Inter-repository Linking of Research Objects with Webtracks

Shirley Crompton (1), Brian Matthews(1), Cameron Neylon(2), Simon Coles (3)

(1) e-Science Centre, Science and Technology Facilities Council

(2) ISIS, Science and Technology Facilities Council

(3) School of Chemistry, University of Southampton

**1. Introduction and Motivation**

Modern research requires the co-ordination of a large number of digital objects and communication mechanisms. A typical research life cycle involves creating experimental or observation data using multiple facilities; refining raw data into derived data to test hypotheses; communicating and presenting finding in various formats. Each stage of this process typically involves various support systems such as repositories, websites and data archives. As each system manages its holdings independently, this often leads to information silos. With the lost of context, the mutual relationship between objects within the original research network is broken. To maximize the value of a piece of scientific work, the connections between its various outputs need to be exposed as they capture the evolution of the research in context. The creation of a linked web of data will enable information providers to improve access and track usage of their contents. Online repositories storing more dynamic objects such as data and grey literature, can also offer a more comprehensive picture of a piece of research via backward and forward citations to inter-link different research outputs in context.

The emergent Linked Data (1) movement provides a mechanism for linking related information on the web through the use of structured, semantic information and HTTP URIs. This and other established techniques such as OAI-MPH (2) provide both frameworks and tools to publish data for linking. Other protocols and technologies like Storelink (3), TrackBack (4), Salmon (5) and Semantic Pingback (6), all offer varying degrees of support for the creation and notification of citation links targeting specific environments. In the research information management sector, prior work by Claddier (7) and Storelink have identified the need for a flexible and light-weight peer-to-peer protocol to enable cross linking of diverse forms of research outputs.

In this paper, we present a Restlet application which facilitates the implementation of the InteRCom protocol (8) developed by the JISC-funded Webtracks project (9). Webtracks can be used to propagate typed links between repositories that:

* captures the relationship between research resources in context;
* supports both forward and backward citation;
* accepts arbitrary annotations for describing links;
* permits the linking of diverse types of digital objects within a discipline;
* has a decentralized model that does not rely on a central registry;
* uses the simple but universal HTTP protocol for notifying and manipulating citation links.

In the remainder of this paper, we give a synopsis of the protocol and describe its usage (Section 2). We then present an architecture overview of the Webtracks application and its key features (Section 3), followed by a walk-through of an exemplar based on the ISIS ICAT Data Catalogue (Section 4).

**2. Inter Repository Communication Protocol (InteRCom)**

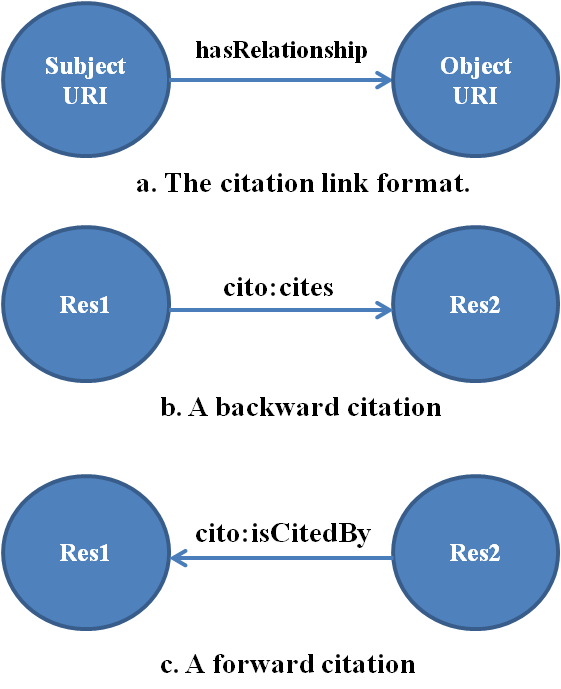


Figure 1 Examples of InteRCom Citation Links.

