

Linked Data tagging with LODr

Alexandre Passant

DERI, National University of Ireland, Galway,
IDA Business Park, Lower Dangan,
Galway, Ireland,
`alexandre.passant@deri.org`

Abstract. LODr is a personal application providing semantic-enrichment features for existing tagged content from various popular Web 2.0 services, such as Flickr, del.icio.us or Twitter. By allowing people to re-tag their content with URIs, rather than simple keywords, it weaves their social data to the Semantic Web. In this paper, we detail the principles of this application, the underlying models, its distributed and collaborative architecture, as well as how it provides new and unforeseen functionalities to Web users. We also compare our approach to existing *augmented tagging* applications and see how, in our opinion, LODr offers an efficient and coherent path between Web 2.0 and the Semantic Web.

Key words: Web 2.0, Linked Data, Tagging, Data Portability, SIOC, MOAT

1 MOAT: A proposal for semantically-enhanced tagging

While tagging is widely deployed on Web 2.0 websites, it raises various issues which have been largely studied and mainly consist in tags ambiguity and heterogeneity, as well as the lack of organisation between them [4]. While it may not be a problem regarding personal tagging, it becomes relevant when trying to discover and retrieve content that have been tagged by others. To solve those issues, some approaches allow people to create relationships between tags (equivalent, narrow ...), that can be then used when retrieving content. From a Semantic Web point of view, those relationships (as well as tags themselves and the tagging actions) can be modeled using the Tag Ontology [6] and its `tags:relatedTag` property.

Yet, our approach is different and consists in taking the meaning of tags into account. Indeed, people may use different tags (for internationalization or personalization purposes) to refer to the same idea, or meaning, but current tagging applications cannot take this into account as for most of them, tags are simply free-text keywords. To take those meanings into account, we extended the usual tripartite model of tagging [5] to a quadripartite one, defining a tagging action as:

$$\textit{Tagging}(\textit{User}, \textit{Resource}, \textit{Tag}, \textit{Meaning}) \quad (1)$$

As our goal is to solve usual tagging problems, the meaning must be represented in a formal way so that user agents can exploit it. While defining

the meaning as free-text will lead to the same issues as before, we consider that it must be represented using URIs of Semantic Web resources. Especially, our goal is to rely on *reference* URIs, such as ones provided in the context of the Linking Data Project [2], while our model also can be used to deal with any URI as for instance corporate ontologies instances. This way of modeling tagging actions leads to facts as "When I tag this picture 'apple', I mean http://dbpedia.org/resource/Apple_Records, i.e. the record label, not the fruit".

Such vision of semantically-enhanced tagging has been recently published through MOAT [7], which consists in (1) a lightweight ontology to represent relationships between tags and resources URIs, extending the Tag Ontology and (2) an open-source and collaborative framework to define and share those relationships within a community and help people to bridge the gap between tagging and semantic indexing, without directly facing RDF modeling¹. Thanks to its alignment with the Tag Ontology, MOAT reuses FOAF to model the *User* part of a tagging action, but considers the use of SIOC [3] as a best practice to model the *Resource* element. Moreover, a direct relationship between the *Resource* and the *Meaning* can be inferred from the tagging actions, and modeled using `sioc:topic`, as seen in Fig. 1.

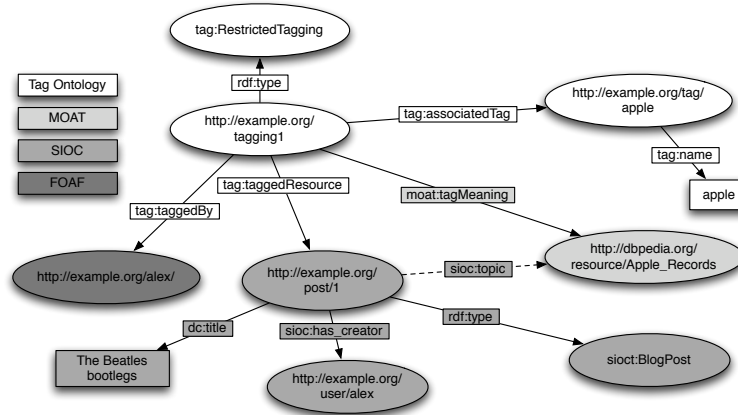


Fig. 1. Semantic tagging using the Tag Ontology, FOAF, MOAT and SIOC

This way of tagging content with URIs offer various advantages, as solving ambiguity and heterogeneity issues by dealing with machine-understandable URIs rather than words. Indeed, it lets user keep their existing tag vocabulary (for instance, different tags according to their preferred language), but links their content to those unambiguous URIs, which makes the content discoverable whatever the original tag is. Most important, it makes tagged data being interlinked with

¹ <http://moat-project.org>

other resources (DBpedia concepts, FOAF URIs ...), that can be used to retrieve and browse related content, following the Linked Data principles [1].

