All About That - A URI Profiling Tool for monitoring and preserving Linked Data

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

ABSTRACT

All About That (AAT) is a URI Profiling tool which allows users to monitor and preserve Linked Data in which they are interested. Its design is based upon the principle of adapting ideas from hypermedia link integrity in order to apply them to the Semantic Web. As the Linked Data Web expands it will become increasingly important to maintain links such that the data remains useful and therefore this tool is presented as a step towards providing this maintenance capability.

1. 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 days of early Hypermedia systems such as Microcosm [3] the integrity of links in the system and the data as a whole has been a problematic issue. There have been many attempts to implement mechanisms for either enforcing link integrity or repairing links in the event of failure such as Ingham et al's W3Objects [7], Phelps and Wilensky's Robust Hyperlinks [11] and Harrison and Nelson's Opal [5]. Work carried out by people like Davis [2] on Microcosm and Kappe [9] on HyperG [10] has shown that is it feasible to implement solutions for small tightly controlled systems such that link integrity can be guaranteed. Despite this none of the various approaches suggested has ever been widely adopted mostly due to the fact that they simply don't scale in the face of massive systems like the World Wide Web or would require a significant re-engineering of the Web's architecture.

Despite this there is still work to be done in link integrity, particularly with regards to its applicability to the Semantic Web. Since the Semantic Web is composed primarily of linked data it is more important than ever that links are reliable and as such link maintenance is one of the active topics of research in the linked data community as discussed by Bizer et al [6]. While many of the large datasets that currently comprise the bulk of the Linked Data Web such as DBPedia[1] 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 becomes mainstream. 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 look into how ideas from link integrity in hypermedia can be taken and applied to the Semantic Web.

2. PROPOSAL

A replication and versioning approach to maintaining link integrity is taken for our initial experiments. As in the work of Veiga & Ferreira [13, 14] integrity is maintained by preserving the linked data that the user is interested in. It is important to note that the integrity of links within the data is not maintained/preserved but rather that the linked data which is of interest to the user is preserved, in essence our prototype is an RDF Versioning tool.

All About That is a prototype tool which implements this concept which we term URI Profiling (see Definition 1). Use of replication means that the data which the user is interested in can be preserved such that even if the source of that information were to be removed from the web the user still has access to that data. Not only do they have access but they have the ability to access versions of the data such that they can work with the data as it was on a particular date and they can monitor how the data has changed over time.

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

2.1 Implementation & Features

All About That (AAT) is implemented as a combination of a Web based user interface and a background service. The Web interface allows users to create and view the profiles of URIs that they are interested in, while the background service periodically retrieves the RDF from the URIs and updates the profiles appropriately. Profiles are named RDF

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

graphs composed of the original triples transformed into annotated triples based upon the RDF reification mechanism, the original RDF is never stored directly since it can be easily recreated from the transformed RDF. The system stores the RDF in a SQL based Triple Store and automatically generates dereferenceable URIs for profiles so that users can access the raw RDF data that comprises the profile if they wish.

Using the annotations that are applied to the triples it is possible to compute the changes that have occurred in the source RDF between the tools periodic retrievals of the RDF. The tool can also compute what was contained in the source RDF on a particular date by extracting the triples from the profile that it knows were present on that date.

Every time a profile is created/updated AAT computes a Changeset[2] for the profile and stores this as a named graph in the Triple Store. Profiles are linked to both their most recent changesets and via a version history graph to all their changesets, all of this data is republished via dereferenceable URIs in order to follow linked data best practises [1].

2.2 Usage Scenario

It is envisaged that AAT will be used as a service on the Linked Data Web with applications being built on top of it just as with many of today's datasets. One prototype application currently under development is using the data being output by the BBC Backstage[3] project and attempting to present it in a useful and interesting way to users. The intention is that by using AAT to monitor the changes in the programmes RDF it is easily possible to spot when new episodes of a given programme are broadcast. A user of the proposed service would be able to visit the website and see at a glance which of their favourite shows has recently broadcast new/repeated episodes and get access to web resources and linked data related to that episode.

3. FUTURE WORK

As it stands All About That is a working platform that can be used to build other applications on top of and our current focus is developing applications like the one described in the previous section to provide a full demonstration of what could be done with such a system.

Further developments to All About That will involve taking further ideas from hypermedia link integrity to expand the capabilities in the system and work towards providing actual link integrity rather than just data preservation and monitoring to the Semantic Web. One of the most promising ideas is to apply the concept of JIT (Just-in-Time) Resolution from systems like Opal [5] to build URI profiles where multiple sources are involved. It is intended that the linked data nature of the Semantic Web be exploited by automatically locating other sources of information about the URI the user is interested in. One way this could be done is JIT-like in that it would involve querying semantic search engines like Sindice [12] and SPARQL endpoints for things which are `owl:sameAs` the URI of interest. Another promising approach is to leverage Coreference Resolution Systems like those of Jaffri et al [8] which are used in systems like RKBExplorer [4] and can be used to find out URIs which are considered equivalent to a given URI. Information from a CRS has more value than that from a pure JIT lookup approach since it will have been computed by well designed algorithms whereas information from search engines will include peoples hand-written RDF documents which may assert invalid equivalences.

[2] Changeset Ontology http://purl.org/vocab/changeset/schema

[3] http://welcomebackstage.com/

4. REFERENCES

[1] 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.

[2] H. Davis. *Data Integrity Problems in an Open Hypermedia Link Service*. PhD thesis, University of Southampton, November 1995. http://eprints.ecs.soton.ac.uk/6597/.

[3] 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.

[4] H. Glaser and I. Millard. Rkbexplorer.com: Anatomy of a semantic web application. In *KISTI Workshop*, December 2008.

[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] T. Heath, M. Hepp, and C. Bizer. Linked data - the story so far. 2009.

[7] D. Ingham, S. Caughey, and M. Little. Fixing the "broken-link" problem: the w3objects approach. *Comput. Netw. ISDN Syst.*, 28(7-11):1255–1268, 1996.

[8] A. Jaffri, H. Glaser, and I. Millard. Uri identity management for semantic web data integration and linkage. In *3rd International Workshop On Scalable Semantic Web Knowledge Base Systems*. Springer, 2007.

[9] F. Kappe. A scalable architecture for maintaining referential integrity in distributed information systems. *Journal of Universal Computer Science*, 1(2):84–104, 1995. http://www.jucs.org/jucs_1_2/a_scalable_architecture_for.

[10] F. Kappe, K. Andrews, J. Faschingbauer, M. Gaisbauer, M. Pichler, and J. Schipflinger. *Hyper-G: A new tool for distributed hypermedia*. Institutes for Information Processing Graz, 1994.

[11] T. A. Phelps and R. Wilensky. Robust hyperlinks: Cheap, everywhere, now. In *Digital Documents: Systems and Principles*, pages 514–549. Springer, 2004.

[12] G. Tummarello, R. Delbru, and E. Oren. Sindice. com: Weaving the Open Linked Data.

[13] 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.

[14] L. Veiga and P. Ferreira. Turning the web into an effective knowledge repository. *ICEIS 2004: Software Agents and Internet Computing*, 14(17), 2004.