However, for the moment, these books are only accessible in printed format. We wish to propose a way of discovering Bremen's architectural landscape that is suited for the tech-savvy tourist.

1.1 Why Semantic Web

The purpose of BauDenkMalNetz is to develop a web portal that publishes online printed text enriched with semantic annotations, as applied to books about Bremen's historical landmarks. There are strong advantages for enriching the contents of our website with semantic metadata. The clear structure of the data provides for a better user experience: browsing based on semantic categories, coupled with enhanced querying options, which operate on the underlying semantic structure of the text and not just string matching. A previously existing website that uses Semantic Web technologies to annotate data on significant buildings is Archiplanet [Arc].

At the same time, from the research perspective, BauDenkMalNetz provides a use case for various natural language processing techniques and Semantic Web standards that could benefit from more exposure. Through our endeavor, we aim to analyze how various Semantic Web languages and applications can be used for our research problem, and thus document a possible methodology for building a successful Web 3.0 application.

Publications usually make use of a concrete set of concepts, that relate to one particular subject area, and thus can be reduced to a strict vocabulary. Identifying this vocabulary was the starting point of the project, a key step in the process of producing a formal representation of the semantic metadata that our web portal needs to store. After we have created a conceptual model of our data, we have analyzed ways of publishing our semantically enriched text online. Finally, we compared and contrasted BauDenkMal-Netz to other cultural heritage web applications, and identified possible directions for further work.

1.2 Technical Terminology

This section will briefly introduce the technical terminology that will be used throughout this document.

1. Semantic Web concepts and principles:
    - **ontology** [BLHL+01]: In the context of information science, an ontology has been defined as "a formal, explicit specification of a shared conceptualization" [SS09], formalized in some logic by defining classes of concepts, together with the relations between them. The underlying structure behind most web ontologies is the taxonomy, a hierarchical representation of terms. For Semantic Web applications, ontologies are used for representing semantic data.
    - **ontology alignment** [ESC07]: Matching and aligning are procedures used in order to interlink various ontologies found across the Web. The purpose of alignment is reducing the heterogeneity of online data, by identifying and merging equivalent, concepts through procedures like pattern matching.
2. Semantic Web standards and languages:
    - **RDF** [Wor]: The Resource Description Framework is a standard endorsed by the W3C, for the purpose of representing semantic data across the Web. Its underlying principle is the RDF triple, a structure comprised of three parts: subject, predicate and object, each of them represented either by a unique URI, or a standard XML data type, defined as a literal.
    - **OWL** [MH04]: The Web Ontology Language is a web language and a W3C standard for defining and instantiating ontologies. OWL provides the vocabulary in which RDF triples are expressed – it defines RDF resources as OWL classes, which are then connected to each other via object properties, or to RDF literals through data properties.
    - **SPARQL** [PS08]: The SPARQL Protocol and RDF Query Language is the standard language for querying the Semantic Web, specifically data stored in

RDF triple format. Queries are usually executed over a SPARQL endpoint, which is a standardized HTTP interface .

- **XSPARQL** [AKK+08]: Combining XQuery, an XML-based query language, with SPARQL, XSPARQL is a novel query language that allows for getting XML results for queries over semantic metadata, and constructing RDF data based on XML documents.

3. Software discussed as part of this project:

- **LaMaPUn** [GJA+09]: The Language and Mathematics Processing and Understanding project, developed by the KWARC research group, is a project dealing with topics such as semantic enrichment, structural semantics and ambiguity resolution, as applied to mathematical documents.
- **Semantic MediaWiki** [Sem]: An extension of MediaWiki (the wiki engine which powers Wikipedia), it provides enhanced features for browsing and organizing its contents via semantic annotations.
- **Drupal** [Dru]: An open-source content management system implemented in PHP, which supports RDF mappings of its content.

2 Building an Ontology

The publications that lie at the basis of our work with BauDenkMalNetz have been made available to us (but not the general public) in simple HTML files. There is a file for each individual building, with pictures associated to each file, and information like the name of the architect being highlighted. Four books have been published thus far [AW09], with more than one hundred buildings being described in total.

In order to enable enhanced browsing and querying, the data on Bremen's historical buildings needs to be organized, and the proper semantic metadata needs to be put in place. For this purpose, we have developed the BauDenkMalNetz ontology, a formal representation of the metadata vocabulary on historical buildings and related concepts, together with the relations among them. As detailed in the following sections, the ontology was engineered in the following stages, which are proposed by the METHONTOLOGY [FLGPJ97] methodology: specifying the requirements, conceptualizing the knowledge domain, formalizing the ontology in OWL, and aligning it to related ontologies.

2.1 Scenario

An example scenario of interacting with a publication backed by the BauDenkMalNetz ontology involves a tourist, working out an itinerary for visiting the city of Bremen. For this purpose, she needs to be able to browse through a particular neighborhood, by filtering the buildings based on their addresses. Suppose she is interested only in visiting those buildings that were built in the 19th century. Then she finds one particular architect that she is familiar with, and she wants to add all of his buildings to her itinerary. Finally, during her visit, she will want to stop at each individual building and read up on its history, like the years between it was built, and what famous people had been living there.

2.2 Requirements

Based on this scenario, we have identified a list of requirements that the BauDenkMalNetz ontology needs to meet in order for the data to be easily accessible:

- *buildings* need to be represented as uniquely identified entities, which will be mapped to individual pages of the website; any knowledge represented using the BauDenkMalNetz ontology needs to be interconnected, with the building entity as the central point of the representation;
- information on the *physical address* and *neighborhood* needs to be available for every building;
- the *architect* and the *architectural style* of a building have to be highlighted when that information is available;
- the *time* and *timespan* over which a building was built has to be specified for individual entries.

A more general requirement that the BauDenkMalNetz website needs to address is browsing from one building to another. This could be supported by information on the buildings' physical location (e.g. they are on the same street), or based on characteristics that they share (e.g. they were built by the same person).

2.3 Text Analysis

Starting from these requirements and based on the original touristic guides, we identified they key concepts of the vocabulary that relates to historical buildings, by employing **n-gram models** to find the most likely occurrences of word groupings. 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 [MS99].

The results of this analysis were used in the conceptualization phase of the BauDenkMal-Netz ontology. The fact that the accuracy of n-gram models increases with the volume of the processed text was an advantage that made us consider this approach.

The first step that enabled us to process the text was removing the unnecessary HTML tags, and stripping it down to a plain-text format. The text is written in German; we needed to normalize it to plain ASCII characters, as the German-specific special characters seemed to interfere with the script used to analyze it. We made use of the LaMaPUn [GJA+09] Perl library for processing the text. We used a list of the most frequent German stop words in order to filter out the information that was not meaningful for the domain vocabulary.

