Semantic information portals[±]

Dave Reynolds, Paul Shabajee[1], Steve Cayzer
Digital Media Systems Laboratory
HP Laboratories Bristol
HPL-2004-67
May 24, 2004*

E-mail: firstname_lastname@hp.com

semantic web

In this paper, we describe the notion of a *semantic information portal*. This is a community information portal that exploits the semantic web standards to improve structure, extensibility, customization and sustainability. We are in the process of developing a prototype directory of environmental organizations as a demonstration of the approach and outline the design challenges involved and the current status of the work.

* Internal Accession Date Only
[1]Institute for Learning Technology Research, University of Bristol, 8-10 Berkeley Sq. Bristol, UK
[±]WWW 2004, May 17-22, 2004, New York, NY USA
Approved for External Publication

Semantic information portals

Dave Reynolds
Hewlett-Packard Laboratories
Filton Road, Stoke Gifford
Bristol, UK
+44 117 3128165
dave.reynolds@hp.com

Paul Shabajee
ILRT, University of Bristol
8-10 Berkeley Sq
Bristol, UK
+44 117 928 7185
paul.shabajee@bristol.ac.uk

Steve Cayzer
Hewlett-Packard Laboratories
Filton Road, Stoke Gifford
Bristol, UK
+44 117 3127056
steve.cayzer@hp.com

ABSTRACT
In this paper, we describe the notion of a *semantic information portal*. This is a community information portal that exploits the semantic web standards to improve structure, extensibility, customization and sustainability. We are in the process of developing a prototype directory of environmental organizations as a demonstration of the approach and outline the design challenges involved and the current status of the work.

Categories and Subject Descriptors
H.4.0 [**Information Systems applications**]: General

General Terms
Design, Experimentation.

Keywords
Semantic Web, Information Portals.

1. INTRODUCTION
Web-based information portals provide a point of access onto an integrated and structured body of information about some domain. They range from very broad domains (e.g. all web pages [1]) to topic-specific domains (e.g. mathematics [2] and fish species [3]).

Community information portals are information portals, which are also designed to support and facilitate the activities of a community of interest. They typically allow members of the community to contribute news and information to the pool, either by submitting information directly to the portal (via some editing or reviewing process) or by posting the information on some associated web bulletin board or other collaboration tool.

The semantic web standards [4,5] enable new approaches to the design of such portals. In particular, they offer standards for how information in portals can be represented. RDF [4] provides a flexible and extensible format for information items and associated metadata; OWL [5] supports explicit representation of the domain ontologies used to classify and structure the items. Together these enable a more decentralized approach to portal architectures, as we discuss in the next section.

2. SEMANTIC INFORMATION PORTALS
There are several advantages to using semantic web standards for information portal design. These are summarised in Table 1, and specific aspects are discussed below.

Table 1. Comparison of design approaches

"Traditional" design approach	**Semantic portal**
Search by free text and stable classification hierarchy.	Multidimensional search by means of rich domain ontology.
Information organized by structured records, encourages top-down design and centralized maintenance.	Information semi-structured and extensible, allows for bottom-up evolution and decentralized updates.
Community can add information and annotations within the defined portal structure.	Communities can add new classification and organizational schemas and extend the information structure
Portal content is stored and managed centrally.	Portal content is stored and managed by a decentralized web of supplying organizations and individuals. Multiple aggregations and views of the same data is possible.
Providers supply data through portal-specific forms. Each portal is supplied and maintained separately.	Providers publish data in reusable form for incorporation in multiple portals. Updates remain under their control.
Portal aimed at human access. Separate mechanisms needed for sharing content with a partner organization.	Information structure is directly machine accessible to facilitate cross-portal integration.

Ontologies: The use of an explicit, shared domain ontology enables multidimensional classification and browsing schemes. A standard format for ontology encoding also facilitates reuse. Several projects have already derived benefits from ontology-driven portal designs [6,7].

Evolution: Requirements change over time leading to extensions to the information model. The semantic web helps in two ways. Firstly, the user interface and submission tools can be generated from the declarative ontology. Secondly, the semi-structured data representation of RDF permits data to be added in a new format, without invalidating existing data, in such a way that both original and extended formats can be used interchangeably.

This suggests an alternative approach to information portal design. Instead a long top-down design cycle, we start from a seed ontology and information structure that we extend incrementally.

Community extensions: Whilst many portals support constrained community annotations, such as comments and ratings, the semantic web approach allows more extensive community customization. For example, during work on a portal for wildlife

WWW 2004, May 17-22, 2004, New York, NY USA.
ACM xxx.

multimedia it became clear that many user communities would like specialized navigation of the data (based on formal species taxonomy or behavior depicted), which is infeasible for the centralized portal. Using the decentralized approach it is possible for communities to develop these specialist navigation structures as a set of external RDF annotations on the portal data. The central site can then aggregate the community-provided enrichments.

