Discovery Informatics: AI Takes a Science-Centered View on Big Data
AAAI Technical Report FS-13-01

Linkitup: Link Discovery for Research Data

Rinke Hoekstra and Paul Groth
Computer Science Department
VU University Amsterdam, The Netherlands
{*rinke.hoekstra,p.t.groth*}*@vu.nl*

Abstract

Linkitup is a Web-based dashboard for enrichment of research output published via industry grade 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. In this paper, we compare Linkitup to the standard workflow of publishing linked data, and show that it significantly lowers the threshold for publishing linked research data.

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.[1] Neelie Kroes, European Commission Vice-President for the Digital Agenda, has called for open access scientific results and data.[2] However, making data available in a sustainable way is still a difficult hurdle for many researchers (Tenopir et al. 2011). Secondly, even though in some domains sharing research data has been shown to correlate with increased citation rate (Piwowar, Day, and Fridsma 2007), 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[3], Dataverse[4], and Figshare[5]. These services share a number of characteristics:

- they make it easy to upload data in most formats;
- they provide a "landing page" for data;
- they provide a citable reference for the data;
- default licensing options (usually creative commons) are given; and,
- they make guarantees about the long-term archival of the data backed by defined business models.

These characteristics have lead to the adoption of these services as recommended practice by a variety of journals including PLoS and Nature.

While these services address the exposure and archival of data, the types of metadata that can be associated with data publications is limited. *Provenance* metadata is mostly limited to authors, title, and publication date, and *content* metadata is supported only through the free text tags and categories we can find in many Web 2.0 applications (e.g. Blogs). Secondly, the metadata in these repositories is 'locked in' and can only be used through the website and APIs they provide.

Good metadata plays an essential role in the proper attribution and discoverability of publications: it explicates information that is often hard to glean from the publication itself. Research papers at least have community-established formatting standards for depicting the most essential metadata. This allows information extraction tools to scan the obvious places. However, such standards do not exist for data publications.

It is widely recognized that Linked Data technology is the most likely candidate to fill this gap.[6] Linked Data relies on four simple, but powerful principles (see also Section):

1. Use web addresses (URIs) as unique identifiers for information resources. These resources can be concrete (documents, entities) or abstract (concepts, relations).
2. Relations (links) between information resources are expressed as *triples*, i.e. $< subject, predicate, object >$, where each of *subject*, *predicate* and *object* are URIs that represent information resources. These triples form the edges of a graph.
3. Every Linked Data URI should be dereferencable via HTTP to a description of the resource.

[1] See http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[2] http://europa.eu/rapid/press-release_SPEECH-13-236_en.htm

[3] http://datadryad.org

[4] http://thedata.org

[5] figshare.com

[6] See http://www.w3.org/DesignIssues/LinkedData.html

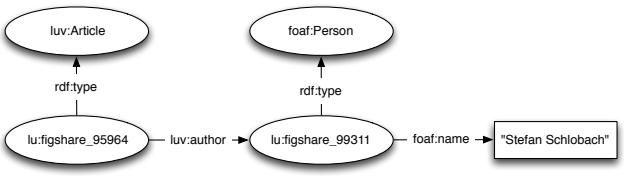

Figure 1: Example Linked Data graph for an article and its author.

4. Reuse existing identifiers as much as possible.

Figure 1 depicts an example Linked Data graph where an article (lu:figshare_95964) is associated with its author (lu:figshare_99311) via a luv:author relation. The URIs abbreviated with the foaf prefix stem from a frequently used vocabulary for expressing agents. The URI that identifies the author can easily be associated with existing author identifiers, such as ORCID, ScopusID or AuthorID.

The web-based architecture of Linked Data, combined with the *reuse* of identifiers across descriptions, allows it to form a semantic network that can span across any number of data repositories. Any reuse of an identifier between the description of two datasets forms a bridge that automatically *links* the datasets together. The LOD2 statistics maintained by (Demter et al. 2012)[7] indicate that the Linked Data cloud currently comprises 2289 interlinked *open* datasets, containing a total of 62 billion relations between resources.

