All About That – A URI Profiling Tool for Monitoring and Preserving Linked Data

Rob Vesse, Wendy Hall and Leslie Carr
Intelligence, Agents & Multimedia Group, School of Electronics & Computer Science, University of Southampton
Southampton, SO17 1BJ
{rav08r,wh,lac}@ecs.soton.ac.uk

Introduction

Link Integrity is a term used to describe whether the links within a system are valid and working. Researchers in this area aim to ensure the validity of links and maintain links over time correcting failures as they are detected. From the early days of hypermedia systems such as Microcosom [1] the integrity of links in the system and the data as a whole has been a problematic issue. As the Semantic Web grows larger we expect that it's focus on linked data will make link integrity one of the main research issues.

While many of the large datasets that currently comprise the bulk of the Linked Data Web such as DBPedia are well maintained it is likely that in the future many small datasets will appear which are poorly maintained and lead to broken links as the Linked Data Web grows. This presents a problem since if we wish to reason across this data and it's no longer there what action do we take? Given this problem we are beginning to explore how ideas from link integrity in hypermedia can be applied to the Semantic Web.

Related Work

There is a large body of existing work on link integrity for hypermedia that we can draw on for ideas and approaches to apply to the Semantic Web. Work by Davis on Microcosom [2] and Kappe on HyperG [3] has shown that it is possible to enforce link integrity in small tightly controlled systems. The limitations of this are that such approaches simply don't scale to Web scale since they typically rely on storing link information separately from data and restricting the ability to modify this. However there are approaches such as Phelps and Wilensky's Robust Hyperlinks [4] and Harrison and Nelson's Opal [5] which show potential for achieving Web scale and applying corrections to links Just-in-Time (JIT).

In our current work we are taking a approach based on the replication and versioning style approach used by Moreau and Gray [6] and Veiga and Ferreira [7,8]. In their work link integrity is maintained by allowing end users to preserve the web pages they are interested in by replication with the system maintaining the replicas as long as users are linking to them. In the Semantic Web field there are systems like Volz et al's Silk [9] which can be used to calculate links according to rules and periodically recalculate these links to ensure they remain valid. The limitations of this is that the link integrity is not maintained on-the-fly but at periodic intervals such that links may become broken between link calculations.

All About That

All About That (AAT) is a URI profiling tool which can be used to monitor and preserve Linked Data that a user is interested in.

Definition 1 - A URIs profile is the transformed and annotated form of the RDF retrievable from the URI such that the temporality and provenance of the triples contained therein are inferable from the profile.

Each triple in the RDF retrieved from the URI is transformed into an annotated form based upon the RDF reification mechanism. Given a triple like the following it is transformed as shown in Figure 1

```
ex:Dog rdf:type ex:Animal
```

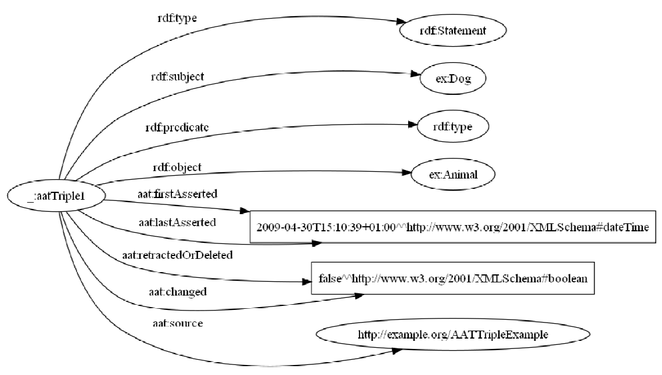

Figure 1 – Transformed and Annotated Triple

This format allows us to preserve the required information and minimise our storage requirements over time as opposed to storing the original RDF directly with some metadata. If the data changes regularly it quickly becomes far more efficient to store data in this form over storing copies of the original data.

A user can use the AAT interface to browse the contents of the profile, see how it has changed over time and to view versions of the RDF as it appeared on a given date. In essence in it's current form AAT is an RDF versioning tool.

