RODIN – An E-Science Tool for Managing Information in the Web of Documents and the Web of Knowledge

Javier Belmonte

Haute Ecole de Gestion 7, rte. de Drize, CH-1227 Carouge, Switzerland. javier.belmonte@hesge.ch

Eliane Blumer

Haute Ecole de Gestion 7, rte. de Drize, CH-1227 Carouge, Switzerland. eliane.blumer@hesge.ch

Fabio Ricci

Haute Ecole de Gestion 7, rte. de Drize, CH-1227 Carouge, Switzerland. fabio.ricci@hesge.ch

René Schneider

Haute Ecole de Gestion 7, rte. de Drize, CH-1227 Carouge, Switzerland. rene.schneider@hesge.ch

Abstract: *RODIN is a tool for user-defined federated search and the simultaneous exploration of the web of documents and the Semantic Web. The system combines a widget aggregation approach for general web resources with an ontology matching approach for Linked Open Data. The project is part of the E-lib.ch-project (www.e-lib.ch), the Swiss initiative for building a single-point-of-access digital library for Switzerland. Within this context, RODIN was basically designed as an innovative and alternative information portal approach for digital libraries that neither depends on indexing as do common search engines nor relies on harvesting approaches as many library information systems do.*

Keywords: *Digital libraries, information architecture and web design for e-science, semantic information management, data driven e-science*

Introduction

RODIN (ROue D'INformation, i.e. information wheel in French) is a system for the user-defined search in preferred web resources and the simultaneous exploration of large-scale ontologies from the Linked Open Data project to allow and facilitate query refinement. RODIN may be of use in any other search intensive environment that goes beyond the simple search paradigm of web engines and that has to make use of heterogeneous data sources and needs ontologies to explore them further.

The implemented system consists of an aggregator facilitating keyword based search and an ontology driven exploration tool, with two ontologies from the Linked Open Data Project connected to the system so far: a) STW (Neubert, 2009), the standard thesaurus for economics and b) DBPedia (Auer et al., 2008), the semantic web offspring of Wikipedia. STW was chosen to prove the feasibility of the system in an economic project; DBPedia was chosen to add encyclopedic and rather open-domain world knowledge to the closed domain of economics. STW contains links to DBPedia, but that interconnectivity is not yet used by the system; it might nevertheless be useful to resolve questions in automatic disambiguation in a later version of RODIN.

Background

RODIN is driven by two main ideas: that of personal knowledge management systems and that of search refinement using simultaneously the web of documents and the web of knowledge and combining these paradigms in a single user interface by using open source software modules and open data sources and developing the appropriate software to bring them together.

