Applying Semantic Web Technology in a Digital Library

Ian Thurlow
BT Group
Adastral Park, Ipswich
United Kingdom
+44 (0) 1473 609585
ian.thurlow@bt.com

Alistair Duke
BT Group
Adastral Park, Ipswich
United Kingdom
+44 (0) 1473 608192
alistair.duke@bt.com

John Davies
BT Group
Adastral Park, Ipswich
United Kingdom
+44 (0) 1473 609583
john.nj.davies@bt.com

ABSTRACT

One of the key aims of the SEKT project is to develop and exploit the semantic technologies that will underlie the next generation of knowledge management systems. A key element of the project is to evaluate and assess the impact of semantic web technology in case study settings. The overall aim of the case study, described here, is to investigate how the semantic web technologies being researched and developed in the project can enhance the functionality of a digital library.

Keywords

Semantic web, digital libraries, case study, SEKT.

1. INTRODUCTION

Semantically enabled technology is expected to bring a number of benefits to the users of corporate digital libraries. In particular, the technology will help people find relevant information more efficiently and more effectively, give better access to that information, and aid the sharing of knowledge within the user community of a digital library.

Section 2.1 gives some background information on the BT digital library. The key requirements for a semantically enhanced digital library are summarised in section 2.2. An overview of the BT digital library architecture is given in section 2.3. The prototype system is described in section 2.4. An outline of the software demonstration is given in section 3.

2. BT's DIGITAL LIBRARY

2.1 The BT Digital Library Today

BT subscribes to approximately 1000 on-line publications, giving end-users access to the full-text of over 900,000 scientific and business articles and papers. In addition, access is provided to over 4 million bibliographic records from the Inspec[1] and ABI[2] databases. A proprietary keyword-based search engine is used to search these information sources. A limited set of advanced search options are provided for the specialist or expert user, e.g. search by author's name, search by title or search by controlled indexing terms. Alternatively, users can browse the contents and abstracts of the library's journals. A prototype knowledge sharing application enables users to annotate web pages of interest. Other users can search these annotations.

2.2 Requirements

An extensive requirements capture exercise identified the following key requirements:

i) large amounts of relevant content are accessible on the Web - the content of the digital library should therefore be extended to include relevant web pages and RSS items,
ii) the bibliographic records from ABI and Inspec should be integrated with data sourced from the web using a common ontology,
iii) bibliographic metadata should be enhanced with richer metadata, e.g. identify named entities within a text.
iv) better search precision is required,
v) users should be able to annotate and share web pages with other registered users of the library,
vi) new applications should be supported by profiles that describe user interests, e.g. to give context to a user's search, or enable relevant information to be pushed to users.

2.3 The BT Digital Library Architecture

The BT digital library case study is based on the 5-layer SEKT[3] architecture [1].

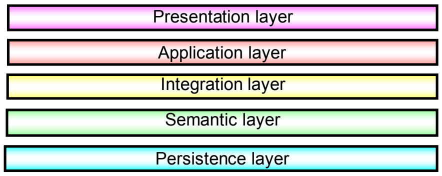

Figure 1. The BT digital library architecture.

The persistence layer comprises the internal sources of information, e.g. the Inspec and ABI bibliographic databases, and external sources of information, e.g. RSS items. The components that draw together relevant content for the digital library, e.g. the focused crawler, the components that populate the database, and the components that build profiles from an analysis of the log files are incorporated into this layer.

The semantic layer provides the components concerned with the creation, enhancement, maintenance, mediation, and querying of ontological information that is linked to the data stored in the persistence layer. Metadata associated with Inspec, ABI and RSS items is transformed into BT digital library ontology-specific

[1] http://www.iee.org/Publish/INSPEC/

[2] http://www.il.proquest.com/products/pt-product-ABI.shtml

[3] http://www.sekt-project.com/

Demos and Posters of the 3rd European Semantic Web Conference
(ESWC 2006), Budva, Montenegro, 11th - 14th June, 2006.

metadata. Named entities are extracted from texts by KIM[4], which employs GATE's[5] ontology-based information extraction module. The KAON2[6] system is used to store and reason over the resulting OWL[7]-based metadata.

The integration layer provides the infrastructure that enables the applications to be built from components in the semantic layer.

The principal case study applications, i.e. a semantic search and browse application, a semantic search agent, and a knowledge sharing application, are provided in the applications layer.

