BauDenkMalNetz – Creating a Semantically Annotated Web Resource of Historical Buildings

Anca Dumitrache and Christoph Lange

Computer Science, Jacobs University Bremen, Germany
{a.dumitrache,ch.lange}@jacobs-university.de

Abstract. 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 analyze how well Semantic MediaWiki and the RDF-aware Drupal 7 content management system satisfy our requirements.

1 Motivation

The architectural landscape of a city is not just made up of well-established landmarks, but of historical buildings with a rich cultural background that lie outside the mainstream touristic circuit. People wanting to explore 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 (Denkmalämter). However, this data has generally not been published in an easily accessible way. Existing databases and form-based search facilities are often tedious to browse through.[1]

In Bremen, an effort to collect this information and present it to the general public was made by the publisher Nils Aschenbeck, who released a series of city guide books [AW09]. However, for the moment, these books have only been published in print. By making use of these books, BauDenkMalNetz (German for "listed buildings web") proposes a way of discovering Bremen's architectural landscape that is suited for the tech-savvy tourist.

[1] See, for example, `http://194.95.254.61/denkmalpflege/index.htm`.

2 Transitioning from Written Text to Digital Media

The purpose of BauDenkMalNetz is to develop a web portal that publishes online printed text enriched with semantic annotations. 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 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 want to analyze ways of publishing our semantically enriched text online. Finally, we want to compare and contrast BauDenkMalNetz to other cultural heritage web applications, and identify possible directions for further work.

2.1 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. The ontology has been formalized and implemented in OWL, and was engineered in the stages specified by the METHONTOLOGY [FLGPJ97] methodology.

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.

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).

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** [2] to find the most likely occurrences of word groupings. The results of this analysis were used in the conceptualization phase of the BauDenkMalNetz 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] A probabilistic model that, given the first $n-1$ words in a sentence, will predict the n^{th} word. [MS99]

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[3]. Most concepts identified during the n-gram analysis were transformed into resources, 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.

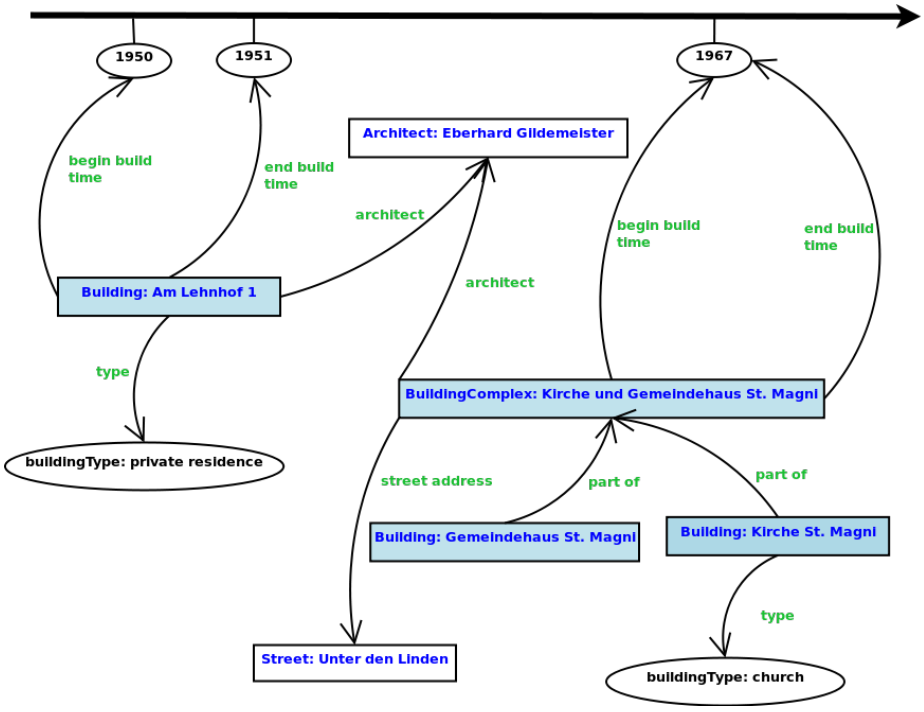

Fig. 1. A fragment of the BauDenkMalNetz ontology

[3] Available at: `http://oaff.info/ontology/bdmn#`.

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.).

2.2 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].

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.

Semantic MediaWiki SMW [Sem] was built as an extension of MediaWiki, the wiki engine which powers Wikipedia. It provides enhanced features for browsing and organizing its contents via semantic annotations. We built the first BauDenkMalNetz prototype using SMW [DLK+10].

Our motivation for using SMW in deploying the initial version of our web portal 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).

