QMAP: AN RDF-BASED QUERYABLE WORLD MAP

Frederik Hogenboom, Alexander Hogenboom, Ricardo van Gelder,
Viorel Milea, Flavius Frasincar, and Uzay Kaymak

Erasmus University Rotterdam
PO Box 1738, NL-3000
Rotterdam, the Netherlands

{287986fh, 287485ah, 287437rg}@student.eur.nl
{milea, frasincar, kaymak}@few.eur.nl

Abstract. In its current state, the Web contains a lot of information which has the potential to be used by other applications, thus addressing different aims than the ones originally intended. In this paper we present QMap, an application that supports the repurposing of existing Web content from originally non-queryable and/or low-level machine-interpretable data. The system architecture provides for the means of extracting the relevant data, querying it based on user interests, and visualizing the results in a user appealing manner. We illustrate the functionally of QMap by employing economic data about countries, data that we extract from the CIA World Factbook Web page. The extracted data is then represented in RDF. Based hereon, the user is able to pose SPARQL-based queries on the extracted data and visualize the results using Google Maps. The generated thematic maps provide intuitive means for performing economic analysis on the targeted countries.

Keywords: RDF(S), SPARQL, visualization, economics, thematic map.

1. Introduction

The current Web comprises tremendous amounts of data, very often presented in a rather subjective way – a result of the attempt to match the user's needs with the developer's vision. Due to the open architecture of the Web, one can obtain this data relatively easy and present it based on personal interests. Very often, the user's needs and the vision of the developers evolve in time and the Web site needs to reflect these changes accordingly. All these aspects require content repurposing (Gerber & Merker 2008), i.e., gathering the existing Web data, restructuring it, and the publication hereof based on new information needs.

One common way to achieve the restructuring of Web sites is by means of building models from existing Web sites and subsequently republishing the models by using a different navigational structure (Obrenovic et. al. 2004, Ricca & Tonella 2001). Initially these models were XML-based, but due to their lack of semantics we have recently witnessed a shift towards semantic rich models specified by using RDF or OWL (Frasincar et. al. 2006, Micu et. al. 2008).

Many of the existing Web sites support simple interaction, where the user is limited to ‘link following’ as the only interaction means with the system. These sites lack advanced forms of interactivity such as, for example, query facilities. Also, the information is often presented as tables, not exploiting the semantic properties of the displayed data. In this sense many Web sites lack advanced ways of information presentation that usually are based on data visualization techniques. The interactivity and visualization needs are additional reasons for republishing existing Web sites so that these aspects lead to an increased user experience with the system.

In the current paper, we present an application that enables the creation of customized thematic maps based on user queries. The focus is on the intuitive visualization of the generated results, achieved by overlays on relevant Google maps. The source we employ for this purpose is the CIA World Factbook, which provides extensive information on countries. The goal is to enable complex queries of this information, while maintaining a high level of intuitiveness on the generated visualizations.

In Section 2, we provide an overview of work related to the current endeavor. Section 3 presents the QMap application through a discussion of the general architecture, the configuration hereof, query design, and finally the generated visualizations. Section 4 concludes the paper and provides some possible focus points for further research.

2. Related Work

Recent research results report on findings regarding the retrieval and semantical interconnection of data from multiple sources. As for retrieving data, a platform for sharing knowledge is proposed in (Kraines et. al. 2006), where data can be enriched using semantic descriptions, thus enabling scientific knowledge to be interpreted by computers. The proposed system requires experts to add semantics to the information stored into the system, instead of relying on natural language processing techniques for interpreting submitted knowledge. On the other hand, in the domain of media & journalism, the focus of the NEWS project (Fernandez et. al 206) is on automatically annotating news messages using natural language processing techniques and mapping these annotations onto an ontology, thus enabling the system to exploit the annotations for dealing with multilingual issues.

Combining data from multiple sources is elaborated in (Chen et. al. 2006). Here, an application development framework is proposed along with a set of semantic

tools (among which: a mapping tool, an ontology-based query interface, and an ontology-based search engine) to facilitate the integration of heterogeneous relational databases. This toolkit is used for an application for traditional Chinese medicine, in which 70 legacy relational databases are semantically interconnected. Another medical application (Sheth et. al. 2006) combines, amongst others, data on patients, procedures, treatments, diagnoses, and insurance plans, with domain knowledge and rules, hereby supporting clinical decisions. The data is mapped onto three ontologies in order to properly represent the aspects of the domain. The data is then accessed through a user interface by using semantic documents.

