Alexander Mehler, Rüdiger Gleim, Alexandra Ernst & Ulli Waltinger (2008):
WikiDB: Building Interoperable Wiki-Based Knowledge Resources for Semantic Databases
Sprache und Datenverarbeitung. International Journal for Language Data Processing 1, 2008.

WikiDB: Building Interoperable Wiki-Based Knowledge Resources for Semantic Databases

Alexander Mehler, Rüdiger Gleim, Alexandra Ernst, Ulli Waltinger
Bielefeld University

1 Introduction

Recently, the need for interoperable releases of the Wikipedia which make its content accessible to machine learning and corpus querying has been stated (Völkel et al., 2006). This research is in the line of efforts to utilize the Wikipedia (Ponzetto and Strube, 2007; Waltinger et al., 2008), wiktionaries (Zesch et al., 2008), wikimanuals and other special wikis (Mehler, 2008c) as large resources of linguistic and encyclopedic knowledge in NLP. The present article follows this approach from the perspective of cognitive interaction technologies. The aim is to enable artificial agents (Kopp and Wachsmuth, 2004; Kopp et al., 2005) to explore *crowdsourced* knowledge resources generated by large communities of web users. Theoretically spoken, this research tackles the grounding problem of cognitive science (Ziemke, 1999) by interfacing artificial agents with social ontologies. That is, object, linguistic and metalinguistic knowledge is exploited in a way that enables virtual agents to identify, label, track and continue the topic of a dialogue to which they participate as the interlocutor of a human user. That way virtual agents become beneficiaries of crowdsourcing so that their human users gain in turn from the increase of their communicative competence.

In order to meet this goal wiki-based knowledge resources have to be preprocessed on three interrelated levels: (i) on the syntactic level of their elementary building blocks (concerning pages and their links), (ii) on the semantic level of the content relations of these building blocks and (iii) on the pragmatic level of (co-)authorship relations. That is, NLP and related approaches demand highly reliable knowledge resources subject to a low effort of preprocessing them as a precondition of their reliability. Thus, fine-grained syntactic, semantic and pragmatic annotations are demanded which make explicit relevant while they filter out irrelevant information. The present article describes an approach to this threefold task of preprocessing, annotating and retrieving data from wiki-based knowledge resources. It addresses the following subtasks:

1. Firstly, the article describes a unified representation format for modeling structure formation on the three semiotic levels.
2. Secondly, it provides algorithms for automatizing the related syntactic, semantic and pragmatic annotations.
3. Finally, the article describes a database in conjunction with an application programming interface which allows maintaining, exploring and further processing these annotations.

As a result of solving these interrelated tasks the present article provides a model of a wiki-based *semantic database* — henceforth called WikiDB — which makes accessible crowdsourced knowledge resources to machine learning and related approaches.

1.1 Utilizing the Wikipedia in Artificial Intelligence and Related Disciplines

Generally speaking, approaches to utilizing the Wikipedia in NLP or AI can be divided into six interrelated areas:

1. *Ontology learning approaches* explore the category graph of Wikipedia in order to extract thematically focused ontologies or thesauri (Milne et al., 2006; Suchanek et al., 2007). Thereby, co-hyponymy links are distinguished, e.g., from purely navigational links (Chernov et al., 2006). Additionally, hyperonymy or ISA relations are separated from other category links (Ponzetto and Strube, 2007). Note that these approaches tend to be greedy as they focus on optimizing local links thereby disregarding the overall picture of the category system. This greedy nature is overcome by Muchnik et al. (2007) who describe an approach to extracting ontologies from the Wikipedia article graph which is sensitive to its topology.
2. *Approaches to enhancing the Wikipedia* combine the notion of the Semantic Web with that of crowdsourcing. Völkel et al. (2006), for example, propose an addition to the WikiMedia syntax which allows users to type hyperlinks by reference to an ontology. This approach presupposes that a large community of users actually makes use of typing so that it is somehow unlikely that corpus analyses will profit from it. See also Schaffert et al. (2006) who describe a system for making semantic annotations of hyperlinks in wiki-based media.
3. *Information extraction approaches:* Auer et al. (2008) propose a database-related approach in the form of the so called *DBpedia* which explicitly represents Wikipedia-based knowledge units by means of RDF. The aim is to provide database functionality for managing and retrieving propositional content from Wikipedia which in the final stage of the project should be queried like a database.
4. *Linguistic network-related approaches* utilize the Wikipedia or Wiktionary as a resource to explore the semantic similarity or relatedness of lexical (Zesch et al., 2008) or textual units (Gabrilovich and Markovitch, 2006, 2007). These relations can be used to infer lexical networks as a knowledge resource of, for example, lexical chaining, topic labeling or named entity recognition (Waltinger et al., 2008; Waltinger and Mehler, 2008b).
5. *Web as corpus approaches:* A general task of machine learning concerns the availability of adequate training and test data. Obviously, the Wikipedia offers such a rapidly growing data set. Consequently, there are approaches which focus on making available this resource in the line of computational and corpus linguistic research (Denoyer and Gallinari, 2006; Zesch et al., 2008).
6. *Agent-oriented approaches:* Multiagent-oriented simulations of the built-up of social tagging systems complement the present field of research from the point of view of AI. Generally speaking, they aim at reconstructing the social-semiotic dynamics of distributed knowledge systems by means of multiagent simulation models (Steels and Hanappe, 2006). See Dellschaft and Staab (2008) and Mehler (2008b) for examples of this line of research.

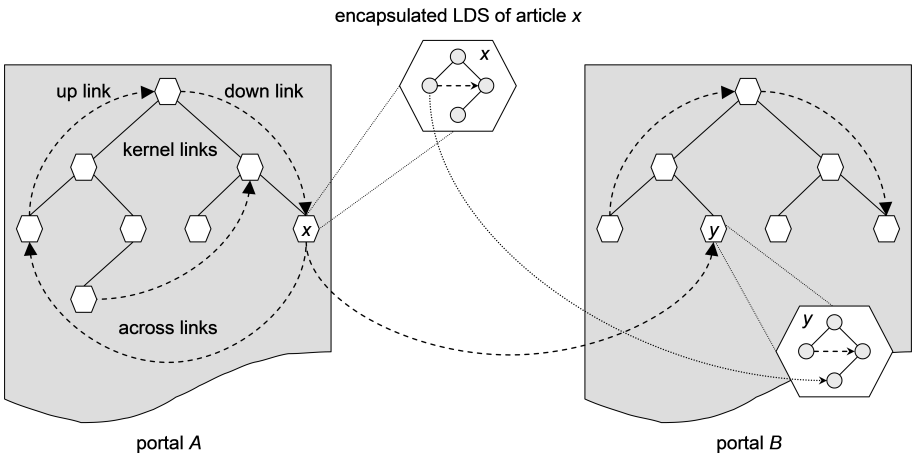


Figure 1: Two-level wiki document networks including the LNS (of article networks) and the LDS (of single articles).

These approaches commonly view the Wikipedia as a broad, widespread and diversified but highly flexible knowledge resource which because of this flexibility requires a high amount of preprocessing. That is, other than directly operating on the Wikipedia and therefore facing data anomalies due to high alteration rates, approaches which buffer and preprocess this resource aim at a higher level of data reliability and persistence. This article offers such an approach. Its basic tenet is that the preprocessing of wikis should occur on three interrelated levels, that is, on the level of syntactic, semantic and pragmatic computing. Beyond that, the article pleads for using database technologies as a prerequisite for mastering the huge amount of data offered by crowdsourced knowledge resources. As mentioned above the article introduces WikiDB as a semantic database which makes accessible crowdsourced knowledge resources to ML and NLP. In a nutshell, WikiDB addresses the gap between knowledge providing resources (i.e. wikis) on the one hand and knowledge demanding algorithms (which require training or testing data) on the other.

The remainder of the article is organized as follows: Section 2 explains the different annotation steps performed by WikiDB and sheds light on the corresponding graph models used to capture these annotations. Section 3 maps these graph models onto the database model of WikiDB. This section also describes an API for managing and retrieving the corresponding data. Next, Section 4 provides several case studies of using WikiDB which demonstrate its time and space complexity. Finally, Section 5 gives a conclusion and prospects future work.

