ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-1017

WissKI: A Virtual Research Environment for Cultural Heritage

Martin Scholz[1] and **Guenther Goerz**[2]

Abstract. In this paper we present the WissKI System, an open source Virtual Research Environment and Content Management System for Cultural Heritage. It promotes semantic enrichment of data on the basis of OWL / RDF using the ontology CIDOC CRM / ISO 21127. The data is rendered in a Wikipedia-like fashion, combining textual, visual and structured information (info boxes) for a documented object on one page. Likewise, data can be acquired using field–based forms and semi–automatically annotated free text, resembling the most common traditional modes of documentation in the cultural heritage domain. This retains a user–friendly visualisation while at the same time providing detailed RDF triple data for automatic processing and data exchange.

1 INTRODUCTION

Research projects in cultural heritage (CH) domains create vast amounts of data in heterogeneous documents and data bases. But in most cases not all of the generated knowledge can be published, and reuse is hindered by the heterogeneous and scattered nature of the data bases. To support the reuse of research data and of the gained knowledge is an important goal of Virtual Research Environments (VREs), which should be designed to assist researchers throughout the "scholarly processing cycle" consisting of four essential steps: First, starting with digital primary sources, their conditioning and augmentation by metadata. Modeling is the second step, leading to annotated linked sources. Formal ontologies together with semantic dictionaries provide the basic building blocks for semantic annotation. Primary data with standardized semantic annotations offer potentials for federation with data from other sources, which obey the same standards, and for interpretation and knowledge generation, including collaborative refinement steps in scholarly communities. At this stage, VREs should provide interfaces to powerful data analysis and inference tools. Finally, the results will be released, presented and published in various formats and hence being turned into a new primary source for future research. For the whole cycle, VREs have to ensure authentication, authorization, and interoperability.

With the WissKI approach, motivated by needs of museum documentation, object-based research, and interoperability, we fostered the design and implementation of a prototypical system architecture of that kind. WissKI's presentation interface and communication facilities are influenced by Wikipedia, but data management relies completely on semantic (web) technologies [1]. The system supports full-text or field-based data acquisition, resembling the most common traditional modes of documentation in the CH domain. Field-based data acquisition is constructed by semantic paths derived from the underlying ontology, whereas full-text data acquisition is supported by a semi-automatic annotation system detecting places, persons, and events. The system supports the use of controlled vocabularies and thesauri.

[1] University of Erlangen-Nuremberg, Germany, email: martin.scholz@cs.fau.de
[2] University of Erlangen-Nuremberg, Germany, email: guenther.goerz@cs.fau.de

2 SYSTEM ARCHITECTURE

The WissKI System is completely web-based and implemented as a modular extension of the very popular open source content management system (CMS) Drupal[3], which already ships with hundreds of features like user management, blogs, etc. For storing the semantically enriched data, we integrated the RDF triple store ARC2[4]. The extensions are open source and can be downloaded from *https://github.com/WissKI*. The system can be easily deployed and maintained on a standard web stack configuration, being completely based on PHP and MySQL; a crucial aspect, as many CH experts are no computer experts. In this paper we will focus on the acquisition and presentation of semantically enriched data, leaving aside aspects of authorship, authenticity, publishing, etc.

2.1 Ontological data schema

WissKI's ontological data schema is illustrated in figure 1. As its logical backbone, WissKI uses the Erlangen CRM[5], an OWL-DL implementation of the CIDOC CRM (ISO 21127)[6], a reference ontology for CH documentation. The CIDOC CRM consists of 86 concepts and 137 properties. A WissKI information space may refine the CRM's concepts and properties in a so-called application ontology, according to the specific needs. Mutual interpretation between WissKI information spaces and other data pools is preserved by the common use of the CRM. Finally, WissKI encourages the use of local and global controlled vocabularies or thesauri for disambiguation and linkage of data sets. While the former are backed from the local data, the latter are used to refer to globally accepted external resources.

For each level of the ontology layer cake, WissKI provides import and export interfaces. It supports well-known exchange formats like Dublin Core and LIDO, the latter being the metadata harvesting format for Europeana[7].

The complexity and detailedness of the CIDOC CRM requires a lot of expertise that cannot be assumed for most practitioners. Consequently, WissKI was designed to offer users a familiar user interface.

