Semantic Wikipedia[*]

Heiko Haller, Markus Krötzsch, Max Völkel, Denny Vrandecic,
Institute AIFB/FZI, Universität Karlsruhe (TH)
76128 Karlsruhe, Germany
{hhaller,kroetzsch,voelkel,vrandecic}@aifb.uni-karlsruhe.de

ABSTRACT

Wikipedia is the world's largest collaboratively edited source of encyclopaedic knowledge. But its contents are barely machine-interpretable. Structural knowledge, e. g. about how concepts are interrelated, can neither be formally stated nor automatically processed. Also the wealth of numerical data is only available as plain text and thus can not be processed by its actual meaning.

We provide an extension to be integrated in Wikipedia, that allows even casual users the typing of links between articles and the specification of typed data inside the articles. Wiki users profit from more specific ways of searching and browsing. Each page has an RDF export, that gives direct access to the formalised knowledge. This allows applications to use Wikipedia as a background knowledge base.

Categories and Subject Descriptors

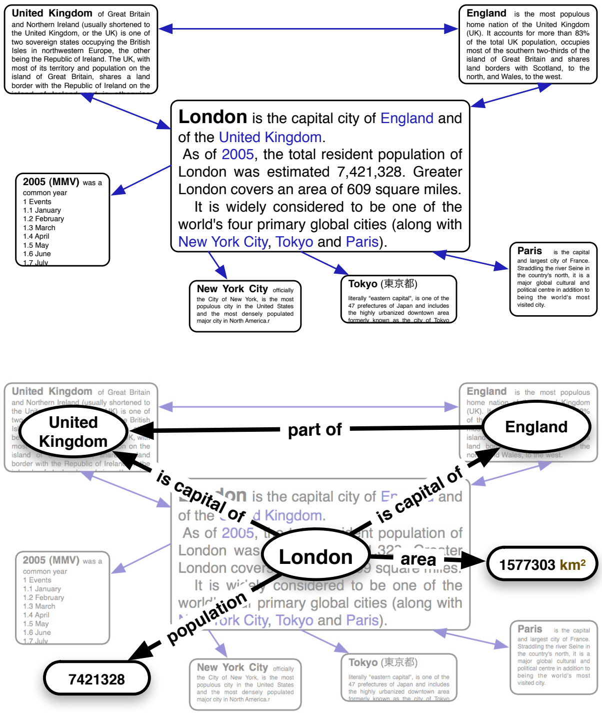
H.3.5 [**Information Storage and Retrieval**]: Online Information Systems; H.5.3 [**Information Interfaces**]: Group and Organization Interfaces—*Web-based interactions*; I.2.4 [**Artifical Intelligence**]: Knowledge Representation; K.4.3 [**Computers and Society**]: Organizational Impacts—*Computer-supported collaborative work*

General Terms

Human Factors, Documentation, Languages

1. INTRODUCTION

This paper describes an extension to be integrated in Wikipedia, that enhances it with Semantic Web [1] technologies. Wikipedia, the free encyclopaedia, is well-established as the world's largest online collection of encyclopaedic knowledge, also being an example of global, self-organising collaboration.



United Kingdom of Great Britain and Northern Ireland (usually shortened to the United Kingdom, or the UK) is one of two sovereign states occupying the British Isles in northwestern Europe, the other being the Republic of Ireland. The UK, with most of its territory and population on the island of Great Britain, shares a land border with the Republic of Ireland on the

England is the most populous home nation of the United Kingdom (UK). It accounts for more than 83% of the total UK population, occupies most of the southern two-thirds of the island of Great Britain and shares land borders with Scotland, to the north, and Wales, to the west.

London is the capital city of England and of the United Kingdom.
As of 2005, the total resident population of London was estimated 7,421,328. Greater London covers an area of 609 square miles.
It is widely considered to be one of the world's four primary global cities (along with New York City, Tokyo and Paris).

2005 (MMV) was a common year
1 Events
1.1 January
1.2 February
1.3 March
1.4 April
1.5 May
1.6 June
1.7 July

Paris is the capital and largest city of France. Straddling the river Seine in the country's north, it is a major global cultural and political centre in addition to being the world's most visited city.

New York City officially the City of New York, is the most populous city in the United States and the most densely populated major city in North America.

Tokyo (東京都) literally "eastern capital", is one of the 47 prefectures of Japan and includes the highly urbanized downtown area formerly known as the city of Tokyo

Figure 1: Currently there are pages and links (above), we feature concepts and data connected by relations (below).

Using Wikipedia currently means *reading* articles—There is no way to automatically gather information scattered across multiple articles, like "Give me a table of all movies from the 1960s with Italian directors". Although the data is quite structured (each movie on its own article, links to actors and directors), its meaning is unclear to the computer, because it is not represented in a machine-processable, i. e. formalised way.

To let the huge and highly motivated community of Wikipedians render the shared factual knowledge of Wikipedia machine-processable, we face several challenges: In addition to technical aspects of this endeavour, the main challenge is to introduce semantic technologies into the established usage patterns of Wikipedia. We propose small extensions to the wiki link syntax and an enhanced article view to show the interpreted semantic data to the user. Powerful inline queries turn parts of a page into a dynamically updated list or table. These queries have the potential to replace the many hand-crafted lists (e. g. cities in Europe).

We expose Wikipedia's fine-grained human edited information

[*] This research was partially supported by the European Commission under contract IST-2003-506826 "SEKT," FP6-027705 "NEPOMUK," and FP6-507482 "Knowledge Web Network of Excellence," and by the German BMBF project "SmartWeb."