We analyzed series of 1 to 4-gram models. The script recognized over 600 possible groupings of words that are likely to occur together. Over 500 of these groups had a likelihood coefficient larger than 2. This coefficient is computed by having the number of incidences of the words in the group together divided by the sum of individual incidences outside of the group.

The text analysis made apparent some clear trends. Most of the likely groups of words that appeared together referred to one of the following categories: *physical buildings* (e.g. Bahnhof (*train station*) Sankt Magnus, Kirche (*church*) Sankt Magni), *personal names* (e.g. Rudolf Alexander Schroeder), *physical addresses* (e.g. Leuchtenburger Strasse (a *street*), Am Bahnhof Sankt Magnus) and *building features* (e.g. Bungalow, Turm (*tower*)). By identifying these categories, we got a first impression of what are the key concepts we need to define for our ontology.

2.4 Conceptualization

Based on this analysis, and according to the requirements identified in the previous section, we conceptualized entities to be represented in the BauDenkMalNetz ontology. Most concepts identified during the n-gram analysis were transformed into resources,

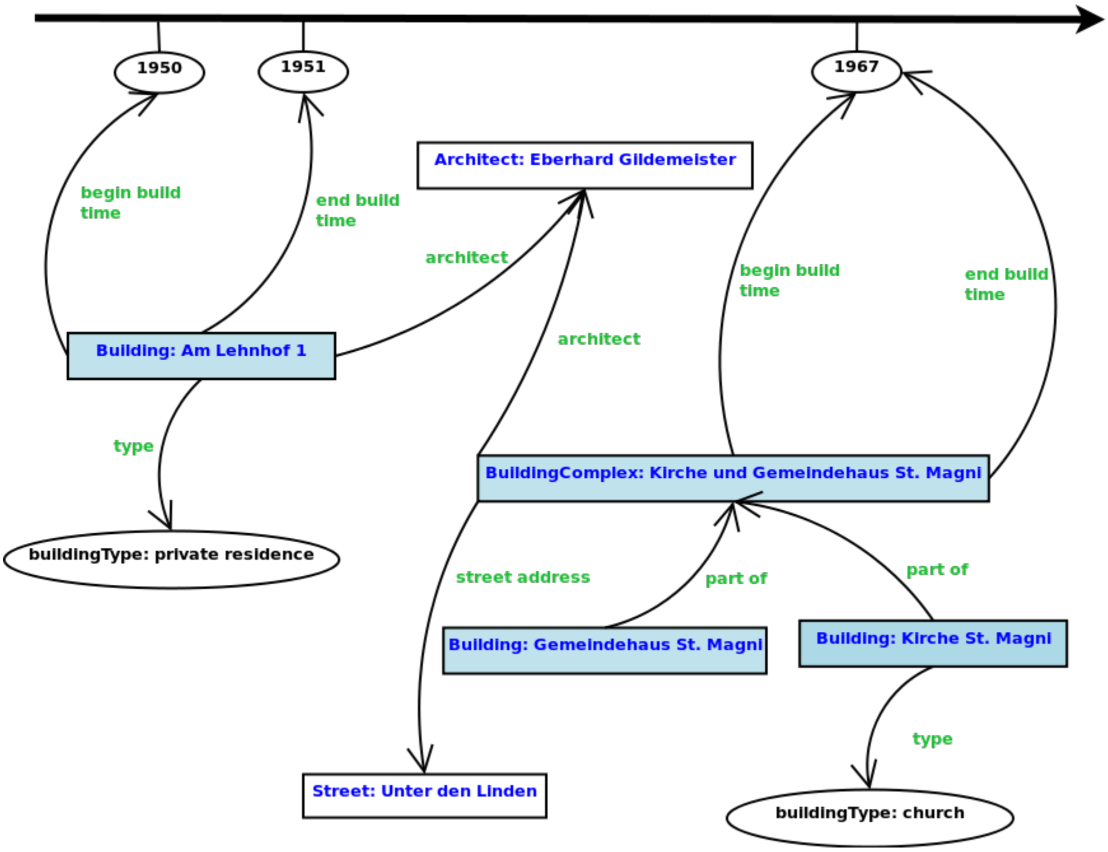

FIGURE 2.1: A fragment of the BauDenkMalNetz ontology

then properties were added to connect them. The core of the BauDenkMalNetz ontology is the following (concepts underlined, relations in *italics*):

- building – a resource identifying a particular building;
- building part – a subconcept of the building entity (e.g. tower, annex);
- building complex – a composite consisting of several building entities;
- building type – different types of constructions (e.g. church, hospital);
- *address* – the physical location of a building;
- *architect* – the person or group of people that have designed the building;
- *inhabitant* – famous person that has lived in that building;
- year – *when a building was built*; can refer to the year when *construction began*, *ended*, or both.

The full specifications of the ontology can be found in Appendix A.

2.5 Alignment to Other Ontologies

The Linked Data community [Hea+] advocates the reuse of knowledge models and vocabularies, in order to achieve interoperability across the Web. Indeed, there already exist various ontologies that model some of the relevant knowledge about historical buildings, out of which we found the following ones relevant for aligning with the BauDenkMalNetz ontology:

- The **GeoNames** [Geo] ontology models geospatial semantic information. In particular, it assigns to individual locations on the globe a unique URI. For our purposes, 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** [Cid] 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.).

3 Publishing in a Semantic Content Management System

For deploying BauDenkMalNetz, we have so far established requirements and analyzed how well two semantic content management systems satisfy these requirements: **Semantic MediaWiki** (SMW [Sem]) and **Drupal 7** [Dru].

3.1 Requirements

Based on the scenario discussed in the previous section, we have also analyzed the requirements that our website needs to provide. Digitally representing publications means that the BauDenkMalNetz web portal needs to build on the use cases of the written text that lies at its core, and enhance them with semantic browsing and querying capabilities that will provide for a better user experience. Therefore, a suitable content management system for deploying BauDenkMalNetz should offer the following functionality:

1. the possibility of integrating RDF triples, and at least a minimum of ontology support;
2. support for querying the RDF content of the website (e.g. by using SPARQL);
3. browsing based on the semantic metadata;
4. extensible publishing support for:
   (a) users, through enabling PDF and HTML exporting;
   (b) machines, by interlinking the publications across the Web, according to linked data principles;
5. the possibility of importing large amounts of text into the system.

