Journal of Computing and Information Technology - CIT 18, 2010, 4, 317–324
doi:10.2498/cit.1001918

Adaptation of NLP Techniques to Cultural Heritage Research and Documentation

Guenther Goerz and Martin Scholz
Department of Computer Science, University of Erlangen-Nuremberg, Erlangen, Germany

The WissKI system provides a framework for ontology-based science communication and cultural heritage documentation. In many cases, the documentation consists of semi-structured data records with free text fields. Most references in the texts comprise of person and place names, as well as time specifications. We present the WissKI tools for semantic annotation using controlled vocabularies and formal ontologies derived from CIDOC Conceptual Reference Model (CRM). Current research deals with the annotations as building blocks for event recognition. Finally, we outline how the CRM helps to build bridges between documentation in different scientific disciplines.

Keywords: WissKI, cultural heritage, Semantic Web, OWL-DL, text analysis, event recognition

1. Museum Documentation: A Challenge for Content Analysis

When we speak of museum documentation, we address a wide variety of document types. There are acquisition and inventory lists or index cards containing more or less detailed records of museum objects. Often these are accompanied by photographs, restoration and archival records. Last but not least, there are documents ranging from short articles over catalogs to scholarly monographs.

With the introduction of information technology in museums and cultural heritage institutions, such records have been stored in (relational) database systems and content management systems. Since the 1990s, many metadata schemata have been proposed for the field of cultural heritage, some with very detailed classification features for specific object types. There is still a discussion about metadata standardisation, as can be seen with recent proposals for LIDO (museumdat/CDWA Lite) [14]. An extensive introduction to metadata standards may be found on Getty Foundation's web site at `http://www.getty.edu/research/conducting_research/standards`, [27/09/2010].

Today, access to museum documentation via the World Wide Web has become a matter of course. Because in most cases the data are too voluminous, only abridged versions are published in print, while the full data are available only in digital form. Web access allows many means to retrieve and publish the data, with very little cost involved. Using controlled language defined in terminologies and formal ontologies, different forms of "intelligent search" come within reach as well as interactive evaluation and visualisation methods.

Most museum documentation is centered around museum objects, i.e. there is a database system or content management system, which contains structured descriptions of museum objects and further information about their creators, provenance, use, and so forth, according to given descriptive and administrative metadata schemata. Besides fields in such data records enforcing (more or less strictly defined) data types, e.g. for inventory numbers, there are free text fields which contain important background information about persons, objects, materials, stylistic features, etc. without any further tagging. Basically, the free text fields are open for any kind of information which cannot be expressed in the strictly defined parts of the schema.

In particular, the free text fields and their relations to other fields indicate a clear need for content analysis. Firstly, named entities must be identified, in particular person and geographic place names. For instance, there may be a data field for the creator of a work of art and another one for the place where it was created, additionally one or more free text fields which speak about the artist's family relations, when he came to the mentioned place, and how long he stayed there, etc. As this example indicates, at least a second type of linguistic expressions, time specifications in a variety of forms, ought to be recognized. Current work adresses the identification and formal representation of event descriptions and how they are related among each other, for which the recognition of named entitites and time specifications is the first step. Hence, the integration of NLP and inferencing tools with a content management system for object documentation becomes indispensable.

With appropriate support by reasoning capabilities, a systematic access to implicit knowledge comes within reach. The technology also provides methods to link the data with external resources, e.g., authority files containing biographical or geographical information. Furthermore, interactivity opens up possibilities for Wiki-style annotation and scholarly communication, as well as forums for the general public. The combination of such techniques including inference with quality assurance could develop into a true "Epistemic Web".

In the following sections we describe our approach to address these problems. The next section outlines the architecture of the software framework we are developing, with the emphasis on language technology. Section 3 will address applied text analysis techniques: We show how the results achieved so far can be used to construct event-based shallow semantic representations based on CIDOC's Conceptual Reference Model (CRM) [3] as a formal reference ontology. The CRM is also the key to transdisciplinary approaches in museum documentation at the crossroads of biology and cultural history, as outlined in the final section.

