RKBExplorer.com: A Knowledge Driven Infrastructure for Linked Data Providers

Hugh Glaser, Ian C. Millard, and Afraz Jaffri

School of Electronics and Computer Science
University of Southampton, UK
{hg,icm,a.o.jaffri}@ecs.soton.ac.uk

Abstract. RKB Explorer is a Semantic Web application that is able to present unified views of a significant number of heterogeneous data sources. We have developed an underlying information infrastructure which is mediated by ontologies and consists of many independent triple-stores, each publicly available through both SPARQL endpoints and resolvable URIs. To realise this synergy of disparate information sources, we have deployed tools to identify co-referent URIs, and devised an architecture to allow the information to be represented and used. This paper provides a brief overview of the system including the underlying infrastructure, and a number of associated tools for both knowledge acquisition and publishing.

1 Introduction

The Linking Open Data Initiative (http://linkeddata.org/) has encouraged the widespread creation and deployment of RDF data using URIs that are both dereferenceable and linked to URIs from other data sources. This initiative led to the creation of DBpedia (http://wiki.dbpedia.org/Datasets), an RDF version of Wikipedia which itself established a base from which other data sets could be linked. There are now several hundred thousand links from DBpedia to other RDF sources and Linked Data sites.

There have been guidelines laid down for the creation of Linked Data [1] but for users not familiar with Semantic Web technology, creating and maintaining a Linked Data site requires a serious amount of time and effort. Such a scenario is encountered in our own ReSIST project, an EU funded Network of Excellence in Resilient Systems. From its early conceptions, it was proposed that the entire project would be supported by a semantically-enabled knowledge infrastructure. The vision was of sets of services and applications for both acquiring and publishing knowledge, working together as a unified coherent resource. Project members and others would be able to explore the knowledge created and acquired from distributed and heterogeneous resources, enabling them to identify relationships and resources that may not previously have been evident.

Many of the participants in the project have very little or no experience or knowledge of the Semantic Web. The aim of www.rkbexplorer.com was to create an infrastructure where knowledge acquired from different data sources could

S. Bechhofer et al.(Eds.): ESWC 2008, LNCS 5021, pp. 797–801, 2008.

become part of 'The Semantic Web' and not just used as a closed part of a static system. The system that has been built provides data providers with a SPARQL endpoint, dereferenceable URIs, data browser, data explorer, consistent reference services and keyword search. The rest of this paper describes these services and highlights the minimal cost of use for non-expert users.

2 Related Work

There have been a number of efforts recently to facilitate the management of linked data. Most noticeable is D2R Server [2] which was produced so that data providers who held their data in relational databases could expose the database via a SPARQL endpoint and HTML browser. This method has the convenience of not having to convert any of the existing data into RDF which involves the execution of custom extraction scripts. A disadvantage of this approach is that additional functionality that can only be performed on RDF data cannot be performed on data exposed through the D2R server. An example of this is the co-reference analysis that is described in Section 4.2.

The individual systems made for exposing and exploring semantic data have been produced on a stand alone basis. A true semantic web site should possess all of the features that are required for Linking Data, and more, and also have the added advantage of being simple in creation and management. The architecture described in the rest of this paper provides just such an infrastructure that is a live system being used daily by many users.

3 Knowledge Acquisition

Ideally, in an active Semantic Web world, we would have simply been able to use existing knowledge sources. These sources would publish their contents against well-known ontologies, both as SPARQL endpoints and resolvable URIs. We would then use them, possibly needing some ontology translation on the way. Unfortunately, this is not yet the case. When the project started in January 2006, there were few such citizens of the Semantic Web, and so we resolved to undertake the bootstrap process ourselves.

We therefore harvested the information from the places we identified, and made it openly available as both resolvable URIs and SPARQL endpoints, against our AKT ontology (`http://www.aktors.org/publications/ontology/`). Since this also involved minting all new URIs, each data source is held in a separate triple-store, on a separate domain, not necessarily running on the same machine.

At present we have acquired RDF data on people, publications and institutions from the 18 partners in the ReSIST project. People and publication data have also been harvested from a number of major metadata resources. We chose the major publishers and aggregators in Computer Science, and have to date harvested some 50 million triples from Citeseer, the ACM, DBLP, NSF and selected IEEE conferences.

The way in which we have structured the system that collects and organises this knowledge means that knowledge from any new data provider can easily be brought into the existing infrastructure, automatically giving all of the benefits that have been built into the system. For example, if a new partner was to join the ReSIST project, information from their site would be converted into RDF. Putting this in a triplestore on the rkbexplorer.com site means that their knowledge has been exposed via SPARQL, tabulator-style browsing, graphical interaction and keyword based searching. The knowledge will also be analysed for any coreferences amongst people and papers and this information will also be stored in a separate knowledge base. All the functionality of Linked Data will have been given to the user for a small investment in knowledge acquisition. They can thus have a fully-functional Linked Data site up and being linked into other sites in minutes or longer, depending primarily on the RDF provided.

4 Information Infrastructure

4.1 Triplestores

We use 3Store [3] as our base repository, with a separate knowledge base representing each data source which we have acquired. The separate knowledge bases facilitate the system scalability, and help to provide the high query performance needed for an application of this sort, while allowing the assertion of the volumes of data (many tens of millions of triples) that we have. Each repository is complemented by a number of services and interfaces, which are openly accessible at `http://<repository>.rkbexplorer.com/`. These resources are additionally being indexed by semantic search services such as `http://sindice.com/`.

