Transwiki:Wikimania05/Paper-IM1

From Meta
< Transwiki:Wikimania05

This page is part of the **Proceedings of Wikimania 2005**, Frankfurt, Germany.

0 MISSING **1** Submitted **2** Editing **3** Author review **4** Final edit **5** DONE

Editing notes:

- Extensive!

WikiOnt: An Ontology for Describing and Exchanging Wikipedia Articles

Contents

Stefan Decker

- **License:** GFDL
- **Slides**: PDF (http://sw.deri.org/~jbreslin/presentations/20050805a.pdf)
- **Video**:
- **Note:** Presentation, 15 minutes

About the author: Andreas Harth is currently studying for his PhD at the Digital Enterprise Research Institute (DERI), National University of Ireland, Galway (NUI Galway). His research interests include information integration on the Semantic Web and RDF storage and querying.

Hannes Gassert is studying Computer Science, Communication and Media Science at the University of Fribourg, Switzerland. He is co-founder of mediagonal Inc., Fribourg. He was a visiting scholar at DERI, NUI Galway in 2005.

Ina O'Murchu is also studying at DERI, NUI Galway. Her research interests include social networks and ontologies for localised communities.

Dr. John G. Breslin received his PhD at NUI Galway. He is a Postdoctoral Researcher at DERI, NUI Galway. His research interests include social networks and online communities. He is co-founder of Ireland's largest bulletin board community, boards.ie.

Dr. Stefan Decker received his PhD at the University of Karlsruhe, Germany. He is a Research Fellow, Adjunct Lecturer at DERI, NUI Galway. His research interests include the Semantic Web and P2P technologies.

Abstract

Ontologies are formal specifications of how to represent the entities in a specific area and the interrelations among them. In the Semantic Web, ontologies can be used to share and reuse knowledge via the Web and they can be seen as a means for knowledge management on a global scale. A specific wiki ontology can be built to integrate Wikipedia (and by extension other MediaWiki-based sites) into the Semantic Web framework and to make Wikipedia machine-processable and -understandable. Through the use of RDF (Resource Description Framework, a W3C recommendation) and URIs, Wikipedia content could be identified, described, linked and combined with other Semantic Web data sources.

Introduction

The free content online encyclopaedia Wikipedia contains approximately 1.5 million articles, more than 500,000 of which are published in English, receiving around 50 million hits a day. It has become one of the most important single knowledge sources on the Web. Wikipedia is currently used mainly by humans who search and browse through its HTML user interface optimised for on-screen display. Web crawlers try to work with this affluent body of content as well.

In contrast to web sites targeting online users, data offered in a machine-understandable format is free from any constraints: it can be processed, integrated, combined and mapped to different system and vocabularies with ease. In contrast to HTML, such data is much more useful to software than it is to humans, but it has the advantage of multiplying the potential of the information it encodes.

What is an Ontology?

At the moment, a documented database scheme is available for MediaWiki-based sites which is sub-optimal for information exchange across sites. Semi-structured data (RDF/XML) can be self-describing and can carry its schema and semantics implicitly in the data, facilitating data exchange and integration. The current data set in Wikipedia is not generally machine-processable, but making the data in Wikipedia machine-processable could open up Wikipedia to a broad range of use cases and data consuming agents. One of these could be the addition of Wikipedia articles to search results, a goal that the "big players" in the search engine game are aiming for as well.

One means of making Wikipedia machine-understandable is by creating a formal ontology. Ontologies are formal specifications of how to represent the entities in a specific domain as well as the various interrelations among them. In the Semantic Web, ontologies can be used to share and reuse knowledge via the Web and they can be seen as a means for knowledge management on a global scale. A specific Wikipedia ontology can be built to integrate Wikipedia into the Semantic Web framework and therefore to make Wikipedia machine-processable and -understandable. Through the use of RDF (Resource Description Framework, a W3C recommendation) and URIs, Wikipedia content could be identified, described, linked and combined with other data sources.

Uses of an Ontologised Wiki

Wikipedia URLs can be used to denote a subject of a document, or to annotate photos: in fact, Wikipedia URLs can become general URIs identifying concepts in the Semantic Web, enabling the Semantic Web community to leverage the structured

knowledge collected and maintained by the Wikipedia. In that sense, ontologising and "RDFising" Wikipedia can build a bridge between these two highly productive communities and allow for various sorts of "cross-pollination" between them.

