Linkitup: Semantic Publishing of Research Data

Rinke Hoekstra[1,2], Paul Groth[1], and Marat Charlaganov[1]

[1] Network Institute,
VU University Amsterdam
rinke.hoekstra@vu.nl, p.t.groth@vu.nl, m.charlaganov@vu.nl
[2] Faculty of Law
University of Amsterdam
hoekstra@uva.nl

Abstract. Linkitup is a Web-based dashboard for enrichment of research output published via data repository services. It takes metadata entered through Figshare.com and tries to find equivalent terms, categories, persons or entities on the Linked Data cloud and several Web 2.0 services. It extracts references from publications, and tries to find the corresponding Digital Object Identifier (DOI). Linkitup feeds the enriched metadata back as links to the original article in the repository, but also builds a RDF representation of the metadata that can be downloaded separately, or published as research output in its own right. We compare Linkitup to the standard workflow of publishing linked data, and show that it lowers the threshold for publishing linked research data.

1 Introduction

Researchers are increasingly faced with the requirement to both archive and share their data in a sustainable way. For example, in 2011, the US National Science Foundation began requiring data management plans for all proposals it considers. [3] Neelie Kroes, European Commission Vice-President for the Digital Agenda, has called for open access scientific results and data. [4] However, making data available in a sustainable way is still a difficult hurdle for many researchers [**?**]. Secondly, even though in some domains sharing research data has been shown to correlate with increased citation rate [5] , this increased impact is hampered by a lack of rich, machine interpretable metadata for data publications.

To address the gap in data sharing and archival, a number of *repository services* have been created to help researchers. Examples include Dryad[5], Dataverse[6], and Figshare[7]. These services are adopted as recommended practice by a variety of journals including PLoS and Nature. Good metadata plays an essential

[3] See http://www.nsf.gov/bfa/dias/policy/dmp.jsp
[4] http://europa.eu/rapid/press-release_SPEECH-13-236_en.htm
[5] http://datadryad.org
[6] http://thedata.org
[7] figshare.com

role in the proper attribution and discoverability of publications: it explicates information that is often hard to glean from the publication itself. It is recognized that Linked Data technology is the a likely candidate for this functionality.[8]

Unfortunately, existing data repositories do not cater for the generation of Linked Data. Furthermore, exposing data as Linked Data is even more difficult for individual researchers. Linked Data publication is often too *complicated* and too *unreliable*. We address these problems through *Linkitup*, a web-based dashboard that leverages existing repository services (currently Figshare.com) to facilitate the publication of Linked Data - available at `http://linkitup.data2semantics.org`. Linkitup helps users find and create links from their data to a variety of existing resources and exposes those links as Linked Data with associated provenance information. We publish the Linked Data produced through Linkitup as a separate data publication within the archive.

Related Work To facilitate data sharing and archival, many data repositories have been created.[9] There is a long history of domain specific data repositories as well as nationally sponsored data repositories. A key aspect of these is that they aim to provide long term hosting and curation of data [4].

The closest work with respect to ours is the work from Gil et al. on Organic Data Publishing [1]. Like our proposal, this work calls for the use of web environments and semantic standards to ease the scientific data sharing process. A key difference is that our work leverages existing repository services, not semantic wikis, and is focused primarily on link creation rather than data curation. We now discuss the architecture and implementation of Linkitup.

2 Linkitup

Linkitup is a Web-based *dashboard* for interacting with a Figshare "article" and the metadata that is already associated with it. A Figshare "article" can be anything from figures, datasets, media files, papers and posters to sets of files. Users can quickly find and select an article to enrich through the article list (top left in Figure 1). All article details are retrieved directly through the Figshare.com API.[10] Linkitup currently does not support publication and enrichment services independently from Figshare, but the two platforms work together seamlessly.