2 A Three-Level Model of Wiki Document Networks

A wiki is more than simply a network of encyclopedic articles. There is a wide range of structure formation based on the type system of wiki pages. Articles, for example, are the smallest self-contained, non-recursively organized content units of wikis which do not include subpages[1] of the same type. On

[1]A subpage of a wiki page is a lower level page extending its name by a slash-separated name suffix (see `http://en.wikipedia.org/wiki/Wikipedia:Subpages`). In the case of a talk, e.g., subpages are used to archive parts of it. Note that the subpages of a wiki page span a tree.

the other hand, portals are complex content units in wikis which may consist of other portals down to the level of articles. Thirdly, portals, articles, and other wiki pages are normally complemented by talk, history and category pages which support collaborative generation of wikis or provide additional semantic information. Thus, a wiki document network is best conceived as a non-uniquely typed or labeled graph in which the *wiki page* is the reference unit of typing or namespace mapping, respectively. Exploring these content units and extracting reliable semantic information thereof requires preprocessing wiki networks on three levels. As explained in the following sections, this concerns syntactic, semantic and pragmatic annotations.

2.1 Syntactic Preprocessing

The syntactic preprocessing of wikis focuses on two reference levels of their structure formation: (i) on the level of the network structure of wikis as a whole and (ii) on the level of the document structure of their constitutive pages:

- The *Logical Network Structure* (LNS) of a wiki is, amongst others, spanned by the hyperlinks of its pages which induce a document graph (Mehler, 2008c). Beyond elementary pages, this graph contains composite units as, for example, portals spanning subgraphs of this document graph (cf. portal A and B in Figure 1). In mathematical terms, the document graph is an attributed, labeled, typed, ordered, directed, hierarchical graph: It is *directed* because of the orientation of the links, *ordered* because of their anchoring, *labeled* by the names of the pages, *typed* according to their namespace mapping, *attributed* by manifold information units (e.g. by the flag 'revision') and *hierarchical* as each vertex encapsulates the logical document structure of the corresponding page which itself is a graph (see, e.g., Article x in Figure 1, see also below). Note that the document graph actually spans as hypergraph as some of its vertices are linked by heterogeneous relations, that is, by hyperedges — see Mehler (2008c) for the corresponding graph model. Thus, we need a database model expressive enough to capture this range of graph structures. This graph model is presented in Section 3.1. Regarding the LNS, syntactic preprocessing means to annotate the *external* structure of wiki pages, that is, to span the document graph. This includes but is not limited to resolving redirect links (Mehler, 2006), unifying the representation of disambiguation pages[2] as well as delimiting portals and other composite units. Because of making explicit these and related aspects of wiki-based networking we speak of the logical *network* structure by analogy to the notion of the logical *document* structure.
- The *Logical Document Structure* (LDS) of a single wiki page is spanned by the constituents of its text structure (i.e., paragraphs, sections, tables, lists etc.). Note that this LDS spans a generalized tree.[3] That is, by analogy to the tree-like structure of portals (cf. Figure 1) the structure of single pages is generalized by graph-inducing (across, up or down) links and other cohesion relations (which are not manifested as hyperlinks). From this perspective, we can reuse the conceptual database model used to map the LNS (see Section 3.1) in order to capture the LDS of single documents. That is, in graph-theoretical terms the LDS provides nothing new. The

[2] Which, for example, in the Wikipedia are non-uniquely manifested by several functional equivalents.

[3] A generalized tree is a graph which consists of a spanning tree in conjunction with additional graph-inducing edges (cf. Dehmer et al. 2007 and Mehler 2008c for a formal definition of this notion).

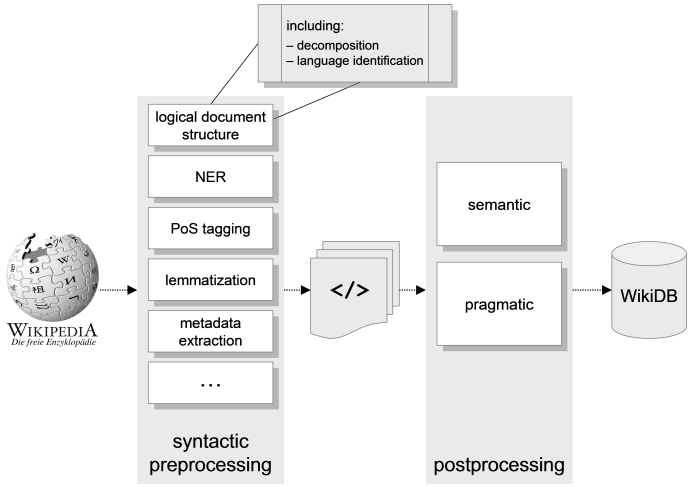


Figure 2: The preprocessing component of WikiDB.

syntactic preprocessing of the LDS results in annotating the *internal* structure of wiki pages. This includes but is not limited to lemmatizing their tokens, tagging parts of speech, recognizing named entities (as, e.g., persons or cities), identifying the language of document segments[4] and delimiting various constituents of their LDS (e.g. lists, paragraphs, sections).[5] Figure 2 summarizes these preprocessing steps. Note that this preprocessing component utilizes TEI P5 (Burnard, 2006) in order to annotate hyperlinks between documents as well as their logical document and lexical structure (partly including decompositions). Table 1 summarizes results on evaluating this preprocessing component.

Preprocessing Step	**Language**	**Parameter**	F**-Score**	**Test Corpus**
Lemmatization	de	888,573 word forms	.921	Negra Corpus (Uszkoreit et al., 1998)
PoS Tagging	de	3,000 sentences	.975	Negra Corpus
	en	5,000 sentences	.956	Penn Treebank (Marcus et al., 1993)
Language Identification	21 lang.	50 chars	.956	corpus extracted from the Wikipedia
	21 lang.	100 chars	.970	corpus extracted from the Wikipedia
Named Entity Recognition	de	full name	.992	newspaper corpus
	de	surname only	.836	newspaper corpus
	de	forename only	.709	newspaper corpus

Table 1: Results of evaluating components of syntactic preprocessing. For details on evaluating the lemmatization, tagging and language identification component see Waltinger and Mehler (2008a). For details on evaluating our named entity recognizer see Waltinger and Mehler (2008b). Note that the latter component mainly focuses on proper nouns.

This two-level model builds on Power et al. (2003) who define the LDS as a description level in-between the level of functional (semantic) and graphical document structure. We extend this notion

[4]Note that wiki pages may cite text in different languages.

[5]Lemmatization is provided for German and English wikis only.

to wiki-based hypertexts which are seen to manifest recurrent structures in the form of elementary wiki pages (e.g. articles), complex wiki documents (e.g. portals) and their graph-inducing hyperlinks. That is, we embed the graph-model of single documents (e.g. articles) into the graph model of the LNS. As a consequence, we get *hierarchical* graphs in which vertices (e.g. articles) denote graphs (i.e. document structures) on their own. These hierarchical graphs are mapped by a single conceptual database model (cf. Section 3).[6] Note that all annotations of the LDS and LNS are made automatically using the representation format described in Section 3.

2.2 Semantic Preprocessing

The semantic preprocessing of wikis concerns two reference points of content-based structure formation: firstly, the hyperlink-based networking of wiki pages and, secondly, the categorization of these pages by means of a social ontology in the form of the corresponding wiki category graph.[7] That is, semantic preprocessing concerns at least the networking of articles and of the corresponding category pages.