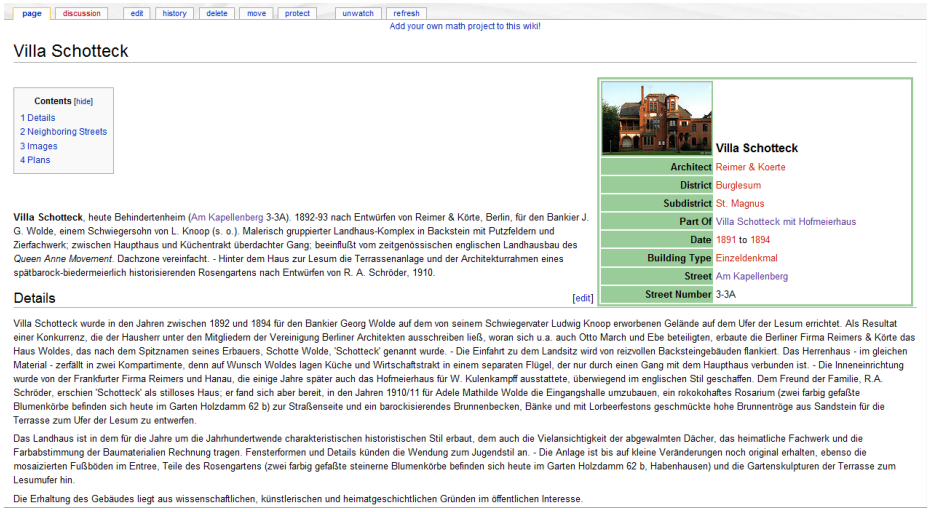

Fig. 2. Screenshot of the SMW prototype

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.

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 so far analyzed Drupal's features w.r.t. the requirements established above. Deploying BauDenkMalNetz in Drupal remains to be done in spring 2011.

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. The keywords pertaining to each resource were then added to the taxonomy of our website, and mapped to the corresponding classes and properties in the ontology. 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.

Requirement 2 is addressed by the SPARQL module for Drupal, which allows us to query our external triple store. The task of building meaningful queries is made even easier by the SPARQL Views [Cla10] module, which supports visual query building and result display.

Results 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[4] 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.

Table 1. Comparison of SMW and Drupal based on the requirements list.

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`

3 Development and Evaluation Plan

During spring 2011, we continued developing the BauDenkMalNetz website in Drupal, by uploading the texts of the tourist guides to our website, with the keywords in the vocabulary highlighted in the resource's pages. We will make semantic browsing available, based on these key concepts, achieved through Drupal's taxonomy feature. Also, for increased functionality, we will add a geospatial aspect to the semantic navigation by utilizing the Google Maps

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

API [Goo]. Finally, resources referring to people (e.g. the architect) will be cross-referenced with Linked Data resources, like DBPedia[5].

For even more advanced querying features, we are considering to make use of the **XSPARQL** [AKK+08] query language. XSPARQL combines the XML query language XQuery with the RDF query language SPARQL, which allows for generating XML-formatted results for queries over the semantic metadata of our website and, in future, interlinked websites. By selecting from a list of available queries, tourists will be able to create personalized guides of historical buildings.

For evaluating the usability of the BauDenkMalNetz website, existing methods for evaluating (semantic) digital libraries [FTA+07; Kru09] are applicable. A group of test-users will navigate through the website, providing feedback based on *usability* (of the content management system with our extensions) and *usefulness* (of the content, in the way our system publishes it). The 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. They will also 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 assessment the user-friendliness of the website we will consider possible improvements. A first release of BauDenkMalNetz, adapted according to the results of an initial evaluation round, is expected in May.

4 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

[5] `http://dbpedia.org`

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** [Fun] 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.

5 Conclusion and Further Work

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 will be developed.

Once finished, the BauDenkMalNetz website 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 our work is not limited to Bremen. We believe that both the ontology and the vocabulary will prove general enough to adapt in order to represent any touristic publication guide on historical landmarks.

Acknowledgments

The authors would like to thank Deyan Ginev for help with the LaMaPUn library, Lin Clark for help with assessing Drupal 7, and the anonymous peer reviewers for their pointers to further related work.

References

[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). Ed. by S. Bechhofer, M. Hauswirth, J. Hoffmann, et al. LNCS 5021. Springer Verlag, 2008.

[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.

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

[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*. Ed. by C. Lange, J. Reutelshöfer, S. Schaffert, et al. CEUR Workshop Proceedings 632. 2010. URL: `http://ceur-ws.org/Vol-632/`.

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

[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). 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.

[Fun] *Case Study: An Ontology of Cantabria's Cultural Heritage*. URL: `http://www.w3.org/2001/sw/sweo/public/UseCases/FoundationBotin/` (visited on 2011-04-12).

[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, Informatik*. 2009. URL: `http://www.kwarc.info/projects/lamapun/pubs/AST09_LaMaPUn+appendix.pdf`.

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

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

[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). Ed. by L. Aroyo, P. Traverso, F. Ciravegna, et al. LNCS 5554. Springer, 2009.

[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.

[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). Ed. by L. Aroyo, G. Antoniou, E. Hyvönen, et al. LNCS 6089. Springer, 2010.

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

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