Linked Data has the advantage over standardized content classification schemes such as e.g. the Universal Decimal Classification (UDC) that it is much more flexible and dynamic.[8] One is free to adopt an existing scheme, such as the Linked Data version of the UDC[9] for describing the content of a dataset, or one can design a new one, or use both at the same time. This feature allows Linked Data-based metadata to be much more fine grained. Compare the UDC subject for 'Anatomy' (UDC code 611),[10] which does not have any child nodes, with the corresponding category in DBPedia – the most popular Linked Data resource available today – which has 23 subcategories.[11]

Unfortunately, existing data repositories do not cater for the generation of Linked Data. And exposing data as Linked Data is even more difficult for individual researchers. The threshold for reaping the potential benefits of Linked Data-based metadata for data publications is currently too high.

At the same time, the reliability of services that allow querying and retrieval of Linked Data is still far from that

[7]These statistics for http://datahub.io are published at http://stats.lod2.eu

[8]The UDC is used by librarians and publishers across the globe as a classification scheme for subject description and indexing of the content of information resources. It is maintained by the UDC Consortium, and is used in over 130 countries for 150-200k document collections, see http://www.udcc.org/index.php/site/page?view=factsheet.

[9]http://udcdata.info/

[10]See http://www.udcc.org/udcsummary/php/index.php?id=37275&lang=en

[11]See http://dbpedia.org/page/Category:Anatomy

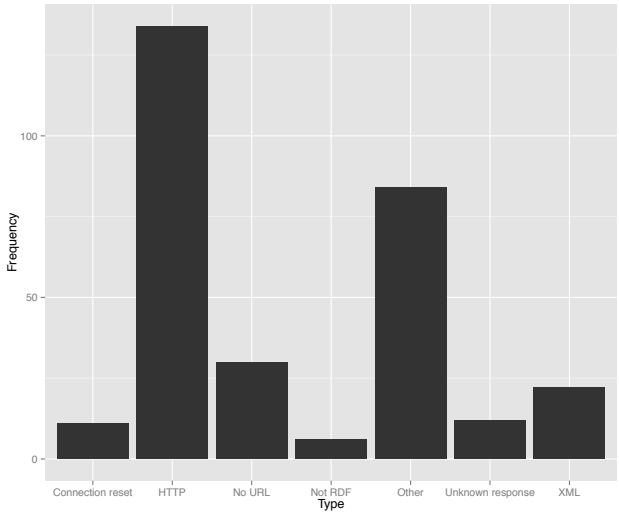

Figure 2: Distribution of error types identified by LODStats.

provided by the data archiving services mentioned before. Our analysis[12] of the LOD2 statistics shows that for the 299 out of 639 Linked Data datasets that have errors, more than half of the errors (157) occur because the server hosting the data cannot deal with the request. For a further 30 cases, the catalogued dataset is no longer available at the listed URL (See Figure 2). This means that roughly half the Linked Data cloud is not accessible.

In short, Linked Data publication is 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. 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.

Structure of this Paper

In this paper, we present the design and implementation of Linkitup and discuss how leveraging repository services changes the Linked Data publication process through technology hiding. We evaluate this process against the current state of art in Linked Data publication as defined in the Linked Data Handbook(Heath and Bizer 2011). We demonstrate its benefits in terms of the ease of meeting Linked Data publication requirements.

The rest of this paper is organized as follows. We begin by reviewing related work in data management and linked data for science. We then describe, in Section , the architecture and implementation of Linkitup. This is followed by the evaluation of the use of Linkitup for linked research data publication. Finally, we conclude with a discussion of future directions for this work.

[12]Available at http://dx.doi.org/10.6084/m9.figshare.695949

Related Work

Data management, archival and sharing has become an increasingly important topic as data sets have grown and many sciences have become more computational in nature (Akil, Martone, and Van Essen 2011). Given its importance, the literature on this topic is large and diverse. Here, our goal is to call attention to examples in the space and the issues that they address. We particularly focus on systems that use Linked Data for science. We begin with a discussion of data archiving in science.

Data Archival in Science A particular important motivation for scientific data archival and sharing has been the requirement for reproducibility in science (Mesirov 2010). Freire et al. highlight the challenges for reproducibility in computational systems (Freire, Bonnet, and Shasha 2012). Systems such as Share (Gorp and Mazanek 2011), iPython (Pérez and Granger 2007) and many workflow systems (Deelman et al. 2009) provide mechanisms for reproducing computational science based on shared data (Goble et al. 2008).