RDF is a language for representing information about resources on the Web. So far, it has mainly been used for representing metadata about Web resources such as title, creator, and date, but as the border between data and metadata is blurring, expressing both the content and structure of an entire encyclopaedia becomes workable and desirable. RDF is particularly intended for software applications rather than being directly displayed to people, and provides a common graph-based data model so that information can be exchanged between applications without any loss of meaning.

People using a Wikipedia ontology could reuse the data in different application scenarios as people can have easy access to Wikipedia for various software programs through the use of an ontology which is extendable, non-proprietary and interoperable across the Internet.

Paper Layout

We propose to use an ontology (WikiOnt) to describe the schema of the Wikipedia / MediaWiki dataset using the Web Ontology Language (OWL). In this paper, we describe the main concepts and relations in our proposed ontology, derived from both the HTML rendering and the original relational data. The authors present methods detailing how to convert the instances into a format adhering to the ontology, alongside a PHP5 implementation of such a converter, relying heavily on regular expressions. After sketching out how the converted data can be integrated with other datasets such as WordNet or the world of FOAF (http://www.foaf-project.org) , we discuss our practical experiences and lessons learned in converting a large-scale interconnected knowledge base such as Wikipedia.

Describing the Ontology

In this section we present the Wikipedia Ontology (WikiOnt). The ontology can be used to share a common understanding of the structure of information within the Wikipedia and between its users. To be able to reuse Wikipedia knowledge, users should be able to write software programs to enable automated processing of the Wikipedia knowledge and the information that is available. The ontology is a machine-interpretable representation of Wikipedia and allows software programs to query the dataset and reuse the data.

The ontology has been created by reverse engineering both the structure of a typical Wikipedia page and the SQL database dump. Our proposed ontology is by no means complete, but tries to capture the essence of a Wikipedia article. We made a conscious design choice to keep the ontology as compact as possible so that users new to the Semantic Web are able to quickly grasp the ontology and are able to reuse the data in own applications.

The ontology comprises of classes and properties. The main classes we identified in Wikipedia are Article, Stub, Category, and Image. Both Stub and Category are subclasses of Article. Instances of classes are connected using properties. We defined some properties, but tried to reuse properties from already existing ontologies such as Dublin Core and SKOS where appropriate.

In the following, we describe the set of classes including the set of properties that can be used to connect instances of a given class. The classes and properties we have defined use the wiki: namespace at http://sw.deri.org/2005/04/wikipedia/wikiont.rdf (in OWL).

Classes

Article

Article is the central concept in our Wikipedia ontology and is defined as a document that holds a description of a concept or thing. An example for an article is http://en.wikipedia.org/wiki/Galway.

Category

A Category is an entry in a taxonomy of concepts. The category instances classify articles into a hierarchy or taxonomy of concepts. A Category is a subclass of an Article. http://en.wikipedia.org/wiki/Category:Cities_in_Ireland is an example for a category.

Stub

A Stub is a subclass of Article. A Stub is an underdeveloped Article or Category that needs further work.

Image

Image denotes a media object that is typically part of an Article. Figure 1 is an example of an image.

Figure 1: Photograph of Galway City

Properties

dc:title

A title of an Article. The title of the article about Galway is "Galway".

wiki:text

The text of an Article in wiki markup. This property contains all information that is needed to display an HTML page with the article content.

dc:creator

The URI of the user or the IP address of the person who last changed the Article. http://en.wikipedia.org/wiki/User:Bastique is an example for a value of a dc:creator property.

dc:date

Date of the last change of the Article, e.g. 30 June 2005 19:45.

wiki:internalLink

A link among two Wikipedia Articles. For example, http://en.wikipedia.org/wiki/Galway has an internal link to http://en.wikipedia.org/wiki/History_of_Galway.

wiki:externalLink

A link to an external web page. For example, the article about Galway links to the Galway Chamber of Commerce and Industry at http://www.galwaychamber.com/.

skos:subject

The skos:subject property is used to relate an Article to a Category. For example, the subject of http://en.wikipedia.org/wiki/Galway is both http://en.wikipedia.org/wiki/Category:Cities_in_Ireland and http://en.wikipedia.org/wiki/Category:Galway.