Aggregation and decentralization: One problem with traditional information portals is that they are often dependent on the responsiveness of the central maintainers, so that if funding disappears, so may the data. In the semantic web approach supplying groups host their own data and the portal becomes an aggregating service. Central organization is still needed (for example, to provide the initial impetus and ensure that appropriate ontologies and controlled vocabularies are adopted). However, once the system reaches a critical mass it can more easily be self-sustaining - anyone can run an aggregator service and to ensure continued access to the data or a new supplier can add data to the pool without a central organization being a bottleneck.

3. APPLICATION EXAMPLE: Directory of Environmental organizations

As part of an EU-funded project, SWAD-E [9], we are putting these ideas into practice by the development of a directory of UK environmental organizations. The idea is that each organization would provide their organization description as RDF data, using a web-based data entry tool, and would then host the data at their own web site (similar in style to FOAF [8]). A portal will aggregate the RDF data and provides a faceted browse interface.

Annotations to this data can be created by third parties and hosted by the suppliers or by an annotation server. These annotations will permit new classification schemes and relational links to be added to the data. In particular, the ability to add new links is seen as opening up exciting opportunities to capture and visualize the complex relationships between environmental organizations.

3.1 Design Issues

The design of this information portal has thrown up a number of challenges with wider implications for semantic web applications:

Moderation and access control: The decentralized portal design enables an interesting security model. In the test implementation the aggregator will have a record of which source URL's are deemed to be authoritative for a given organization. Each organization can then impose its own access and validation rules governing the update of that data. Some central administration is needed to moderate this "white list" of acceptable information sources. A semantic web crawler approach, which supports dynamic addition of new sources, is one possible approach but does not in itself address the problems of discovering "unsuitable" material.

Navigation: The rich classification of portal items is only useful if the interface complexity is kept under control. Our current experience suggests that a faceted browse approach modeled after the Flamenco project [12] offers a good balance between expressiveness and simplicity.

Provenance: The ability to mix community extensions and annotations with organizations' own data is a powerful feature of the approach. However, it is important that when a user is navigating the site they are able to clearly separate authoritative data from third party data, and in the latter case find where it came from in order to decide how much to trust it. This raises design issues for efficient recording of provenance, trust model issues (delegation and so forth) but also user interface issues of how to make the provenance of items clear.

Open-ended data model: We wish to support the open-ended nature of the RDF data model so that new properties and classes (whether authoritative or third party) can be incrementally added. The visualization engine needs to adapt to such changes without requiring new rendering templates to be created at each stage.

4. STATUS AND CONCLUSIONS

An early prototype of the environmental directory is being developed using existing organization databases and building on the Jena framework [10] and our semantic blogging tools [11].

Our architecture includes interesting features. We use a template-driven rendering approach that addresses the structured display of open-ended data models by using both the type and property lattices to guide template selection and layout. Many extensions require no template modification. We represent the navigation terms using the SWAD-E SKOS thesaurus proposals [9] and use the Jena rule engines to provide the required transitive closure and other inferences over that representation.

The aim of this demonstrator is to show practical applications of all of the core aspects of the semantic web (decentralization, ontologies, semi-structured data) working together. In contrast, prior projects such as [6][7] tend to exploit, for example, ontologies while retaining a centralized, top down paradigm.

5. REFERENCES

[1] The Open Directory project, http://www.dmoz.com/

[2] The Math-Net initiative, http://www.math-net.de/

[3] FishBase - a global information system on fishes, http://www.fishbase.org/

[4] Klyne, G., and Carroll, J.J. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Proposed Recommendation 15 December 2003.

[5] Dean, M. and Schreiber, G. OWL Web Ontology Language Reference. W3C Proposed Recommendation 15 Dec 2003.

[6] Maedche, A., Staab, S., Stojanoic, N., Studer, R., Sure, Y., Semantic portal - The SEAL approach, in Creating the Semantic Web. Fensel, D et al (eds.) MIT Press, MA, Cambridge, 2001.

[7] Kavournarakis, G., Christophides, V., Plexousakis, D., and Alexaki, S., Querying Community Web Portals (2000) http://citeseer.nj.nec.com/karvounarakis00querying.html

[8] The Friend of a Friend (FOAF) project. http://www.foaf-project.org/

[9] SWAD-Europe Project. http://www.w3.org/2001/sw/Europe/

[10] Jena 2 - A Semantic Web Framework http://www.hpl.hp.com/semweb/jena.htm

[11] Semantic Blogging for Bibliographies. http://www-uk.hpl.hp.com/people/steve_cayzer/semblog.htm

[12] Flamenco Project. http://bailando.sims.berkeley.edu/flamenco.html