InteRCom is a general application layer protocol built on HTTP REST for linking digital data resources of any type. It permits the use of arbitrary RDF to capture the relationship between two resources. Each citation link is an RDF triple (Figure 1a). The link propagator may use any vocabulary or term to describe the context of a link. Thus, it is possible to describe backward (b) and forward (Figure 1c) citations through the judicious choice of link properties. Building on this concept, a set of semantically annotated links between research resources identified by their unique HTTP URIs can be aggregated to form a graph of citation and provenance that traverses multiple repositories.

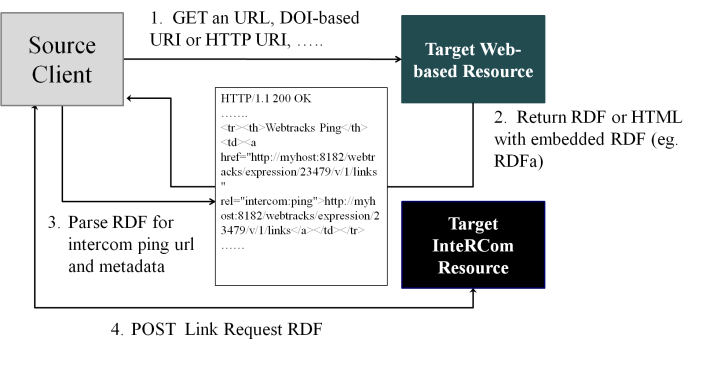


Figure 2 A Simple Use Case Illustrating Usage of the InteRCom Protocol.

Similar to StoreLink which extends Trackbacks, InteRCom () is a two stage peer-to-peer protocol that does not require a centralised registration service. It has the notion of a Sender and Receiver components which are responsible for the posting and receipt of citation notifications or ‘pings’. Stage 1 (steps 1-3) of the protocol involves getting Target metadata and discovering its InteRCom ping endpoint for receiving link notification. Stage 2 (step 4) is sending the link RDF to the Target’s advertised endpoint using a HTTP Post request which may optionally include metadata on the Source object. InteRCom does not constraint the format of the RDF post message as long as it is valid and well-formed.

**3. The Webtracks Application**

shows the general architecture of the Webtracks Restlet Java application. It provides a flexible foundation with extension points for developers to fast track custom implementation of the InteRCom protocol using the rich features provided by the underlying Restlet Framework (10). The design is independent of and can work with different security regimes. An implementation can leverage existing Restlet components/plugins or implement its own custom security filter to comply with the data host’s security regime. Restlet supports popular security protocols such as HTTP Basic and Digest, Amazon S3, OAuth, HTTPs etc.

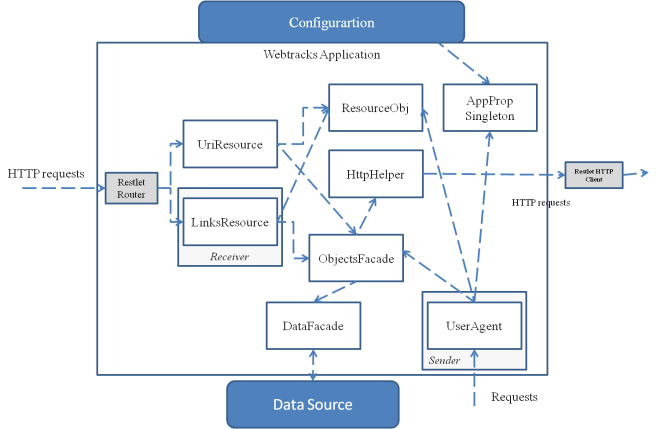
****

Figure 3 Architecture View of the Webtracks Restlet Application.