Exploring Wiki-Based Document Spaces A first task of semantic preprocessing concerns the evaluation of hyperlinks according to the semantic similarity of the articles linked by them. This task relates to identifying hyperlinks which lead to thematically unrelated pages and, thus, interfere with knowledge crawlers following these links. A widespread example of this *Link Abuse Problem* (LAP) relates to the usage of *dates* as anchors of semantically underspecified links. Look, for example, at the snapshot of the Wikipedia article on the composer Wolfgang Rihm in Figure 3: evidently, the first sentence solely contains hyperlinks (anchored by `March 13`, `1952`, `German` and `Karlsruhe`) to thematically unrelated articles far away from the topic *music*. Following such links the user is distracted from her focal thematic interest. As a result of this LAP users are offered too many hyperlinks only a fraction of which continues the focal article thematically. In order to tackle the LAP we use a simple algorithm which classifies hyperlinks as either *thematically related* or *unrelated* from the point of view of the anchor page. More specifically, for the directed article graph $\mathcal{A}(X) = (V, A)$ of input wiki X which may contain multiple or parallel arcs as well as loops this can be done in five steps:

1. We start with building an undirected graph $\mathcal{A}^*(X) = (V, E)$ such that $\forall e \in E \exists a \in A : \mathrm{in}(a) \neq \mathrm{out}(a) \wedge e = \{\mathrm{in}(a), \mathrm{out}(a)\}$. Note that $\mathcal{A}^*(X)$ does not contain multiple edges so that maybe $|E| < |A|$. Next, we map each article $v \in V$ onto a corresponding vector representation $\|v\| = \vec{v} \in \mathbb{R}^n$ which locates this article in vector space or semantic space, respectively, (Landauer and Dumais, 1997).
2. Based on this model we compute the similarities of all vector representations of all pairs $v, w \in V$ of documents. This is done by means of the cosine measure $\sigma : \mathbb{R}^n \times \mathbb{R}^n \to [-1, 1]$. Applying this measure we build a weighted undirected completely connected graph $\mathcal{A}'(X) = (V, E', \mu')$ where for all $e = \{v, w\} \in E' = [V]^2$: $\mu'(e) = \sigma(\|v\|, \|w\|) = \sigma(\|w\|, \|v\|)$.[8]

[6] See Mehler et al. (2007) for details on the graph model of the LDS and Mehler (2008c) for details on the graph model of the LNS.

[7] Note that in Wikipedia category pages belong to the namespace *category*.

[8] $[Y]^k$ is the set of all subsets of k elements of Y.

Figure 3: The first segment of the Wikipedia article on Wolfgang Rihm (composer) from the English Wikipedia (downloaded at July 11, 2008).

3. Next, we delete as many edges $e \in E'$ of lowest weight $\mu'(e)$ from $\mathcal{A}'(X)$ until we get a graph $\mathcal{A}''(X) = (V, E'', \mu'')$ which has as many edges as $A^*(X)$, that is, $|E''| = |E|$ — μ'' is the restriction of μ' on $E'' \subseteq E'$. Note that $\mathcal{A}''(X)$ is computed as an approximation of $\mathcal{A}^*$ irrespective of any knowledge about interlinked articles.
4. Now we can compute the F-score of the overlap of E and E'' as an indicator of the degree of approximation of $\mathcal{A}^*(X)$ by $\mathcal{A}''(X)$. That is $F(E'', E) = \frac{2}{1/\mathrm{P}(E'',E)+1/\mathrm{R}(E'',E)}$ where $\mathrm{P}(E'', E) = \frac{|E''|\cap|E|}{|E|}$ and $\mathrm{R}(E'', E) = \frac{|E''|\cap|E|}{|E'|}$ are the corresponding precision and recall values. Obviously, in the present case $F(E'', E) = \mathrm{P}(E'', E) = \mathrm{R}(E'', E)$.
5. Finally, for the best performing model used to compute $\mathcal{A}''(X)$ (see below) we call each hyperlink $a \in A$ *semantically supported* for which $\{\mathrm{in}(a), \mathrm{out}(a)\} \in E \cap E''$. Otherwise, if $\{\mathrm{in}(a), \mathrm{out}(a)\} \notin E \cap E''$, a is called *semantically unsupported*. Finally, we call any edge $e \in E'' \setminus E$ *semantically desirable* (as it would be semantically supported if a corresponding link would appear).

Table 2 reports results of computing the F-measure value for three models of $\mathcal{A}''(X)$:

1. As a baseline scenario we generate *Random Graphs* (RG) by uniformly randomly linking as many pairs of articles from V as there are edges in E. Further, we repeat this procedure 100 times and average the corresponding F-scores (see the fourth column in Table 2).
2. Secondly, we apply the *Vector Space Model* (VSM) (Salton, 1989). In this case those articles are linked which share many of more important lexical units, that is, units which are highly weighted in terms of the TFIDF scheme. This gives so called *Vector Space Graphs* (VSG) to be compared with $\mathcal{A}^*(X)$ (see the fifth column in Table 2).
3. Thirdly, we utilize an association measure which is sensitive to the order of texts to be processed (Mehler, 2008b). It builds on the vector space model except that it uses a weighting scheme without logarithmic damping. As a result we get so called *Association Graphs* (AG) to be compared with $\mathcal{A}^*(X)$ (see the last column in Table 2).

Wiki	URL	#Articles	#Links	RG	VSG	AG
Ameisen Wiki	ameisenwiki.de	636	5,748	.13	.34	**.35**
Firefox Wiki	www.firefox-browser.de	416	9,254	.1	**.43**	.40
OpenOffice Wiki	www.ooowiki.de	639	11,542	.06	.35	**.43**
Glottopedia	www.glottopedia.org	1,634	16,584	.11	.24	**.39**
InfoWissWiki	server02.is.uni-sb.de/courses/wiki	536	3,740	.16	.52	**.56**
Gentoo (Linux) Wiki	de.gentoo-wiki.com	696	10,069	.08	.19	**.21**
Statistik Wiki	statwiki.wiwi.hu-berlin.de	583	5,282	.12	.30	**.41**

Table 2: Results of evaluating the measurement of document network similarities regarding seven special wikis. Note that only articles have been taken into account as wiki pages.

Table 2 shows that with a sole exception association graphs perform best. We also see that the baseline scenario of random graphs is definitely outperformed: only about 10% of the hyperlinks are expectable to be set correctly by the principle of chance. However, we also see that small F-scores predominate irrespective of the wiki and of the model under consideration: even in the case of the best performing model about 50% of the links are classified as being *semantically unsupported* though being set in the underlying wiki. This also means that there are about 50% of edges within the corresponding association graphs which are classified as being *semantically desirable*: they do not exist in the original wiki though the corresponding articles are somehow related, at least on the level of their lexical organization. These findings can be explored by algorithms in NLP and ML to bypass the LAP. That is, while semantically unsupported links are left unprocessed, semantically desirable edges offer an alternative to exploring knowledge if data is sparse or otherwise lacking. Note that this approach also opens the possibility to enhance text linkage in existing wikis.

Re-aligning the Category Graph Any approach to NLP which presupposes, e.g., by analogy to the kernel hyperonymy hierarchy of WordNet (Fellbaum, 1998) a tree-like structure is doomed to fail when directly operating on the Wikipedia category graph (Voss, 2006). The reason is that the latter is not a tree, but a disconnected cyclic digraph. Thus, in order to facilitate social ontologies as *hierarchical* classifiers (Stein and Meyer zu Eißen, 2007) we need to annotate and separate hyperlinks between categories which span tree-like structures by analogy to hyperonymy relations from other graph-inducing relations. The idea behind this approach is that connected components of social ontologies span generalized trees (as done, e.g., by web documents — cf. the structure of portals in Figure 1). Such a generalized tree can be inferred from a social ontology given by a category graph of input wiki X in three steps:

1. Firstly, we have to explore the similarity relations of interlinked categories. These similarity relations can be computed by means of the VSM. More specifically, let $\mathcal{C}'(X) = (C', H')$ be the category network of wiki X and $[c] = \{v \in V \mid c \in C(v)\}$ be the set of all articles directly assigned to category $c \in C'$ — $C'(v)$ is the set of all categories to which v is assigned.[9] Then, for any categories $c, d \in C'$ for which there is an arc $a \in H'$ such that $\mathrm{in}(a) = c$ and $\mathrm{out}(a) = d$ we

[9]That is, $\mathcal{C}'(X)$ is spanned by the dominance relation among the category pages of X.

compute

$$\kappa(c,d) = \frac{1}{|[c]| \cdot |[d]|} \sum_{v \in [c], w \in [d]} \frac{\sigma(\vec{v},\vec{w}) + 1}{2} \quad \in [0,1] \tag{1}$$

as the average similarity of documents v, w assigned to c and d, respectively, where $\sigma\colon \mathbb{R}^n \times \mathbb{R}^n \to [-1,1]$ is the cosine measure operating on the vector space representations of v, w.