2 LODr principles

2.1 Objectives

While our first experiments with MOAT have been done in a corporate context², mainly using dedicated ontologies instances as references for the *Meaning* part of the tagging actions, we decided to extend the approach and apply the MOAT principles in a wider context. To achieve this goal, we implemented LODr – `http://lodr.info` –, a personal application that allows one to re-tag his existing Web 2.0 content and weave it into the Semantic Web thanks to the previous principles. Its main objective is to provide a simple-way to create RDF and interlinked content from existing Web 2.0, so that queries like *"Please list all slideshare items tagged with a topic related to the Semantic Web"* can be answered.

One other important motivation is that we did not want to create another tagging service, but a system that gives users a way to semantically enrich existing tagged data that have been created thanks to their favourite tools, since we wanted to avoid *social network fatigue* and let users keep their existing tagging habits. We believe that this approach can enable a smooth transition between Web 2.0 and the Semantic Web.

LODr is an open-source application, written in PHP5 using an Object-Oriented model and while it is completely RDF-based, simply requires the generic Apache / PHP / MySQL setup thanks to the ARC2³ framework.. More technical details regarding the implementation will be given in the rest of this paper.

2.2 Aggregating and storing distributed tagged content

The first step to achieve our goals is to provide for each user a single interface where he can access his complete tagged data, so that he can then semantically-enrich it. To complete this step, after having installed the LODr application on his webserver, the user just need to enter his own URI. This URI must be dereferencable, returning RDF or RDFa description of himself, and must include his online accounts using the `foaf:holdsAccount` property and `foaf:OnlineAccount` instances with their related `foaf:accountName` and `foaf:accountServiceHomepage`, as follows:

```
<http://apassant.net/alex> a foaf:Person ;
  foaf:holdsAccount <http://flickr.com/people/terraces> .

<http://flickr.com/people/terraces> a foaf:OnlineAccount ;
  foaf:accountName "33669349@N00" ;
  foaf:accountServiceHomepage <http://www.flickr.com/> .
```

² <http://www.w3.org/2001/sw/sweo/public/UseCases/EDF/>

³ <http://arc.semsol.org>

From this FOAF file, as all the user accounts can be discovered, the process of aggregating the data can start. To do that, LODr relies on a set of wrappers, where each wrapper is a single PHP class and corresponds to an existing service. Each of them inherits from a main `LODrWrapper` class which contains all the actions common to any wrapper, so that wrappers themselves are really lightweight and people can easily write their own for new services. For instance, the Flickr wrapper is only 24 lines of code. The wrappers currently use RSS feeds to retrieve content (thanks to MagpieRSS⁴), which unfortunately limits the number of items. In the future, and since the current source-code modeling permits it, we may rely on services APIs to retrieve the complete data for a service (but generally asks for an API key, which makes the process less intuitive) or on SIOC exporters for some of them.

While we use RSS, we yet noticed that services do not use an uniform way to define tags, so that we must deal with heterogenous content modeling. For instance Flickr use a `media:category` attribute, bibsonomy uses `dc:subject`, while on Twitter, we must rely on regular expressions parsing, as tagging is not natively supported but has been spontaneously added by users, using the `#tag` syntax in their posts. So, this first step of LODr translates any RSS item to unified instances of `sioc:item`, using the Tag Ontology to model the tagging actions, and stores each item in the backend triple store, having each item in its own graph. Then, the user must rely on a crontab script to regularly retrieve new data.

As soon as the data has been aggregated from those distributed sources, a general tagcloud (i.e. involving all tagged data from various services) is generated while for each tag a faceted interface can be used to browse content, thanks to Exhibit⁵, as we will depict later in Fig. 4.