To facilitate data sharing and archival, many data repositories have been created.[13] Beyond the repositories discussed in the introduction, there is a long history of domain specific data repositories as well as nationally sponsored data repository. For example, in social science, there is the Inter-university Consortium for Political and Social Research[14]. In linguistics, there is the Linguistic Data Consortium[15]. In biomedicine, there are databases such as the Protein Data Bank[16] or the Reference Sequence Database[17]. There are also national scientific data repositories for example the Australian National Data Management Service[18] and the Dutch DANS EASY archive[19]. A key aspect of these data repositories is that they aim to provide long term hosting and curation of data (Marcial and Hemminger 2010).

Linked Data for Science Data preparation forms the bulk of work done in scientific workflows (Garijo et al. 2012). Metadata and semantics is seen as key for leveraging scientific data (Gray et al. 2005). A number of disciplines use Semantic Web and Linked Data for sharing data. For example, in neuroscience data is exposed as Linked Data via the Neuroscience Information Framework.[20] The biomedical community shares its terminologies using Semantic Web standards through BioPortal(Whetzel et al. 2011). The systems biology community shares their data using similar standards (Wolstencroft et al. 2011). Nature also exposes their article metadata as Linked Data.[21].

Beyond the usage of Linked Data standards for exposing data in various communities, a number of authors have pointed out how Linked Data can be leveraged to improve science. Kauppinen et al. describe Linked Science – the use of Linked Data for interconnecting scientific data to enable new science to be done (Kauppinen, Baglatzi, and Keßler forthcoming 2012). The need for Linked Data is bolstered by Wynholds who calls for an infrastructure for scientific data set linking (Wynholds 2011). Linked Data can also enable the connection between data and scientific models (Mäs et al. 2011). Finally, a number of authors have introduced approaches for encapsulating data sets for attribution and credit using Linked Data (Mons et al. 2011; Bechhofer et al. 2013).

The closest work with respect to ours is the work from Gil et al. on Organic Data Publishing (Gil, Ratnakar, and Hanson 2012). 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.

Figure 3: The Linkitup dashboard interface

Architecture & Implementation

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. Article metadata typically consists of the *title*, all *authors*, a *description*, one or more *categories* (field of research) to which the article belongs, several *tags* and any number of (hyper)*links* to relevant external resources, such as an author's homepage, project website, etc. An author can mark

[13] See http://databib.org for a comprehensive listing.
[14] http://www.icpsr.umich.edu/
[15] http://www.ldc.upenn.edu
[16] http://www.rcsb.org/pdb/home/home.do
[17] http://www.ncbi.nlm.nih.gov/refseq/
[18] http://www.ands.org.au/index.html
[19] http://www.dans.knaw.nl/en
[20] http://www.neuinfo.org/
[21] http://data.nature.com

Figure 4: Suggested links to Wikipedia (DBPedia, actually).

Figure 5: Suggested links from authors to DBLP

an article as *private* or *public*. Once an article is made public, it is assigned a DOI, and an official citation text.

To get to the dashboard, a user has to login and allow Linkitup access to its articles on Figshare.com. Users can quickly find and select an article to enrich through the article list (top left in Figure 3). All article details are retrieved directly through the Figshare.com API.[22] Linkitup currently does not support publication and enrichment services independently from Figshare, but the two platforms work together seamlessly.

Linkitup Plugins

Figure 3 shows a screenshot of the Linkitup dashboard for a paper about a prototype system for clinical decision support (Hoekstra et al. 2012). The standard Figshare metadata is shown on the right ("Article Details"), and linking services

[22] See http://api.figshare.com.

Figure 6: Suggested links to the NIF Registry

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 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 4 shows candidate links from our paper to DBPedia; selected links lit up in green. DBPedia is a Linked Data version of the so-called info boxes on Wikipedia and works as a switchboard for the web of data: URIs coming from DBPedia are reused across virtually all Linked Data sets currently available. 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.