4.2 Consistent Reference Service (CRS)

The way in which the RKB explorer and other applications give a unified view of tens of triplestores (knowledge bases) with tens of millions of triples, requires a well-founded method of allowing URIs to bridge between the triplestores, when they are considered to refer to the same concept.

The ReSIST activity embraces this. It includes in its architecture the deployment of a number of CRSes, which are knowledge bases of URI equivalences for the application being considered, according to appropriate criteria.

We chose to keep this knowledge separately from the main data. One reason is simply that of good engineering practice. It is easier to maintain knowledge that is being created by the CRS builder separate from the knowledge that is being created by the information provider. Indeed, different CRS providers will exist for the same information in an open Semantic Web world. A second reason is that a CRS is designed for a purpose, or set of purposes. Some applications might wish to consider that two concepts are the same, while this may not be the case for another application over the same knowledge. For example, in undertaking citation analysis, a paper with the same title and text that appeared both as a journal article and technical report should be considered as two separate papers. In an

application such as ours, where we are considering who works with whom on what topic, it might well be more appropriate to consider that they should be treated as one resource, while still representing the separate details in a consistent fashion. As information providers of the basic information, we include a `coref:hasCRS` to the associated CRS in the RDF for a resource, so that it can be easily found, although there can be more than one CRS, corresponding to different policies.

Thus, the CRS is essentially an open service, which gives a view of URI equivalence: when presented with a URI, it returns all the URIs it considers equivalent. Note that it aims to avoid the mistake of creating a new URI; such an action would simply add further to the problem by being a new authority. It was also decided not to use `owl:sameAs`, since this is a much stronger assertion than the CRS is making. Of course, the knowledge bases themselves may still be using it where appropriate.

To generate knowledge for the CRS, the system uses the expected heuristic of string similarity (very conservatively), confirming the identification with the other relational uses, such as publication place (for papers), funding body (projects) and place of work (people), where these have already been the subject of equivalence identification. This means that to begin the generation of knowledge for the CRS, there is a 'cold start' problem, as there are almost no real URIs that are in common. This is achieved by string analysis of the titles of publication, and hence spreading to authors.

4.3 The RKB Explorer

For the user interface for exploring the knowledge bases we settled on a simple and very static presentation, with one window. This was primarily because a previous attempt based on a more dynamic style, for example allowing a choice of topics, was criticised as being non-intuitive by many of our users, few of whom had any knowledge of Semantic Web technologies.

A coherent and unified view of the underlying data sources is presented to the user, through the application of co-reference resolution algorithms and community of practise analysis to consolidate duplicate references and to identify related resources. Users may search and browse through the information available based around the four core themes which are present within the Explorer interface, namely People, Resilience Mechanisms, Publications and Projects. At any one time the top half of the interface window details a given instance of one of these types of resource, while the lower half lists those resources of each type which are related to the currently selected item.

This enables 'opportunistic' browsing, allowing a user to discover further information related to that which they are viewing, with associations determined by detailed domain specific analysis over the underlying knowledge bases. Utilising this interface, end users may be presented with data that is of interest or related to their field, potentially including work from areas or communities of which they were previously unaware.

Since the general problem of distributed queries remains unsolved, the system has to implement querying as appropriate for its environment. To gather all the

RDF related to a particular URI, it firstly resolves the URI (which includes any `owl:sameAs` in that store). It then looks up the URI in the associated CRS, which can be identified from the `coref:hasCRS` that was provided, and finds other, equivalent URIs. These can now be looked up in their CRSes, and the process continues, essentially to the fixed point. There is also the provision for other CRSes which are not directly associated with information sources to be consulted. It would be possible to consult all CRSes, but we consider this unnecessary. The CRSes we choose to trust for equivalence in these applications are either the original information providers, or ones we have chosen ourselves.

5 Conclusion

In order to ensure data providers take advantage of the benefits of linked data, the creation and maintenance processes that accompany the publishing of such data must not be a hindrance to those not willing to invest much time or effort.

We have presented a real-world Semantic Web application that is based on large-scale information from independent sources, using an ontology to mediate between them and rank resources when presenting consolidated results to users.

It provides a number of related applications, including the RKB Explorer, which gives an accessible and functional user interface. This, along with the usefulness of the knowledge resources have been extensively validated by the ReSIST Project partners, as reported in [4].

Since the system does not function by harvesting information into a common store, it is thus truly web-based. By employing resolvable URIs and distributed repositories to which queries can be fielded, we have created a real-world and scalable solution. Our system empowers linked data providers by providing a complete infrastructure for the curation and integration of large sets of RDF. The system will be shown to work with the existing sources of data that we have acquired and the various features that have been described in this paper will also be demonstrated. In addition, we will be happy to be provided with interesting RDF resources and use the system to provide a Linked Data site for them.

This work is supported by the ReSIST Network of Excellence, funded by FP6 under contract IST 4 026764 NOE.

References

[1] Bizer, C., Cyganiak, R., Heath, T.: How to publish linked data on the web (2007)
[2] Bizer, C., Cyganiak, R.: D2r server - publishing relational databases on the web. In: Proceedings of the 5th International Semantic Web Conference (2006)
[3] Harris, S., Gibbins, N.: 3Store: Efficient bulk RDF storage. In: Proceedings of the 1st International Workshop on Practical and Scalable Semantic Systems (2003)
[4] Glaser, H., Millard, I.C., Anderson, T., Randell, B.: ReSIST Project Deliverable D10: Prototype knowledge base. Tech. Rept., University of Southampton. Technical report (2007)