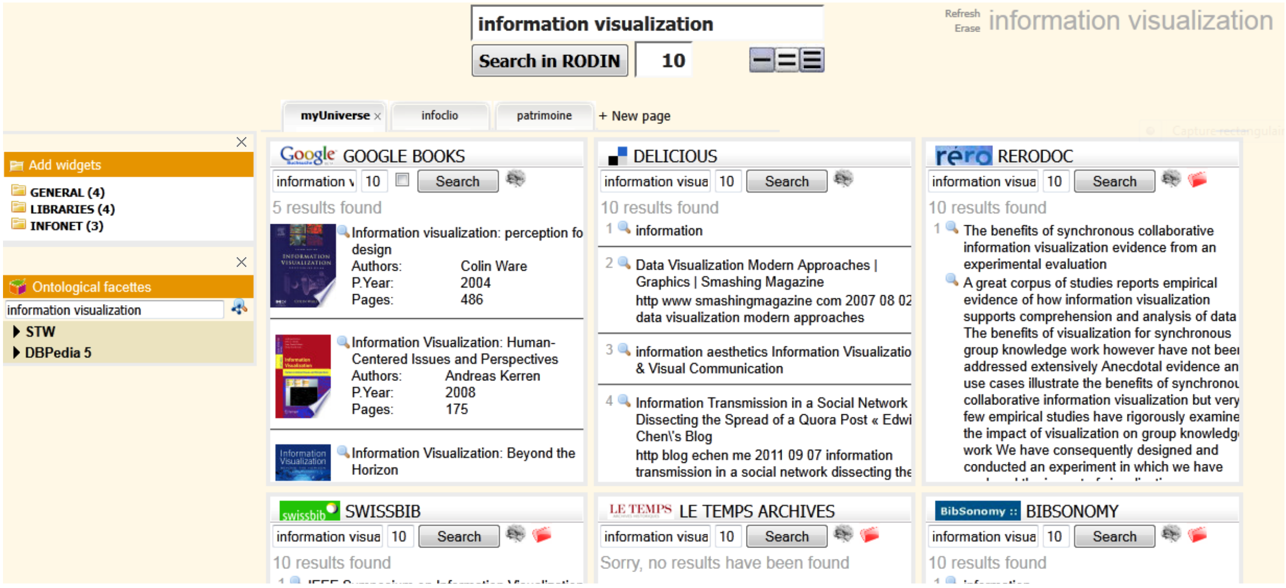
The general idea behind RODIN is based on the hypothesis that – in scientific or any search done by experts – the user generally makes use of a limited number of resources that are visited regularly: web engines, digital libraries, catalogues etc. and that can be aggregated or mashed (Hoyer, Stanoevska-Slabeva, Janner, & Schroth, 2008). Every extended search process, being a mixture of browsing and searching (Olston & Chi, 2003) usually takes several steps and along the way, new keywords arise out of the documents found and are used to refine the search and make it more precise, a method described thus: "browsing as berrypicking" (Bates, 1989) or "subject pearl growing" (Morville & Callender, 2010). Nowadays, the user has two continuously growing information pools that support search intensive processes: the web of documents, generally explored through the help of search engines and the web of knowledge or semantic web (Shadbolt, Berners-Lee, & Hall, 2006), unfortunately still being less user-friendly in terms of exploration. Nevertheless, the latter is growing steadily and is more and more augmented by already developed thesauri and taxonomies, although the linking between Linked Open Data and the web (Cyganiak & Bizer, 2008) and the matching between search queries and ontologies remains a challenging issue (Euzenat & Shvaiko, 2007).

In some cases, these taxonomies were developed not only over decades but rather over centuries as a result of a subtle intellectual reflection of information specialists. Unsurprisingly, libraries do nowadays realize that they cannot merely benefit from the semantic web, but that they also have a considerable contribution to make; thus more and more of these knowledge representations are made intelligible for the Semantic Web, often using the SKOS-data model (Miles & Brickley, 2005).

SKOS (Simple Knowledge Organization System) was published by the World Wide Web Consortium (W3C) in 2009 as a new standard for web-based controlled vocabularies. It serves as a data model to publish thesauri and taxonomies within the semantic web and – as a consequence – to make them semantically interoperable. This will lead, on the one hand, to a unification process of the heterogeneous library resources and, on the other, make (if not already making) large volumes of intellectual classifications useful for the Linked Open Data Project. In RODIN, they are used as support for the user while browsing and searching the web.

Using RODIN

The core of the user interface consists of a widget aggregator that allows the user to select an appropriate number of information resources under the form of widgets from a box on the left, by simply adding them to the main part of the user interface. By doing this, the user creates his own meta search engine using the organization form as an aggregator.



GOOGLE BOOKS

5 results found

Information visualization: perception fo design
Authors: Colin Ware
P.Year: 2004
Pages: 486

Information Visualization: Human-Centered Issues and Perspectives
Authors: Andreas Kerren
P.Year: 2008
Pages: 175

Information Visualization: Beyond the Horizon

DELICIOUS

10 results found

1 information

2 Data Visualization Modern Approaches | Graphics | Smashing Magazine
http www smashingmagazine com 2007 08 02 data visualization modern approaches

3 information aesthetics Information Visualizatio & Visual Communication

4 Information Transmission in a Social Network Dissecting the Spread of a Quora Post « Edwi Chen\'s Blog
http blog echen me 2011 09 07 information transmission in a social network dissecting the