WikiSym'06, August 21-23, 2006, Odense, Denmark.
ACM 1-59593-413-8/06/0008.

in a machine-readable way by using the W3C standards on RDF, XSD, RDFS, and OWL. This opens new ways to improve Wikipedia's capabilities for querying, aggregating, or exporting knowledge, based on well-established Semantic Web technologies. We hope that Semantic Wikipedia can help to demonstrate the promised value of semantic technologies to the general public.

The primary goal of this project is to supply an implemented extension to be actually introduced into Wikipedia in the near future. The implementation is rapidly developing, and the software can be tested online at `http://wiki.ontoworld.org`.

2. IDEA

Our primary goal is to provide an extension to MediaWiki which allows to make important parts of Wikipedia's knowledge machine-processable with as little effort as possible[3]. Since our system is conceived as an extension of MediaWiki it adheres to these core wiki principles—often referred to as the "wiki way" [2]—with all the advantages and disadvantages that this brings.

We designed the following key elements for our annotations:

- *categories*, which classify articles according to their content,
- *typed links*, which classify links between articles according to their meaning, and
- *attributes*, which specify simple properties related to the content of an article.

Categories already exist in Wikipedia, though they are mainly used to assist browsing. Typed links and attributes are novel features that are explained below and detailed in subsequent sections.

We restricted the annotations to have as their subject always the topic of the current page. Thus it is not possible to make statements about a topic elsewhere then on the topic's page. This helps e. g. to locate erroneous statements.

2.1 Relating Concepts with Typed Links

Typed links are obtained from normal links by slightly extending the way of creating a hyperlink between articles, as illustrated in Figure 1. As for the Web in general, links are arguably the most basic and also most relevant markup within a wiki, and their syntactic representation is ubiquitous in the source of any Wikipedia article. The introduction of typed links thus is a natural consequence of our goal of exploiting existing structural information. Through a minor, optional syntax extension, we allow wiki users to create (*freely*) *typed links*, which express a *relation* between two pages (or rather between their respective subjects).

In order to explicitly state that London is the capital of England, in the "London" article one just extends the existing link to `[[England]]` by writing `[[is capital of::England]]`. This states that a relation called "is capital of" holds between "London" and "England." Typed links stay true to the wiki-nature of Wikipedia: Every user can add an arbitrary type to a link or change it. Of course existing link types should be used wherever applicable, but a new type can also be created simply by using it in a link. To make improved searching and similar features most efficient, the community will have to settle down to re-use existing link types. As in the case of categories, we allow the creation of descriptive articles on link types to aid this process.

Note how typed links integrate seamlessly into current wiki usage. Semantic MediaWiki places semantic markup directly within the text to ensure that machine-readable data agrees with the human-readable data of the article. The notation we have chosen makes the extended link syntax largely self-explicatory.

In the Semantic Wikipedia, even very simple search algorithms would suffice to provide a precise answer to the question "What is the *capital of England*?" In contrast, the current text-driven search returns only a list of articles for the user to read through. Details on how the additional type information can be added in an unobtrusive and user-friendly way are given in the next section.

2.2 Data Values as Concept Attributes

Attributes provide another interesting source of machine readable data, which incorporates the great number of data values in the encyclopedia. Typically, such values are provided in the form of numbers, dates, coordinates, and the like. For example, one would like to obtain access to the population number of London. It should be clear that it is not desirable to solve this problem by creating a typed link to an article entitled "7421328" because this would create a unbearable amount of mostly useless number-pages whereas the textual title does not even capture the intended numeric meaning faithfully (e.g. the natural lexicographic order of titles does not correspond with the natural order of numbers). Therefore, we introduce an alternative markup for describing attribute values in various datatypes.

In order for such extensions to be used by editors, there must be new features that provide some form of *instant gratification*. Semantically enhanced search functions improve the possibilities of finding information within Wikipedia. Additionally, Wikipedia's machine-readable knowledge is made available for external use by providing an RDF export of each page. This enables the creation of additional tools to leverage Wikipedia contents and re-use it in other contexts. Thus, in addition to the traditional usage of Wikipedia, a new range of services is enabled inside and outside the encyclopaedia. Experience with earlier extensions, such as Wikipedia's category system, assures us that the benefits of said services will lead to a rapid introduction of typed links into Wikipedia.

2.3 Inline Queries

Semantic MediaWiki offers inline queries. In edit mode, the user can specify the query using a wiki-like syntax. In normal view-mode, the results of the query are displayed. The expressivity is less than SPARQL and the current implementation uses MySQL 4.1 queries, as we could not find a scalable, 100% open-soure (i. e. not Java) triple store with SPARQL and inferencing support. As an example, we show a query asking for all actors born in Boston: `<ask>[[Category:Actor]] [[born in::Boston]]</ask>`.

3. CONCLUSIONS AND OUTLOOK

We have demonstrated that the system provides many immediate benefits to Wikipedia's users, such that an extensive knowledge base might be built up very quickly. The emerging pool of machine accessible data presents great opportunities for developers of semantic technologies who seek to evaluate and employ their tools in a practical setting. In this way, Semantic Wikipedia can become a platform for technology transfer that is beneficial both to researchers and a large number of users worldwide, and that really makes semantic technologies part of the daily usage of the World Wide Web.

4. REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, (5), 2001.

[2] W. Cunningham and B. Leuf. *The Wiki Way. Quick Collaboration on the Web*. Addison-Wesley, 2001.

[3] M. Völkel et al. Semantic wikipedia. In *Proc. of the WWW 2006, Edinburgh, Scotland, May 23-26, 2006*, MAY 2006.