2. The WissKI Approach and System Architecture

WissKI ("Wissenschaftliche KommunikationsInfrastruktur" – "Scientific Communication Infrastructure", `http://wiss-ki.eu`) is a DFG-sponsored project that addresses these questions. Our approach comprises the construction of a general web-based information portal, which at the same time offers facilities of a knowledge-based workstation and a moderated tool for science communication and different forms of publication. In general, the system will support scholarly communication and a new way of documentation in memory institutions, provide long-term availability of research results, assure the identity of authorship and the authenticity of information, enable persistence of citations, offer quality management tools, and support the preparation of scientific publications. For data input, it is compliant with formats familiar to its user community, such as index card fields and free text.

WissKI is being developed as an open source project, based on the open source Content Management System Drupal (`http://drupal.org/`; [08/01/2010]). The functionality required for WissKI is developed as new modules or as extensions to existing ones, including modules for vocabulary control, semantic text analysis, and data in/export. In particular, a module has been developed for accessing the semantic data in a set of triple stores separate from Drupal's relational database.

WissKI is ontology-based; for knowledge representation and reasoning, i.e., as an ontology language, we use Description Logics (DL) in its Semantic Web variant OWL-DL to implement the common generic reference ontology CIDOC CRM as well as some formal domain ontologies derived from it. The WissKI Base Ontology, which serves as a shared reference ontology, is an extension of the Erlangen CRM (`http://www.erlangen-crm.org`) – an OWL-DL implementation of the CIDOC CRM [6]. The CRM has been designed as a reference ontology for the cultural heritage (CH) domain to support data exchange and interoperability between CH systems. As opposed to most CH data schemata, it is truly event- and property-centered instead of a descriptive, bibliographic

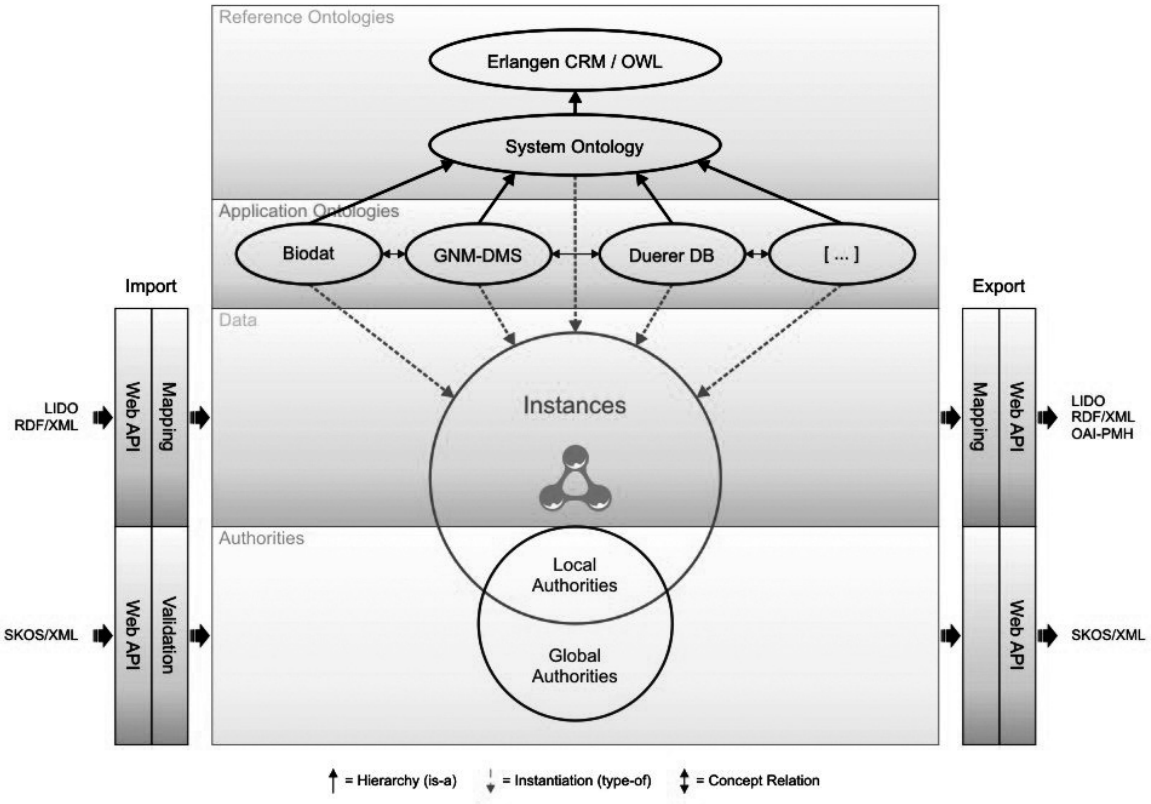

Figure 1. The ontology setup of the WissKI system.