3.2 Semantic MediaWiki

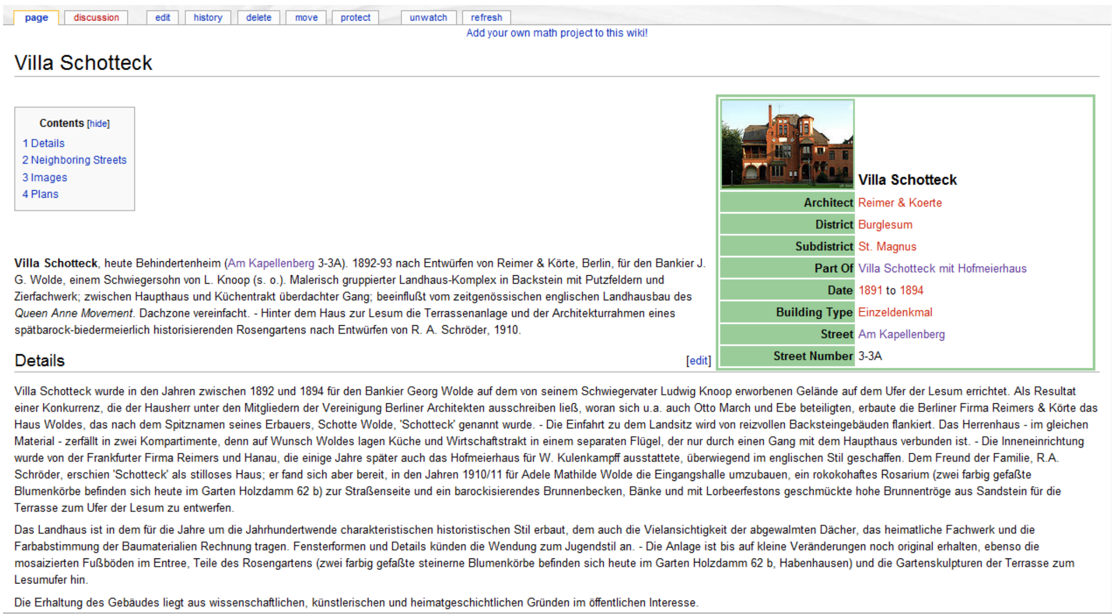

FIGURE 3.1: Screenshot of the SMW prototype.

Our motivation for using **Semantic MediaWiki** [Sem] (SMW) in deploying the initial protoype of BauDenkMalNetz [DLK+10] was its suitability for rapidly creating a working prototype (cf. [BDH+09]). SMW allows for easily adding and editing of the necessary data and metadata available on historical buildings, in keeping with requirements 1 and 3. New information could be easily incorporated and linked to the already existing data via SMW's page creation and editing tools. At the same time, the metadata vocabulary (i.e. the ontology) could be easily modified, simply by adding in-text annotations.

Requirement 2 is addressed by a simple query language included in SMW. The SMW querying functionality does not operate directly on RDF, and instead uses a syntax that addresses RDF triples based on the names with which they are declared in the wiki pages. While it provides basic functionality for querying RDF data, which includes selecting pages in the wiki, together with what properties of the pages to display, the SMW query language lacks the complexity of SPARQL (e.g. querying within a particular namespace).

When further assessing requirement 1, we found that the conceptual model of our metadata was less obvious and never explicitly formalized, as the ontology, to which the texts adhere, is not necessarily specified explicitly in SMW, but rather implied from the annotations done directly on the text. 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.

Requirement 5 was also not 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. The volume of data that needs to be processed makes it almost impossible to have the texts annotated manually, like we did for building the prototype, while also making BauDenkMalNetz rather suited for the employment of natural language processing techniques in order to get the needed semantical annotations.

3.3 Drupal 7

As our goal is to publish existing content, rather than creating new content in a collaborative way, we also considered **Drupal** [Dru], a rather traditional content management system. Given the BauDenkMalNetz documents collection and our ontology, we have analyzed Drupal's features with regards to the requirements established above.

Requirement 1 is satisfied as the latest version 7 of Drupal provides an RDF API [CDC+09] that is integrated in the Drupal core. This enabled us to easily upload our OWL ontology into the website, by using the RDF vocabulary import feature. For printed media, where a particular text usually does not undergo much change after being published, the advantage that Drupal brings is that, as the structure of the text is already known, its conceptualization can be set as the core of the website via the RDF API even before the website is deployed.

In order to set up the semantic metadata, first a PHP-compatible was installed, for the purpose of indexing the local RDF mappings of our website. A special content type for representing building resources from the BauDenkMalNetz ontology was created through

FIGURE 3.2: Screenshot of the Drupal website.

a custom Drupal module (the implementation details can be found in section B.1), and fields were declared for the RDF properties that have the building resource as subject. Finally, a geospatial component was integrated into individual building pages, by making use of the Google Maps API [Goob].

Requirement 2 is addressed by the SPARQL module for Drupal, which enables the use of semantic queries for our website. After the RDF mapping was set for individual pages and the website was populated, a SPARQL endpoint was set up over the triple store, in order to query the local RDF data. A visual interface for building meaningful queries, which can be manipulated both for the website administrator and by its users, is provided by the SPARQL Views [Cla10] module, which supports visual query building and result display.

For a detailed overview of the Drupal website, see Appendix B.

```
PREFIX bdmn: <http://oaff.info/ontology/bdmn>

SELECT ?name
WHERE {
    ?construction bdmn:hasName ?name .
    {
      { { ?construction bdmn:hasArchitect "Hermann Brede" .
        } UNION
        { ?building bdmn:hasPart ?construction .
          ?building bdmn:hasArchitect "Hermann Brede" . }
      } UNION
      { ?building bdmn:isPartOf ?construction .
        ?building bdmn:hasArchitect "Hermann Brede" .
      }
    }
}
```

FIGURE 3.3: Sample SPARQL query for constructions of all types whose architect is Hermann Brede.

3.4 Comparison

When comparing SMW to Drupal, we have encountered some drawbacks of SMW that led us to reconsider our approach. The flexibility and agility of SMW were not of a particular advantage in our setting. The publication sources are imported from external sources, and therefore we are not interested in MediaWiki's collaboration support. The ontology and its connections to other ontologies are, for now, created just by us, but they are not evolved or extended dynamically by a community – therefore we are not interested in giving write access to the ontology via the content management system. We rather prefer having a clear conceptual model of the metadata from the beginning. Drupal supports the initial import of such an ontology before importing the content and thus is suited for managing annotations to publications that have already existed before.

Also, we have concluded that using SPARQL to power our query engine would provide more flexibility for our queries, while also making them portable, as SPARQL is not platform dependent. While SMW is currently working to integrate SPARQL[1] functionality in its core, for the moment, the support it provides is limited, whereas Drupal provides SPARQL support through the modules discussed in the previous section.

For a point-to-point comparison between Drupal and SMW, see table 3.1.