Figure 1 shows a screenshot of the Linkitup dashboard for a paper about a prototype system for clinical decision support. The standard Figshare metadata is shown on the right ("Article Details"), and linking services are accessible on the left ("Plugins"). As mentioned in the introduction, the Figshare metadata is *internal* to that service. Linking services essentially tie Figshare specific identifiers to Linked Data URIs. A verbatim Linked Data version of the Figshare metadata may use the right format, but does not reuse existing URIs, and therefore does not *link* to any other datasets or descriptions thereof. The

[8] See `http://www.w3.org/DesignIssues/LinkedData.html`
[9] See `http://databib.org` for a comprehensive listing.
[10] See `http://api.figshare.com`.

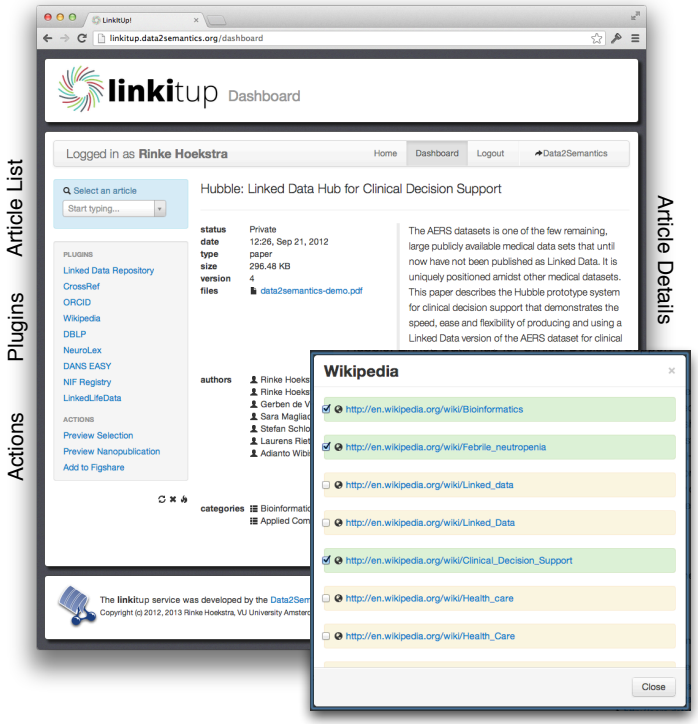

Fig. 1. The Linkitup dashboard interface

linking services are separate modules that implement the interaction between an article's metadata, and third party services.

A plugin typically uses a selection of article metadata (tags, categories, authors) to query a remote service, and returns a list of candidate matches. The results are rendered to a dialog using a *standard* UI template. This allows users to select links they deem correct using an interface that is independent of the plugin used. Crucial in this process is that the *user* is in control of which links are added to the dataset. Figure 1 shows candidate links from our paper to DBPedia; selected links lit up in green. The DBPedia plugin retrieves the URIs of resources from DBPedia for which the label matches that of any *tag* or *category* associated with the article through Figshare. Linkitup is equipped with nine plugins that serve to demonstrate the range of services we can connect to. Four plugins call a REST service, three use a SPARQL endpoint, one uses a custom scraper and one is based on the *content* of the Figshare article (Table 1).

Linkitup publishes the results of the enrichment process in two ways: 1) the *links* section of the original article on Figshare is updated with the newly found links to external resources, and 2) it generates a Linked Data representation

Name	Service	Source	Links to
Elsevier LDR	REST	Tags & Categories	Funding agencies
ORCID	REST	Authors	ORCID Author IDs
NIF Registry	REST	Tags & Categories	Datasets
LinkedLifeData	REST	Tags & Categories	Entities & Concepts
DBPedia	SPARQL	Tags & Categories	Entities & Concepts
DBLP	SPARQL	Authors	Authors
NeuroLex	SPARQL	Tags & Categories	Concepts
DANS EASY	Custom	Tags & Categories	Datasets
Crossref	Custom	Citations	DOIs

Table 1. Overview of Linkitup plugins

of all metadata as a *nanopublication* [6] that is made available both as separate article on Figshare, and to a triple store. Where possible, we reuse existing vocabularies in the RDF generation. Since Linkitup nanopublications are essentially *annotations* of other publications, we intermix the nanopublication format with both PROV [2] and the Open Annotation (OA) specification.[11] All PROV and Open Annotation statements are contained in the *provenance* part of the publication. Users can inspect the provenance trace of their enrichment process through a visualization provided by the PROV-O-Viz tool.[12]