(Dublin Core) style [4]. The WissKI Base Ontology also provides concepts and properties to support compatibility to another leading CH data schema, LIDO [14]. Optionally, specific domain ontologies deriving from the base ontology may be added. The object data are described as individuals instantiated from the ontology concepts, linked by ontology properties, and represented as RDF triples. Although two knowledge systems may not share the domain ontologies, the base ontology still ensures a minimum import and export compatibility for individuals. Figure 1 illustrates the system's ontological setup.

The first prototype of the WissKI system was presented at the conference "Museums and the internet 2010" (`http://www.mai-tagung.de`; [27/09/2010]) and is available as a demonstration system. Within the Drupal framework, it contains the reference ontologies CRM and WissKI Base in OWL-DL, various import-export interfaces, and the key components explained below, such as local and global name authorities (thesauri in SKOS format), tools to configure masks for data input and querying and last, but not least, the linguistic annotator tool plus a morphological analyzer embedded in a general web-based text editor. Further domain ontologies have been implemented for our applications in biology (Biodat) and cultural heritage (Nuremberg goldsmith art and the Early Duerer research database).

2.1. Reasoning with CRM

There are several reasons in favor of an OWL-DL implementation of the reference model. First of all is that for description logics there exist several reasoners providing powerful sound and complete inference services. Automatic classification and consistency checking of concepts and instances are well supported as well as dealing with semistructured data, since instances of the same class may have different attributes. (In OWL-DL a class definition describes both the necessary and sufficient conditions for membership). Also, data base queries can be expressed and executed to reveal implicit information, i.e.,

that is not directly retrievable from the database structure, e.g.

- "Get all drawings made by persons with location 'Nuremberg'."
- "Get all persons that could have met X."

WissKI makes use of consistency checking to identify contradictions and underspecifications automatically by means of the DL inference engine RACER (`http://www.racer-systems.com`; [15/01/2010]). Current work addresses the provision of predefined templates for queries such as those in [2], which cover most actual query types. Furthermore, due to DL's open world assumption, ontologies are more flexible w.r.t. changing the data schema than relational systems.

However, the WissKI approach also imposes major challenges to automatic reasoning, such as huge data bases (millions of triples, as is the case with our Biodat database, or even billions) and reasoning with distributed, access-restricted data. These are current research topics in DLs; eventually, specialized reasoning services need to be developed, cf. [13]. For the time being, we have made just one experiment with a specialized reasoner to precompile the transitive closure of the subsumption relation.

2.2. Using controlled vocabularies

In science, controlled vocabularies are a means to achieve accuracy and clarity. WissKI supports the use of established vocabularies and thesauri – "global name authorities" (GNA) for proper names –, as well as the creation of new vocabularies – "local name authorities" (LNA).

Each user group has to decide which GNAs shall be imported and to define a ranking over the vocabularies representing preferences.

Furthermore, the type of terms in a vocabulary must be specified – person names, place names, biological taxonomic terms, etc. – by means of a (predefined) template that defines the usage of the terms in the context of the base ontology. Support for user-defined templates is under way, though. The template type system helps for flexible and uniform term usage in forms (content hints) as well as for text analysis.

WissKI components may query GNA and LNA entries through an abstraction layer that provides a common functionality for all vocabularies. Import of GNAs is generally possible in SKOS [10], a generic and flexible RDF/OWL-based format for encoding all kinds of vocabularies. Alongside, interfaces may be implemented and installed to directly access vocabularies in cases where an SKOS import is not feasible. Currently, there is a proof-of-concept API for the gazetteer web service `geonames.org`.

