UNIVERSITY OF Southampton
School of Electronics and Computer Science

Preserving Linked Data on the Semantic Web by the application of Link Integrity techniques from Hypermedia

Rob Vesse, Wendy Hall and Les Carr
{rav08r,wh,lac}@ecs.soton.ac.uk
27 April 2010

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Link Integrity

Aims to ensure that a Link is valid

Link is dereferenceable and goes to the intended content

Semantic Web introduces additional issues

Co-reference

Identity & Meaning

Two main types of Solution

- Prevention & Maintenance
- Recovery

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Link Integrity in Hypermedia

Open Hypermedia

Robust Hyperlinks (Phelps & Wilensky 2004)

Opal (Harrison & Nelson 2006)

Replication & Versioning

Community of Agents (Moreau & Gray 1998)

RepWeb (Veiga & Ferreira 2003)

Memento (Sompel et al 2009)

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Link Integrity for the Semantic Web

Co-reference/Identity

CRS (Jaffri et al 2007) – Compute co-references and republish

Okkam (Bouquet & Stoermer 2008) – Standardise URIs across applications

Maintenance

Silk Framework (Volz et al 2009) – Compute links between datasets based on similarity metrics

DSNotify (Haslhofer & Popitsch 2009) – Monitors datasets to spot and repair broken links

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Applying Recovery to the Semantic Web

Useful data sources for recovery already available

Sindice Cache

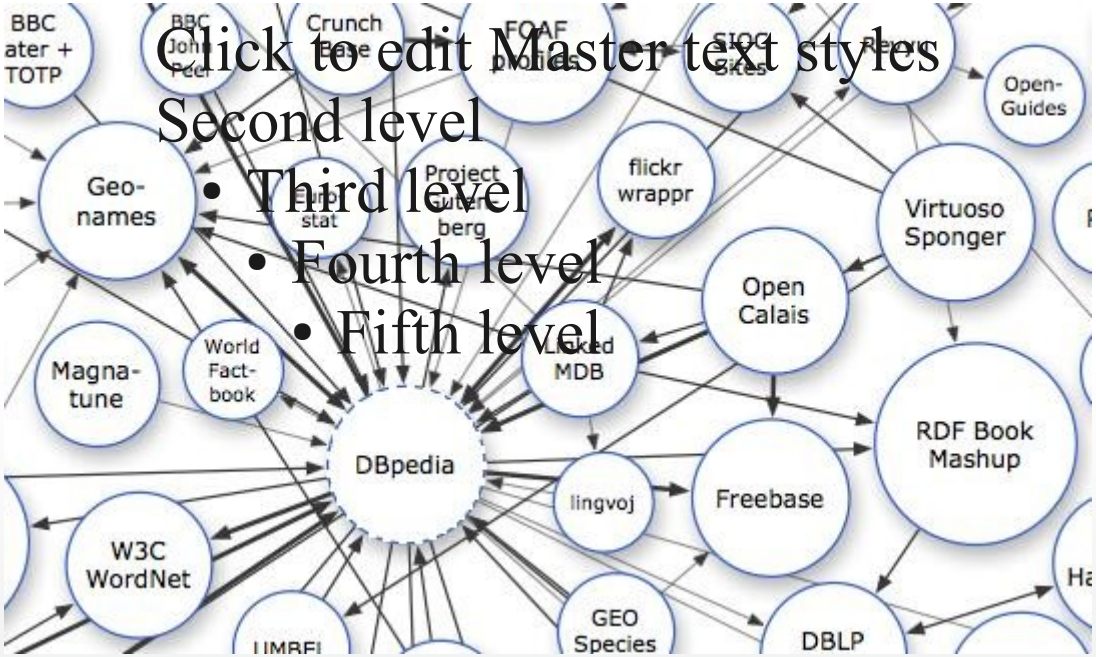
Data Warehouses e.g. LOD Cloud, Uberlic.org

'Authoritative' linking hubs e.g. DBPedia

Co-reference services e.g. SameAs.org

Possible to exploit the heavy interlinking of the Semantic Web

Exploiting Interlinking



- Lots of other datasets refer to its URIs
- Use these linkages to find relevant data to replace the lost data

Exploiting Interlinking – What if DBPedia disappeared?