The Webtracks design also leverages the Linked Data principle of identifying web resources with “cool” (11) opaque HTTP URIs which can be de-referenced to different representations of the linked resources. Webtracks has an HTTP Resource-oriented Architecture (ROA) that exposes data rather than internal algorithms as in other Pingback protocols. Instead of having a dedicated Receiver to processes pings, Webtracks abstracted the functionalities into the Link container resource object (LinksResource, see Figure 3). gives a model of the resources or domain data exposed by the Webtracks REST web services. In the model, a link is defined as the child of a digital research object. As the links are first class objects on the web, they can be assessed and manipulated directly via uniform HTTP methods. There are several advantages of using a ROA approach. Firstly, the hierarchy can be easily accommodated within a repository’s existing resource model (as in the case of our exemplar). Secondly, it simplifies client interactions with the application and improves connectivity of citation resources on the web. For instance, repositories can provide hypermedia representations that contain HTTP links to the referenced resources; other applications can create mashups of the links resource and for link propagators to update a link directly using HTTP methods. Thirdly, repositories may also delegate the generation of RDF representation of the digital research objects to existing link data services via server re-direct or URL-rewrite mechanisms.

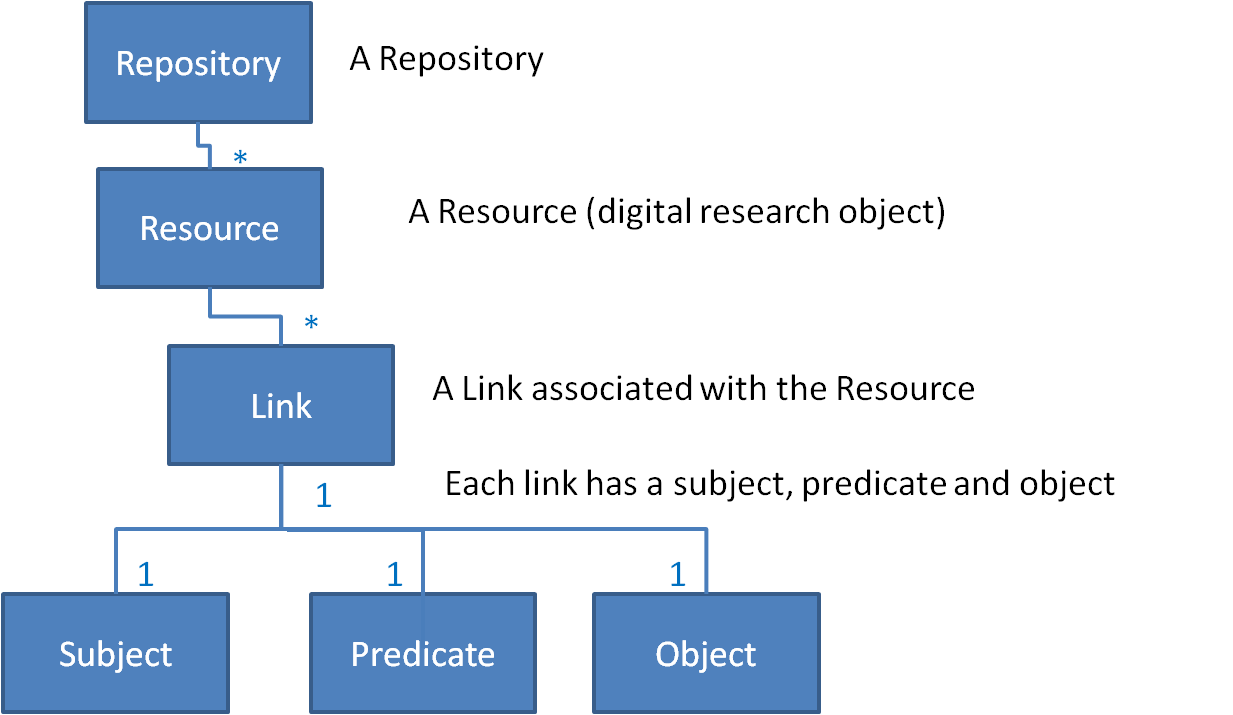


Figure 4 A Simple Resource Hierarchy for Digital Research Objects and Citation Links.