Other recent applications combine information from one or multiple sources with geographical information visualized on Google Maps. Google Maps is an online application for viewing maps and satellite images. Through the Google Maps API[1] one is able to create custom maps with user-defined functionality. With the London Profiler, the University College London Centre for Advanced Spatial Analysis shows the applicability of Google Maps to thematic maps[2]. With the London Profiler, one can request thematic maps of London based on geo-demographics of the city from data on population attributes such as ethnicity, deprivation and crime. Each thematic map is placed on a layer on top of the Google Map and can be switched on or off. The added value of these type of maps is that certain relations might become visible, which are hard to discover without employing similar visualization techniques. A shortcoming of the London Profiler is that a user is not able to specify his or her own query on a certain dataset, but can only execute a pre-programmed query. WikiMapia[3] is another example of representing data using the Google Maps API. WikiMapia tries to describe all locations in the world by enabling users to define and describe areas.

Finally, we relate to some Semantic Web technologies that we consider relevant for the current context. The Resource Description Framework (RDF) (Klyne & Caroll 2004) focuses on the representation of information about resources. Even though these resources were originally envisioned in a Web context, the simple data model of RDF enables the application of this language even when the resources in question are not directly retrievable from the Web. Its power comes from the fact that RDF is intended as a common framework for exchanging data between applications. One of the most successful applications of RDF on the Web

---

[1] Available at: http://www.google.com/apis/maps/
[2] Available at: http://www.londonprofiler.org/
[3] Available at: http://www.wikimapia.org/

are already present in the form of RDF Site Summary (RSS 1.0), a web feed format used mostly by news sites and weblogs.

RDF Schema (RDFS) (Brickley & Guha 2004) is a semantic extension of RDF and a first important step towards towards the definition of ontologies. RDFS adds several constructs to the RDF vocabulary and extends the underlying semantics.

3. QMap

In this section we give a comprehensive overview of the QMap application. Upon presenting the architecture of the application in a general setting, we illustrate the envisioned functionality of QMap in a particular case.

3.1. Architecture

A general overview of the information flow as designed for the QMap application is shown in Figure 1, at an abstract level. It should be noted that the architecture presented here is source-independent, although in later sections we shall rely on the CIA World Factbook[4] as a source for exemplifying the functionality of the presented application.

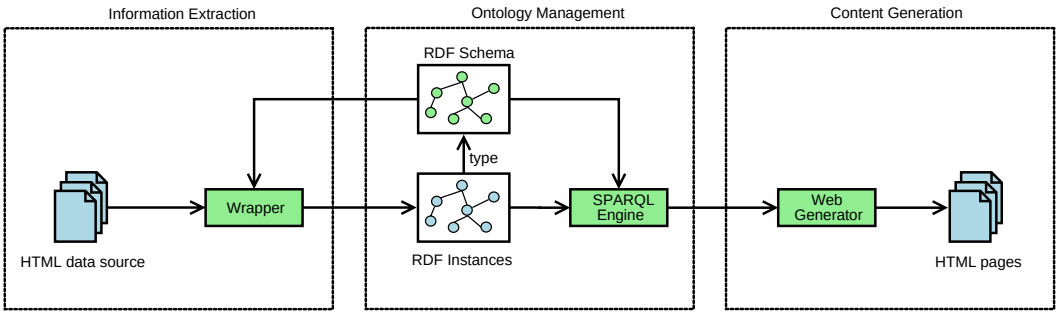

Figure 1. QMap information flow.

At an abstract level, we distinguish between 3 main components of QMap: i) Information Extraction, ii) Ontology Management, and iii) Content Generation. In what follows, we describe the main functionality hereof and discuss the interactions between components. Finally, we describe how this framework can easily be extended to incorporate various data sources.

[4] Available at: https://www.cia.gov/library/publications/the-world-factbook/index.html

The *Information Extraction* component is responsible for retrieving data from the selected datasource(s). This information is assumed to come from HTML pages, and thus semantic annotations are considered absent. Besides knowledge regarding the general structure of the HTML source, information of the concepts and relations contained herein is assumed. This can be seen in Figure 1 through the interaction between the Wrapper module and the RDF Schema. The RDF Schema should thus provide an abstract structure of the entities that are relevant for the scope of the application.