2. Secondly, as category graphs such as $\mathcal{C}'(X) = (C', H')$ are usually disconnected or have multiple sources we introduce a virtual root category $\top$ which secures connectivity by a single source.[10] More specifically, let $CC(\mathcal{C}'(X))$ be the set of all connected components of $\mathcal{C}'(X)$, $\mathcal{C}''(X) = (C'', H'') \in CC(\mathcal{C}'(X))$ and $rt(\mathcal{C}''(X))$ the set of all vertices $r \in C''$ such that for every $c \in C'' \setminus \{r\}$ there is a path in $\mathcal{C}''(X)$ from r to c. Note that we assume that for each $\mathcal{C}''(X) \in CC(\mathcal{C}'(X))$, $rt(\mathcal{C}''(X)) \neq \emptyset$. Then, we build an extended weighted category graph $\mathcal{C}(X) = (C, H, \kappa)$ where $C = C' \cup \{\top\}$ and $H = H' \cup \{a \,|\, \mathrm{in}(a) = \top \wedge \mathrm{out}(a) = r \in rt(\mathcal{C}''(X)) \wedge \mathcal{C}''(X) \in CC(\mathcal{C}'(X))\}$.
3. Thirdly, let $\mathrm{mst}(\mathcal{C}(X)) = (C, H^*, \top, \kappa^*)$ be the directed minimum spanning tree of $\mathcal{C}(X)$, $P_{rv} = (v_{i_0}, a_{j_1}, v_{i_1}, \ldots, v_{i_{n-1}}, a_{j_n}, v_{i_n})$, $v_{i_0} = \top, v_{i_n} = v, a_{j_k} \in H, \mathrm{in}(a_{j_k}) = v_{i_{k-1}}, \mathrm{out}(a_{j_k}) = v_{i_k}, k \in \{1,\ldots,n\}$, be the unique path in $\mathrm{mst}(G)$ from $\top$ to $v \in C$ and $V(P_{rv}) = \{v_{i_0}, \ldots, v_{i_n}\}$ be the set of all vertices of P_{rv}. Further, let κ^* be the restriction of κ to H^*. Then, each arc $a \in H \setminus H^*$ is classified as an[11]

$$\begin{aligned}
\textit{up arc} \quad &\text{iff} \quad a \in H_{[2]} \subseteq H \setminus H^* \cap H_{\mathrm{u}} = \{b \,|\, \mathrm{in}(b) = v \in V \ \wedge \ \mathrm{out}(b) = w \in V(P_{rv}) \setminus \{v\}\} \\
\textit{down arc} \quad &\text{iff} \quad a \in H_{[3]} \subseteq H \setminus H^* \cap H_{\mathrm{d}} = \{b \,|\, \mathrm{in}(b) = w \in V \ \wedge \ \mathrm{out}(b) = v \in V(P_{rw}) \setminus \{w\}\} \\
\textit{reflexive arc} \quad &\text{iff} \quad a \in H_{[4]} \subseteq H \setminus H^* \cap H_{\mathrm{r}} = \{b \,|\, \mathrm{in}(b) = \mathrm{out}(b) = v \in C\} \\
\textit{across arc} \quad &\text{iff} \quad a \in H_{[5]} \subseteq H \setminus H^* \cap C^2 \setminus (H^* \cup H_{\mathrm{u}} \cup H_{\mathrm{d}} \cup H_{\mathrm{r}})
\end{aligned}$$

Finally, we get a *weighted directed generalized tree*

$$\mathrm{GT}(\mathcal{C}(X)) = (C, H_{[1]}, H_{[2]}, H_{[3]}, H_{[4]}, H_{[5]}, H_{[6]}, H_{[7]}, \top, \kappa) = (C, H_{[1..7]}, \top, \kappa)$$

where $H_{[1]} = H^*$, $H_{[6]} = H_{[7]} = \{\}$ as the graph model of the category graph of input wiki X.

The basic idea of this algorithm is that kernel arcs $a \in H^*$ are spanned between more similar categories while up, down and across links are spanned between less similar ones. In this sense we assume the separability of a kernel hierarchy as the skeleton of the given social ontology. Of course, this can be seen to be disputable. However, it secures the accessability of social ontologies by means of the powerful apparatus of tree-based operations.

Figure 4 shows the interface of the wiki category viewer (Gleim et al., 2007) integrated into WikiDB: Having activated the category *Musik* (*music*) within the download of the German Wikipedia the user gets information about dominated as well as sibling categories. Note that across and down arcs are indicated by icons to the left of the corresponding category name.

The two-level approach to semantic preprocessing described in this section goes beyond related approaches to utilizing wikis as knowledge resources in NLP. The reason is that it does not only evaluate the semantic similarity of interlinked pages (whether articles or categories) but also provides

[10] Note that the category graphs of the Wikipedia tend to have cycles (Mehler, 2008a).

[11] Note that the MST of a directed graph cannot be computed by an algorithm adapted from its correspondent operating on undirected graphs. See, e.g., Edmonds (1967) for an algorithm for computing the MST of a directed graph.

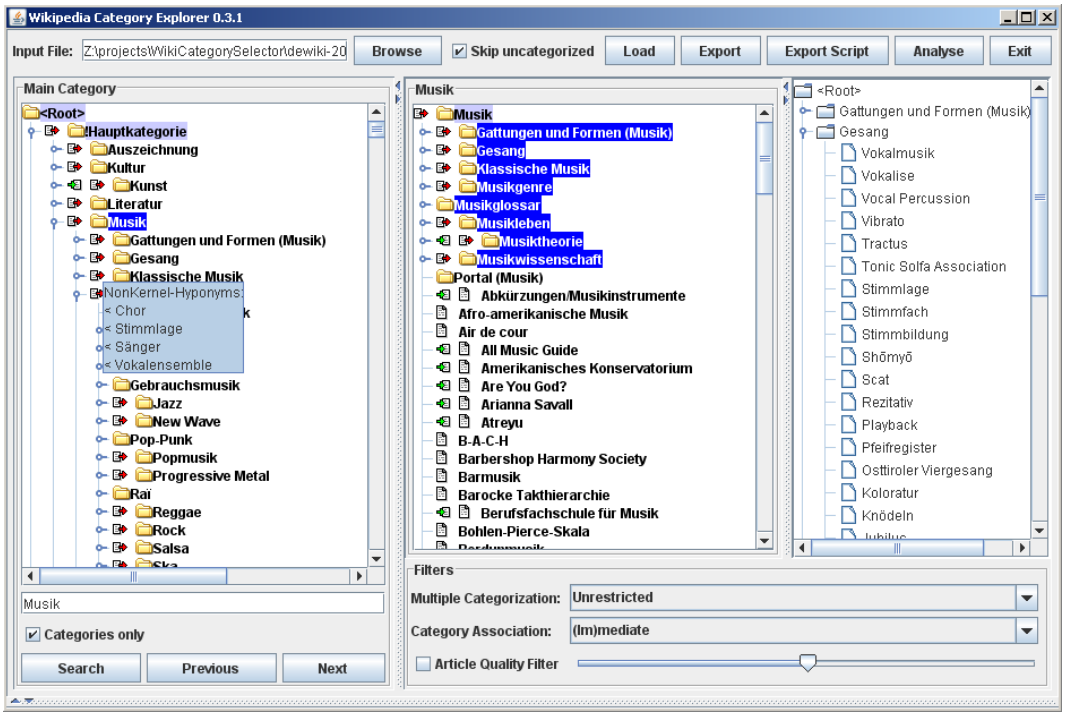


Figure 4: The wiki category viewer integrated into WikiDB by example of the category network of the German Wikipedia (Gleim et al., 2007).

means to filter out those hyperlinks which fall below the expected value of the corresponding similarity value. This approach tackles the LAP which affects the usability of wikis as resources of, e.g., lexical chaining (Mehler et al., 2007). It also helps to scale the Wikipedia category system in order to identify well-defined subtrees of the kernel hierarchy of the category graph from loosely spanned subgraphs. In a nutshell, semantic preprocessing is an indispensable step to augment the reliability of wiki-based knowledge resources. However, *pragmatic* preprocessing offers an additional reference point of judging the reliability of these resources. In the case of the category graph we may, for example, delete additions made by less reliable authors. The prerequisites of identifying such authors are explained in the next section.

2.3 Pragmatic Preprocessing

A central point missed in approaches to wiki-based knowledge resources concerns the process of their generation. We prevent this ignorance by including representations of the social network of those authors who have edited the wikis used to generate the final semantic database. More specifically, *pragmatically preprocessing* a given wiki concerns the mapping between its document units on the one hand and their authors on the other. We retain a bigger part of the information who has edited which segment of which page at which time. This information can be used to identify, for example, author communities who edit some thematically delimited subnetwork of the article graph. In this section we distinguish document-author relations, span collaboration networks based thereon and outline how to explore author communities.