[1] http://semantic-mediawiki.org/wiki/SPARQL_and_RDF_stores_for_SMW

Table 3.1: Comparison of SMW and Drupal based on the requirements list 3.1.

Req.	SMW	Drupal	Results
1.	inline RDF triples declaration, no explicit ontology support	RDF part of the core, Evoc module for ontology import	*Drupal* for better ontology support
2.	SMW query language	SPARQL, SPARQL Views modules	*Drupal* for advanced querying possibilities
3.	wiki pages mapped to resources and categories	RDF mapping for content types	draw
4a.	third-party plugin, not well documented	Printer, e-mail and PDF versions module in development[a]	*Drupal*
4b.	synchronizing with vocabularies supported by SMW through export[b] and import[c]	Evoc external vocabulary support	draw
5	through page creation, with manual semantic annotations	through page creation, but with specialized content types	*Drupal*

[a] http://drupal.org/project/print
[b] http://semantic-mediawiki.org/wiki/Help:RDF_export
[c] http://semantic-mediawiki.org/wiki/Help:Import_vocabulary

4 Evaluation

4.1 Methodology

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 [FTA+07] and semantic digital libraries [Kru09]. Out of the three concepts discussed by Fuhr: performance, **usability** and **usefulness**, we focused on the last two for our evaluation.

Usability refers to the ease with which users navigate the content. The test users provided feedback on how easy/difficult it was to navigate the system. Their reaction time was also measured.

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

4.2 Evaluation Setup

We have performed a task-based evaluation of our website, by devising a scenario that involved making use of the basic functionality of our website, which we then asked our test users to execute. The test users were 4 Computer Science students from Jacobs University, and each user got the same scenario. They had a limited time of 5 minutes to complete their assignment, and had to interact with the system without any additional help from the evaluator. This method was based on an evaluation setup for semantic wikis [HHM+09].

The main tasks that the users were supposed to perform were formulated based on the user interaction scenario discussed in section 2.1:

1. Suppose that you are currently visiting Sankt Magnus. Starting from the "Welcome" page of the BauDenkMalNetz website, browse to the section listing all districts of Bremen, and then browse through the buildings in the district of *Sankt Magnus*.

2. Find out how you can reach the building you are interested in by viewing its physical location on a map.

3. Suppose that you are particularly interested in the architect *Hermann Brede*, who designed many of the buildings in Sankt Magnus. Browse buildings in this district until you encounter one that he has designed, and then find out what other buildings he has designed in the area.

Finally, the users were asked whether they found it easy/difficult to navigate the website in order to perform their assignment, and whether the contents of building pages were sufficient and informative. These tasks were constructed in order to test the custom queries for navigating the semantic contents of the website, but also to test whether the underlying structure of the website, the ontology, supports all the necessary information and meets the user requirements.

4.3 Results Analysis

We will discuss the results of the user evaluation by referring to the usability of the system, usefulness of its contents, as described in subsection 4.1.

Usability has been assessed by measuring the number of tasks that were successfully completed in the assigned time. The success rate in this case was 100%, as all users managed to work out the 3 tasks provided to them without any extra assistance. Furthermore, when asked about how difficult it was to find their way navigating the system, the unanimous answer was that the website was informative enough to render its browsing intuitive. However, one of the users preferred using the keyword search module offered by Drupal over the predefined semantic queries embedded into pages, motivating his decision by the fact that keyword queries give him give him more freedom for browsing through the content. This issue will be adressed as part of the usefulness discussion.

Usefulness was assessed by inquiring the participants on whether the contents of building pages were consistent with what they were expecting to find out. All of them agreed that the RDF property fields embedded in building pages (e.g. architect, architectural style, build years) provided sufficient information about the key concepts that should be highlighted for buildings. However, they unanimously agreed that the map component should be better integrated, for example by providing the possibility to create custom maps based on a query. A further assessment of extended mapping functionality will be discussed in section 6, detailing directions for further development. When assessing querying usefulness, even though we will be keeping keyword search as an alternative means of querying for the BauDenkMalNetz website, the SPARQL queries we implemented could nevertheless benefit from providing extended information in the query results (e.g. when querying for all the buildings built by one particular architect, the outputted result could also include information on the geolocation of the building). Finally, one user expressed the desire to see the website integrated with social media platforms such as Facebook. While integrating a social-network repost component into our website would not involve making use of the semantic metadata, the concept of cross-website linking is consistent with Linked Data principles of interconnectivity [Gru08], and might also increase exposure of our web portal.

5 Related Work on Cultural Heritage

There exist a number of projects that process data about cultural heritage using Semantic Web technologies. Most approaches encountered gather the information from a wide array of sources (e.g. historical documents, archaeological excavation reports etc.), and consequently one of their main issues is developing an ontology that serves as a common medium for these different types of texts. In contrast, the BauDenkMalNetz ontology was developed from a singular source – published texts written in the same style, by the same author, on the topic of cultural heritage. Therefore, the ontology's intended use is not to provide a universal definition of the vocabulary describing historical buildings, but to define the vocabulary used by this particular series of publications. By studying the related work on cultural heritage we were able to shed some light on how we could improve our data model in order to represent a greater pool of sources, therefore enabling the reusability of our core ontology. For this purpose, the following applications have been assessed:

- **MANTIC** [MPV10] is a project similar to BauDenkMalNetz, that represents data on cultural heritage sites of the city of Milan, that was gathered from historical sources and publications. At its core, it uses the CIDOC CRM ontology for storing 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 work.

  Unlike BauDenkMalNetz, MANTIC deals with historical sources, which comprise a great variety of publications, written in different styles and over a long period of time. MANTIC provides a good example of how CIDOC CRM can be reused for representing historical landmarks, however, since the sources MANTIC deals with are so disjointed, identifying a common vocabulary for them is more difficult, and therefore no special ontology that deals primarily with historical buildings was devised.

- The **Fundación Marcelino Botín** [HRC+08] worked on a similar project that aimed to gather information on eleven cultural heritage sites of Cantabria, a region of Northern Spain. Like MANTIC, the Cantabria project had to reconcile information from a heterogeneous set of sources, by adapting the CIDOC CRM ontology to suit their dataset. However, most of the data populating the ontology had already been preprocessed (as spreadsheets, web pages etc.), and adding content to the project website was done in a semi-automated way. Therefore, unlike BauDenkMalNetz, the Cantabria project is intended as a community portal, where experienced users can modify or add new data to the website and to the ontology. Aside from providing another example of how to reuse existing standards, this project is relevant for us because of the way it makes use of the various benefits brought by using semantic metadata: a semantic search engine, an interactive map based on geoposition metadata, and interoperability with other cultural heritage repositories.