Based on the RDF Schema as described in the previous paragraph, the wrapper creates instances that are fed into the *Ontology Management* component. In other words, the ontology that describes the HTML data source is populated with instances from that source. A SPARQL engine module ensures communication between the Ontology Manager and the Content Generator module. Both the RDF Schema and the extracted instances are employed here for answering queries.

Finally, the *Content Generator* component feeds user generated queries into the SPARQL engine, which provides the query results. HTML pages are generated based on these results as layers above the relevant Google map.

It should be noted that, upon properly configuring the wrapper for some envisioned source and providing an appropriate RDF Schema for that source, any type of country related data may be queried. Additionally, due to the dependence between the wrapper module and the RDF Schema, extensions of the latter are supported. In the following sections, we illustrate the functionality of the general architecture presented here by relying on the CIA World Factbook as data source.

3.2. Configuration

A number of configuration settings may be modified through a form, which presents an overview of the different QMap variables: *ciawfb*, *ciawfbOnt*, *ciawfbSrc*, *ciawfbRdf*, *writeRdf*, *timeOut*, *devKey*, *startCoords*, *startZoom*, and *debug*. The *ciawfb* variable is a flag indicating whether the CIA World Factbook should be employed as datasource for the application. The location of the ontology describing this datasource can be indicated through the *ciawfbOnt* variable. The root file of the CIA World Factbook is indicated by the *ciawfbSrc* variable. As this source may be represented in other formats besides RDF, the *ciawfbRdf* flag indicates whether the content is indeed RDF. Whether the models created by the application should be in RDF is specified through the *writeRdf* switch.

The *timeOut* variable indicates the number of tries to access a source QMap should attempt before timing out. For the interaction with the Google Maps API, the required key is stored in the *devKey* variable. Through *startCoords* and *start-Zoom* the user may indicate the default coordinates of zoom level, respectively, for the generated Google Map. For development purposes, the application can be set to run in debug mode by choosing the appropriate setting for the *debug* flag. Upon submission of this form, the configuration file is saved and the server is reset.

3.3. Query Design

Upon starting the server and, optionally, adjusting the configuration settings, QMap is accessed through the main query page, that is set by default in the *basic* mode, as displayed in Figure 2. QMap provides two different modes for designing queries: the *basic* and the *pro* mode, respectively.

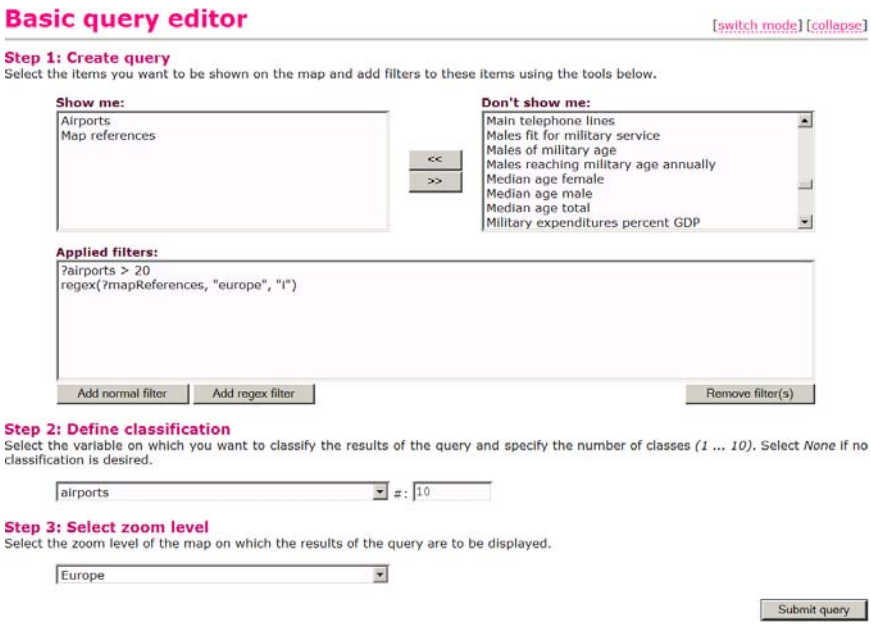

Figure 2. The basic query editor.