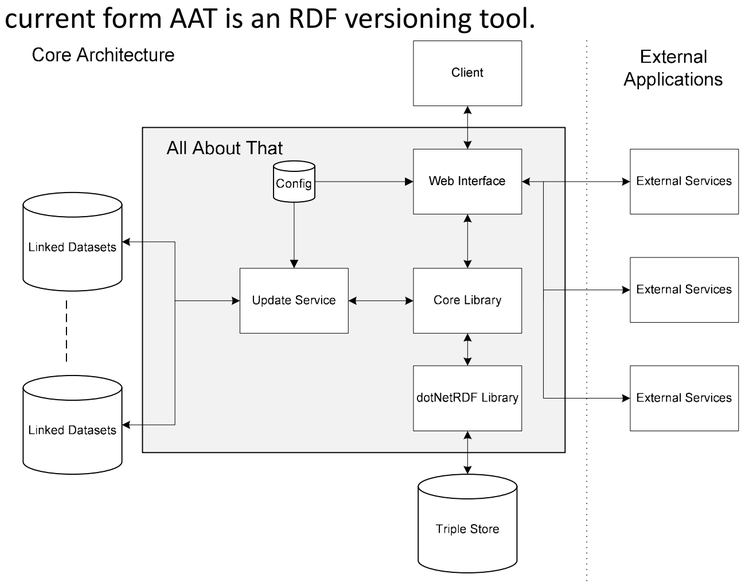

Figure 2 – All About That System Architecture

As can be seen in Figure 2 one of the key components of AAT is the Update Service which regularly retrieves the RDF available at a given URI and updates the local profile. This regular update of the profile allows us to detect changes in the RDF and to retain sufficient information to version over the RDF. AAT attempts to detect four types of changes in RDF:

1. New triples – completely new triples introduced into the RDF
2. Changed triples – triples where the object of the triple has changed and the subject-predicate pair of the triples allows only one value
3. Missing triples – triples which are no longer present in the RDF but have been seen in a recent update
4. Retracted/Deleted triples – triples which are no longer present in the RDF and haven't been seen for some time

Users can view these changes in the AAT interface or they can retrieve the changes as RDF encoded using the Talis ChangeSet ontology [10]

Figure 3 – Change Report in Web Interface

Following Linked Data best practises [11] all data gathered by AAT is republished via a variety of dereferenceable URIs which allow the retrieval of the following information:

- Profile Contents
- Profile Export (the triples in their original form)
- Profile Versions (an export as of a specific date and time)
- Profile Changes
- Profile Change History

BBC Programmes Demo

As can be seen in Figure 2 it is envisaged that data from AAT will be consumed by external services to provide rich Semantic Web applications. Currently we are developing a demonstration application which uses an instance of AAT configured to monitor the BBC Backstage Programmes data. This data is composed of descriptions of programmes that are broadcast by the BBC, as new episodes of programmes appear AAT detects these changes and reports them in it's change reports. The demonstration application consumes this data to provide a feed of information about new episodes of programmes.

A faceted browsing interface is provided to the user to allow them to browse through this feed of information and access Linked Data.

Future Work

In the future we plan to look at making AAT preserve linked data in a much more linked data oriented way. When a user asks for the profile of a given URI it should be easily possible to leverage semantic data sources such as SPARQL endpoints, Sindice's URI lookup and Cache API [12] and SameAs.org [13] to find other sources of information about that URI. From this a profile composed of multiple sources could easily be created and would allow you to preserve much more information about a URI.

The fundamental aim of this research is to be able to provide a service that can be used to maintain the integrity of linked data by allowing anybody to request a given URI from the service and have it return the RDF for that URI regardless of whether that URI is still directly accessible on the web. It should be able to be deployed in a distributive fashion in order to scale sufficiently to make it viable for widespread use on the Web and for Semantic Web applications.

References

[1] A. M. Fountain, W. Hall, I. Heath, and H. C. Davis. Microcosm: an open model for hypermedia with dynamic linking. In Hypertext: concepts, systems and applications, pages 298-311, New York, NY, USA, 1992. Cambridge University Press.
[2] H. Davis. Data Integrity Problems in an Open Hypermedia Link Service. PhD thesis, University of Southampton, November 1995.
[3] F. Kappe. A scalable architecture for maintaining referential integrity in distributed information systems. Journal of Universal Computer Science, 1(2):84-104, 1995.
[4] T. A. Phelps and R. Wilensky. Robust hyperlinks: Cheap, everywhere, now. In Digital Documents: Systems and Principles, pages 514-549. Springer, 2004.
[5] T. L. Harrison and M. L. Nelson. Just-in-time recovery of missing web pages. In HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia, pages 145-156, New York, NY, USA, 2006. ACM.
[6] L. Moreau and N. Gray. A Community of Agents Maintaining Links in the World Wide Web (Preliminary Report). In The Third International Conference and Exhibition on The Practical Application of Intelligent Agents and Multi-Agents, pages 221-235, London, UK, Mar. 1998. http://www.ecs.soton.ac.uk/~lavm/papers/gcWWW.ps.gz.
[7] L. Veiga and P. Ferreira. Repweb: replicated web with referential integrity. In SAC '03: Proceedings of the 2003 ACM symposium on Applied computing, pages 1206-1211, New York, NY, USA, 2003. ACM.
[8] L. Veiga and P. Ferreira. Turning the web into an effective knowledge repository. ICEIS 2004: Software Agents and Internet Computing, 14(17), 2004.
[9] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Silk – A Link Discovery Framework for the Web of Data. In 2nd Linked Data on the Web Workshop (LDOW2009), 2009.
[10] S. Tunnicliffe & I. Davis, Changeset Ontology, http://purl.org/vocab/changeset
[11] C. Bizer, R. Cyganiak, and T. Heath. How to publish linked data on the web, 2007. http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial
[12] G. Tummarello et al, Sindice.com, http://www.sindice.com
[13] H. Glaser and I. Millard, SameAs.org, http://sameas.org