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 Microcosm [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 Microcosm [2] and Kappes on HyperG [3] has shown that it is possible to enforce link integrity in small tightly coupled systems. The distinction here is 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 and versioning. As long as the URI is a link to some resource it is in these pages that the links are maintained and the pages are regularly replicated and updated. This allows a user to browse pages of linked data and retrieve the changes as RDF encoded using the Talis Changeset ontology [10].

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 retrieveable from the URI such that the temporality and provenance of the triples contained therein are inferable from the profile.

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 it’s current form AAT is an RDF versioning tool.

There are many RDF datasets that are current maintained and can be visualised as RDF in the AAT interface. The following RDF data sets that were downloaded from AAT are given below:

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. Unchanged Triples – triples which are no longer present present in the RDF and haven’t been seen for some time.

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 to users.

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, Sindic’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 returned 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 distributed fashion in order to scale sufficiently to make it viable for widespread use on the Web and for Semantic Web applications.

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

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. 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 it’s current form AAT is an RDF versioning tool.

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.

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 and versioning. As long as the URI is a link to some resource it is in these pages that the links are maintained and the pages are regularly replicated and updated. This allows a user to browse pages of linked data and retrieve the changes as RDF encoded using the Talis Changeset ontology [10].

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 retrieveable from the URI such that the temporality and provenance of the triples contained therein are inferable from the profile.

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 it’s current form AAT is an RDF versioning tool.

There are many RDF datasets that are current maintained and can be visualised as RDF in the AAT interface. The following RDF data sets that were downloaded from AAT are given below:

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. Unchanged Triples – triples which are no longer present present in the RDF and haven’t been seen for some time.

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 to users.

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, Sindic’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 returned 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 distributed fashion in order to scale sufficiently to make it viable for widespread use on the Web and for Semantic Web applications.

References