- owl:sameAs and rdfs:seeAlso are useful links to follow
- DESCRIBE against other datasets SPARQL endpoints also useful for recovering data

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Expansion Algorithm

In essence a crawler which follows links and uses user definable data sources to discover linked data about a URI

Works even if the URI itself is unresolvable

User can define data sources and services to use using simple RDF vocabulary

- voID with a couple of additions to control the algorithm
- Otherwise defaults to Sindice Cache, DBPedia and SameAs.org

Trivially parallel => easily scalable

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Expansion Algorithm

Returns an RDF dataset, each URI we retrieve data from has a corresponding named graph in the dataset

Means consuming applications can discard data from sources they don't trust/unaware of

Allows consuming applications to determine how many sources assert a particular statement

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Applying Preservation to the Semantic Web

Provide end users the means to preserve the Linked Data they are interested in

Allow them to monitor it over time to preserve changes in the data

View change history of data over time

Republish the data so other people can use it

UNIVERSITY OF Southampton
School of Electronics and Computer Science

All About That (AAT)

Uses the expansion algorithm to retrieve an RDF dataset about the URI the user wants to preserve

'Smushes' the dataset to a single graph while preserving data about the sources which assert each triple

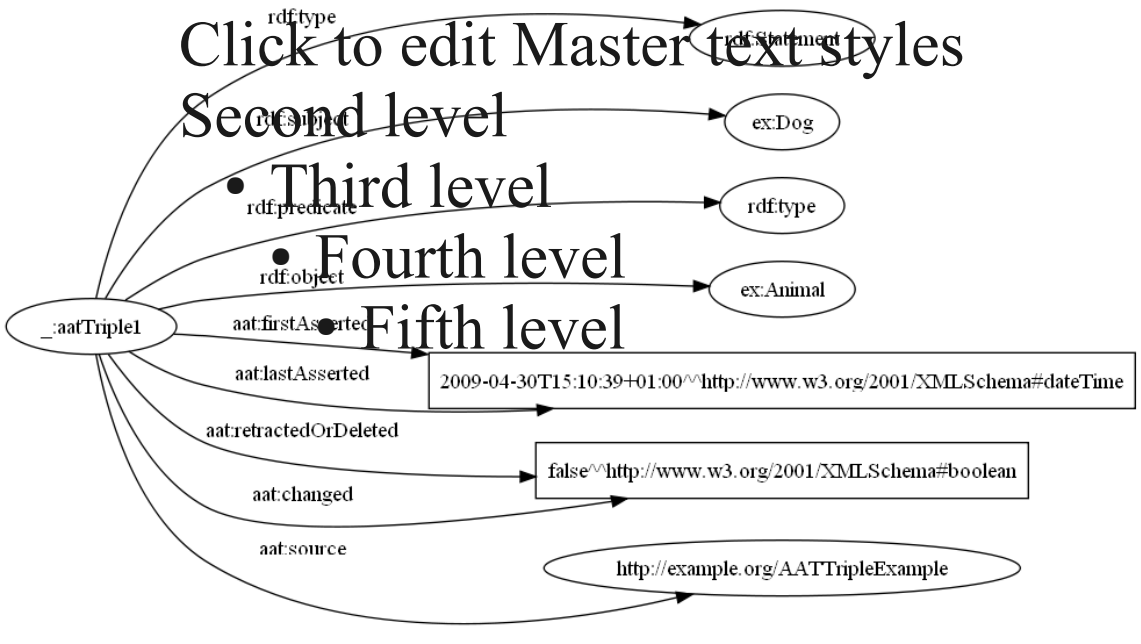
Preserves graphs by transforming the original graph into an annotated form

Use this as opposed to named graphs as want to annotate at the triple rather than graph level

Initial data bloat is a trade off against decreased storage needs over time

UNIVERSITY OF Southampton
School of Electronics and Computer Science

All About That (AAT)



Triple transformed and annotated using the AAT Schema

· Reification is the basic unit of preservation
·Store when we first and last asserted each triple
·Store source(s) for each Triple

·Each triple in the RDF Graph to be preserved is transformed into this form
·Transformations of all Triples in a Graph form a named graph in AATs Triple Store

UNIVERSITY OF Southampton
School of Electronics and Computer Science

All About That (AAT)

Data is monitored over time allowing Change Reporting and Versioning

Regularly retrieve the linked data for a URI and compare against local annotated data and update

Compute the changes and express using Talis ChangeSet Ontology

End users can ask to see the data as the system perceived it to be at a given date and time

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Future Work

Produce larger set of experimental results

Detailed analysis of the effectiveness of the expansion algorithm i.e. precision and recall

Improving the expansion algorithm

Integration with term based search

Integration with other link maintenance frameworks e.g. Silk, DSNotify

Investigate distributing the algorithm for improved scalability

UNIVERSITY OF
Southampton
School of Electronics
and Computer Science

Questions?