3. Semantic Analysis of Free Text

In WissKI, semantic analysis of free text is oriented toward the needs of museum documentation. Following the conceptual model of event-driven documentation, semantic analysis focuses on the detection of events relevant for documentation and the involved actors and objects, as well as the circumstances, in particular place and time.

Because of our actual text corpus, NLP techniques have been designed for German, although an essential part of algorithms is language neutral. The system has been made ready for processing English texts with little effort. The goal is to augment the texts with high quality semantic annotations suited for scientific publishing. As state-of-the-art text analysis algorithms cannot guarantee such a level of accuracy, the whole annotation process is designed as semi-automatic.

3.1. The annotation system

Key feature of the semantic annotation of free text (fields) in WissKI is the embedding of TinyMCE (`http://tinymce.moxiecode.com`; [15/01/2010]), an online WYSIWYG editor providing rich text editing ("Word-like look-and-feel") for XHTML as a convenient writing environment. The editor is extended by plug-ins to support semantic annotation, comprising GNA and LNA terms and events as well as links between them. The annotation process is live: While typing, the user is immediately shown recognized items and is encouraged to revise the automatically found annotations or to choose among alternatives via dialogues. The

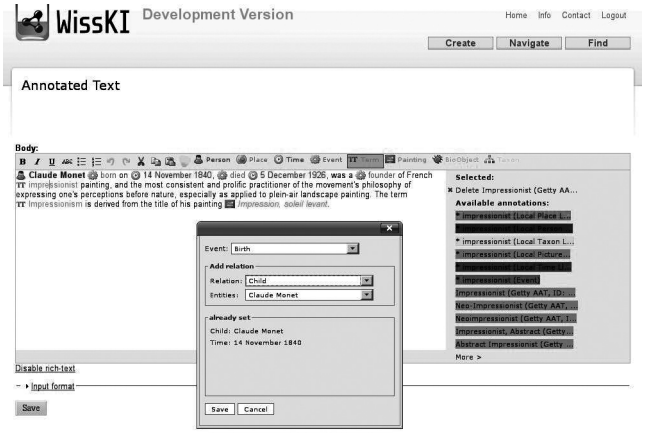

Figure 2. The enhanced WYSIWYG editor user interface. It shows the first sentences of the English Wikipedia article on Claude Monet, parsed and annotated with the WissKI system. Though the system is designed especially for German texts, most of the functionality is language neutral.

editor uses icons, colours and different fonts to distinguish term types and annotation states such as "proposed" and "approved". Further studies regarding the appropriate graphical interface and representation of annotations are underway to optimize users cooperation in the annotation process w.r.t. high quality semantic markup. Figure 2 shows the editor GUI to-date.

In order to keep the processing complexity on client side as low as possible, actual semantic analysis is executed on the server. The editor exchanges data with the server in an AJAX-like way, so that the annotation backend has a synchronized copy of the text which allows for sophisticated semantic analysis.

3.2. Semantic analysis techniques

Generally, semantic annotation in WissKI falls into three parts (cf. Figure 3). In a preprocessing step, the text is tokenized, tagged and lemmatized using standard open source software: As tagger, we use the Stuttgart Tree Tagger (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`; [15/01/2010]), augmented with our own morphology tool MORPH [8].

Then, semantically relevant text chunks containing the discourse referents – introduced by name authority entries as well as event markers – are detected and annotated. At last,

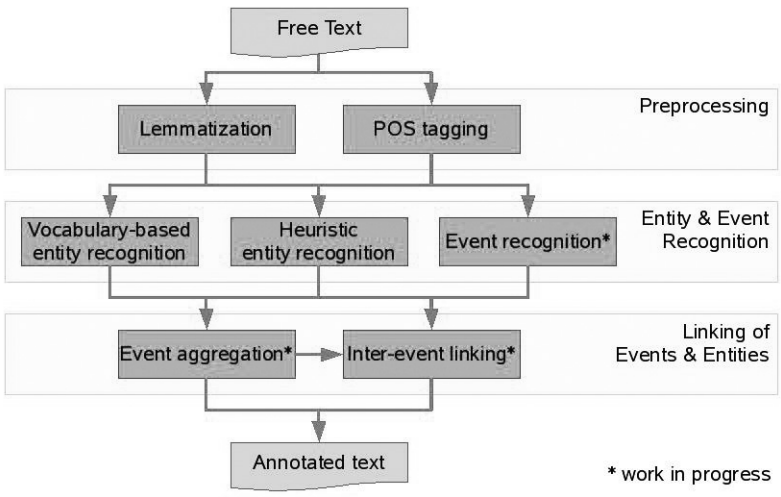

Figure 3. The three steps of the free text annotation unit.

links between these chunks, especially between events and their properties, are established. In any case, shallow processing methods are preferred. While vocabulary term recognition is completed, detection of events and relations is work in progress.