**4. The Webtracks-ICAT Data Catalogue**

ICAT[[1]](#footnote-1) is an open source metadata management system designed for large scientific facilities to manage its holding of experimental data. It also provides a mechanism to link all aspects of the research chain from proposal through to publication. At ISIS[[2]](#footnote-2), all experimental data are produced, captured and catalogued into ICAT along with the metadata about the experiment. ICAT uses the Core Scientific Metadata Model (12) which describes each Investigation associated with a particular Experiment on a sample generating data, and the associated data are then mapped to the Investigation with the appropriate parameters. ISIS registers a DOI (13) for each experiment and encourages researchers to cite in their publications as the facility is keen to capture forward citations to track usage of its data resources. Since Investigation is the core entity which already has a unique HTTP URI, this can be used as the parent resource that contains our Webtracks Link resource.

In the implementation, the Investigation resource generates RDF and HTML with RDFa representations in response to the HTTP Get operation and the requested content type. This supports the InteRCom Stage 1 Get metadata operation. As described in Section 3, the Links resource is the *de facto* InteRCom Receiver. It has the dual functions of providing representations of its state in response to HTTP Get; it also processes InteRCom (Stage 2) pings and link updates requests via the HTTP Post, Put and Delete operations. The Webtracks Architecture documentation gives detailed information on the operations (14).

In addition to the ICAT exemplar, we have also built a second exemplar based on the STFC ePublications Archive[[3]](#footnote-3) (14). Using the two exemplars, we were able to demonstrate the posting of citation links between the two repositories. Figure 5 shows HTML representation of an ICAT Investigation resource (a) and its subordinate citation Links resource (b) containing a citation link to an Epubs Expression (c) or publication. The Expression’s subordinate Links (d) resource also contains a citation which refers back to the Investigation resource (a). As the resources are identified by HTTP URIs, it is possible to seamlessly access representations of the related resources using the hyperlinks.

**5. Conclusion and Future Work**

Linking on-line research objects between a loosely coupled federation of digital objects repositories enables data management by tracking the heterogeneous components of research. Webtracks offers a simple and neutral mechanism based on the Linked Data environment for annotating typed links between resources maintained by different repositories. There is no requirement for a centralised service or common policies on vocabularies used for link annotation. Furthermore, repositories are free to implement the Links resource in line with their custom business policies on link creation.

Our exemplars have proven the concept of restoring context to dispersed research outputs. Our immediate task is to fully integrate Webtracks-ICAT into the ISIS data processing pipeline to automate Webtracks link creation and capture provenance in context when new data are generated. In addition, the knowledge built up by link triples captured by individual repository can be mixed and matched to provide alternative research networks based on the specific relationships of interest. We are working on an aggregation service in the form of a SPARQL query builder to support the self assembly of research object containers that constitute selective paths through the linked web of data.

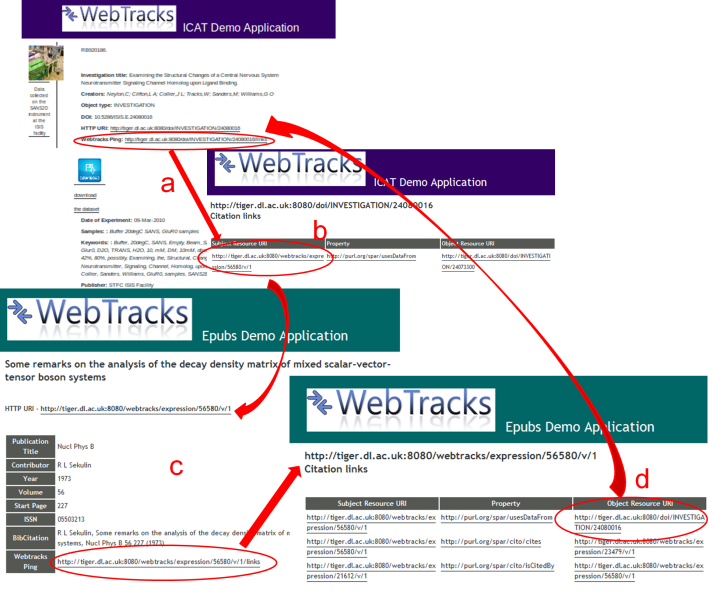


Figure 5 DOI Landing Page for Investigation 24080016 Showing the Linked Resources.