This basic editor allows the creation of simple queries in three steps. In Step 1, the relevant properties (from the perspective of the user) can be selected from the list of available properties, and added to the *Show me* list. Additionally, normal (comparison) or regular expression (regex) filters may be added to these properties, thus restricting the results list based on the constraints defined by the user.

Step 2 involves choosing a classification for the results, based on one of the available properties. Finally, in step 3, a zoom level may be selected for the Google

map created by the application. A number of default zoom levels are defined by default: the world (most general), Africa, Asia, Australia, Europe, and North and South America.

Figure 2 also presents our first query example. The query selects countries with more than 20 airports and located in Europe. The associated SPARQL query is given below:

```
PREFIX ont: <http://www.daml.org/2003/09/factbook/factbook-ont#>
SELECT ?airports ?mapReferences
WHERE {
      ?airport > 20 .
      FILTER(regex(?mapReferences, “europe”, i) .
       }
```

Internally QMap inserts additional variables in the query for representing the geographical coordinates of the central point of countries. These coordinates are used for visualizing the generated results. These countries should be classified in 10 groups based on the number of airports that they have. The zoom level is defined to be Europe.

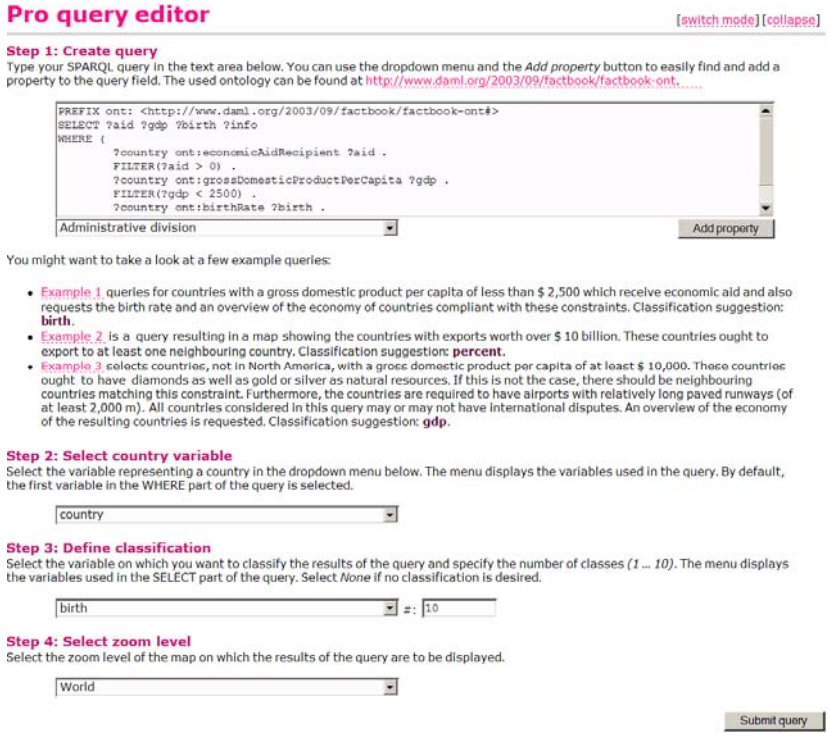

Figure 3. The pro query editor.

More complex queries may be created through the *pro editor* interface. The user can switch to this mode through the *switch mode* command, present in the top-right corner of the page. This mode is displayed in Figure 3.

Creating a query in the pro editor comprises 4 steps. The first step consists of entering a query directly using SPARQL. Although this can be done in the regular fashion, three examples of possible queries are provided, together with a textual explanation of the result of these queries. A more extensive presentation of two of the three default queries is given is Section 3.4. The second step of creating a query in the pro editor consists of specifying which variable represents a country that should be marked on the generated Google map. The final two steps are identical to the last two steps of the basic query editor.