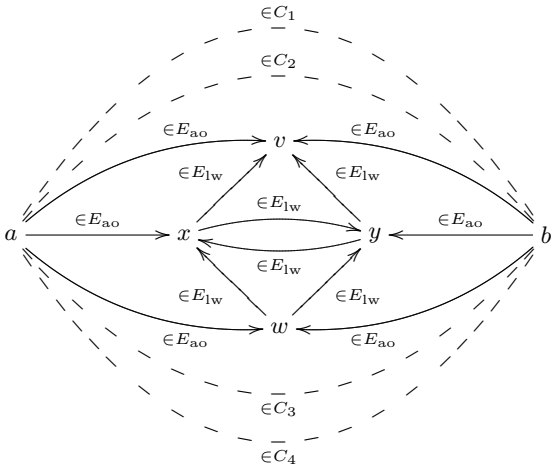


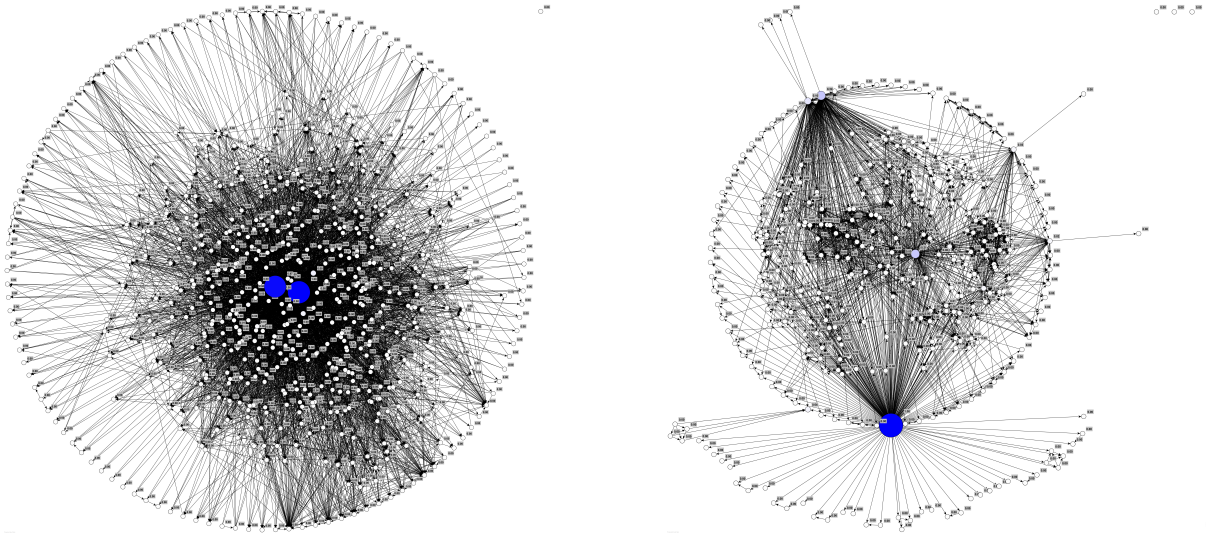
Figure 5: Fundamental authorship relations in wikis. For simplicity reasons we abstract from the direction of edges between authors $a, b \in V_1$ and therefore show undirected edges only. As edges between authors are implied from document relations they are visualized by dashed lines.

Authorship Relations and Collaboration Networks Wikis are well known for their history function as they store information about the editing process per article, portal, etc. (Wattenberg et al., 2004). That way, one can reconstruct which author has made which contribution to which page at which time. This information can be explored to study pragmatic aspects of wiki generation (Wattenberg et al., 2007). One reason might be to identify contributions of tendentious authors (e.g. tendentious editing) in order to exclude them from the final semantic database. A candidate source of identifying such types of authors which goes beyond single pages — and, thus, is of interest for our graph-theoretical approach — relates to collaboration networks whose vertices denote agents linked by a function of their commonly edited pages.

Based on the information which authors $a, b \in V_1$ have edited which pages $x, y \in V_2$ we can extract a variety of graphs which capture this information to varying degrees of explicitness. In this paragraph we give a short account of such networks as being derivable from wiki history pages. More specifically, we consider four such graphs. We do that by analogy to the bibliometric and webometric relations of (co-)*citation* (and bibliographic coupling) on the one hand and *sitation* on the other (cf. Fang and Rousseau, 2001; Rousseau, 1997; Björneborn, 2004). Let $E_{\mathrm{au}} \subseteq V_1 \times V_2$ be the *author of* relation defined over the set of authors V_1 and the set of wiki pages (or documents) V_2 and $E_{\mathrm{lw}} \subseteq V_2 \times V_2$ the *linked with* relation.[12] Starting from the setting of Figure 5 we call two documents $x, y \in V_2$ *co-siting* if they are commonly linked with a third document $v \in V_2$ and *co-referenced* if some $w \in V_2$ is linked with both of them. Further, if x, y are linked with each other they are called *mutually linked*. Finally, documents $v \in V_2$ for which $(a, v), (b, v) \in E_{\mathrm{au}}$ are called *co-authored* while their authors a, b are called *co-authoring*. Based on these distinctions we introduce four different collaboration graphs:

- *Co-authoring graphs* (V_1, C_1) are spanned by means of co-authored documents. That is, $\forall\{a, b\} \in$

[12] That is, $(a, x) \in E_{\mathrm{au}}$ iff agent a is author of page x, whereas $(x, y) \in E_{\mathrm{lw}}$ if there is a hyperlink starting from x and targeting to y. For more details on these and related relations in multilevel graphs spanned over agents, documents and lexical items used to generate these documents see Mehler (2008c).



(A) (B)

Figure 6: The article graph (A) together with its co-authorship graph (B) of the OpenOffice.org Wiki (`www.ooowiki.de`; download: January 2008).

$C_1 \exists v \in V_2 \colon (a,v),(b,v) \in E_{\text{au}}$. Mehler (2008c) shows that these networks have the small-world property (Newman, 2003).

- *Co-linkage author graphs* (V_1, C_2) are spanned by means of mutually linked documents. That is, $\forall \{a,b\} \in C_2 \exists x,y \in V_2 : (a,x),(b,y) \in E_{\text{au}} \wedge (x,y),(y,x) \in E_{\text{lw}}$. Graphs of this sort can be explored to identify groups of authors linking to each others' articles.
- *Co-sitation author graphs* (V_1, C_3) are spanned by means of co-siting documents. That is, $\forall \{a,b\} \in C_3 \exists x,y,v \in V_2 \colon (a,x),(b,y) \in E_{\text{au}} \wedge (x,v),(y,v) \in E_{\text{lw}}$. Graphs of this sort can be explored to identify groups of authors which tend to commonly link to the same reference articles, that is, authors who trust the same information sources.
- *Co-reference author graphs* (V_1, C_4) are spanned by means of co-referenced documents. That is, $\forall \{a,b\} \in C_4 \exists x,y,w \in V_2 : (a,x),(b,y) \in E_{\text{au}} \wedge (w,x),(w,y) \in E_{\text{lw}}$. Graphs of this sort can be explored to identify groups of authors referred to by the same referencing articles, that is, authors whose articles are trusted by the same information sources.
- Finally, we might consider *authorship graphs* $(V_1, \cap_{i=1}^{4} C_i)$. Note that maybe $C_i \cap C_j \neq \emptyset, i,j \in \{1,\ldots,4\}$.
- *Weighting and orientating edges:* A further source of building collaboration graphs relates to weighting their edges. In the case of co-authorship graphs this might be done, for example, by a function of the number of co-authored documents, the proportion of co-authored document segments etc. These and related information can be explored directly from wiki history pages. They might also be a source for spanning oriented authorship links, that is, arcs.

The graph-theoretical means to capture this sort of graphs are moderate: (un-)directed, weighted, labeled graphs (due to the membership of edges in C_1, C_2, C_3 or C_4). Thus, we can reuse our database model of document networks (cf. Section 2.1 and Section 3.1).