2.3 Semantic-enrichment of tagged content

When the tagged data is translated to SIOC and stored locally, users can start its semantic-enrichment. For each item, users can edit it to get the list of related tags and give a meaning to each of them, being suggested URIs that have already been assigned by the LODr community for this tag, as seen in Fig. 2. When no URI have been previously defined within the community or when existing ones do not correspond to the meaning of the tag in the current context, a new URI can be added, directly or using the Sindice search widget⁶. Furthermore, to ease the process of choosing the right meaning, human-readable labels can be displayed instead of URIs. Yet, this approach can slow the process as a SPARQL query must be run to get the label of each URI. As soon as the meaningful URIs have been selected, the content is interlinked to these URI (as depicted previously in Fig. 1), saved locally and can be browsed as we will see in the next section. Moreover, users can decide to automatically re-tag future incoming content to ease the process.

⁴ <http://magpierss.sourceforge.net/>

⁵ <http://simile.mit.edu/exhibit/>

⁶ <http://sindice.com/developers/widget>

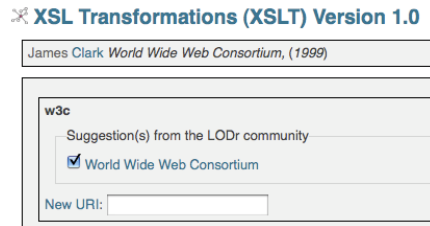


Fig. 2. Re-tagging a Bibsonomy item with LODr

As LODr is based on the MOAT principles, it requires interaction with a dedicated tag server that stores the relationships between tags and URIs for the community that uses it⁷. While a default public server is available, that stores all the relationships defined within the complete LODr community, a community can decide to use the tool with its own tag server which might be useful, for instance, in a corporate intranet (as in our first MOAT experiment). Moreover, various servers can be used within the same LODr client, so that URI can be suggested from various communities.

Following the principles described in detail in [7], when a new relationship between a tag and a URI is defined by a user, it is stored in the server so that other people from the community can reuse it. Here, we combine the Web 2.0 principles of architecture of participation with Semantic Web modeling, by providing a collaborative framework to define machine-readable meanings for tags: while each user define a meaning for its own use, the whole community benefits from it, which enhances the global process of semantic-enrichment of tagged content.

3 Browsing data

3.1 Browsing data locally

Once data have been re-tagged with those URIs, new browsing interfaces are available within the local LODr service. First, the tagcloud is completed by a *conceptcloud* which consists in a weighted list of used URIs. Once again, instead of URIs, labels can be displayed, as seen on Fig. 3. As different properties can be used to model labels (e.g. `foaf:name`), we wrote a lightweight reasoning engine to deal with `rdfs:label` subproperties using ARC triggers and SPARQL CONSTRUCT.

As for the tagcloud, each concept provides an hyperlink to the related items. More than a list of items, the system also displays (Fig. 4): (1) a description of the concept, using the `rdfs:comment` property (or subproperties); (2) a list of

⁷ <http://moat-project.org/server>



Fig. 3. Tagcloud and conceptcloud in LODr

related concepts by co-occurrence; (3) a list of related concepts that share a direct-relationship with the current one and (4) a list of related concepts that share a common property together. As Fig. 4 shows, SPARQL and XSLT are related since they share the `dbpedia:Category:WorldWideWebConsortiumstandards` value for `skos:subject` in DBpedia. For each cloud, we limit the list of URIs to the ones that are used in other tagging actions, to avoid information overload.

Moreover, this interface can be easily internationalized. The configuration file lets the user define an ordered list of preferred languages, that is then used to create a complex SPARQL query, involving `FILTER` and `OPTIONAL` clauses, which retrieves the label (or description) in the given language (by preference order). Yet, when no label can be found, the URI of the concept is displayed.

Finally, one important thing to notice is that the complete XHTML template embeds RDFa⁸ to describe the item (using SIOC) and the tagging actions (using the Tag Ontology and MOAT), so that content can be easily discovered and crawled by dedicated semantic search engine, or plug-ins as Semantic Radar⁹.

3.2 Browsing data from the community

As LODr relies on a distributed architecture, each tagged content is spread – and user-owned – on the network. Yet, the community server stores the re-tagged items in its centralized triple store, when receiving clients pings, so that one can use it to browse the complete tagged dataset. A faceted interface similar to the individual ones can be used, and includes an additional facet to browse the data by user. Moreover, it features a SPARQL endpoint – as, actually, any individual LODr instance – where advanced users can run their own queries, as the one introduced in the first section of this paper.

