Research on Linked Data and Co-reference Resolution

Hugh Glaser and Ian C. Millard
School of Electronics and Computer Science, University of Southampton, UK
{ hg, icm }@ecs.soton.ac.uk

Won-Kyung Sung, Seungwoo Lee, Pyung Kim and Beom-Jong You
Korea Institute of Science and Technology Information, South Korea
{ wksung, swlee, pyung, ybj }@kisti.re.kr

Abstract

This project report details work carried out in collaboration between the University of Southampton and the Korea Institute of Science and Technology Information, focussing on an RDF dataset of academic authors and publications. Activities included the conversion of the dataset to produce Linked Data, the identification of co-references in and between datasets, and the development of an ontology mapping service to facilitate the integration of the dataset with an existing Semantic Web application, RKBExplorer.com.

Keywords: linked data, Semantic Web, co-reference resolution, ontology mapping.

1. Introduction

Prior to this project, The Korea Institute of Science and Technology Information (KISTI) had generated an RDF representation of metadata relating to some 26,000 publications and 160,000 authors from CiteSeer and several Korean domestic IT-related conferences, and defined an ontology identifying key concepts such as people, publications, journals, conferences, organisations and the relationships between them (Kang, 2006). However, in order to leverage maximal benefit from this resource, KISTI sought to make the data available and easily accessible to a wide range of tools, and to attempt to integrate this dataset with existing sources by identifying equivalent or similar identifiers in other repositories.

The project built on previous work within the School of Electronics and Computer Science at the University of Southampton (ECS) where there is a strong background in Semantic Web technologies, both at an infrastructure level and in creating tools to facilitate the visualisation and exploration of RDF datasets by end users. The CS AKTive Space (Shadbolt, 2004) and more recently RKB Explorer (Glaser, 2008) applications utilise underlying semantic datasets to assist users in the navigation of an information domain, identifying related resources and enabling the opportunistic discovery of relationships which may not have previously been known.

Of particular importance is the ability to perform co-reference resolution in the context of Linked Data. The work undertaken has centred around four main challenge areas –

1. Conversion of the KISTI dataset into a format suitable for publishing as Linked Data
2. Investigate issues relating to the interoperation of different ontologies
3. The identification of co-referent or duplicate identifiers within and between datasets
4. Integration of KISTI resources within the RKB Explorer application

These challenges are addressed in the following sections.

2. Creating Linked Data

The Open Linked Data initiative has in recent years provided a key focus on producing easily accessible resources on the Semantic Web. A number of significant datasets have been published, including numerous cross-linkages which enables the integration of these resources to form the emerging "Web of Data". By publishing information in line with Linked Data guidelines (http://linkeddata.org/docs/how-to-publish), the value and usefulness of that data can be greatly

enhanced through interlinking with other data sources, and can readily be consumed by a wide variety of tools and services.

Best practice prescribes that all non-information resources (eg real-world entities such as people, places or publications) are given URI identifiers which are resolvable using HTTP. When dereferencing such an identifier, the user or client application is redirected as appropriate to an information resource which provides a detailed description of that entity, either in a structured data format such as RDF, or in HTML for human interpretation.

Researchers at ECS have previously created a platform on which RDF datasets can easily be hosted, publishing information in a Linked Data compliant manner, in addition to providing SPARQL endpoints along with basic search and triple-browser facilities. Semantic metadata is imported from RDF/XML or Turtle documents into a 3store repository, providing the back-end storage and inference capabilities, while libraries written in PHP deal with publishing human and machine readable representations of the data for each URI identifier.

The KISTI dataset was received by ECS in the form of an ntriples dump from their repository. Firstly, triples which formed parts of the ontology were removed, as the ontology itself is already held in an OWL document. In order to be able to publish the information as Linked Data, all identifiers must be from a domain which can be resolved via HTTP. To achieve this, the ntriples dump was processed into Turtle, offering smaller file size and easier manipulation via @prefix statements, and checks performed to ensure that no blank nodes or hash-fragment identifiers remained. Finally, all triples relating to an identifier representing a concept of 'unknown' were removed, as these would create false linkages between a large number of resources, and the Semantic Web operates under an open world assumption. After these changes were made the dataset was loaded into the ECS hosting platform, with all URIs in the form http://kisti.rkbexplorer.com/id/....

Depending on the Accept headers passed as part of an HTTP request to resolve a kisti.rkbexplorer.com/id/xyz URI, a browser or client application is automatically redirected via an HTTP 302 response to either a human readable HTML rendering at /description/xyz, or RDF/XML semantic markup at /data/xyz. The ECS platform uses the SPARQL endpoint of the underlying 3store repository to dynamically generate and cache the concise bounded description of the URI which has been requested, returning a representation of the data in the appropriate format.

As a result, the KISTI dataset is now published in line with the Linked Data guidelines, providing easy access to the information contained within the repository. By simply making an HTTP request for a given identifier, a description containing all knowledge regarding that resource is returned. As all identifiers are resolvable in this way, one can navigate through and traverse between datasets in the Web of Data, in a manner analogous to that of navigating the World Wide Web by following hyperlinks between documents, using a variety of tools and applications.