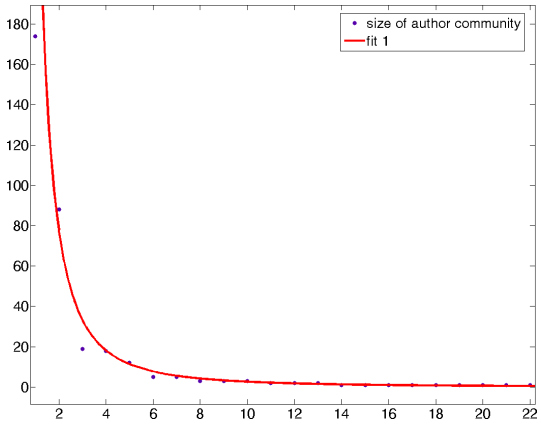


Figure 7: The distribution of the size of author communities within the weighted co-authoring graph of the `OpenOffice.org` Wiki.

Exploring Author Communities Roughly speaking, an author community is a group of authors who collaborate on writing some texts to a higher degree than members of different communities. This notion can be directly translated into cluster analysis where clusters are separated by a high degree of intra-cluster and a low degree of inter-cluster relatedness or similarity. As we do not deal with feature vector spaces but weighted graphs, we have to apply a graph cluster algorithm in order to detect subgraphs as community models. This can be done, for example, by means of the *Chinese Whispers Algorithm* (CWA) (Biemann, 2006). Figure 7 shows the distribution of the sizes of the author communities extracted by the CWA from the collaboration network of the `OpenOffice.org` wiki (cf. Figure 6). This distribution is Zipfian as can be seen from fitting the power law $p_x = Cx^{-\gamma}$ with $\gamma = 2.105$. This model gives an adjusted coefficient of determination of $.9879$: obviously, a very good fit in support of a very skewed distribution of the size of author communities. This and related measurements get possible by preprocessing pragmatic structures and integrating them into WikiDB. It opens the door beyond commonplace usages of the Wikipedia as it allows to explore authorship relations as a reference point of knowledge extraction. Note that from a graph-theoretical point of view clusters induce heterogeneous relations which span hypergraphs. The next section explains the data model of WikiDB used to capture such hypergraphs.

3 The Physical Data Model: An API for WikiDB

WikiDB captures the logical document structure of wiki pages and the logical network structure spanned by them. That is, wikis are modeled as labeled typed graphs whose vertices denote ordered hierarchies of content objects where leafs are labeled by tokens and lemmata, respectively. This graph model is serialized by means of WikiDB which uses TEI P5 (Burnard, 2006) for mapping intra- and intertextual document structures of a wide range of complexity. In this section we explain the database model underlying WikiDB. More specifically, we introduce `Hydra` as a database management system in conjunction with an *Application Programming Interface* (API) — written in C++ and Java — which enables users to explore many different linguistic data structures including the LDS of wiki

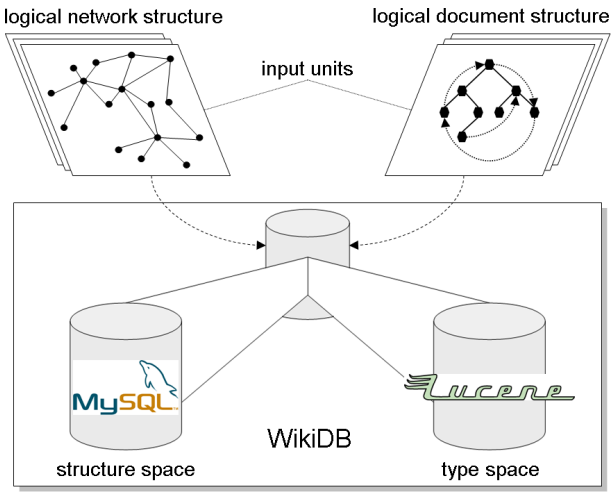


Figure 8: Managing the LNS of wikis, the LDS of their pages and the vocabulary of these pages by means of WikiDB which integrates a relational database and Lucene.

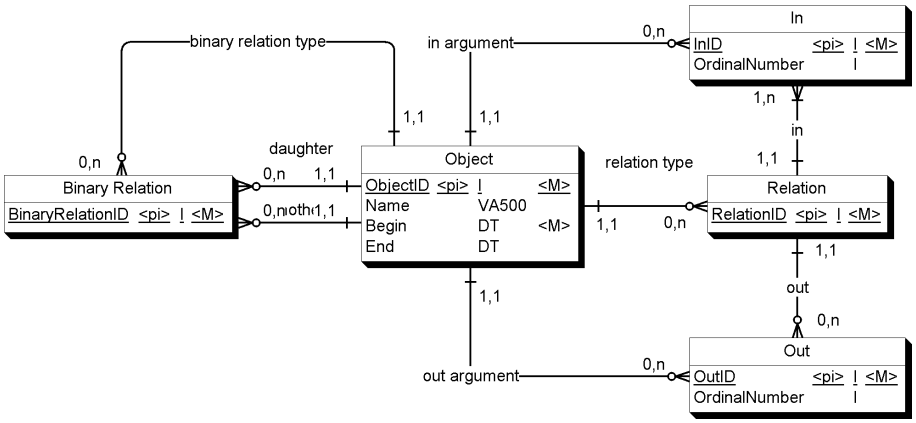


Figure 9: Outline of the conceptual model of WikiDB.

pages and the LNS of wiki document networks. A specific instantiation of `Hydra` which represents a certain wiki (e.g. a language-specific release of the Wikipedia) according to the three-level model of Section 2 is called a *Wiki DataBase* or WikiDB for short. The next section describes the conceptual data model underlying `Hydra` to capture intra- and intertextual document structures while Section 3.2 explains the API of `Hydra`.

3.1 Conceptual Database Model

In Section 2 we introduced a graph model for representing and further processing wiki-based data. In this section we complement this discussion by describing a conceptual data model used to capture these graphs by a relational database (see Figure 9). The central class of this data model is the class `Object` which represents graphs (whether directed or undirected) and their vertices. The different types of objects managed by this class as well as the membership relation of vertices to graphs and various types of arcs and edges are classified by `Object`, too. In other words: `Object` covers the set of

object structures as well as the ontology used to type these structures within the same conceptual unit. Next, directed and undirected edges are captured by the class `BinaryRelation` while hyperedges, that is, heterogeneous relations are mapped by the class `Relation`. Note that the distinction between `In` and `Out` vertices of hyperedges relates to the distinction between directed and undirected hyperedges (cf. Figure 9): While in the case of directed hyperedges both classes are instantiated, only `In` is instantiated in the case of undirected hyperedges (which denote, e.g., unordered clusters of authors as models of author communities — cf. Section 2.3). As all graph models described so far are mapped by these relational database constructs we can represent the LDS of wiki pages as well as the LNS of wiki document networks by the same database model. Note that the lexical organization of pages is not captured by this relational model but by an additional Lucene[13]-based index which is sensitive to linguistic annotations of texts. Consequently, WikiDB integrates a relational database as a model of document structures — the so called *structure space* — with a Lucene-based model of lexical structures — the so called *type space* (cf. Figure 8). As the syntactic preprocessing component generates XML code based on TEI P5, WikiDB integrates an API for managing and retrieving XML data broken down into heterogeneous (i.e., relational and index-based) resources. This API is described in the following section.

3.2 Application Programming Interface

Modeling linguistic resources by means of an XML-based language allows for an intuitive and document oriented representation. In the area of text representation the Text Encoding Initiative's TEI[14] format as well as the XCES[15] standard have earned acceptance in the community (Ide and Suderman, 2004; Rahtz, 2003). With XQuery a powerful query language to operate on such data is available. But despite the extensive use of XML in many applications a major drawback remains: Working on large XML documents is rather awkward. Today there is quite a number of XML Database Management Systems (DBMS) available. However, all suffer from the problem of inefficiency when it comes to querying big documents. In such cases XML is nonetheless very useful in order to have a formal description of a linguistic resource. But in order to actually work on the data another form of representation is needed which better fits the requirements of efficient access. Often it is acceptable to extract only specific parts of the contained information and then insert and query them in a relational database. However, this means a loss of information or — when keeping the original XML document — at least asynchronous data management. Finding a balance between a good performing representation and an accurate one which covers the expressiveness of the original XML document is a hard task.