Moreover, this global entry point is also used to deliver RSS feeds for each URI (containing the related tagged items) and to provide a Ubiquity¹⁰ command

⁸ <http://rdfa.info>

⁹ <http://sioc-project.org/firefox>

¹⁰ <https://wiki.mozilla.org/Labs/Ubiquity>

The screenshot shows the LODr interface. At the top, it says 'LODr' and 'Tagging, Aggregating, Interlinking, The LOD-way'. There are navigation links: 'clouds | all items | orphan items | LODed items | conflicted items'. Below is a 'SPARQL' section with a text box containing a SPARQL query. Underneath are three sections for 'Related URIs': 'Related URIs (co-occurrence)' with a box for 'Extensible Stylesheet Language Transformations'; 'Related URIs (direct relationships)' with a box saying 'Nothing yet'; and 'Related URIs (shared properties)' with a box listing 'Resource Description Framework', 'Web sémantique Web 2.0', and 'Extensible Stylesheet Language Transformations'. The main 'Items' section shows '6 item' and a list of two items: '1. Named graphs, provenance and trust [details]' and '2. Scalable Querying Service over Fuzzy Ontologies [details]'. There are also filters for 'BLOCS', 'CARTE', and 'LIGNE DE TEMPS', and a 'Source' box with 'http://bibsonomy.org'.

Fig. 4. Browsing items tagged with a particular URI

so that a user can, by browsing Wikipedia, simply retrieve all the items tagged to the underlying concept¹¹. Here, we rely on the DBpedia SPARQL endpoint to identify the concept related to the page (using the `wikipedia-*` properties), and then redirect to the browsing interface for that particular URI. Moreover, this command which gives more visibility, serendipity, and fun to the process, can be used by people who are not actively yet publishing data using LODr. We also hope it can convince them to install a LODr client and start the process of semantically indexing their data.

4 Conclusion and future works

In this challenge description, we introduced LODr, a system providing semantic-enrichment features for existing tagged data, combining Web 2.0 collaborative principles and Semantic Web technologies. While semantic indexing can be a complex process, we tried to make it as easy as possible, by completely hiding the Semantic Web modeling principles and underlying architecture to users.

While the main aim of LODr is to enhance Web 2.0 content discovery using the Linked Data principles, we hope it can be used to improve algorithms that extract or link to ontologies from tags, and plan to include such algorithm in future versions. Indeed, by letting people voluntary link tags to URIs in the context of particular tagging actions, we think it can provide an interesting, and validated, traineeset for such approaches.

¹¹ <http://lodr.info/tools>

Finally, as various *augmented-tagging* applications have been recently published, we think the originality of LODr resides in: (1) its way of linking tags and tagged data to existing Semantic Web resources, and not only relating tags together as in Gnizr¹², which makes the application live in its own closed-world, (2) its ability to use any URI (e.g. FOAF profiles, Semantic Web conference corpus URIs) and not only DBpedia ones as in Faviki¹³, (3) its integration with existing Web 2.0 content, which does not require to subscribe to a new independent tagging application, avoiding *social network fatigue* and (4) its complete Semantic-Web based interface and especially its RDFa output and SPARQL endpoint, which makes easy to integrate its data into other applications. Especially, regarding this latest point, we think that it makes LODr a nice Linked Data citizen, as it helps to provide open and interlinked RDF data from any Web 2.0 tag-based service to any Semantic Web dataset, thus creating an efficient and coherent path between those two worlds.

Acknowledgments

This material is based (in part) upon works supported by the Science Foundation Ireland under Grant No. SFI/02/CE1/I131. We would like to address special thanks to Benjamin Nowack for providing the ARC2 library and to the whole Linking Open Data community for their efforts regarding RDF data availability on the Semantic Web.

References

1. Chris Bizer, Richard Cyganiak, and Tom Heath. How to Publish Linked Data on the Web. <http://sites.wiwiw.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/>, 20 July 2007.
2. Ayers D. Raimond Y Bizer C., Heath T. Interlinking open data on the web. In *Poster, 4th Annual European Semantic Web Conference (ESWC2007), Innsbruck, Austria*, 2007.
3. J.G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards Semantically-Interlinked Online Communities. *2nd European Semantic Web Conference*, May 2005.
4. Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata, December 2004.
5. Peter Mika. Ontologies are us: A unified model of social networks and semantics. In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, *The Semantic Web - ISWC 2005, Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer, 2005.
6. Richard Newman, Danny Ayers, and Seth Russell. Tag ontology, December 2005.
7. Alexandre Passant and Philippe Laublet. Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China, Apr 2008*.

¹² <http://gnizr.com>

¹³ <http://faviki.com>