Our second illustrative example is the first example of the pro query editor given in Figure 3. The query asks for countries with a gross domestic product per capita of less than $ 2 500 which receive economic aid and also requests the birth rate and an overview of the economy of countries compliant with these constraints. These countries should be classified in 10 groups based on the observed birth rate. The SPARQL query designed for this purpose is shown below:

```
PREFIX ont: <http://www.daml.org/2003/09/factbook/factbook-ont#>
SELECT ?aid ?gdp ?birth ?info
WHERE {
        ?country ont:economicAidRecipient ?aid .
        FILTER(?aid > 0) .
        ?country ont:grossDomesticProductPerCapita ?gdp .
        FILTER(?gdp < 2500) .
        ?country ont:birthRate ?birth .
        ?country ont:economyOverview ?info .
      }
```

3.4. Visualizations

In this section we provide the visualization of the two queries presented above. For this purpose we employ Google Maps, and present the query results as an overlay on these maps.

The generated Google map for the first query, with the appropriate overlay, is displayed in Figure 5. The resulted countries are classified based on the number of airports. An initial analysis shows that the number of airports in Eastern Europe is lower than in the Western part of the continent.

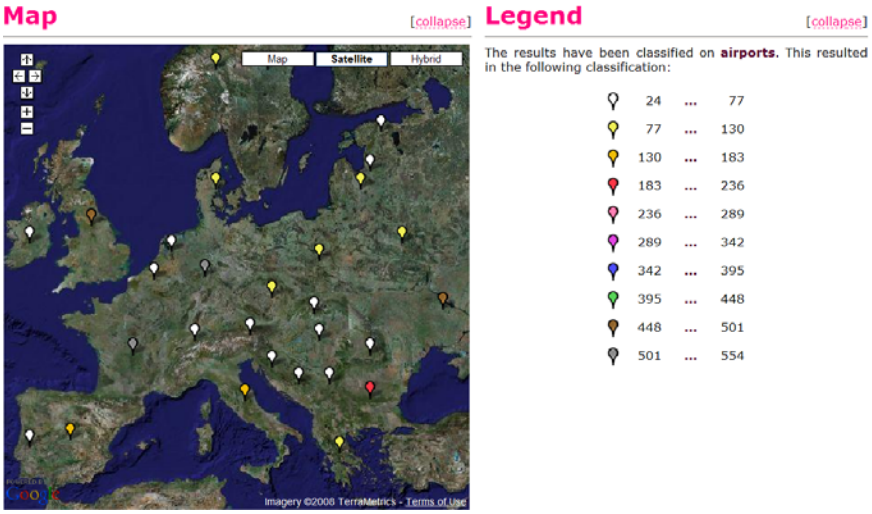

Figure 4. Results for the first query.

The generated Google map for the second query, with the appropriate overlay, is displayed in Figure 5. The countries selected through the query are classified based on birth rate, i.e., the number of births per 1000 people. The generated visualization shows that the birth rate for underdeveloped countries is highest in Central Africa.

QMap, as presented in this paper, works with the Microsoft Internet Explorer 6.0 or higher, Mozilla Firefox 2.0 or higher and Opera 9.0. The Java Runtime Environment 1.5 or higher is needed in order to be able to run QMap. Apache Tomcat 4.1 or 5.5 Web server must be present on the client machine.

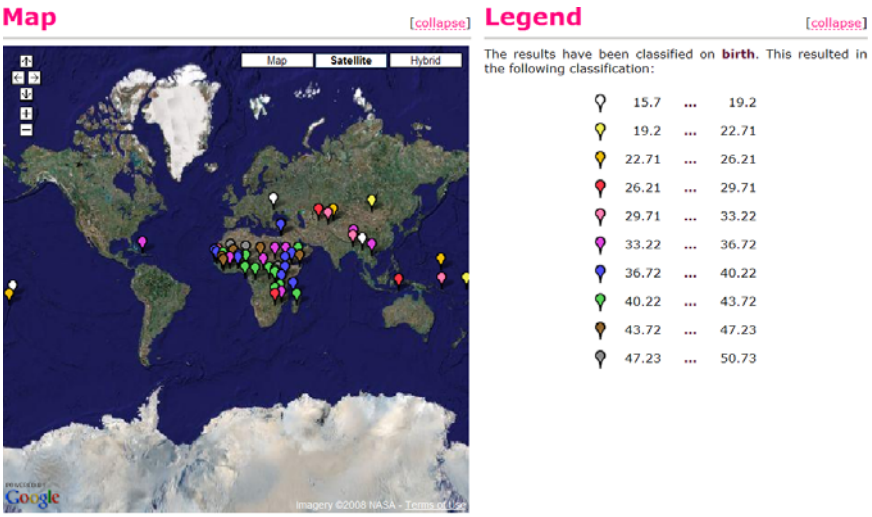

Figure 5. Results for the second query.