**References**

1. *Linked Data - The Story So Far.* **Chris Bizer, Tom Heath & Tim Berners-Lee.** 3, s.l. : International Journal on Semantic Web and Information Systems, 2009, Vol. 4, pp. 1-22.

2. **Open Archives Initiative.** Open Archives Initiative Protocol for Metadata Harvesting. *Open Archives Initiative .* [Online] 2011. http://www.openarchives.org/pmh/.

3. *A Protocol for Exchanging Scientific Citations.* **Matthews, Brian, et al.** s.l. : Fifth IEEE International Conference on e-Science, 2009. e-science. pp. 171-177.

4. **Sixapart.com.** TrackBack Technical Specification. *Trackback.* [Online] 2011. http://www.sixapart.com/pronet/docs/trackback\_spec.

5. **Panzer, John.** Salmon Protocol. *Salmon.* [Online] 2011. http://www.salmon-protocol.org.

6. **Tramp, Sebastian.** Semantic Pingback. *Semantic Pingback Project.* [Online] August 2011. http://aksw.org/Projects/SemanticPingBack.

7. **Brian Matthews, Katherine Portwin, Catherine Jones, Bryan Lawrence.** Recommendations for Data/Publication Linkage, CLADDIER Project Report III. *BADC.* [Online] Nov 2007. http://claddier.badc.ac.uk/trac/attachment/wiki/WikiStart/Report\_III\_RecommendationsForDataLinking-final.doc.

8. **Shirley Crompton, Brian Matthews, John Casson, Arif Shaon, Mark Borkum.** The Inter-repository Communication Protocol. *InteRCom Specification.* Warrington : Science and Technology Facilities Council E-Science Centre, 2011.

9. **STFC.** Webtracks. *Webtracks.* [Online] JISC, 2011. http://www.jisc.ac.uk/whatwedo/programmes/mrd/clip/webtracks.aspx.

10. **The Restlet Project.** Restlet. *The Restlet Project.* [Online] 2.0.8, Noelios Technologies, 2011. http://www.restlet.org.

11. **Leo Sauermann, Richard Cyganiak, Danny Ayers, Max Volkel.** Cool URIs for the Semantic Web. *W3C.* [Online] 2008. http://www.w3.org/Provider/Style/URI.

12. *Using a Core Scientific Metadata Model in Large-Scale Facilities.* **B. Matthews, S. Sufi, D. Flannery, L. Lerusse, T. Griffin, M. Gleaves, K. Kleese.** s.l. : IDCC, 2009. 5th International Digital Curation Conference.

13. **The International DOI Foundation.** DOI System and Linked Data. *DOI System.* [Online] 20 April 2011. http://www.doi.org/news/DOINewsApr11.html#2.

14. **Shirley Crompton, Brian Matthews, Simon Coles.** *Webtracks Architectue.* [pdf] Warrington : STFC E-Science Centre, 2012.

15. **Bilder, Geoffrey.** Content Negotiation for CrossRef DOIs. *Cross Ref.* [Online] 19 April 2o11. http://www.crossref.org/CrossTech/2011/04/content\_negotiation\_for\_crossr.html.

1. http://code.google.com/p.icatproject/ [↑](#footnote-ref-1)
2. http://www.isis.stfc.ac.uk/ [↑](#footnote-ref-2)
3. http://epubs.cclrc.ac.uk [↑](#footnote-ref-3)