skos:narrower

Two Categories can be related to each other using this property. Consider our running example again: http://en.wikipedia.org/wiki/Category:Galway is a narrower category than http://en.wikipedia.org/wiki/Category:Cities_in_Ireland. The skos:narrower links form a hierarchy of concepts which can be browsed. For example, to reach the Article about Ireland, you can browse the path Geography-Europe-European countries-Ireland.

wiki:redirectsTo

Sometimes, multiple Articles appear that cover the same concept. To be able to express that one article is a redirect of another article, we use the wiki:redirectsTo property. E.g. International_Business_Machines redirects to IBM.

sioc:views

The sioc:views property is used to capture the number of page views for a given Article.

img:width, img:height

The width and height of an image.

wiki:contentType

The content type to denote the content type of the image.

Example Instance

While the ontology provides the schema or vocabulary, the main parts of Wikipedia are the actual instances that are described using the ontology.

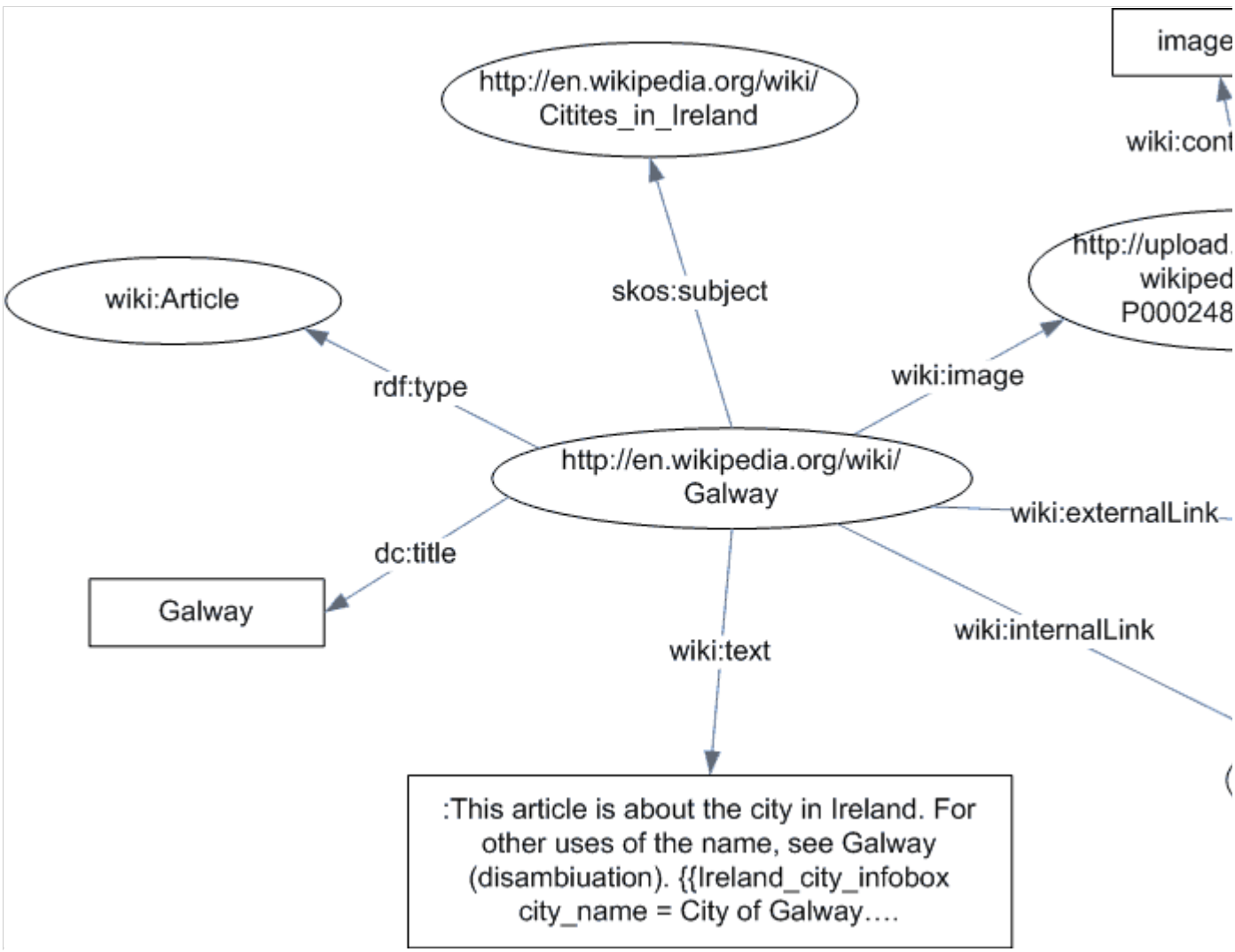

Figure 2: Graphical form of example instance

Figure 2 shows an example instance of the running example Galway in graphical form, and the following listing shows the corresponding RDF/N3 serialization.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix wiki: <http://sw.deri.org/2005/04/wikipedia/wikiont.owl#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<http://en.wikipedia.org/wiki/Galway> rdf:type wiki:Article ;
          dc:title "Galway" ;
          skos:subject <http://en.wikipedia.org/wiki/Cities_in_Ireland> ;
          wiki:image <http://upload.wikimedia.org/wikipedia/en/e/e3/Galway.jpg> ;
          wiki:externalLink <http://www.galwaychamber.com/> ;
          wiki:internalLink <http://en.wikipedia.org/wiki/Connacht> ;
          wiki:text ":This article is about the city in Ireland..." .