We propose `Hydra`, a system which allows the lossless import and export of arbitrary XML documents. Note that WikiDB is a specific application of `Hydra` to the area of wiki-based data. Its strengths are efficient "browsing" through the document structure, quick lookups of simple XLinks and fast exports of arbitrary sub trees. Beside these core functionalities further extensions are available. Out of the box schemas are constrained to represent a specific class of documents. This includes the degree to

[13] Cf. `http://lucene.apache.org/`.
[14] Cf. `http://www.tei-c.org/Guidelines/P5`.
[15] Cf. `http://www.xces.org`.

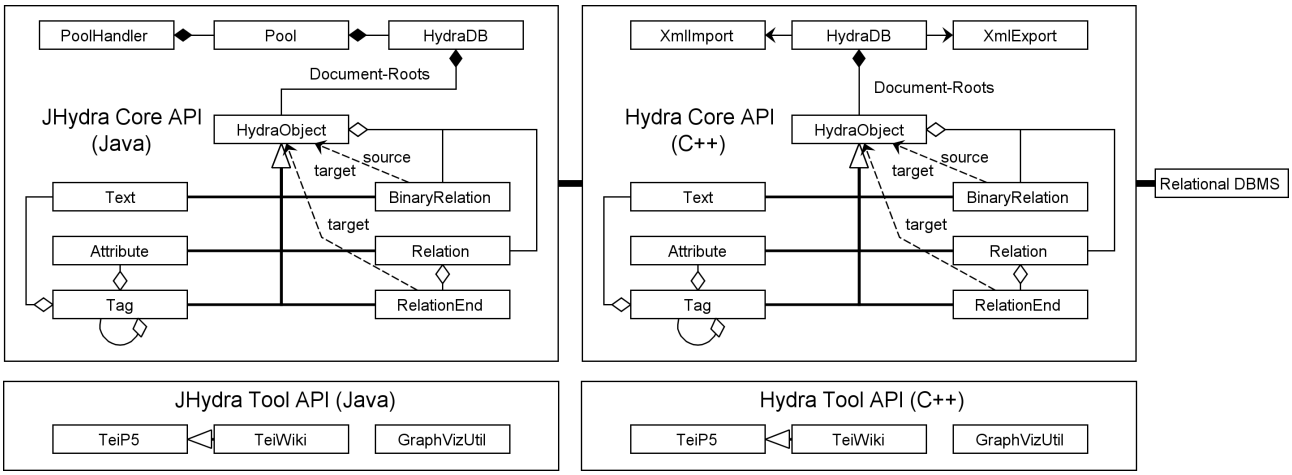


Figure 10: The architecture of `Hydra`.

which references between elements can be expressed. If additional references are to be annotated the underlying schema has to be extended — however this may not always be possible without rendering existing documents invalid. To overcome this problem `Hydra` offers — in addition to standard XML representation — the insertion of binary as well as n-ary relations between *any* XML element. The original XML data stays untouched but within the database the relations can be used for queries and further analysis.

Our approach is an optimized mapping of XML structures onto a relational database schema. We currently use MySQL, but the concept can be transferred to virtually any SQL-compliant DBMS. Figure 10 depicts the architecture of `Hydra` which can be grouped into three layers: The relational DBMS as backend, a C++ based API for managing, browsing and querying the XML representation and finally a Java API is offered which allows the C++ API-functionality to be used in Java-based client/server systems or applications. Both APIs share a common object-oriented class model to work on XML structures in the database — similar to the so called *Document Object Model* (DOM). The Java API is extended with a connection pool management to improve performance in concurrent data access.

The `Hydra` Core API is extended by the `Hydra` Tool API. It offers functions which are optimized for working on specific classes of XML documents — as for example TEI P5 documents. By means of this API, amongst others, the document network structure, the logical document structure and the coauthorship network can be extracted and directly accessed. These functions are used in the evaluation of the system in Section 4.

4 Processing and Retrieving Data from Wikis

In this section we finally present empirical data on extracting, annotating and retrieving data from three wikis based on the WikiDB. We consider the following wikis of varying size:

- the *Firefox Wiki* as an example of wiki-based technical documentation,
- the German release of the *Glottopedia* as an example of wiki-based knowledge communication and

Wiki Network	#Pages	#Links	Download	Preproc.	Import	Export	Select LNS	Collocate
Firefox Wiki	1,843	19,544	1338s	226s	111s	10s	3.1s	1.407ms
Glottopedia (German)	2,693	46,866	1712s	327s	185s	9s	4.6s	.379ms
Wiktionary (German)	100,555	3,582,382	274260s	11232s	10844s	272s	213.7s	11.412ms

Table 3: Estimating the time effort of using the special purpose WikiDB for typical database operations: *Download* (elapsed time for corresponding HTML-based web crawl), *Preprocessing* (time needed for tagging and preprocessing), *Import* (time of inserting the completely preprocessed wiki download and annotating the LNS in the database), *Export* (elapsed time for generating a database dump), *Select LNS* (time to select document graph and export into dot file) and *Collocate* (the time of generating a list of collocations for a single lexical item — averaged over 1,000 items).

Wiki Network	Tags Min	Tags Avg	Tags Max	Sel. Min	Sel. Avg	Sel. Max	Chars Min	Chars Avg	Chars Max
Firefox Wiki	46	95	2,684	110ms	176ms	2,660ms	4,039	10,450	303,031
Glottopedia (German)	59	110	1,550	78ms		1,135ms	5,231	10,413	237,216
Wiktionary (German)	57	175	11,273	81ms	124ms	6,451ms	5,069	16,038	1,252,357

Table 4: *Tags Min* (number of tags of the smallest article), *Tags Avg* (number of tags of the article whose size is most close to the average size of articles), *Tags Max* (number of tags of the largest article), *Sel. Min* (time to select the smallest article and to export its LDS into XML), *Sel. Avg* (time to select the median sized article and to export its LDS into XML), *Sel. Max* (time to select the largest article and to export its LDS into XML), *Chars Min* (character size of the smallest XML instance article), *Chars Avg* (character size of the average sized XML instance article), *Chars Max* (character size of the largest XML instance article).

- the German release of the *Wiktionary* as an example of wiki-based dictionaries.

The space and time effort for mapping these wikis are specified together with an estimation of handling wikis as a function of their size measured in the number of pages, links and authors. See Table 3, 4 and 5 for numeric results of analyzing these three special wikis. The numeric results reported in these tables give an impression of the complexity needed to manage and retrieve data by WikiDB. Note that because of the very rapid method of a complete database download, the whole data can be further processed by a specialized API which is independent from the underlying database management system. That way, many more optimizations are thinkable which help to boost the utilization of Wikipedia and special wikis as knowledge resources.

5 Conclusion

In this article we have described an API for exploring the logical document and the logical network structure of wikis (Section 2.1). It also provides an algorithm for the semantic preprocessing, filtering and typing of these building blocks (Section 2.2). Further, the article models the process of wiki generation based on a unified format of syntactic, semantic and pragmatic representations (Section 2.3). This three-level approach to make accessible syntactic, semantic and pragmatic aspects of wiki-based structure formation is complemented by a corresponding database model — called WikiDB —

Wiki Network	Articles	HTML Source	TEI P5 XML (incl. LDS)	TEI P5 XML (incl. LDS+PoS)	Database
Firefox Wiki	1,843	16.4MB	9.5MB	40.9MB	170.2MB
Glottopedia (German)	2,693	32.8MB	16.4MB	57.4MB	270.2MB
Wiktionary (German)	100,555	1989.7MB	861.4MB	2238.6MB	14064.2MB

Table 5: Space complexity of wiki representation in preprocessing and database representation: *Articles* (number of articles), *HTML Source* (size of the HTML document collection), *TEI P5 XML (incl. LDS)* (size of the TEI P5 representation including the LDS of the articles), *TEI P5 XML (incl. LDS+PoS)* (size of the TEI P5 representation including the LDS of the articles and their part of speech tagging), *Database* (size of the MySQL database files).

and an API operating thereon (Section 3). Finally, the article provides an empirical study of using the three-fold representation format in conjunction with the WikiDB (Section 4).