2. Interoperation between Multiple Ontologies

The KISTI dataset is expressed in accordance with the 'KISTI Research Reference Ontology', whereas existing resources hosted by ECS are predominantly created utilising the AKT Portal Ontology. While both ontologies are fit for purpose and cover similar concepts of people, publications, organisations and similar entities, there are a number of structural differences beyond simple concept translation. One example of this is the level of indirection between a publication and it's authors: within AKT they are directly linked with the predicate akt:has-author, whereas KISTI has an intermediary 'CreatorInfo' object representing information about each author, with properties identifying a Person resource along with details of their affiliation and the ordering of authors for that given document.

The problems of interoperating with multiple ontologies are prevalent throughout the Semantic Web, as commonly agreed ontologies have been slow in their creation and uptake, and are yet to

achieve widespread adoption. As a result, ontology mapping technologies have been the focus of much attention within research communities, with many schemes attempting to create automatic analysis and translation tools. However, given a mapping schema, tools to actually perform translation between two ontologies are less developed.

To overcome these issues, one option would have been to rewrite the KISTI instance data so that it conformed with the other existing data expressed in the AKT Ontology. However, this is contrary to the spirit of the Semantic Web, and would have created a consistency problem with the maintenance and publishing of the KISTI data.

Instead, during this project we built on and extended an experimental mapping service created by researchers at ECS. This service, available at http://www.rkbexplorer.com/mapping/, takes an XML configuration document prescribing the steps required to translate from one ontology format to another. Simple predicate and class mappings can be defined, supporting entity re-writing and triple inversion as required within the standard service implementation. More advanced translations, such as dealing with the level of indirection outlined previously, are handled by custom functions implemented in PHP, called dynamically by the service as defined by the XML configuration.

Client applications can use this service to automatically resolve KISTI linked data URIs and translate their results into the AKT Ontology to achieve seamless interoperation with existing AKT datasets. The process of URI resolution, RDF parsing, and Ontology Mapping does incur additional overheads compared to directly querying a repository via a SPARQL endpoint, however we hope to make improvements to the performance of this prototype mapping service in due course.

3. Management and Identification of Co-reference information

One of the most overlooked problems to date is that of co-reference, or the multiplicity of identifiers, which can occur in two different ways on the Semantic Web. Firstly, when a single URI is incorrectly used to identify more than one resource, and secondly when multiple URIs identify the same resource. Both situations occur frequently when studying scenarios in which multiple datasets are combined or accessed in conjunction.

For an example of the first situation, a URI in a document repository may be used to identify a single author when, in fact, there are a number of people with the same name who are being incorrectly conflated into a single individual.

The second situation occurs much more frequently, as different datasets use their own URIs to identify the same resource. The success of the Semantic Web vision largely relies on the availability of large volumes of well curated and coherent data, over which software processes can perform analyses to evaluate data, form decisions, and base their actions. Clearly there is likely to be overlap and duplicity of information between repositories, particularly with people and publications, and hence there is a need for careful management of such equivalences.

The most prevalent way of dealing with 'duplicate' URIs that are deemed to be the same is to use the owl:sameAs predicate to link between them. However, the semantics of owl:sameAs dictate that all the URIs linked with this predicate have the same identity, implying that the subject and object must be the same resource. In addition to the widespread misuse of this predicate, the major disadvantage with this approach is that the two URIs become indistinguishable, even though they may refer to different entities according to the context in which they are used, for example a person who has changed institution.

The team at ECS have taken an alternative approach to the management of co-referent URIs within their semantic datasets. A unified view over several different knowledge bases with tens of millions of triples has been achieved by utilising a number of distinct, distributed 'Co-reference Resolution Service' (CRS) instances to separately maintain knowledge of URI synonymity (Glaser, 2009).

There are several benefits in keeping this knowledge separate 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 separately from the knowledge that is being created by the information provider. Indeed, different CRS providers may 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, and the policies used to populate it will be appropriate to the purposes. Some applications might wish to consider that two concepts are the same, while this may not be the case for another application using the same knowledge in a different context. 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, whereas in another application concerned with the textual output of an individual it may be thought of as same resource appearing in two different publication formats. Applications are free to utilise one or more CRSes as appropriate for the context in which they are operating.

Thus, the CRS is essentially an open and distributed service, which gives a view of URI equivalence: when presented with a URI, it returns all the URIs that it considers to refer to the same concept or resource. Methods are provided for CRS maintainers to easily add new identifiers, merge existing ones if they are found to be equivalent, or to split equivalence bundles where erroneous assertions have been made.

Having developed appropriate means to handle the representation and management of co-reference, we are still faced with an extremely challenging problem in the automatic determination of whether two URIs are referring to the same concept under any given context. Indeed, even human users find co-reference identification tasks difficult. For example, the DBLP publication repository holds information regarding Computer Science publications, and yet despite careful manual curation, inconsistencies can often be found through both the conflation of authors and the existence of duplicate or alternative representations of the same individual.