At the time of writing, 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):

REST-Based Plugins The Elsevier *Linked Data Repository* (LDR)[23] is a "set of services and APIs [. . .] to store and retrieve content enhancements and other forms of semantic metadata about both Elsevier content and content available in other resources published on the Web". Although cur-

[23] See http://data.elsevier.com

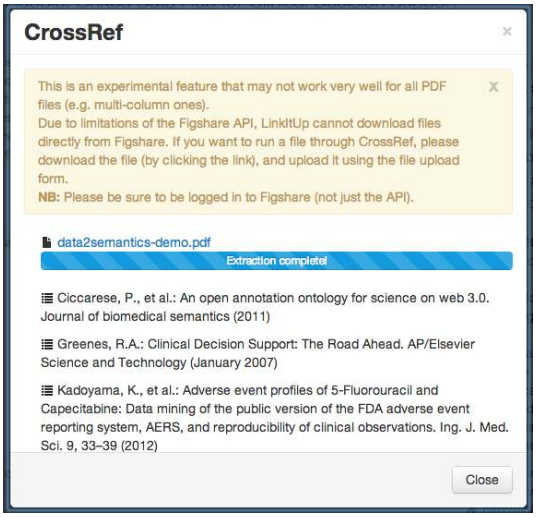

Figure 7: Extracted references presented by the CrossRef plugin

rently the LDR exposes only a single vocabulary (the SciVal-Funders vocabulary)[24] it hosts many more, and it is expected that the number of publicly accessible information resources will grow. The LDR plugin matches tags and categories to concepts in the LDR via the LDR REST interface.

The *ORCID* plugin allows users to unambiguously identify the authors of their article by relating the Figshare author identifier to an ORCID.[25] ORCID is an open platform for issuing persistent digital identifiers for researchers, and allows for automatic linking to other author identifiers, such as the Scopus Author ID of Elsevier and the ResearcherID of Thomson Reuters.

NIF Registry is the data registry of the Neuroscience Information Framework (Akil, Martone, and Van Essen 2011).[26] The NIF registry catalogs "electronic resources that have been selected by NIF curators, or contributed by the community, as valuable tools for researchers and students in the field of neuroscience. [...] It contains datasets, software tools, brain atlases, granting agencies, tissue banks, and many others". The plugin uses categories and tags associated with an article to find potentially relevant entries in the registry. Figure 6 shows an example results list for this plugin.

LinkedLifeData (LLD) is an integrated repository of major information resources in the health care and life sciences developed by the LarKC project.[27] URIs in LLD are based on UMLS identifiers, with labels associated from integrated thesauri, datasets and publications such as MeSH, UniProt and PubMed. The LLD plugin again uses the tags and categories of an article against the autocomplete REST service

[24] See http://www.funding.scival.com/.

[25] See http://orcid.org

[26] See http://www.neuinfo.org/products/nif_registry.shtm

[27] See http://linkedlifedata.com and http://larck.eu respectively.

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 LLD to retrieve matching resources.[28]

SPARQL-Based Plugins SPARQL is, analogous to SQL in database land, the standard query language and protocol for Linked Data.[29] The *DBPedia* plugin uses the DBPedia SPARQL endpoint to lookup resources with a label corresponding to one of the tags and categories of the current Figshare.com article. A Linkitup user will be presented the more familiar "Wikipedia" results.

The *DBLP* plugin relates author names to authors of publications in the DBLP Computer Science Bibliography[30] through the SPARQL endpoint hosted by the L3S Research Center[31] Figure 5 shows the DBLP identifiers that match the author names for our paper.

NeuroLex is a Wiki-based lexicon of important terms in neuroscience.[32] Like the NIF registry, it is part of the Neuroscience Information Framework, and it is constructed to "help improve the way that neuroscientists communicate about their data". The NeuroLex plugin uses a SPARQL endpoint to retrieve resources that have a label that corresponds with one of the tags or categories associated with the Figshare article.

Custom Plugins *DANS EASY* is the electronic archiving system for research data of Data Archiving and Network Services (DANS),[33] an institute of the Netherlands Royal Academy (KNAW) and the Netherlands Research Organisation (NWO). EASY stores a huge number of datasets from the humanities and social sciences, and is growing in importance with the emphasis on data publication by the major funding agencies. The DANS EASY plugin uses tags and categories of a Figshare.com article to find related datasets in EASY, and returns a title, description and persistent identifier for those datasets. DANS EASY does not offer a search API and we had to resort to a simple scraping-solution on top of its HTML-form based search interface.

Finally, the *CrossRef* plugin stands out from the pack in that it is not based purely on metadata, but *extracts* citations

[28] The autocomplete service runs directly against a full text index of the LLD corpus, making it perform significantly faster than using the LLD SPARQL endpoint.

[29] See http://www.w3.org/TR/sparql11-overview/

[30] See http://dblp.uni-trier.de.