As WissKI focuses on the use of controlled vocabularies and ontologies, the recognition of discourse referents makes heavy use of the available GNAs and LNAs. A language-independent lexicon search algorithm identifies occurrences of vocabulary entries in the text. Apart from lexicon search, heuristics are used for detecting calendar dates and, additionally, for detecting unknown person and place names. The parsers make use of lexico-syntactic rules optimized for German.

A more detailed description and an evaluation of the annotator performance with different name authorities cf. [7]. The annotations offered to the user for selection are given in the vocabulary ranking order.

Instances describing the semantic information implied in a term are created using the vocabulary type templates. The occurrence of a term may trigger the creation of several connected instances according to CRM modelling. E.g., each calendar date (e.g. "01.01.2010") produces four instances: One for the label, one for the time interval corresponding to that label, and one each for the beginning and end of the interval. The time interval instance is linked by specific properties to each of the other instances.

3.3. Encoding semantic structure

While the text is encoded in XML and presented with XHTML, the information gained from semantic analysis is expressed in OWL-DL. Instead of directly storing the data into the system's triple store, the OWL statements are interwoven into the text via RDFa, i.e., the OWL statements are encoded "inline" in the XHTML document using XHTML language elements [1]. This allows for both, a human- and a machine-readable representation, in one place. Even fine-grained linkage of text passages and corresponding OWL-DL statements is easily possible, as text and OWL-DL are encoded at the same place. There are numerous scenarios for exploiting this linkage: e.g., an application inferring contradictions from the OWL-DL statements may easily show the respective text passages. RDFa is also a convenient way to allow for divergent opinions and marking authorship. As the semantic data are stored only locally in individual author's texts, the community knowledge base will remain unaffected. Curators may include a subset of an author's assertions to the knowledge base, if consistent, turning it into community consensus.

3.4. On event recognition

Recognition of events in free text is a current research topic. Parsing results annotated with person and place names and time specifications provide a first partial semantic representation, on which event hypotheses with actors, objects, instruments, temporal-spatial locations, etc. can be built. In the composition of structured semantic representations, we follow a Neo-Davidsonian approach, in which the event is represented by a discourse referent, being the only argument of event (type) predicates.

Discourse referents for events are triggered by certain keywords. Again, a controlled vocabulary together with a template for instance creation may serve as a source for basic event detection. Because experiments with automatic vocabulary generation from scope notes of the event concepts in the CRM definition showed disappointing results, a handcrafted event vocabulary is currently under development.

To cope with some linguistic peculiarities, special parser features are required:

- Event descriptions are usually assumed to be associated with verbs. However, scientific writing goes hand in hand with an extensive use of nominalisation while the verb is semantically generic (e.g., "to happen"). In particular, a frequent phenomenon in German, event recognition is shifted towards the identification of event-bearing nouns.
- In German, many verbs have prefixes which occur discontinuously, sometimes with the main part of the verb at the front of the sentence and the other at the very end. Correct recognition and assembly of the parts is crucial, as the parts themselves may have completely different meanings from the compound.

At the time of writing, we are considering either to implement a special-purpose parser or to use a publicly available resource with broad coverage such as Gerold Schneider's dependency parser Pro3Gres using a German language model [12]. Semantic representations of events will contain type-independent information about participants, affected objects, the temporal-spatial setting, and also type-specific attributes like instruments, results, etc. The latter ones are derived from the properties associated with the event concept in the base ontology. For the identification of fitting discourse referents and anaphora resolution in general, we are going to re-implement previous work on Discourse Representation Theory [5]. Future work will address inter-event relationships, temporal, causal and mereological, building a bridge to plan construction.