RERODOC

10 results found

1 The benefits of synchronous collaborative information visualization evidence from an experimental evaluation
A great corpus of studies reports empirical evidence of how information visualization supports comprehension and analysis of data The benefits of visualization for synchronous group knowledge work however have not beer addressed extensively Anecdotal evidence an use cases illustrate the benefits of synchronou collaborative information visualization but very few empirical studies have rigorously examine the impact of visualization on group knowledg work We have consequently designed and conducted an experiment in which we have

SWISSBIB

10 results found

LE TEMPS ARCHIVES

Sorry, no results have been found

BIBSONOMY

10 results found

Figure 1. RODIN general interface

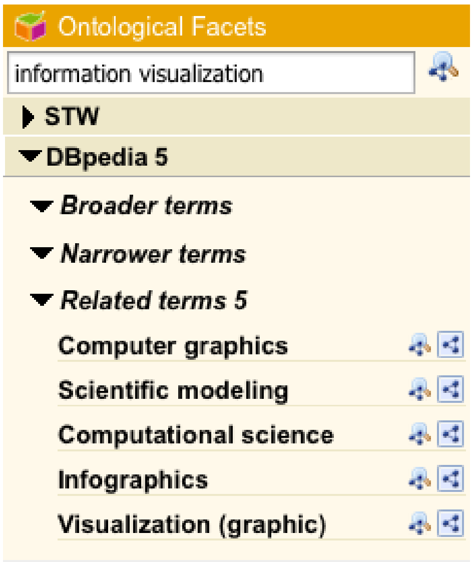
Compared to these systems, RODIN's widget box contains only resources that allow searching via an adequate search field. Most of these resources are in connection with the realm of digital libraries, due to the main project E-lib.ch. The widgets can be organized under different tabs, allowing the user to define several distinct search environments, according to the number of projects or extensive searches s/he is currently dealing with. So far, the results are given in the relevant widgets. In a future version, the user may decide between a mashed and a widget-based representation of the results given.

Besides the widget aggregation box, the user interface is dominated by a prominent search field above the aggregator. Any simple keyword based search may be initiated from this search field (see Figure 1). The results of this search are represented in the corresponding widgets as received by the corresponding servers.

Ontology Exploration for Query Expansion

Simultaneously with this simple search, the keywords are transferred to the ontology module and appositely translated into SPARQL-Queries to check for full or partial match against available concepts in the connected ontologies. Due to the large amount of data, disambiguation issues and the need to control and optimize the ontological search process (tokenization, compound analysis, disambiguation, sequencing, search, token collection), this request may take quite a long time and in some cases is pruned by RODIN after a suitable configurable and user-friendly interval of 15 seconds; it is interrupted or restarted whenever a new search is initiated in the main search field or in the ontological search field.

Following the SKOS Data organization, RODIN displays each ontology concept using first its preferred label(s) followed by the alternative label(s). If the matching between the search terms and any ontological entries is successful, the semantic context, i.e. the narrower, larger, and related terms are displayed in the ontology box on the right hand side of the window to be browsed by the user (see Figure 2).



Ontological Facets

information visualization

▶ STW

▼ DBpedia 5

▼ *Broader terms*

▼ *Narrower terms*

▼ *Related terms 5*

Computer graphics

Scientific modeling

Computational science

Infographics

Visualization (graphic)

Figure 2. Ontological facets

For this context, we coined the term ontological facets, since the three different semantic extensions of a term can be seen as facets of a semantic concept.

As can be seen in Figure 2, every term appears with two icons on the right, the first one (with an RDF-icon like symbol and a loupe) is used to further explore the ontology; the other one (with the RDF-icon like symbol surrounded by a rectangle) is used for the visualization of the term in its ontological context. A left click on the term itself will add it to the refinement bar, as described below. This functionality allows a direct query expansion of the initial search.

These two icons allow the user to explore the ontology in two manners, the folder-like structure of the ontological facets or a graphic visualization, as described in the following section. Both may guide the user along his/her search to find alternative search terms depending on the path (narrower, larger, and related) taken and give him/her an overview of the complex scope of the semiological relationships.

