An ontology-based system for interactive exploration of information sources

Jeroen Wester
jeroen@aduna-software.com

Christiaan Fluit
chris@aduna-software.com

Jeen Broekstra
jeen@aduna-software.com

Herko ter Horst
herko@aduna-software.com

Aduna
Prinses Julianaplein 14-B
3817 CS Amersfoort
The Netherlands
+31 (0)33 465 99 87
http://www.aduna-software.com/

Arjohn Kampman
arjohn@aduna-software.com

ABSTRACT

We outline an ontology-based architecture on three levels that enables users to explore information sources. On the data level Aperture is a Java framework for extraction of data and metadata. On the model level extracted metadata is stored in an RDF framework called Sesame. On the presentation level a faceted navigation engine called Spectacle and Cluster Map visualization enable ontology-based interaction with the data.

Keywords

Sesame, Aperture, Spectacle, Cluster Map, facet navigation, information visualization, browsing.

1.INTRODUCTION

This poster presents an ontology-based architecture that enables users to explore information sources (Figure 1). The architecture has three levels: data, model and presentation. The core of the system is a RDF database for storage and querying of information sources. On the data level, data and metadata are extracted from information sources. Information visualization and faceted navigation components, on the presentation level, enable the user to interact with the data and metadata. In the following sections, we briefly introduce each level and the software components that enable it in our proposed architecture.

2.DATA LEVEL

Aperture 6 is a Java framework for extracting and querying full-text content and metadata from various information systems (e.g. file systems, web sites, mail boxes) and the file formats (e.g. documents, images) occurring in these systems. The framework enables crawling heterogeneous information sources and extracting metadata from these information sources. Aperture makes use of RDF graphs to communicate information between components. Aperture uses a dedicated RDF vocabulary for describing properties of documents, files, and e-mails that is a specialization/extension of the Dublin Core metadata vocabulary.

3.MODEL LEVEL

At the model level, information retrieved by Aperture's metadata extractors is stored and made available to higher-level, presentation components, using a Sesame RDF repository 6. The Sesame framework provides the system with a flexible, scalable way of storing and manipulating large RDF models, and can provide reasoning support for RDF Schema- or OWL-based ontologies.

Using Sesame's support for declarative querying in SeRQL 6 or SPARQL 6, information expressed in Aperture's own RDF vocabulary can be mapped or transformed to other vocabularies.

In the case of the current architecture, SeRQL querying is used to map the information to a vocabulary that is especially suited for use by the Spectacle faceted navigation engine.

4.PRESENTATION LEVEL

At the presentation level two components, a faceted navigation engine called Spectacle 6 and an information visualization component called the Cluster Map library 6, enable the user to interact with the retrieved data and metadata by querying the RDF database.

The Spectacle engine allows users to navigate an information space by progressively selecting desired facet values of information objects.

The Cluster Map is an information visualization technique for sets of classified objects. Its main purpose is to show if and how these sets overlap, very similar in nature to Venn diagrams and Euler diagrams.

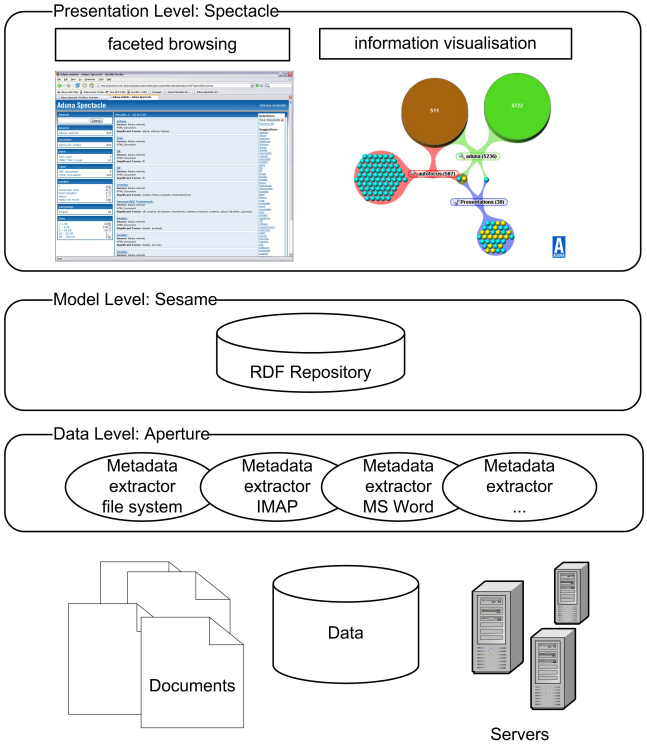

Figure 1: overview of the system architecture

5.THE ROLE OF ONTOLOGY

The ontology-based nature of this architecture is based on the following choices:

a. The crawling of information sources on the data level is ontology-driven. This means that Aperture will crawl for metadata that is part of a defined vocabulary. For example, it will crawl Dublin Core-like properties such as title and author and expose that information as an RDF graph.

b. The interaction on the presentation level is ontology-driven. This means that users interact with the model and the data while searching, browsing or exploring. For example, the Spectacle interface can present the user a 'Document type' facet that corresponds to the class 'Document Type' in the underlying ontology, listing known instantiations of that class as navigation steps.

6.REFERENCES

[1] Seaborne, S. and Prud'hommeaux, E. *SPARQL Query Language for RDF*. W3C Candidate Recommendation, April 6 2006. See http://www.w3.org/TR/rdf-sparql-query/.

[2] Becket, D. and Broekstra, J. *SPARQL Query Results XML Format*. W3C Candidate Recommendation, April 6 2006. See http://www.w3.org/2001/sw/DataAccess/rf1/.

[3] Broekstra, J., Kampman, A. and van Harmelen, F. Sesame: An Architecture for Storing and Querying RDF and RDF Schema. In *Proceedings of the First International Semantic Web Conference (ISWC 2002)*, Sardinia, Italy, June 9-12 2002, p. 54-68. Springer-Verlag Lecture Notes in Computer Science (LNCS) no. 2342. See also http://www.openrdf.org/.

[4] Broekstra, J.: SeRQL: A Second-Generation RDF Query Language. Chapter 4 in *Storage, Querying and Inferencing for Semantic Web Languages*. PhD Thesis, Vrije Universiteit Amsterdam (July 2005). ISBN 90-9019-236-0. See also http://www.openrdf.org/doc/SeRQLmanual.html.

[5] The Aperture Framework. See http://aperture.sourceforge.net/

[6] The Aduna Cluster Map Library. See http://aduna.biz/products/technology/clustermap/index.html.

[7] Aduna Spectacle Manual. Technical Report. See http://aduna.biz/products/spectacle/docs/Aduna-Spectacle-2005.1-manual.pdf.