- **CultureSampo** [HMK+09] is an application that publishes cultural heritage information about Finland. Like BauDenkMalNetz, CultureSampo builds on existing standards for conceptualizing cultural items, and then extends them with domain specific information. However, as it covers a larger content (history, folklore, artifacts etc.), CultureSampo integrates a wide array of domain specific ontologies, that were developed in a semi-automatic fashion based on existing thesauri. While the development methodology of CultureSampo is relevant and can be adapted for BauDenkMalNetz, the scope of the project is too wide to enable us to reuse their data model.

6 Further Development

A possible idea of further expanding BauDenkMalNetz would be by better integrating the map component into the Drupal website. As discussed in section 3.3, the BauDenkMalNetz website integrates with Google Maps through a Drupal module, adding a map component with markers for each building page. However, as of now, Drupal does not provide the possibility of mapping its location fields to RDF. Therefore, the location data of buildings will not be semantically annotated, and SPARQL queries cannot be performed on it. Also, physical locations of buildings cannot be connected to other Linked Data resource directories (e.g. DBPedia [ABK+07] entries that connect geolocations to historical events, famous people etc.).

Mapping the semantic metadata relating to geolocation to a map instance can be achieved by using **XSPARQL** [AKK+08] as an alternative query language. One of the use cases of XSPARQL is creating queries on RDF that output results formatted in XML [PKC+09]. Although it is not yet integrated into Drupal, XSPARQL could provide additional possibilities of visualising results of semantic queries, as the XML output allows for easier manipulation.

A use case of XSPARQL proposed by the W3C is outputting RDF queries in Keyhole Markup Language [SP08] (KML). KML is an XML notation for representing geolocation information, developed to be used by Google Earth [Gooa]. KML can also be adapted for use by Google Maps.

```
prefix bdmn: <http://oaff.info/ontology/bdmn#>
prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>

<kml xmlns="http://www.opengis.net/kml/2.2">
{
for $building $name $long $lat
from <buildings_geodata.rdf>
where {
$building a bdmn:Building; bdmn:hasName $name;
        a [ a geo:SpatialThing; geo:long $long; geo:lat $lat ]
}
return <Placemark>
        <name>{fn:concat("Location of ", $name)}</name>
        <Point>
            <coordinates>
                {fn:concat($long, ",", $lat, ",0")}
            </coordinates>
        </Point>
    </Placemark>
}
</kml>
```

FIGURE 6.1: Sample XSPARQL query for creating KML data for buildings.

Through the reuse of concepts from the GeoNames [Geo] ontology, resources representing physical locations in the BauDenkMalNetz ontology are mapped to unique URIs representing their geolocation. The XSPARQL query exemplified in figure 6.1 can then select the RDF resources representing buildings that also map to meaningful geolocation information, and return the result in pre-formatted KML format, which can then be used for visualization in Google Earth.

While XSPARQL has yet to become a definitive W3C standard, the BauDenkMalNetz project, and arguably the Drupal community, would definitely benefit from having a PHP module for creating XSPARQL queries integrated into Drupal.

7 Conclusion

After assessing in which ways traditional printed publications on historical landmarks can be enhanced by transposing them in a digital format and enriched with semantic annotations, we devised the BauDenkMalNetz ontology, by analyzing its requirements and processing the texts that were made available to us by using text mining techniques. In keeping with linked data principles, we aligned our ontology to other existing representations that relate to our specific domain, like CIDOC CRM and GeoNames. Once we determined the structure of our metadata, we compared how different content management systems (SMW and Drupal 7) satisfy the requirements for deploying the BauDenkMalNetz website. As Drupal provides a more rigorous way of declaring a conceptual model, which is more suitable for digital publications, we have chosen it as the medium in which our web portal was developed. Finally, the website was implemented with the aid of a custom Drupal module for the BauDenkMalNetz ontology, and a series of other external modules used for working with RDF.

The BauDenkMalNetz website aims to provide a comprehensive and easy-to-use guide to the city of Bremen, and possibly even help boost the touristic appeal of the city. A possible enhancement to the web resource will be creating the map rendering of historical landmarks with the aid of semantic querying with XSPARQL. However, the scope of our work is not limited to Bremen. We believe that the ontology we devised will prove general enough to adapt in order to represent any touristic publication guide on historical landmarks.

A BauDenkMalNetz Ontology

The scope of the BauDenkMalNetz ontology is to formalize the main concepts behind a series of books [AW09] about Bremen's architecture. However, through alignment to related ontologies and an analysis of similar projects, we aim to create a reusable data model that could be adapted for other similar cultural heritage projects.

The main concepts of the BauDenkMalNetz ontology will be declared in this appendix. Concepts that have been reused from other ontologies are marked in *italics*, with the prefix marking the provenance ontology. For this purpose, the following prefixes will be employed:

- **cidoc:** CIDOC-CRM ontology [Cid];
- **geo:** GeoNames ontology [Geo].

The BauDenkMalNetz ontology is available for download at its namespace URI: `http://oaff.info/ontology/bdmn`.

A.1 Class Declarations

1. **Building Type**
    - *Properties:* 11 isOfType (range)[1];
    - *Description:* This class comprises the types that can be attributed to a 3 Building class, either by relating to the building's intended use, or to the particular way in which it was built.
    - *Examples:* church, tower, pharmacy.
2. ***cidoc:E24.Physical_Man-Made-Thing***
    - *Superclass of:* 3 Building, 4 BuildingComplex, 5 BuildingPart;
    - *Description:* This class is used to connect the local concepts relating to physical buildings to the CIDOC-CRM general concept of an object constructed by man.

[1]This class is in the range of the *isOfType* property (number 11 in the *Property Declarations* section).

3. **Building**

   - *Subclass of:* 2 cidoc:E24.Physical_Man-Made-Thing, 11 geo:Feature;
   - *Properties:* 7 hasPart (domain)[2], 8 isPartOf (domain);
   - *Description:* This class comprises all individual buildings.
   - *Examples:* Sankt Magnus church, Zum Birkenhof 5 house.

4. **BuildingComplex**

   - *Subclass of:* 2 cidoc:E24.Physical_Man-Made-Thing, 11 geo:Feature;
   - *Properties:* 7 hasPart (range);
   - *Description:* This class comprises sets of 3 Building classes that are part of an ensemble or connected through their intended use.
   - *Example:* Sankt Magnus church and parish hall complex.

5. **BuildingPart**

   - *Subclass of:* 2 cidoc:E24.Physical_Man-Made-Thing, 11 geo:Feature;
   - *Properties:* 8 isPartOf (range);
   - *Description:* This class comprises all auxiliary constructions that are somehow connected to a 3 Building class.
   - *Example:* Sankt Magnus church tower.