Visualizing the Ontology

The visualization of the semantic scope (see Figure 3) is especially helpful to the user whenever the number of ontological relations is high, as is often the case in bibliographical catalogues, with their large numbers of synonyms, hyper- and hyponyms, being a result of the long tradition of thesauri and taxonomies they represent.

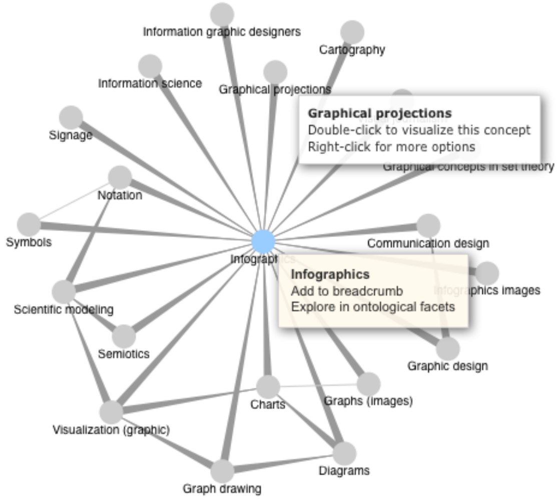


Figure 3. Semantic scope visualization

The ontological visualization itself is positioned above the widgets, offering the user a simultaneous view of the selected part of the web of documents (i.e. the search result in the widgets) and the corresponding knowledge of the semantic web.

The graphical visualization allows several user interactions:

- by giving a new representational view by rearranging the elements on the screen,
- by adding one of the graphical represented terms to the refinement bar,
- by allowing the start of a new search in the ontological facets,
- by enlarging the view of the concepts by displaying the scope of concepts around a selected peripheral term,
- by navigating the SKOS part of the ontology.

Search Refinement

Besides exploration and browsing of the ontological knowledge, the search results of the widgets as well as the terms of the ontologies themselves can be used for search refinement, i.e. query expansion. The user may collect these terms and initiate a new search with the collected terms. The terms are collected in a breadcrumbs-like list and represented horizontally in a bar between the main search field and the area containing the other elements of the user interface, i.e. the widget box and the ontological facets, the widget results and, if activated, the ontological visualization.

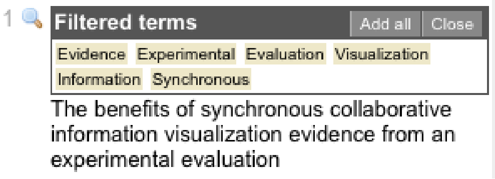
The breadcrumb list is enlarged with any new term added as a result of one of the actions described in the following subsections.

Keyword Extraction via DBPedia Spotlight

By clicking on a small loupe or magnifier positioned on the left of every single result, a sub module calculating the key terms of the result is activated. The sub module represents an integration of DBPedia-Spotlight[1] (Mendes, Jakob, García-Silva & Bizer, 2011), a tool that provides DBPedia concept recognition in text documents.

If the DBPedia-Spotlight service fails to recognize concepts in the chosen text, we perform a search for concepts with similar labels directly in the DBPedia ontology. Should our search for similar labels fail, our last resort is to simply select the most common words in the text. Because the user is only interested in the key terms, the strategy used to extract them from the text can be hidden from him/her. Once the keywords are found, a small window showing them opens above the single result's text (see Figure 4).

[1] http://dbpedia.org/spotlight



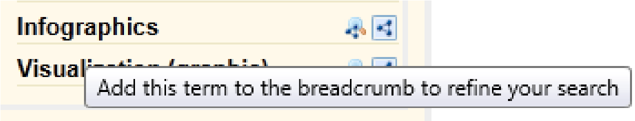
The benefits of synchronous collaborative information visualization evidence from an experimental evaluation

Figure 4. Keyword extraction from DBPedia Spotlight

By the activation of this module the user is once again guided in the use and the view of a connected ontology, by getting related keywords which link the document to DBPedia. The terms may therefore be used to browse the ontology whenever no match between the terms of the initial search and the ontologies themselves was found. On the other hand, one, several or all terms filtered by DBPedia Spotlight may be used to refine the search, either by clicking on them or by adding all of them to the refinement bar (see Figure 7).