4. Conclusions & Future Work

QMap enables complex queries on economic data relating countries, and presents the results of such queries in the form of intuitive visualizations by employing Google Maps. One of the foundational features of the application consists of the ability to extract knowledge from free text, as available in the CIA World Factbook, and representing this in RDF.

Following on the same lines, QMap also provides a level of flexibility that enables the addition of different sources other than the CIA World Factbook. An interesting focus area for further research could consist of further extending the data employed for queries, allowing for multiple, inter-related sources.

Finally, a comment on the different query design interfaces is in place. Currently, as previously discussed, queries may be designed in two different modes: the basic and the pro mode. The basic query editor offers a high level of intuitiveness when designing queries, but only allows for a limited representational power when doing so. The pro query editor is aimed at employing the full representational power of SPARQL, but achieving this limits the level of intuitiveness from the perspective of the non-technical user. In this context, we envision an extension of QMap that allows for the processing of queries expressed in natural language. In this manner, one could express complex queries without the necessity for advanced knowledge of the SPARQL query language.

References

Brickley, D. & Guha, R.V. (2004). *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation, 10 February 2004.

Chen, H. & Wang, Y. & Wang, H. & Mao, Y. & Tang, J. & Zhou, C. Yin, A. & Wu, Z. (2006). *Towards a Semantic Web of Relational Databases: a Practical Semantic Toolkit and an In-Use Case from Traditional Chinese Medicine*. Proceedings of the Fifth International Semantic Web Conference (ISWC 2006), pp. 750-763, Springer.

Dean, M. (2003). *CIA World Factbook Ontology*. Available at: http://www.daml.org/2003/09/factbook/factbook-ont . Last visited: January 2007.

Fernandez, N. & Blazquez, J.M. & Fisteus, J.A. & Sanchez, L. & Sintek, M. & Bernardi, A. & Fuentes, M. & Marrara, A. & Ben-Asher, Z. (2006). *NEWS: bringing Semantic Web Technologies into News Agencies*. Proceedings of the Fifth in International Semantic Web Conference (ISWC 2006), pp. 778-791, Springer.

Frasincar, F. & Houben, G.J. & Barna, P. (2006). *HPG: the Hera Presentation Generator*. Journal of Web Engineering 5(2):175-200.

Gerber, N. & Merker, L. *Tackling the Problem of Repurposing Web Content*. EDUCAUSE Quarterly 31(1):62-65.

Klyne, G. & Caroll, J.J. (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax*. W3C Recommendation, 10 February 2004.

Kraines, S. & Guo, W. & Kemper, B. & Nakamura, Y. (2006). *EKOSS: A Knowledge-User Centered Approach to Knowledge Sharing, Discovery, and Integration on the Semantic Web*. Proceedings of the 5th International Semantic Web Conference (ISWC 2006), pp. 833-846, Springer.

di Lucca, G.A. & Fasolino, A.R. & Pace, F. & Tramontana, P & de Carlini, U. (2002). *WARE: A Tool for the Reverse Engineering of Web Applications*. Proceedings of the Sixth European Conference on Software Maintenance and Reengineering (CSMR '02), pp. 241-250, IEEE Computer Society.

Micu, A. & Mast, L. & Milea, V. & Frasincar, F. & Kaymak, U. (2008). *Financial News Analysis Using a Semantic Web Approach.* In Semantic Knowledge Management: an Ontology-based Framework. Idea Group.

Obrenovic, Z. & Starcevic, D. & Selic, B. (2004). *A Model-Driven Approach to Content Repurposing.* IEEE MultiMedia 11(1):62-71.

Ricca, F. & Tonella, P. (2001). *Understanding and Restructuring Web Sites with ReWeb*. IEEE MultiMedia 8(2):40-51.

Sheth, A. & Agrawal, S. & Lathem, J. & Oldham, N. & Wingate, H. & Yadav, P. & Gallagher, K. (2006). *Active Semantic Electronic Medical Record.* Proceedings of the Fifth International Semantic Web Conference (ISWC 2006), pp. 913-926, Springer.