6. ***cidoc:E39.Actor***

   - *Superclass of:* 7 cidoc:E21.Person, 8 cidoc:E74.Group;
   - *Properties:* 4 hasArchitect (range), 6 hasInhabitant;
   - *Description:* This class comprises people or groups of people that are connected to a 3 Building class, either as the building's architects or its inhabitants. The E39.Actor subclasses represent that, even though they execute the same role and therefore are declared together, people should be declared differently from groups of people.

7. ***cidoc:E21.Person***

   - *Subclass of:* 6 cidoc:E39.Actor;
   - *Description:* This class comprises all individual people that are connected to a 3 Building class, either as the building's architects or its inhabitants.

[2]This class is in the domain of the *hasPart* property (number 7 in the *Property Declarations* section).

8. ***cidoc:E74.Group***

    - *Subclass of:* 6 cidoc:E39.Actor;
    - *Description:* This class comprises all groups of people that are connected to a 3 Building class, either as the building's architects or its inhabitants.
    - *Example:* Garrels family.

9. ***cidoc:E4.Period***

    - *Properties:* 9 hasPeriod (range);
    - *Description:* This class comprises all stylistic time periods to which a 3 Building class can ascribe. It should be noted that this not refer to exact years, but rather to culturally defined time periods.
    - *Examples:* Bauhaus, neoclassicism.

10. ***cidoc:E52.Time-Span***

    - *Properties:* 5 hasBuildTime (range), 12 beginYear (domain), 13 endYear (domain);
    - *Description:* This class refers to the time frame during which a 3 Building class was constructed. The exact years are declared through its data properties.

11. ***geo:Feature***

    - *Superclass of:* 3 Building, 4 BuildingComplex, 5 BuildingPart;
    - *Description:* This class is used to connect the local concepts relating to physical buildings to the GeoNames concept of geolocation, in order for each individual construction to receive a unique URI based on its physical location.

12. **Location**

    - *Superclass of:* 13 District, 14 Street;
    - *Properties:* 3 hasAddress (range);
    - *Description:* This class comprises all names by which a physical location can be called. Its subclasses detail more specific address components, such as streets or districts. The class does not refer specifically to a geolocation, but rather to its human-readable name.

13. **District**

    - *Subclass of:* 12 Location;

- *Description:* This class comprises the set of all urban districts or neighborhoods.
- *Examples:* Bremen Sankt Magnus, Upper East Side New York.

14. **Street**
    - *Subclass of:* 12 Location;
    - *Description:* This class comprises all street names.
    - *Example:* Zum Birkenhof.
15. **Photo**
    - *Properties:* 10 hasPhoto;
    - *Description:* This class comprises all images representing a 3 Building class.

A.2 Property Declarations

1. **hasDescription**
    - *Range:* plain RDF literal (with optional language information);
    - *Description:* This property connects an ontology class to its main description text string.
2. **hasName**
    - *Range:* plain RDF literal (with optional language information);
    - *Description:* This property connects an ontology class to its unique name text string.
3. **hasAddress**
    - *Range:* 12 Location;
    - *Description:* This property describes the human-readable address of a construction.
4. **hasArchitect**
    - *Range:* 6 cidoc:E39.Actor;
    - *Description:* This property describes the architect of a construction.
5. **hasBuildTime**
    - *Range:* 10 cidoc:E52.Time-Span;

- *Description:* This property describes the exact time frame during which a construction was built.

6. **hasInhabitant**

    - *Range:* 6 cidoc:E39.Actor;
    - *Description:* This property describes the person or group of people that have, at one point in time, inhabited a construction.

7. **hasPart**

    - *Domain:* 3 Building;
    - *Range:* 5 BuildingPart;
    - *Description:* This property describes construction annexes to a 3 Building class.

8. **isPartOf**

    - *Domain:* 4 BuildingComplex;
    - *Range:* 3 Building;
    - *Description:* This property describes a complex of buildings of which multiple 3 Building classes may be a part.

9. **hasPeriod**

    - *Range:* 9 cidoc:E4.Period;
    - *Description:* This property describes the stylistic period to which a construction ascribes.

10. **hasPhoto**

    - *Range:* 15 Photo;
    - *Description:* This property describes images that represent physical entities in the ontology.

11. **isOfType**

    - *Range:* 1 BuildingType;
    - *Description:* This property describes the 1 BuildingType characteristic of a construction.

12. **beginYear**

    - *Domain:* 10 cidoc:E52.Time-Span;

- *Range:* plain RDF literal;
- *Description:* This property describes the year in which the building of a construction began.

13. **endYear**

- *Domain:* 10 cidoc:E52.Time-Span;
- *Range:* plain RDF literal;
- *Description:* This property describes the year in which the building of a construction ended.

B BauDenkMalNetz Drupal 7 Website

The BauDenkMalNetz website was deployed by enhancing Drupal 7 [Dru] with a series of modules providing various RDF functionalities, and the BDMN custom module designed specifically for integrating our ontology into the website. The main references for implementing the website were a series of IBM developerWorks articles on using Drupal 7 together with Semantic Web technologies [Cla11; CC11]. This appendix will provide the documentation of the implementation details for the custom module, together with an overview of the most relevant modules used during development. In the final section, we will detail how content is imported to the website.

B.1 BDMN Module Documentation

The BDMN module for Drupal 7 was implemented with the purpose of providing an automated way of setting up a content type (i.e. a type of page that can be published on the website) for the 3Building class from our ontology. Specifically, the PHP module creates a "Building" content type, such that Drupal pages of this content type are published as instances of the Building class in our ontology, and all RDF properties that have this class as a domain are mapped to fields in the page, where objects can be declared either as the URI of another resource, or as literals. The module is comprised of 3 files: `bdmn.info` (detailing general information about the module, like its name and module dependencies), `bdmn.install` and `bdmn.module`.

The `bdmn.install` PHP file defines our custom content type via the `bdmn_install` function, next it defines the fields which will be embedded in pages of this type via the `_bdmn_installed_fields` function, and then it adds the fields to the content type via the `_bdmn_installed_instances()` function. Finally, the content type is added to the internal structure of Drupal.

The `bdmn.module` PHP file maps standard prefixes to the URI of our ontology, together with the ontologies we reused, through the `bdmn_rdf_namespaces` function, and then proceeds to map the content type and fields defined in the previous file to RDF entities in the `bdmn_rdf_mapping` function.

B.2 External Modules Overview

Drupal 7 comes with an RDF module as part of its core, which needs to be enabled in order for RDF mappings in nodes to be activated, and a triple store like ARC2[1] needs to be installed as part of the website libraries.