4. Transdisciplinarity

In a long-term perspective, the federation of cultural heritage and science data is a prerequisite for finding answers to really hard research problems by modelling complex systems such as the medieval city, the globalization of knowledge, climatic change, biodiversity, etc. However, interdisciplinary practice has shown that this goal is often hindered by differing systems of terminology and classification that the disciplines bring in. This may result in experts misunderstanding each other and data sets not matching properly.

Recent efforts showed that there is in fact a way to a solution, indicated by the term "transdisciplinarity"; first results have been presented at workshops of the CIDOC working group on "Transdisciplinary Approaches in Documentation" at the CIDOC 2008 and 2009 conferences (online materials are available via `http://www8.informatik.uni-erlangen.de/IMMD8/Services/transdisc`; [27/09/2010]). Originating from philosophy of science [11], transdisciplinarity concentrates on problems which cannot be solved within a single disciplinary framework. It takes a new view on the unity of science, focussing on scientific rationality, not systems. Taking into account that transdisciplinarity addresses the practice of research, this framework supports an action and event perspective on a generic level, i.e. for the tasks of classification, representation, annotation, linking, etc. Thus, a key to transdisciplinarity is the unity of language for modelling, argumentation and justification that ensures mutual understanding and thus is a prerequisite for the creation of an added value.

As a simple example, in one of our databases on Nuremberg goldsmith art we recognize clues in the data, which point beyond the domain of cultural history: there are goblets and centerpieces (epergnes) showing sculptered animals, such as lizards and beetles. Two of the documented objects exhibit a beautiful stag beetle, which induces interesting questions about those insects, not only on their iconographic significance, but also on their determination and classification in biology, the distribution of species, etc. So, there is a need to connect with further knowledge sources, such as resources from biology, biodiversity research, etc. Whereas the stag beetle in the foot of the goblet is described in terms of art history and metallurgy, we find a completely different description of a pinned stag beetle in the BIODAT data base. We may be lucky to identify it there if we know the precise species name in advance, but in many cases the matching task will fail. Here, the CIDOC CRM can play the role of such a transdisciplinary framework; at least for the stag beetle on goblets and still life paintings, some other insects and also birds on drawings and paintings, the modelling task has been performed successfully. Furthermore, for the birds – hooded crows in Dutch winter scenes in Brueghel paintings – our transdisciplinary modelling effort provided a nice result for biodiversity research as a side effect: During the "little ice age" hooded crows lived in Western Europe, whereas today they can only be found east of the Elbe river. Thus, the CRM as a bridge to connect different description systems, also contributes to approach the long-term goal of an "Epistemic Web" [9].

Of course, appropriate tools are needed to support the expert in transdisciplinary research. The WissKI system is one effort to ease information finding and linking and the task of integrating data into the transdisciplinary framework. At least for the latter, the annotation engine and GUI play a key role as a means to access the semantics of free text and to obtain a formal representation of human knowledge that can be used for further automated processing.

5. Acknowledgments

The authors are grateful for valuable hints and discussions to Siegfried Krause, Georg Hohmann, Karl-Heinz Lampe, Mark Fichtner, and Bernhard Schiemann and to the anonymous reviewers for valuable suggestions.

References

[1] B. ADIDA, M. BIRBECK, S. MCCARRON, S. PEMBERTON, *RDFa in XHTML: Syntax and Processing*, World Wide Web Consortium, Oct. 2008.

[2] P. CONSTANTOPOULOS, V. DRITSOU, E. FOUSTOUCOS, *Developing query patterns*, in Research and Advanced Technology for Digital Libraries, vol. 5714 of Lecture Notes in Computer Science, Berlin etc., 2009, Springer Verlag, pp. 119–124.

[3] N. CROFTS, M. DOERR, T. GILL, S. STEAD, M. STIFF, *Definition of the CIDOC Conceptual Reference Model*, ICOM/CIDOC Documentation Standards Group, CIDOC CRM Special Interest Group, 5.02 ed., January 2010. `http://www.cidoc-crm.org/official_release_cidoc.html`.

[4] M. DOERR, The CIDOC conceptual reference model: an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24 (2003), pp. 75–92.