We have deployed a number of algorithms and heuristics which aim to identify co-referent identifiers in and between our datasets, within the experimental domain of modelling academic publications, projects, and related research activities. The general approach is two-fold. Firstly, various methods are used to identify co-reference candidates, which are pairs of URIs that are thought to potentially refer to the same resource. Secondly, a number of different co-reference analysis techniques are applied as appropriate to the lists of candidates to evaluate whether they are indeed equivalent.

Typical heuristics for finding candidates may include publications or organisations with similar titles or names, common co-authorship of academic publications, more complex graph matching, or specific sub-graph inspection around already known co-referenced entities. These can then be analysed using techniques such as direct equality of normalised strings, 'fuzzy' matching of specific predicate values, specific comparison of person names, or post-analysis of graph analyses. These different approaches may be applied as appropriate to the context of the information being processed, as prior knowledge of the domain and ontology or ontologies is required. Some techniques are particularly applicable to the 'cold-start' scenario, where no existing co-reference resolution has been performed, whereas others are more suited to an iterative or incremental on-demand application.

It should be noted that when performing co-reference analysis it is important to be cautious, and to use algorithms in such a way that there is high confidence that the co-reference is correct. The repercussions of asserting incorrect co-references may be significant, as other analyses or applications may build upon these and produce further false deductions.

There are also potential problems when encountering 'dirty' data, in which resources have been incorrectly conflated at source, or when values are encountered which are not as expected given the ontology. Conflations are extremely difficult to resolve as it requires modifying the original data to separate incorrectly merged properties from the two or more different entities.

4. Integration with RKBExplorer.com Application

The RKB Explorer Application is an interface which has been developed to provide a synthesised and coherent view over various underlying Linked Data repositories. Featuring an intentionally simple display and interaction model, the RKB Explorer is designed for non-expert users, and does not expose any of the internal semantic representations. The main focus is assisting users to explore an information domain, highlighting related resources and other interesting links. At any given time, the upper half of the display provides details of the resource currently being viewed, while the lower half identifies additional resources which have been deemed relevant by means of ontologically informed analyses performed on the dataset.

In order to integrate the disparate Linked Data sources, many instances of the Co-reference Resolution Service are used to store and represent knowledge concerning equivalent identifiers between different data sets. Internally, given a URI for a resource to be displayed, the RKB Explorer application queries the CRSes as required to find all duplicate identifiers. Information for each of these equivalents is retrieved, either via a direct SPARQL query to the relevant endpoint if from a domain known within the system configuration, or by HTTP resolution of the Linked Data URI. The resulting information is combined, before being processed for display to the user.

As outlined in Section 2 above, we have created an ontology mapping service capable of translating information represented in one vocabulary into another as required 'on-the-fly'. By extending the configuration options within the RKB Explorer application, we have been able to simply define the KISTI dataset as an additional resource, to be accessed via HTTP URI resolution, but additionally passed through the mapping service to convert the data returned into the AKT Ontology.

The resulting effect is that knowledge from within the KISTI dataset is seamlessly integrated with that of other existing datasets, permitting users to (unknowingly) traverse these resources within the RKB Explorer application as if they were one coherent information source. The same mechanisms can now be used to integrate knowledge from others datasets and other ontologies, such as DC Terms and SKOS. Furthermore, the synthesis and combination of datasets often provides a more comprehensive representation of a given person, publication, or related entity, resulting in a view which is greater than the sum of the constituent parts.

5. Conclusions

This project has addressed a number of issues relating to both Linked Data and Co-reference Resolution. The team at ECS have introduced and demonstrated the topics covered in this report to researchers at KISTI, facilitating technology and knowledge transfer wherever possible.

Through the provision of easy access to information published as Linked Data, and the application of co-reference analysis and CRS utilities, a wide variety of disparate and previously disconnected datasets can be used in unison. The KISTI dataset is no longer a stand-alone resource, and the exploitation of co-referent identifiers enables the traversal between and interoperation of information within that dataset to additional data about a given concept or entity from other external datasets.

References

Glaser, Hugh, Ian C. Millard, and Afraz Jaffri. (2008) RKBExplorer.com: A Knowledge Driven Infrastructure for Linked Data Providers. European Semantic Web Conference, 2008.

Glaser, Hugh, Afraz Jaffri, and Ian C. Millard (2009). Managing Co-reference on the Semantic Web. LDOW 2009.

Kang, In-Su, Hanmin Jung, Seungwoo Lee, Pyung Kim, and Won-Kyung Sung. (2006) Semantic Web Ontology and Inference for Research Community, Korea Computer Congress 2006 (in Korean).

Shadbolt, Nigel R., Nicholas Gibbins, Hugh Glaser, Stephen Harris and m.c. schraefel (2004) CS AKTive Space or how we stopped worrying and learned to love the Semantic Web. IEEE Intelligent Systems, 19 (3). pp. 41-47