Linkitup transforms the process of publishing linked research data by *hiding* the underlying technology. Technology hiding allows researchers to enrich their data without having to go through the steps typically associated with linked data publishing . From the Linked Data Handbook [3, Chapter 4], we identify six considerations in the publishing chain: decide how to *mint Cool URIs*, decide on *triples to include* in the description of a resource, *describe the dataset* itself, choose appropriate *vocabularies*, if necessary define additional *terms*, and *make links* to and from external data sources. Linkitup facilitates each of these steps:

1. Linkitup uses its own slash-based URI scheme for minting URIs.
2. Linkitup Linked Data will be hosted through an adapted Pubby[13] interface that returns an HTML description of the resource that contains both *incoming* and *outgoing* links.
3. Linkitup describes each dataset in terms of what it *is* about, e.g. using the 'voiD' vocabulary,[14] how it *came* about, using the PROV vocabulary, and how it can be *used* in terms of licensing, waivers and norms.
4. Linkitup uses a small selection of well known vocabularies for publishing enriched data (DCTerms, FOAF, SKOS, PROV, OA and Nanopub).
5. Every Linkitup plugin tries to put the Figshare article into context by mapping its rudimentary metadata to richer descriptions from (linked) data sets. These plugins – and thus data sets – represent the *external linking targets*

[11] The Open Annotation model is defined by the W3C Open Annotation community group, and is subject to change. Linkitup uses the community draft of February 2013, `http://www.openannotation.org/spec/core/20130208/index.html`.

[12] `http://provoviz.org`

[13] Pubby is a standard front end for triple stores, that implements the basics of content negotiation for Linked Data, see `http://github.com/cygri/Pubby`.

[14] voiD: vocabulary of interlinked datasets, see `http://www.w3.org/TR/void/`.

described in [3, Section 4.3]: Linkitup takes care of identifying and selecting appropriate targets for linking research data.

3 Conclusion

Linkitup is a dashboard enabling the discovery and publication of linked research data using an existing repository service, Figshare.com. Importantly, Linkitup provides crucial benefits over existing Linked Data publication practices in terms of easy of use (technology hiding) and persistence (i.e. relying on the archives guarantees). We are working to expand the integration of Linkitup with other commonly used services, e.g. by publishing directly from Dropbox into Figshare via Linkitup, and by supporting other repositories (e.g. DANS EASY).

We intend to expand the number of services that Linkitup supports, in particular, through deeper content analysis. Finally, we aim to provide richer notifications to let users track how their data is being interlinked. While Linkitup is focused on science, it also serves as a model for the integration of user facing Web 2.0 services with Linked Data publication, which potentially help us build a richer Web of Data.

Acknowledgments This publication was supported by the Dutch national program COMMIT.

References

1. Yolanda Gil, Varun Ratnakar, and Paul C. Hanson. Organic data publishing: A novel approach to scientific data sharing. In Tomi Kauppinen, Line C. Pouchard, and Carsten Keßler, editors, *LISC*, volume 783 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.
2. Paul Groth and Luc Moreau. PROV-Overview: An Overview of the PROV Family of Documents. Working group note, W3C, April 2013. `http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/`. Latest version available at `http://www.w3.org/TR/prov-overview/`.
3. Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011.
4. Laura Haak Marcial and Bradley M. Hemminger. Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science and Technology*, 61(10):2029–2048, 2010.
5. Heather A Piwowar, Roger S Day, and Douglas B Fridsma. Sharing detailed research data is associated with increased citation rate. *PloS one*, 2(3):e308, January 2007.
6. Eric Schultes, Christine Chistester, Kees Burger, Paul Groth, Spyros Kotoulas, Antonis Loizou, Valery Tkachenko, Andra Waagmeester, Sune Askjaer, Steve Pettifer, Lee Harland, Carina Haupt, Colin Batchelor, Miguel Vazquez, Jose Maria Fernandez, Jahn Saito, Andrew Gibson, and Louis Wich. The Open PHACTS Nanopublication Guidelines. Technical report, March 2012.