Query Expansion

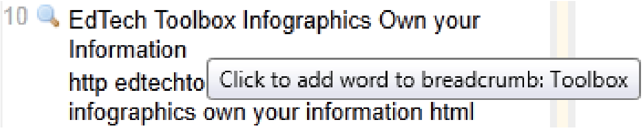
In the meantime, every click on the ontological terms found and represented in the ontological facets will add these terms to the refinement bar (see Figure 5).



Add this term to the breadcrumb to refine your search

Figure 5. Ontological term addition

Finally, the user may click on any word in the documents found, i.e. their representation in the result list and add it by clicking on it to the breadcrumb list (see Figure 6).



EdTech Toolbox Infographics Own your Information
http edtechto
infographics own your information html

Click to add word to breadcrumb: Toolbox

Figure 6. Simple word pickup

The search refinement process itself is started with the “Refine Search” button at the end of the breadcrumb list of refining terms (see Figure 7). The activation of this button initializes a new search in the widgets selected with the initial search term (still visible in the major search field) and the terms selected from the documents and the ontologies.

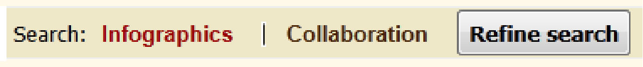


Figure 7. Search refinement

Technical Details

RODIN combines several functionalities for browsing and search refinement and linking the visually perceptible web of documents and the web of knowledge under the form of selected ontologies.

In this context it has always been one of the major goals of RODIN to build a twofold infrastructure that would combine as many technical issues as possible: a project-specific platform serving as a portal for the main project E-lib.ch and its subprojects with a dedicated approach for the world of (digital) libraries and the corresponding ontologies and a rather generic approach being able to integrate if possible any information resource of the web of documents on the one hand and any ontology from the web of knowledge on the other.

The system is based on the open source version of POSH. The software runs under an APACHE2/PHP on a window server. It makes extensive use of AJAX for ontology processing and stores the data results in a relational database for quick recall of relevant portions of data. The main architecture is based on POSH (Portaneo Open Source Homepage, www.portaneo.net), an open source aggregator that allows the administration of widgets and their combination under a common interface. The framework built around POSH permits a structured integration of a wide variety of widgets under a unified Look & Feel for all widgets.

Every application related to a widget is integrated using the systems API and – if possible –a RESTful interface. Unfortunately so far rather few programming interfaces make use of this technology and the integration of APIs is far from being standardized. This led in several cases to a rather unexpected amount of handcrafting which we reduced to a minimum by adding a suitable generalization to the implemented widget. The generalization consisted of the implementation of a framework with a programming interface fulfilling the needs of each particular data source while allowing the same behavior for any widget.

As regards ontologies and the semantic web, we integrated – as mentioned above – two ontologies using the SKOS-Data Model: STW and DBPedia. Furthermore, RODIN gives a use case for DBPedia Spotlight with a further interest in exploring ontologies and using the results of this process in combination with the result of general web searching for a refined continuation of both processes.

Ontological data sources are integrated using an object oriented interface. The latter allows the easy integration of further sources by using a common programming interface and reusing available sub components.

Since RODIN was mainly designed and implemented as a so-called "web 3.0", i.e. a semantic web tool, it does not contain any collaborative functionalities like recommending a search environment or sharing widgets or the mutual manipulation of a working environment. Since these "web 2.0" functionalities are already contained in the POSH framework, we deliberately decided to omit them from the current system since we believe they add no value to a "web 3.0" tool. Nevertheless, if the user community considered them useful, they could easily be added. For the first release of RODIN, it was decided to put a focus on the single user experience.

Usability Issues

Due to the complexity of the user interface and in order to assure user friendliness and a maximum of user acceptance, several usability tests have been conducted so far: a first one after the implementation of the widget aggregator with a test group of seven practicing information specialists, a second one after the integration of the Linked Open Data resources, i.e. after the implementation of all major functionalities.