[3] *http://drupal.org/*
[4] *https://github.com/semsol/arc2/wiki*
[5] *http://erlangen-crm.org* and [3]
[6] *http://cidoc-crm.org* and [2]
[7] *http://www.europeana.eu*

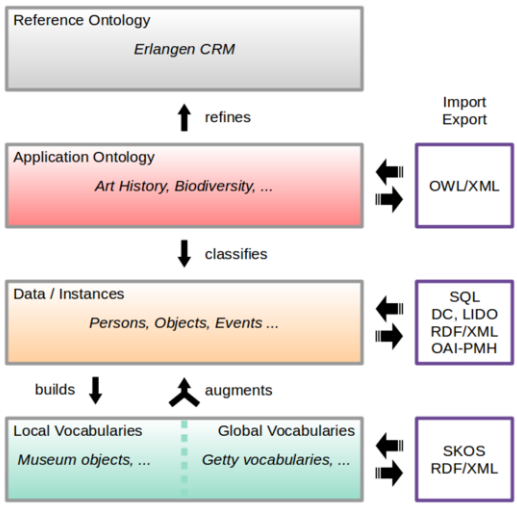

Figure 1. The ontological data schema applied in WissKI

A key feature is the introduction of so-called ontology paths, often recurring modelling patterns with a specific meaning. By defining and grouping such patterns, the complexity can be boiled down to — from the user's perspective — sets of key-value pairs for each category of the domain, like museum objects, persons or places. These sets are used in WissKI for data input, presentation and querying and allow the balancing act between compact and human-understandable data rendering and deep semantic modelling.

As an example, while the acquisition form for a museum object offers a simple "creator" field, the deep semantic modelling involves the museum object that was created in a production event that was carried out by an actor which had a naming with the data value of the field attached to it.

2.2 Data input

WissKI supports field–based and text–based data acquisition.

Forms for field–based data input are compiled using the defined ontology paths. Data will be stored as RDF triples according to the path definitions. The system will detect and display possible references to controlled vocabulary entries (like persons, places or objects) and automatically link them appropriately, eventually building a knowledge graph.

WissKI encourages writing free text and annotating occurences of named entities like persons, places and calendar dates and relations between the entities. WissKI assists the user by presenting annotation proposals based on an automatic text analysis. Currently, German and English are supported. The analysis process involves a preprocessing phase with lemmatisation and POS-tagging. Afterwards lexicon-based and heuristic algorithms are applied for named entity detection and disambiguation as well as relation detection.[8] As we aim at high-quality annotations, the user always has the possibility to manually revise the annotations proposed by the system. From the annotations, RDF triples will automatically be generated and added to the triple store.

To lower the acceptance threshold, the system extends the WYSIWYG web editor TinyMCE[9] for text input, which has a look-and-feel of common text processors. Text and annotation are encoded using (X)HTML.

Apart from manual input, WissKI provides a tool for automatic conversion of SQL databases and their import into the WissKI System to facilitate migration from legacy systems.

2.3 Data presentation

Like Wikis, WissKI preferably presents data on web pages, each describing one object or topic of discourse. This naturally goes together with traditional object-centered CH documentation. Each page may contain free text, images and structured information boxes. The structured information is compiled from data in the triple store according to the defined ontology paths.

Furthermore, the system provides alternative visualisations of the triple data like triple tables and several interactive graph representations. Here, the user may "look behind the ontology paths" and explore the full depth of the triple data.

Whenever possible, mentions of other object instances in the text or structured information will be be rendered as web links pointing to the linked object.

2.4 Data querying

Apart from following the links on the web pages, WissKI allows three ways of searching the local data pool. First, one can browse listings of object instances sorted by predefined categories. Second, the system provides a search form similar to those of library search facilities. Last but not least, the system implements a full-featured SPARQL [5] endpoint for advanced user queries or automatic processing.

3 CONCLUSION AND OUTLOOK

We presented an easy-to-use, web-based VRE for cultural heritage that orientates on Wikis for data presentation while relying on semantic technology. Our further development on the system aims at richer text annotation and analysis, integrating reasoning facilities and support in answering complex scientific questions.

ACKNOWLEDGEMENTS

We would like to thank our project partners Mark Fichtner, Georg Hohmann and Siegfried Krause. This project is funded by the German Research Council (DFG).

REFERENCES

[1] Tim Berners-Lee, James Hendler, and Ora Lassila, 'The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities', *Scientific American*, **284**(5), 34–43, (May 2001).

[2] N. Crofts, M. Doerr, T. Gill, S. Stephen, and M. Stiff. Definition of the CIDOC Conceptual Reference Model. Version 5.0.4, November 2011.

[3] Guenther Goerz, Martin Oischinger, and Bernhard Schiemann, 'An Implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL', in *CIDOC 2008 — The Digital Curation of Cultural Heritage.*, pp. 1–14, Athen, (September 2008). ICOM CIDOC.

[4] Guenther Goerz and Martin Scholz, 'Adaptation of NLP Techniques to Cultural Heritage Research and Documentation', in *32nd International Conference on Information Technology Interfaces*, pp. 1–8, Cavtat/Dubrovnik, (June 2010).

[5] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. W3C Recommendation, January 2008. *http://www.w3.org/TR/rdf-sparql-query/.*

[8] A more detailed description of the analysis process can be found in [4].

[9] *http://www.tinymce.com/*