Acknowledgment

Financial support of the German Research Foundation (DFG) through the Excellence Cluster 277 *Cognitive Interaction Technology* (via the Project *Knowledge Enhanced Embodied Cognitive Interaction Technologies* (KnowCIT) — `http://ariadne.coli.uni-bielefeld.de/knowcit/`), the SFB 673 *Alignment in Communication* (via the Project X1 *Multimodal Alignment Corpora: Statistical Modeling and Information Management* — `http://ariadne.coli.uni-bielefeld.de/sfb/projects/X1/`), the Research Group 437 *Text Technological Information Modeling* (via the Project A4 *Induction of Document Grammar for Webgenre Representation* — `http://ariadne.coli.uni-bielefeld.de/indogram/`) and the LIS-Project *Entwicklung, Erprobung und Evaluation eines Softwaresystems von inhaltsorientierten P2P-Agenten für die thematische Strukturierung und Suchoptimierung in digitalen Bibliotheken* (`http://ariadne.coli.uni-bielefeld.de/agpib/`) at Bielefeld University is gratefully acknowledged.

References

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives (2008). DBpedia: A nucleus for a web of open data. pp. 722–735.

Biemann, C. (2006). Chinese whispers — an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of Graph-based Methods for Natural Language Processing (TextGraphs-1) at the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2006), Rochester, New York*.

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. Ph. D. thesis, Royal School of Library and Information Science, Department of Information Studies, Denmark.

Burnard, L. (2006). New tricks from an old dog: An overview of TEI P5. In L. Burnard, M. Dobreva, N. Fuhr, and A. Lüdeling (Eds.), *Digital Historical Corpora*, Volume 06491 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

Chernov, S., T. Iofciu, W. Nejdl, and X. Zhou (2006). Extracting semantic relationships between wikipedia categories. In *1st International Workshop: SemWiki2006 — From Wiki to Semantics (SemWiki 2006), co-located with ESWC 2006 in Budva, Montenegro, June 12*.

Dehmer, M., A. Mehler, and F. Emmert-Streib (2007). Generalized trees. In *Proceedings of the 2007 International Conference on Machine Learning: Models, Technologies & Applications (MLMTA'07), June 25-28, 2007, Las Vegas*.

Dellschaft, K. and S. Staab (2008). An epistemic dynamic model for tagging systems. In *HYPERTEXT 2008, Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, June 19-21, 2008, Pittsburgh, Pennsylvania, USA*.

Denoyer, L. and P. Gallinari (2006). The wikipedia xml corpus. *SIGIR Forum 40*(1), 64–69.

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards 71B*, 233–240.

Fang, Y. and R. Rousseau (2001). Lattices in citation networks: An investigation into the structure of citation graphs. *Scientometrics 50*(2), 273–287.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Gabrilovich, E. and S. Markovitch (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence, Boston, MA*.

Gabrilovich, E. and S. Markovitch (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), Hyderabad, India, January 6-12, 2007*, pp. 1606–1611.

Gleim, R., A. Mehler, M. Dehmer, and O. Pustylnikov (2007). Aisles through the category forest — utilising the wikipedia category system for corpus building in machine learning. In J. Filipe, J. Cordeiro, B. Encarnação, and V. Pedrosa (Eds.), *3rd International Conference on Web Information Systems and Technologies (WEBIST '07), March 3-6, 2007, Barcelona*, Barcelona, pp. 142–149.

Ide, N. and K. Suderman (2004). The american national corpus first release. In *Fourth international conference on language resources and evaluation (LREC 2004)*, pp. 1681–1684.

Kopp, S., N. C. Gesellensetter, L. Krämer, and I. Wachsmuth (2005). A conversational agent as museum guide – design and evaluation of a real-world application. In T. P. et al. (Ed.), *Intelligent Virtual Agents*, LNAI 3661, pp. 329–343. Berlin: Springer.

Kopp, S. and I. Wachsmuth (2004). Synthesizing multimodal utterances for conversational agents. *Journal of Computer Animation and Virtual Worlds 15*, 39–52.

Landauer, T. K. and S. T. Dumais (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics 19*(2), 313–330.

Mehler, A. (2006). Text linkage in the wiki medium – a comparative study. In J. Karlgren (Ed.), *Proceedings of the EACL Workshop on New Text – Wikis and blogs and other dynamic text sources, April 3-7, 2006, Trento, Italy*, pp. 1–8.

Mehler, A. (2008a). A graph model of social ontologies. In preparation.

Mehler, A. (2008b). On the impact of community structure on self-organizing lexical networks. In A. D. M. Smith, K. Smith, and R. Ferrer i Cancho (Eds.), *Proceedings of the 7th Evolution of Language Conference (Evolang 2008), March 11-15, 2008, Barcelona*, pp. 227–234. World Scientific.

Mehler, A. (2008c). Structural similarities of complex networks: A computational model by example of wiki graphs. *Applied Artificial Intelligence 22*.

Mehler, A., U. Waltinger, and A. Wegner (2007). A formal text representation model based on lexical chaining. In *Proceedings of the KI 2007 Workshop on Learning from Non-Vectorial Data (LNVD 2007) September 10, Osnabrück*, Osnabrück, pp. 17–26. Universität Osnabrück.

Milne, D., O. Medelyan, and I. H. Witten (2006). Mining domain-specific thesauri from wikipedia: A case study. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, pp. 442–448. IEEE Computer Society.

Muchnik, L., R. Itzhack, S. Solomon, and Y. Louzoun (2007). *Physical Review E 76*, 016106.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review 45*, 167–256.

Ponzetto, S. and M. Strube (2007, July). Deriving a large scale taxonomy from wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*, Vancouver, B.C., pp. 1440–1447.

Power, R., D. Scott, and N. Bouayad-Agha (2003). Document structure. *Computational Linguistics 29*(2), 211–260.

Rahtz, S. (2003). Building tei dtds and schemas on demand. In *Paper presented at XML Europe 2003, London, March 2003*, Taganrog, Russia.

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics 1*(1).

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading, Massachusetts: Addison Wesley.

Schaffert, S., D. Bischof, T. Bürger, A. Gruber, W. Hilzensauer, and S. Schaffert (2006). Learning with semantic wikis. In *Proceedings of SemWiki2006 Workshop "From Wiki to Semantics"*, pp. 109–123.

Steels, L. and P. Hanappe (2006). Interoperability through emergent semantics. a semiotic dynamics approach. *Journal on Data Semantics VI*, 143–167.

Stein, B. and S. Meyer zu Eißen (2007). Topic identification. *Künstliche Intelligenz (KI) 3*, 16–22.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, pp. 697–706. ACM.

Uszkoreit, H., T. Brants, D. Duchier, B. Krenn, L. Konieczny, S. Oepen, and W. Skut (1998). Studien zur performanzorientierten Linguistik. Aspekte der Relativsatzextraposition im Deutschen. *Kognitionswissenschaft 7*(3), 129–133.

Völkel, M., M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer (2006). Semantic wikipedia. In *Proceedings of the 15th International Conference on World Wide Web (WWW '06), Edinburgh, Scotland, May 23 – 26*, pp. 585–594. New York: ACM Press.

Voss, J. (2006). Collaborative thesaurus tagging the wikipedia way.

Waltinger, U. and A. Mehler (2008a). Web as preprocessed corpus: Building large annotated corpora from heterogeneous web document data. In preparation.

Waltinger, U. and A. Mehler (2008b). Who is it? context sensitive named entity and instance recognition by means of Wikipedia. Submitted.

Waltinger, U., A. Mehler, and G. Heyer (2008). Towards automatic content tagging: Enhanced web services in digital libraries using lexical chaining. In *Fourth International Conference on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal*. Barcelona.

Wattenberg, M., F. B. Viégas, and K. Dave (2004). Studying cooperation and conflict between authors with history flow visualization. In *Proceedings of the 2004 conference on Human factors in computing systems*, New York, pp. 575–582. ACM.

Wattenberg, M., F. B. Viégas, and K. J. Hollenbach (2007). Visualizing activity on wikipedia with chromograms. In M. C. C. Baranauskas, P. A. Palanque, J. Abascal, and S. D. J. Barbosa (Eds.), *Human-Computer Interaction — INTERACT 2007, 11th IFIP TC 13 International Conference, Rio de Janeiro, Brazil, September 10-14, 2007, Proceedings, Part II*, Volume 4663 of *Lecture Notes in Computer Science*, pp. 272–287. Springer.

Zesch, T., C. Müller, and I. Gurevych (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech (Morocco)*.

Ziemke, T. (1999). Rethinking grounding. In A. Riegler, M. Peschl, and A. von Stein (Eds.), *Understanding Representation in the Cognitive Sciences. Does Representation Need Reality?*, pp. 177–190. New York/Boston/Dordrecht: Kluwer/Plenum.