[5] I. FISCHER, B. GEISTERT, G. GOERZ, Incremental semantics construction and anaphora resolution using lambda-drt. In *Proceedings of DAARC-96 – Discourse Anaphora and Anaphor Resolution Colloquium*, S. Botley and J. Glass, eds., Lancaster, July 1996, pp. 235–244.

[6] G. GOERZ, M. OISCHINGER, B. SCHIEMANN, An Implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL. In *Proceedings CIDOC 2008 – The Digital Curation of Cultural Heritage*, Athen, Benaki Museum, 15.-18.09.2008, Athen, September 2008, ICOM CIDOC, pp. 1–14.

[7] G. GOERZ, M. SCHOLZ, Content analysis of museum documentation in a transdisciplinary perspective. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage*, Social Sciences, Humanities, and Education (LaTeCH – SHELTR 2009), Athens, March 2009, Association for Computational Linguistics, ACL, pp. 1–9.

[8] G. HANRIEDER, Morph: Ein moulares und robustes morphologieprogramm für das deutsche in common lisp. In *LDV-Forum*, vol. 11 (1994), pp. 30–38.

[9] M. D. HYMAN, J. RENN, Toward an epistemic web. Unpublished manuscript prepared for the 97th Dahlem Workshop on Globalization of Knowledge and its Consequences, `http://archimedes.fas.harvard.edu/mdh/epistemic-web.pdf`, November 2007.

[10] A. MILES, S. BECHHOFER, SKOS Simple Knowledge Organization System, World Wide Web Consortium, Aug. 2009.

[11] J. MITTELSTRASS, Transdisciplinarity – New Structures in Science, in Innovative Structures in Basic Research. Ringberg-Symposium, 4-7 October 2000, no. 5 in Max Planck Forum, München, 2002, pp. 43–54.

[12] R. SENNRICH, G. SCHNEIDER, M. VOLK, M. WARIN, A new hybrid dependency parser for german. In *Proceedings of GSCL Conference, Potsdam*, 2009.

[13] K. SRINIVAS, OWL Reasoning in the Real World: Searching for Godot, in Description Logics, 2009. *Proceedings of the 22nd International Workshop on Description Logics (DL 2009)*, Oxford, UK, July 27-30.

[14] R. STEIN, E. COBURN, CDWA Lite and museumdat: New Developments in Metadata Standards for Cultural Heritage Information. In *Proceedings of the 2008 Annual Conference of CIDOC*, Athens, September 15-18 2008.

Received: June, 2010
Accepted: November, 2010

Contact addresses:

Guenther Goerz
Department of Computer Science
University of Erlangen-Nuremberg
Haberstr. 2
91058 Erlangen
Germany
e-mail: `goerz@informatik.uni-erlangen.de`

Martin Scholz
Department of Computer Science
University of Erlangen-Nuremberg
Haberstr. 2
91058 Erlangen
Germany
e-mail: `martin.scholz@informatik.uni-erlangen.de`

GUENTHER GOERZ was born in 1947 in Nuremberg; Studied mathematics, computer science and philosophy (with a focus on logic, philosophy of language, philosophy and history of science) in Erlangen; PhD in computer science (natural language processing); 1972–1987 researcher at the University of Erlangen-Nuremberg; 1981 visiting assistant professor UCLA, Los Angeles; 1985, 2004 visiting scholar at CSLI, Stanford University; 1999 Visiting Scholar at ICSI, UCB, Berkeley. 1989–1991 professor of computer science, University of Hamburg; since 1991 professor of computer science (artificial intelligence), University of Erlangen-Nuremberg; also member of the School of Humanities Research; Interests: natural language processing; applied logic, knowledge representation and processing; digital media and cultural heritage documentation. Further research topics include: logic and philosophy of language, ontologies, philosophy and history of science.

MARTIN SCHOLZ has been a research assistant at the Department of Computer Science of University of Erlangen-Nuremberg, Artificial Intelligence Division since 2009. In 2008 he finished his studies in computer science at the Univ. of Erlangen-Nuremberg with a diploma thesis on named entity recognition. His research interests are natural language processing and knowledge representation, esp. knowledge extraction from documentational texts and semantic web technology.