The first test's results led to some changes in the general look and feel of the system's interface as well as some adjustments concerning the design, arrangement and proportion of the distinct components. Since the system did not contain all functionalities yet, only two severe and six other usability issues were detected. Several users had problems finding the most important elements, such as the search field and the widget box, while others had problems understanding the meaning of the icons for displaying and refining the search results. All of these problems were solved and changed before the integration of the ontological search began.

The second usability test was done as a combination of a heuristic evaluation by two external usability experts and three user acceptance tests including eye-tracking with information specialists, conducted by the same experts. Generally speaking, the interface was now considered "clearly laid out" but nevertheless "challenging in its complexity". Test participants mainly criticized the lack of a structured "Help" explaining the proper use of the major functionalities, but considered the system itself as helpful for search refinements. The test members found it difficult to understand the term "ontology" as well as the terms "broader", "narrower" and "related" and others such as "breadcrumbs".

In the end, the usability experts suggested several modifications in the design, such as location of the refinement terms next to the search field, further simplifications in the overall design and a reduction of the system features to some core functionalities which will be realized before the release of the system.

Conclusions

In this paper we describe RODIN, a tool for simultaneous browsing and searching in the web of documents and the semantic web. The system offers a number of functionalities for search refinement and is designed for the use of information specialists and users in the context of digital libraries, but may be helpful in any similar context.

In the near future, RODIN will be installed in several Swiss scientific libraries to be tested by information specialists as a support for their work (cataloguing, searching) and by their patrons, as a searching device in scientific work. Besides that, the development of a mobile version has started and will be realized within the next development cycle.

Acknowledgements

This work is funded by the Rector’s Conference of the Swiss Universities of Applied Sciences and the Competence Network “Information Services Network” of the University of Applied Sciences and Arts Western Switzerland RCSO-ISNet.

Thanks to Sonja Hamann and Daniel Streiff for their contribution in visualizing the ontologies.

References

Auer, S., Bizer, C., Kobilarov, G., & et al., (2008). *DBpedia: A Nucleus for a Web of Open Data*. Springer Lecture Notes in Computer Science. Berlin: Springer.

Bates, M.J. (1989). The design of browsing and berrypicking techniques for the on-line search interface. *Online Information Review* 13, 5, 407-431. DOI: 10.1108/eb024320

Cyganiak, R. & Bizer, C. (2008). *Pubby - A Linked Data Frontend for SPARQL Endpoints*. Retrieved June 14, 2009 from http://www4.wiwiss.fu-berlin.de/pubby/

Euzenat, J. & Shvaiko, P. (2007). *Ontology matching*. Heidelberg: Springer.

Hoyer, V., Stanoevska-Slabeva, K., Janner, T., & Schroth, C. (2008). Enterprise mashups: Design principles towards the long tail of user needs. *IEEE International Conference on Service Computer (SCC’08)*, 2, 601-602.

Mendes, P., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: Shedding light on the web of documents. *7th International Conference on Semantic Systems*, 7–9 September 2011, Graz, Austria.

Miles, A. & Brickley, D. (2005). *SKOS Core Vocabulary Specification*. W3C working draft, W3C, November 2005. Retrieved June 26, 2012 from http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20050510/

Morville, P. & Callender, J. (2010). *Search patterns. Design for discovery.* Farnham: O’Reilly.

Neubert, J. (2009). Bringing the "Thesaurus for Economics" on to the Web of Linked Data. In: *Linked Data on the Web (LDOW2009)*. Retrieved June 11, 2012 from http://events.linkeddata.org/ldow2009/papers/ldow2009_paper7.pdf

Olston, C. & Chi, E. (2003). ScentTrails: Integrating browsing and searching on the Web. *ACM Transactions on Computer-Human Interaction*, 10(3), 177-197. DOI=10.1145/937549.937550

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The Semantic Web revisited. *IEEE Intelligent Systems*, 21(3), 96-101.