[31] See http://www.l3s.de and the endpoint at http://dblp.l3s.de/d2r/sparql.

[32] See http://neurolex.org.

[33] See http://dans.knaw.nl.

from the Figshare article,[34] and returns potential matches in the CrossRef DOI registry by calling its REST service (see Figure 7). This allows authors to provide explicit links between their Figshare article and the DOIs of cited papers.

Publishing Linkitup Linked Data

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 of all metadata as a *nanopublication* (Schultes et al. 2012) that is made available both as separate article on Figshare, and to a triple store. The two Figshare articles are connected by key-value tags of the form "RDF=$\langle$nanopub_id$\rangle$$" and "about=$\langle$article\_id$\rangle$$", respectively.

A nanopublication as defined in (Schultes et al. 2012) consists of three parts: an *assertion*, containing the contents of the publication, *supporting* evidence for the claims in the assertion, and *provenance* describing the processes that led to the publication. Since Linkitup nanopublications are essentially *annotations* of other publications, rather than independent assertions, we intermix the nanopublication format with both the standard Linked Data for provenance, PROV (Groth and Moreau 2013), and a schema for expressing annotations as Linked Data, the Open Annotation (OA) specification.[35]

All PROV and Open Annotation statements are contained in the *provenance* part of the publication. In addition to the basic assertions shown here, this part contains author details and a PROV description of the generating activity (e.g. lu:linkitup_20130501T113510) including timestamps. The *assertion* combines a verbatim Linked Data representation of the Figshare article metadata, with the links found through the Linkitup application. Depending on the plugin, a link may be represented as a 'match' with a Figshare tag or category, or as a resource that is 'related' to the Figshare article itself. For instance, links found to LinkedLifeData and DBPedia resources are matches, whereas datasets found in DANS EASY or the NIF Registry are merely related resources.

Evaluation: Compliance with Linked Data Publishing

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 (Figure 8).

In this section we evaluate the advantages and inevitable tradeoffs inherent in this approach by comparing Linkitup

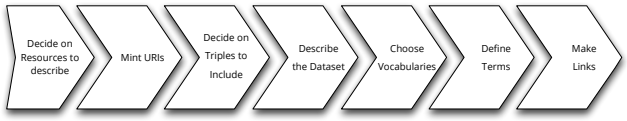

Figure 8: Steps in Linked Data publishing, as described in (Heath and Bizer 2011).

to the process of linked data publishing described in the Linked Data Handbook (Heath and Bizer 2011, Chapter 4). From (Heath and Bizer 2011), we identify six considerations in the publishing chain: decide how to *mint Cool URIs* (section 4.1), decide on *triples to include* in the description of a resource (section 4.2), *describe the dataset* itself (section 4.3), choose appropriate *vocabularies* (sections 4.4.1-4.4.6), if necessary define additional *terms* (section 4.4.6), and *make links* to and from external data sources (section 4.3). We briefly discuss each consideration in the context of Linkitup.

Mint Cool URIs Resource URIs should follow the Linked Data principles, and be *cool*.[36] In short, they should follow the HTTP URI scheme, and be dereferencable to a description of the resource they identify. The format of the description depends on the client that uses the URI as URL. (Heath and Bizer 2011) summarize these rules as: keep out of namespaces that you do not control, abstract away from implementation details, and use natural keys within URIs (section 4.1).

Linkitup naturally meets all of these requirements. First of all, it uses a standard *slash*-based URI scheme, that is used for all resources published by Linkitup. The form of the $< article\_uri >$ depends on whether the Figshare article is *private* or *public*. For the former, we use the form:

```
http://linkitup.data2semantics.org/
          resource/figshare_[Article ID]
```

and for the latter we add an equivalence relation (owl:sameAs) with a standard DOI-based URI. URIs for *tags*, *categories* and *authors* only use the latter method. Figshare identifiers are shared across articles between literally equivalent tags. Categories and their identifiers are global by nature.

Triples to Include (Heath and Bizer 2011) go in some detail to discuss what triples should be returned to a HTTP request for the description of a resource. For now, Linkitup Linked Data will be hosted through an adapted Pubby[37] interface that returns an HTML description of the resource that contains both *incoming* and *outgoing* links. Another possible approach is to use the Linked Data API[38] to provide REST web services access to the underlying data as done by the OpenPHACTS project (Gray et al. 2012).