The main Semantic Web related modules that were employed for BauDenkMalNetz are:

- **RDF Extensions**[2] – This module pack contains the **Evoc** module for importing our external RDF vocabulary through its namespace URI, but also the **RDF UI** module, providing a visual interface for editing RDF mappings.
- **SPARQL**[3] – This module pack provides the **SPARQL endpoint** module, used for setting up a semantic query endpoint over our triple store, which was then registered for use via the **SPARQL Endpoint Registry** module.
- **SPARQL Views**[4] [Cla10] – A plugin for the **Views**[5] module for smart query building and visualization, SPARQL Views implements an interface for visually building semantic queries. This module was used for displaying the relations between different resources on our website as links on their respective pages.

Some general purpose modules were also installed, in order to provide extended functionality:

- **GMap**[6] – This module provides the possibility to integrate a Google Map component marking a building's location into individual building pages.
- **Drush**[7] – A command line shell and scripting interface for Drupal, this module was used for the website installation and maintenance.

[1] https://github.com/semsol/arc2/wiki
[2] http://drupal.org/project/rdfx
[3] http://drupal.org/project/sparql
[4] http://drupal.org/project/sparql_views
[5] http://drupal.org/project/views
[6] http://drupal.org/project/gmap
[7] http://drupal.org/project/drush

B.3 Populating the Website

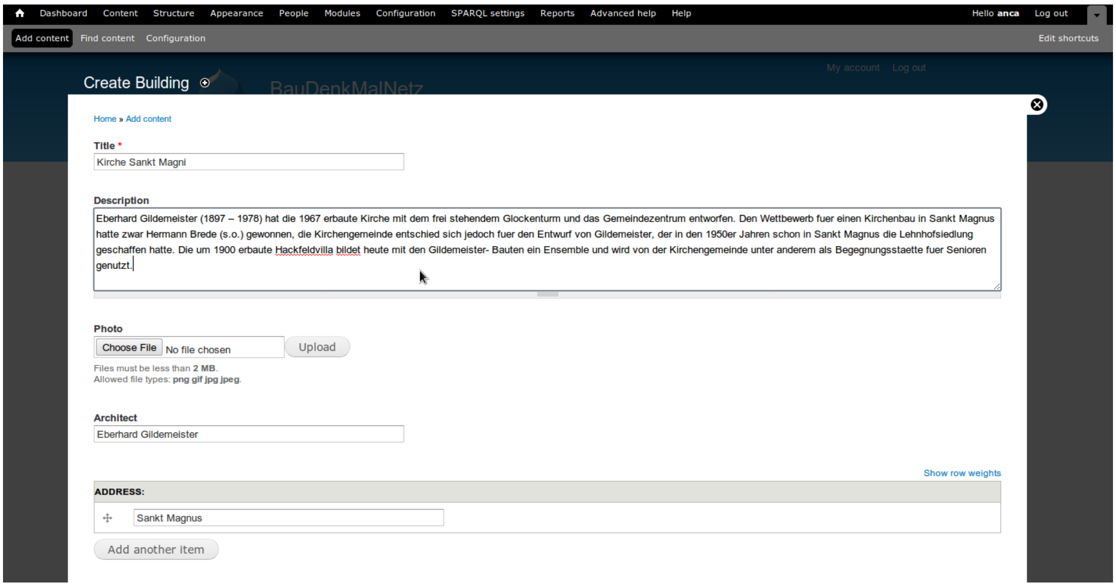

FIGURE B.1: Screenshot of the form for adding a “Building” page.

For creating individual building pages, Drupal provides an input form for the content type created by the BDMN module, as seen in figure B.1. The text in the description field is taken from the individual HTML file of the building. The fields used to connect the building to other RDF resources (e.g. architect, address) are populated with the keywords selected from the building file by the n-gram analysis in section 2.3.

In order for the SPARQL Views module to show the link to RDF resources connected to the building page by some property (as seen in figure 3.2), these resources also need to be created as pages, so as to receive a unique URI. For this purpose, specific content types for the other classes in the ontology were created (manually this time, as no other RDF properties need to be represented except for the ones linking them to building classes).

For example, a “Person” content type (representing the 7 E21.Person ontology class) was created, and each person that has designed a building is assigned an individual page of this content type. The predefined semantic queries, implemented with the aid of SPARQL Views, will then automatically provide a link on each building page to the page of its architect, and the individual page of an architect will list links to all the buildings that were designed by him. The criteria for matching a building page to the page of an architect is that the text string in the “architect” field from the building page must match the title of the architect’s page.

Bibliography

[ABK+07] S. Auer, C. Bizer, G. Kobilarov, et al. "DBpedia: A Nucleus for a Web of Open Data". In: *ISWC/ASWC*. 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007 (Busan, Korea, Nov. 11–15, 2007). Ed. by K. Aberer, K.-S. Choi, N. F. Noy, et al. Lecture Notes in Computer Science 4825. Springer Verlag, 2007. ISBN: 978-3-540-76297-3.

[AKK+08] W. Akhtar, J. Kopecký, T. Krennwallner, et al. "XSPARQL: Traveling between the XML and RDF worlds – and avoiding the XSLT pilgrimage". In: *The Semantic Web: Research and Applications*. 5th European Semantic Web Conference (ESWC) (Tenerife, Spain, June 1–5, 2008). Ed. by S. Bechhofer, M. Hauswirth, J. Hoffmann, et al. Lecture Notes in Computer Science 5021. Springer Verlag, 2008.

[Arc] *Archiplanet*. URL: http://www.archiplanet.org (visited on 2010-03-07).

[ATC+09] L. Aroyo, P. Traverso, F. Ciravegna, et al., eds. *The Semantic Web: Research and Applications*. 6th European Semantic Web Conference (ESWC) (Hersonissos, Crete, Greece, May 31–June 4, 2009). Lecture Notes in Computer Science 5554. Springer Verlag, 2009.

[AW09] N. Aschenbeck and I. Windhoff. *Landhäuser und Villen in Bremen*. Bremen: Aschenbeck Verlag, 2009.

[BDH+09] J. Bao, L. Ding, R. Huang, et al. "A Semantic Wiki based Light-Weight Web Application Model". In: *Proceedings of the 4th Asian Semantic Web Conference*. 2009, pp. 168–183.

[BLHL+01] T. Berners-Lee, J. Hendler, O. Lassila, et al. "The semantic web". In: *Scientific american* 284.5 (2001), pp. 28–37. ISSN: 0036-8733.

[CC11] S. Corlosquet and L. Clark. "The Semantic Web, Linked Data and Drupal, Part 2: Combine linked datasets with Drupal 7 and SPARQL Views". In: *developerWorks, IBM* (May 2011). URL: http://www.ibm.com/developerworks/web/library/wa-datasets/.