<http://upload.wikimedia.org/wikipedia/en/e/e3/Galway.jpg> wiki:contentType "image/jpg" .
```

Instance Conversion

Currently, the conversion of the original Wikipedia content into its RDF representation is carried out on the base of an imported MySQL dump of the two tables ``cur *and* ``*categorylinks*. The actual conversion is performed by a set of PHP scripts that can form the basis both for mass export as well as for web service-style query interfaces.

The PHP scripts split up the the extraction and conversion task into the following three components:

- *class WikipediaArticle*: Loads a wiki Article from the database and extracts information from the text, the wiki markup and the database fields. This class is responsible for all the extraction tasks involved, which are mainly performed by a series of pattern matches against the entry using Perl-style regular expressions. This step is entirely independent from RDF or a specific ontology.
- *class N3WikipediaArticle extends WikipediaArticle*: Based on the information extracted in the previous step, N3WikipediaArticle produces an RDF/N3 representation of an article. Using the capabilities of its parent class and the extracted metadata, N3WikipediaArticle forms RDF triples and serialises them using N3 syntax. The result obtained is a series of triples in the form "<subject> <predicate> <object>" describing the wiki article.
- *QueryRunner*: Initiates the conversion process, either for one term only, for a specific set of articles, or for the entire encyclopedia.

As expected, the Wikipedia dataset, while being extremely rich and interesting, poses challenges for both the tools and methods in use: the generated output had to be split up into a large number of individual files, one per article (thanks to RDF/N3, combination and concatenation of such files is trivial), and when trying to convert the entire dataset at once, PHP produces a segmentation fault.

The PHP scripts are released under the GNU GPL and to be found online at the web site of the DERI Semantic Web cluster (http://sw.deri.org/2005/04/wikipedia/) . The authors plan to make the entire metadata corpus available as well.

Our conversion is based on the 2005-06-23 version of the SQL dump of Wikipedia.

Exchange of Wiki Instances

In the following section we argue how the ontology can be used for data exchange between wiki sites.

Why is an Ontology Useful for the Wiki Community?

In order to interlink (http://wikifeatures.wiki.taoriver.net/moin.cgi/InterLink) wiki communities, some common data format is necessary for the import and export of wiki data and metadata. Apart from the Semantic Web direction, some work is being carried out to enable interconnections between wiki sites. #Schroeder2005 discusses the use of "near linking" and "near searching" to make pages on remote wikis as accessible as local pages.

However, since there are many varieties of wiki software, there is currently no standard format for exchanging data (e.g. remote pages) between wikis. Using the proposed ontology will enable the harvesting of information from remote wikis and integrating them in the local repository. Similarly, lists of "Recent Changes" (a popular feature on wiki sites, whereby the most recently changed pages are listed along with their modification dates) could be exchanged between wiki sites using the ontology.

How is Data Currently Exchanged?

The ontology can provide instances of data between wiki sites in a standardised format. At the moment, there is no standardised method of exchanging data between sites, either in terms of (A) the marked-up text (since markup tags can differ from wiki system to wiki system) or (B) the underlying data storage system (ranging from relational database systems to flat file storage).

(A) Wiki systems can use varying sets of markup to format text and link to other articles. A common method of exchanging wiki articles between wiki sites is to "copy and paste" the article from an editor interface in one site to that on another, and saving. However, such a duplication can only be useful if the same wiki markup language is used by both sites, as otherwise the copied article may lose formatting and links, or in extreme cases even break the target wiki system.