[34] Linkitup uses pdfminer, https://github.com/euske/pdfminer/, together with regular expressions based on standard LateX bibliography styles listed at http://amath.colorado.edu/documentation/LaTeX/reference/faq/bibstyles.html.

[35] 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.

[36] See http://www.w3.org/DesignIssues/LinkedData.html and http://www.w3.org/TR/cooluris/

[37] 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.

[38] http://code.google.com/p/linked-data-api/

Describe the Dataset A dataset can be described in terms of what it *is* about, e.g. using the 'voiD' vocabulary,[39] how it *came* about, using the PROV vocabulary, and how it can be *used* in terms of licensing, waivers and norms. Figshare publishes figures, media, poster, papers and filesets under a CC-BY license. Datasets are published under CC0. Since Linkitup nanopublications are published as dataset on Figshare, they automatically fall under the CC0 license. The Linkitup metadata contains an assertion to that effect, using the Dublin Core license property (dcterms:license). Linkitup currently does not publish a voiD description of the nanopublication, but it does include provenance information of the linking process using the PROV vocabulary. Future versions will incorporate the provenance of individual plugins runs.

Vocabularies & Terms It is important that Linked Data publishers express their data as much as possible in terms of existing vocabularies. As described in (Heath and Bizer 2011), it may be hard to choose the right vocabulary to use since not all vocabularies are equally stable (maintenance), widespread, broad or expressive. Linkitup uses a small selection of well known vocabularies for publishing enriched data (DCTerms, FOAF, SKOS, PROV, OA and Nanopub). These vocabularies have broad user bases, and strong community backing. Linkitup also uses its own vocabulary alongside these, to maintain a mapping with the terminology used in the Figshare API.

Make Links This step lies at the heart of what Linkitup and Linked Data are about. 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* described in (Heath and Bizer 2011, Section 4.3): Linkitup takes care of identifying and selecting appropriate targets for linking research data. We foresee that a growing number of plugins will make it necessary to categorize them to a specific *profile*, inspired by the categories assigned to the Figshare article. A historian is not very likely to have much interest in linking a publication or dataset to a Neuroscience lexicon.

Linkitup supports users to *manually* select *automatically* identified candidates for linking and expresses these links with predefined *predicates* from the SKOS vocabulary (again (Heath and Bizer 2011, Section 4.3)). This arguably restricts the freedom of users to create links fully manually, but gives them more control over the *quality* of links than fully automatic link creation.[40] Even though most plugins work on the basis of a literal string match between a Figshare author name, tag or category and a term hosted by the external service, the results from non-SPARQL based services are typically returned with some degree of confidence (a score). Linkitup plays the modest role of broker, and lets the weighting of the quality of results to the original service provider and the user.

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

[40] Advanced users can add their own links by editing the resulting nanopublication by hand, and uploading it as a new version to Figshare.

Conclusion

In this paper, we described Linkitup, a dashboard enabling the discovery and publication of linked research data by leveraging 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). Going forward, 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 already have a prototype implementation to that effect that analyzes, extracts and visualizes information from the data along the way.[41] Additionally, we will 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

Akil, H.; Martone, M. E.; and Van Essen, D. C. 2011. Challenges and opportunities in mining neuroscience data. *Science* 331(6018):708–712.

Bechhofer et al., S. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29(2):599 – 611. Special section: Recent advances in e-Science.

Deelman, E.; Gannon, D.; Shields, M.; and Taylor, I. 2009. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems* 25(5):528–540.

Demter, J.; Auer, S.; Martin, M.; and Lehmann, J. 2012. Lodstats – an extensible framework for high-performance dataset analytics. In *Proceedings of the EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer. 29

Freire, J.; Bonnet, P.; and Shasha, D. 2012. Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In *Proceedings of the 2012 international conference on Management of Data*, 593–596. ACM.

Garijo, D.; Alper, P.; Belhajjame, K.; Corcho, O.; Gil, Y.; and Goble, C. 2012. Common motifs in scientific workflows: An empirical analysis. In *8th IEEE International Conference on eScience*. USA: IEEE Computer Society Press.

Gil, Y.; Ratnakar, V.; and Hanson, P. C. 2012. Organic data publishing: A novel approach to scientific data sharing. In Kauppinen et al. (2011).