[CDC+09] S. Corlosquet, R. Delbru, T. Clark, et al. "Produce and Consume Linked Data with Drupal!" In: *The Semantic Web – ISWC 2009*. 8th International Semantic Web Conference (ISWC). Ed. by A. Bernstein, D. R. Karger, T. Heath, et al. LNCS 5823. Springer Verlag, Oct. 2009.

[Cid] *The CIDOC Conceptual Reference Model.* URL: http://cidoc.ics.forth.gr (visited on 2010-03-07).

[Cla10] L. Clark. "SPARQL Views: A Visual SPARQL Query Builder for Drupal". In: *Poster and Demo Proceedings of the 9th International Semantic Web Conference (ISWC).* 2010. URL: http://iswc2010.semanticweb.org/pdf/518.pdf.

[Cla11] L. Clark. "The Semantic Web, Linked Data and Drupal, Part 1: Expose your data using RDF". In: *developerWorks, IBM* (Apr. 2011). URL: http://www.ibm.com/developerworks/web/library/wa-rdf/.

[DLK+10] A. Dumitrache, C. Lange, M. Kohlhase, et al. "Prototyping a Browser for a Listed Buildings Database with Semantic MediaWiki". In: *5th Workshop on Semantic Wikis.* (Hersonissos, Crete, Greece, May 31, 2010). Ed. by C. Lange, J. Reutelshöfer, S. Schaffert, et al. CEUR Workshop Proceedings 632. Aachen, 2010. URL: http://ceur-ws.org/Vol-632/.

[Dru] *Drupal.org – Community plumbing.* web page at http://drupal.org. URL: http://drupal.org.

[ESC07] J. Euzenat, P. Shvaiko, and E. Corporation. *Ontology matching.* Springer, 2007. ISBN: 3540496122.

[FLGPJ97] M. Fernández-López, A. Gómez-Pérez, and N. Juristo. "METHONTOLOGY: from Ontological Art towards Ontological Engineering". In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI-97.* (Stanford, USA, Mar. 1997). MIT Press, 1997, pp. 33–40.

[FTA+07] N. Fuhr, G. Tsakonas, T. Aalberg, et al. "Evaluation of digital libraries". In: *International Journal of Digital Libraries* 8 (2007), pp. 21–38.

[Geo] *GeoNames.* URL: http://www.geonames.org (visited on 2010-04-23).

[GJA+09] D. Ginev, C. Jucovschi, S. Anca, et al. "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.

[Gooa] *Google Earth.* URL: http://earth.google.com (visited on 2011-05-01).

[Goob] *Google Maps.* URL: http://maps.google.com (visited on 2011-01-10).

[Gru08] T. Gruber. "Collective knowledge systems: Where the social web meets the semantic web". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 6.1 (2008), pp. 4–13. ISSN: 1570-8268.

[Hea+] T. Heath et al. *Linked Data – Connect Distributed Data across the Web.* URL: http://linkeddata.org (visited on 2010-06-11).

[HHM+09] P. Haase, D. Herzig, M. Musen, et al. “Semantic Wiki Search”. In: *ESWC.* 6th European Semantic Web Conference (ESWC) (Hersonissos, Crete, Greece, May 31–June 4, 2009). Ed. by L. Aroyo, P. Traverso, F. Ciravegna, et al. Lecture Notes in Computer Science 5554. Springer Verlag, 2009, pp. 445–460.

[HMK+09] E. Hyvönen, E. Mäkelä, T. Kauppinen, et al. “CULTURE SAMPO – A National Publication System of Cultural Heritage on the Semantic Web 2.0”. In: *ESWC.* 6th European Semantic Web Conference (ESWC) (Hersonissos, Crete, Greece, May 31–June 4, 2009). Ed. by L. Aroyo, P. Traverso, F. Ciravegna, et al. Lecture Notes in Computer Science 5554. Springer Verlag, 2009.

[HRC+08] F. Hernández, L. Rodrigo, J. Contreras, et al. “Building a cultural heritage ontology for Cantabria”. In: *Proceedings of the 2008 Annual Conference of the International Documentation Committee of the International Council of Museums (CIDOC), Athens, September.* 2008, pp. 15–18.

[Kru09] S. R. Kruk. “Semantic Digital Libraries. Improving Usability of Information Discovery with Semantic and Social Services”. PhD thesis. National University of Ireland, Galway, 2009.

[Lan11] Landesamt für Denkmalpflege Bremen, ed. *Denkmaldatenbank.* Jan. 10, 2011. URL: http://194.95.254.61/denkmalpflege/index.htm (visited on 2011-03-05).

[MH04] D. L. McGuinness and F. van Harmelen. *OWL Web Ontology Language Overview.* W3C Recommendation. World Wide Web Consortium (W3C), Feb. 10, 2004. URL: http://www.w3.org/TR/2004/REC-owl-features-20040210/.

[MPV10] G. Mantegari, M. Palmonari, and G. Vizzari. “Rapid Prototyping a Semantic Web Application for Cultural Heritage: The Case of MANTIC”. In: *The Semantic Web: Research and Applications (Part II).* 7th Extended Semantic Web Conference (ESWC) (Hersonissos, Crete, Greece, May 30–June 3, 2010). Ed. by L. Aroyo, G. Antoniou, E. Hyvönen, et al. Lecture Notes in Computer Science 6089. Springer Verlag, 2010.

[MS99] C. D. Manning and H. Schütze. “Statistical Inference: n-gram Models over Sparse Data”. In: *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: MIT Press, 1999. Chap. 6.

[PKC+09] A. Passant, J. Kopecký, S. Corlosquet, et al. *XSPARQL: Use Cases*. W3C Member Submission. World Wide Web Consortium (W3C), Oct. 20, 2009. URL: http://www.w3.org/Submission/xsparql-use-cases/.

[PS08] E. Prud'hommeaux and A. Seaborne. *SPARQL Query Language for RDF*. W3C Recommendation. World Wide Web Consortium (W3C), Jan. 15, 2008. URL: http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

[Sem] *Semantic MediaWiki*. URL: http://semantic-mediawiki.org (visited on 2010-03-04).

[SP08] E. Stefanakis and K. Patroumpas. "Google Earth and XML: Advanced Visualization and Publishing of Geographic Information". In: *International Perspectives on Maps and the Internet* (2008), pp. 143–152.

[SS09] S. Staab and R. Studer, eds. *Handbook on Ontologies*. International Handbooks on Information Systems. Berlin/Heidelberg: Springer Verlag, 2009. DOI: 10.1007/978-3-540-92763-3.

[Wor] World Wide Web Consortium (W3C), ed. *Resource Description Framework (RDF)*. URL: http://www.w3.org/RDF/.