(B) Wikipedia (which uses the MediaWiki PHP/MySQL-based software system) provides regular database dumps of their entire database which can be imported by anyone who has installed the MediaWiki software. It is also possible to extract articles on a particular topic using supplementary tools or even by just "grepping" for SQL insert article statements that contain certain keywords or that belong to a certain category number. While such an import is technically feasible (but still not convenient) if everyone uses one particular wiki system, there are in fact many wiki systems for which this will not be possible.

Duplicating Articles Across Wiki Sites

Duplicating article or tables of contents across wiki sistes can thus be facilitated using features of the WikiOnt ontology. The copying of articles described in (A) and (B) above is useful, but may need to be repeated when content is changed on the source site. Providing timestamped instances of article data and methods of checking these timestamps against cached or imported data would allow the automatic update of articles originally composed on remote sites.

For example, let us say that site 2 has an article about "oranges" that has a sub-article about "satsumas" that has been imported from site 1. The sub-article is defined as being unchanged from its source version or not, and if unchanged, its timestamp on site 2 can be checked against that on site 1. If site 1 has a newer version, site 2 can be updated (if the article passes site 2's spam check). Even if an article is not duplicated regularly (and barring the infrastructure of a "pingback" system similar to blogs), a simple link to the original article can provide the most up-to-date version.

As articles are imported, local links on site 1 need to be converted to remote links from site 2 to site 1 (this is achieved by specifying whether an export of instance data is for local or remote use: local links in an article are then represented using the WikiOnt InternalLink or ExternalLink classes respectively). Similarly, when an article is imported to site 2, external links to site 2 from the original article on site 1 may need to be converted to local links in site 2.

How to Provide for the Exchange of Data

Since one of the original precepts of wiki systems was the free exchange of informational articles, much of the wiki software that exists has been developed under an open source or creative commons license. As such, it is possible to develop importers and exporters for WikiOnt instance data that can be freely distributed from

the corresponding wiki system download sites.

Another precept of wikis is that any page can be edited by anyone: while this is still true in the main (although private wikis with ACL-controlled areas or locked wiki pages are becoming more common), it would be preferable to restrict the amount of data a normal user could import and to permit administrators (termed "bureaucrats", "developers" or "sysops" on MediaWiki) full import privileges, thus avoiding spam or bias from a single user or organisation.

A RESTful Web Service Interface to Access Wikipedia

Based on the Wikipedia dataset in RDF, it is quite straightforward to provide a web service interface for the dataset. The approach we chose is to load the converted dataset into an RDF store, YARS in our case, and then pose queries via HTTP to retrieve the pieces of Wikipedia information we want. Because the data is already available in a machine-readable format, namely RDF, it is straightforward to read in and reuse the data in software programs.

We have provided an online service at http://sw.deri.org/2005/04/wikipedia/store that allows to retrieve and query Wikipedia articles via HTTP.

For example, to query for the wiki text of the article about Galway, you can pose the following query:

```
@prefix : <http://sw.deri.org/2004/06/yars#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ql: <http://www.w3.org/2004/12/ql#> .
@prefix wiki: <http://sw.deri.org/2005/04/wikipedia/wikiont.owl#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<> ql:select {
       (?text) .
}; ql:where {
       ?x rdf:type wiki:Article .
       ?x dc:title "Galway" .
       ?x wiki:text ?text .
} .
```

The default result set returned is RDF/N3. To change the result format to XML, an Accept: text/xml header should be sent with the HTTP request and the result will then be formatted in XML, which facilitates postprocessing with standard XML-based tools and infrastructure.

Wikipedia and the Rest of the Web

By converting Wikipedia into RDF it can be used semantically for searches, for categorising, for filtering, and for linking information.

In this section we explain how Wikipedia relates to other community-built datasets and discuss some methods on how to make associations between these different datasets.

The full potential of the Semantic Web can only be achieved when linking different datasets together and therefore creating associations between previously unrelated items.

Wikipedia is complementary to some traditional datasets, some community-built such as dmoz, some that are output of research efforts such as WordNet, some

government-funded projects such the CIA World Factbook, and some build on Web standards such as weblogs.

In this section we discuss possible connections between these datasets and Wikipedia and discuss how RDF and related Semantic Web technologies help to interoperate between these datasets.

Category Information in dmoz

The dmoz open directory project is a human-edited directory of the Web and is community-built by a large group of volunteer editors. At the core of dmoz is a very large taxonomy which is used to categorise a large number of web resources. The dmoz dataset is made available as an RDF dump and a number of sites reuse the dataset.

Wikipedia focuses on providing content, whereas dmoz tries to categorise existing pages. URLs of Wikipedia's external links and URLs of web resources categorised in dmoz may overlap. For example, http://www.galwaychamber.com/ occurs in both the Wikipedia page about Galway and in the dmoz category "Regional: Europe: Ireland: Galway: Business and Economy". Inferring connections based on URLs makes it possible to link Wikipedia articles to dmoz categories and vice versa.

Definitions in WordNet

WordNet is a collection of words in the English language, organised into synonym sets and linked together. WordNet is used in linguistics and cognitive sciences projects. The WordNet dataset is made available as Prolog dumps but RDF versions exist (http://www.semanticweb.org/library/ and http://www.w3.org/2001/sw/BestPractices/WNET/).

WordNet has definitions and synonyms for almost every word in the English language. Description of concepts are typically very brief in a dictionary-style format. For example, the entry for Galway is "a port city in western Ireland on Galway Bay". Providing brief information to virtually any topic can help to fill Stub pages in Wikipedia and might be helpful for people who want to lookup only a very concise definition of a concept.

Although there are various efforts to assign URIs to WordNet concepts and therefore make these concepts adressable, to date no dominant scheme for addressing WordNet concepts evolved.

Country Facts in CIA World Factbook

The CIA World Factbook provides structured information about the countries of the world, included geographical, political, historical, and statistical information. For example, the Wikipedia article about Ireland mainly talks about history, but lacks for example facts about Ireland's economy (e.g. GDP).

The CIA World Factbook is available in RDF and therefore could be combined with the RDF Wikipedia version to provide a more comprehensive coverage of country information.

Discussion

The previous sections showed examples of datasets that are complementary to Wikipedia and can provide additional and related information to Wikipedia articles.

A large number of datasets are currently being converted to RDF. By providing an RDF serialization of Wikipedia these other datasets can be combined and linked into Wikipedia, resulting in tighter linked pieces of information that are more valuable to the user.

Conclusion

This paper proposes an ontology (WikiOnt) to describe articles from the Wikipedia free encyclopaedia using the Web Ontology Language (OWL). This ontology will allow the production of wiki data and metadata using the Semantic Web format RDF. Making available Wikipedia metadata as RDF is key to enable more powerful datamining such multidimensional, semantic analyses. We described the main concepts and relations in our proposed ontology, and presented a PHP converter to transform existing dataset instances into an RDF representation adhering to the ontology. Both conversion scripts and a partially converted Wikipedia dataset are available online. The paper also discussed how the Wikipedia ontology and dataset relates to other freely available datasets and how associations between these datasets can be built to create a web of interrelated data sets.

Acknowledgements

This work was supported by Science Foundation Ireland under the DERI Lion project (SFI/02/CE1/I131).

References

- Schroeder, A., "Divide and Conquer: New Approaches to Scaling in Wiki Communities", Proceedings of the IADIS International Conference on Web-Based Communities 2005, pp. 177-182.
- Deborah L. McGuinness, Frank van Harmelen, "OWL Web Ontology Language Overview", W3C Recommendation 10 Feb 2004, http://www.w3.org/TR/owl-features/
- Frank Manola, Eric Miller, "RDF Primer", W3C Recommendation 10 Feb 2004, http://www.w3.org/TR/rdf-primer/