[41] See http://data2semantics.github.io/#goldendemo

Goble, C.; Stevens, R.; Hull, D.; Wolstencroft, K.; and Lopez, R. 2008. Data curation + process curation=data integration + science. *Briefings in Bioinformatics* 9(6):506–517.

Gorp, P. V., and Mazanek, S. 2011. Share: a web portal for creating and sharing executable research papers. *Procedia Computer Science* 4(0):589 – 597. Proceedings of the International Conference on Computational Science, {ICCS} 2011.

Gray, J.; Liu, D. T.; Nieto-Santisteban, M. A.; Szalay, A. S.; DeWitt, D. J.; and Heber, G. 2005. Scientific data management in the coming decade. *CoRR* abs/cs/0502008.

Gray, A. J.; Groth, P.; Loizou, A.; Askjaer, S.; Brenninkmeijer, C.; Burger, K.; Chichester, C.; Evelo, C. T.; Goble, C.; Harland, L.; et al. 2012. Applying linked data approaches to pharmacology: Architectural decisions and implementation. *Semantic Web*.

Groth, P., and Moreau, L. 2013. PROV-Overview: An Overview of the PROV Family of Documents. Working group note, W3C. http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/. Latest version available at http://www.w3.org/TR/prov-overview/.

Heath, T., and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool.

Hoekstra, R.; Magliacane, S.; Rietveld, L.; de Vries, G.; Wibisono, A.; and Schlobach, S. 2012. Hubble: Linked Data Hub for Clinical Decision Support. In *Post conference demo proceedings of ESWC 2012*.

Kauppinen, T.; Baglatzi, A.; and Keßler, C. forthcoming 2012. Linked Science: Interconnecting Scientific Assets. In Critchlow, T., and Kleese-Van Dam, K., eds., *Data Intensive Science*. USA: CRC Press.

Kauppinen, T.; Pouchard, L. C.; and Keßler, C., eds. 2011. *Proceedings of the First International Workshop on Linked Science 2011, Bonn, Germany, October 24, 2011*, volume 783 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Marcial, L. H., and Hemminger, B. M. 2010. Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science and Technology* 61(10):2029–2048.

Mäs, S.; Müller, M.; Henzen, C.; and Bernard, L. 2011. Linking the outcomes of scientific research: Requirements from the perspective of geosciences. In Kauppinen et al. (2011).

Mesirov, J. P. 2010. Accessible reproducible research. *Science* 327(5964):415–416.

Mons et al., B. 2011. The value of data. *Nature Genetics* 43(4):281–283.

Pérez, F., and Granger, B. E. 2007. IPython: a System for Interactive Scientific Computing. *Comput. Sci. Eng.* 9(3):21–29.

Piwowar, H. A.; Day, R. S.; and Fridsma, D. B. 2007. Sharing detailed research data is associated with increased citation rate. *PloS one* 2(3):e308.

Schultes, E.; Chistester, C.; Burger, K.; Groth, P.; Kotoulas, S.; Loizou, A.; Tkachenko, V.; Waagmeester, A.; Askjaer, S.; Pettifer, S.; Harland, L.; Haupt, C.; Batchelor, C.; Vazquez, M.; Fernandez, J. M.; Saito, J.; Gibson, A.; and Wich, L. 2012. The Open PHACTS Nanopublication Guidelines. Technical report.

Tenopir, C.; Allard, S.; Douglass, K.; Aydinoglu, A. U.; Wu, L.; Read, E.; Manoff, M.; and Frame, M. 2011. Data sharing by scientists: Practices and perceptions. *PLoS ONE* 6(6):e21101.

Whetzel, P. L.; Noy, N. F.; Shah, N. H.; Alexander, P. R.; Nyulas, C.; Tudorache, T.; and Musen, M. A. 2011. Bioportal. *Nucleic acids research* 39(suppl 2):W541–W545.

Wolstencroft, K.; Owen, S.; du Preez, F.; Krebs, O.; Mueller, W.; Goble, C.; and Snoep, J. L. 2011. The seek: A platform for sharing data and models in systems biology. In *Methods in Systems Biology*, volume 500 of *Methods in Enzymology*. Academic Press. 629 – 655.

Wynholds, L. 2011. Linking to scientific data: Identity problems of unruly and poorly bounded digital objects. *International Journal of Digital Curation* 6:214–225.