2.4 The BT Digital Library Prototype System

Squirrel, a tool for searching and browsing semantically annotated information, combines free text and semantic search with ontology-based browsing. Natural language summaries of ontological information are presented to the user. Search results are ranked, taking into account user profiles.

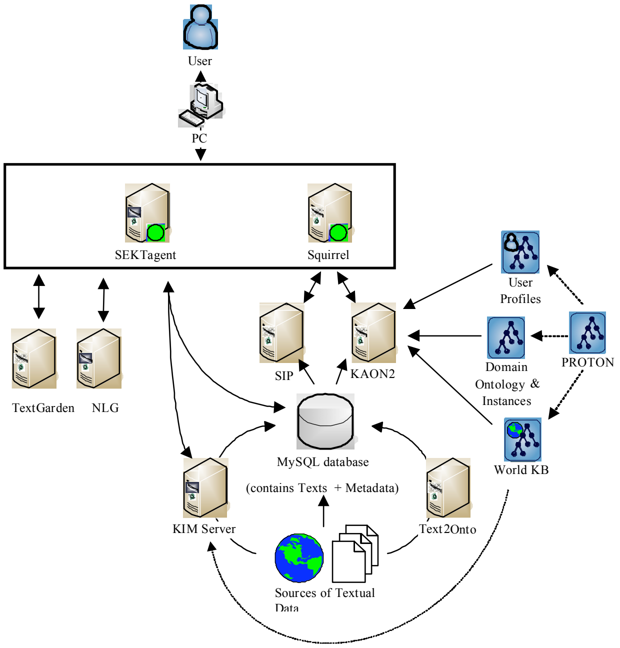

Figure 2. Squirrel and SEKTagent architecture.

SEKTagent, a semantic search agent facility, enables semantic queries to be specified, scheduled, and then invoked periodically over the digital library's pre-indexed documents. Relevant knowledge is delivered proactively to users. *Squidz*, a knowledge sharing application, enables a community of users to annotate Web pages of interest, share those annotations, and explore the interests and topics of other users. The tool builds upon current ideas of user tagging and community folksonomies and links user tags to a more formal ontology.

The applications are supported by a number of server side components that analyse textual documents and generate ontological information using the PROTON[8] ontology. A profile construction component, which is integrated with a web browser, enables profiles of users' interests to be constructed. A focused crawler enables relevant Web content to be added to the digital library. A classifier classifies Web content against topics in the BT digital library ontology (using vectors of co-occurring topics).

The BT digital library ontology is based on the PROTON ontology, which includes defines the top-level generic concepts required for semantic annotation, indexing and retrieval. This base ontology is extended with the additional classes and properties that are required to facilitate the SEKT-specific and case study-specific applications. The ontology has been populated with ABI and Inspec bibliographic data, along with Web content under a unified topic hierarchy. In addition, a world knowledge base (originally developed for the KIM platform) has been expressed in PROTON. This knowledge base comprises more than 200,000 entities, including around 36,000 locations, 140,000 companies and organisations, politicians, business leaders, technologists, etc.

The following components are used, either directly or indirectly, by the applications: PROTON, digital library extensions to PROTON, KIM, KAON2, semantic annotation, knowledge generation, knowledge repurposing, and the SEKT Integration Platform (SIP). Integration occurs at three levels: i) at the ontological level using a single overarching ontology on heterogeneous information sources, ii) at the component level using SIP (which allows SEKT technology components to be configured into data processing pipelines), and iii) at the application level, where applications are integrated into a portal.

3. SOFTWARE DEMONSTRATION

The demonstration is based on a typical usage scenario, i.e. a user views a set of *SEKTagent* results, configures a new search agent, navigates to the *Squirrel* tool, invokes some semantic queries and browses some meta-results, and finally uses *Squidz* to share a page with the digital library community.

4. ACKNOWLEDGMENTS

The work described in this paper was developed as part of the SEKT project, under EU funding (IST IP 2003-506826).

5. REFERENCES

[1] Semantic Web Technologies, Trends and Research in Ontology-based Systems, John Wiley & Sons, 2006.

[4] http://www.ontotext.com/kim/

[5] http://gate.ac.uk/

[6] http://kaon2.semanticweb.org/

[7] http://www.w3.org/TR/owl-features/

[8] http://proton.semanticweb.org/