

COVER

This is the author-version of article published in the Springer "Lecture Notes in Computer Science" series:

F. Abbattista, F. Calefato, D. Gendarmi and F. Lanubile, "Shaping personal information spaces from collaborative tagging systems", In PB. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part III, LNAI 4694, pp. 728–735, 2007. © Springer-Verlag Berlin Heidelberg 2007

Shaping Personal Information Spaces from Collaborative Tagging Systems

Fabio Abbattista, Fabio Calefato, Domenico Gendarmi, Filippo Lanubile

University of Bari,
Dipartimento di Informatica,
Via Orabona, 4, 70126 - Bari, Italy
{fabio,calefato,gendarmi,lanubile}@di.uniba.it

Abstract. The appearance of powerful tools for lightweight metadata creation, such as collaborative tagging systems, is harnessing the power of online communities, although such metadata are limited to human consumption only. In this paper we first propose an abstract model for representing a generic collaborative tagging system which uses RDF as the underlying technology to store metadata created by different online communities. Then, we present a scenario with the purpose of illustrating how a service able to retrieve tags from different folksonomies can support users in the organization of their personal information spaces within the context of a digital library.

Keywords: collaborative tagging, folksonomy, personal information space, RDF.

1 Introduction

Ontologies play a central role in the Semantic Web vision because they establish common vocabularies and semantic interpretations of terms accessible by machines [1]. While centralized controlled systems can significantly profit by expressivity of ontology languages such as OWL, lightweight ontologies have spread over the loosely controlled, distributed environment of the Web. This tendency towards lightweight, easily accessible and extensible metadata is evidenced by the appearance of RDF-based technologies such as RSS and FOAF, which currently represent the majority of public available metadata on the World Wide Web [8].

Powerful tools for lightweight metadata creation, such as collaborative tagging systems, harness the power of the community and have been shown effective in creating large amounts of metadata quickly, albeit, so far, this metadata are limited to human consumption only [5].

Collaborative tagging systems allow people to organize a set of resources, annotating them with tags via a web-based interface. The activity of labeling is called tagging, as it consists of attaching one or more tags to the resource. This activity is accomplished individually, as each user of the system is free to choose the tags he wishes, with no restrictions. However, while using the system every one can see who

else is participating to it by observing others' tagging activities. This tight feedback loop [12] brings that asynchronous and asymmetrical collaboration which makes these systems social. The result of such a social activity is a collection of annotations, also called folksonomy.

Existing collaborative tagging systems can be discriminated according to the kind of resources they allow to annotate. If the objects to annotate are bookmarks these applications are also called social bookmarking systems (e.g. del.icio.us¹). If tagging systems are used to organize and share scientific publications they are also referred to social reference management applications (e.g. CiteULike²). Tagging applications which allow users to share media resources, such as photo, video or audio content are instead generally named social media sharing systems (e.g. Flickr³, YouTube⁴, LastFm⁵). Moreover, there are systems which allow to tag abstract things which do not represent a resource on the web but they are anyhow univocally identifiable through web-based mechanisms. These kinds of systems usually are just labeled as social networking sites as they allow to link people sharing common interest (e.g. 43 Things⁶).

Another distinctive feature of collaborative tagging systems is the opportunity for users to upload resources besides tagging them. Within such categorization there are two kinds of tagging systems: those where users can annotate just references to resources already available on the web and systems which allow their users to upload new resources. In the latter case, the system has to manage the univocal identification of the uploaded resource as well as the information about who created it. Making a comparison with this categorization of tagging systems and the previous one, typically social bookmarking and social reference management systems belong to the former category, while other kind of systems allows users to upload new resources.

Finally, among systems where users are able to upload resources, another classification can be made, distinguishing between narrow and broad folksonomies [13]. In narrow folksonomies only owner of the resources can tag them, whereas in broad folksonomies every user can tag any resource, regardless of its creator.

In this paper we formalize a generic model to represent any collaborative tagging system, regardless of its distinguishing features. We propose RDF as the enabling technology to store all the annotations performed by users in a collaborative tagging system. Exposing tagging data in RDF could speed up the process of sharing metadata across live communities, leading to both collective and individual benefits in the information organization.

The remainder of the paper is organized as follows. Section 2 presents the model and its implementation in RDF. Section 3 depicts an application scenario in the context of a digital library. Finally in Section 4 we draw conclusions and identify directions for further work.

¹ <http://del.icio.us/>

² <http://www.citeulike.org/>

³ <http://www.flickr.com/>

⁴ <http://www.youtube.com/>

⁵ <http://www.last.fm/>

⁶ <http://www.43things.com/>

2 An abstract model of Collaborative Tagging Systems

Despite of the different kind of collaborative tagging systems available on the web, a generic conceptual model can be conceived to formalize the tagging activity in all the above presented systems. Some attempts of formally describing folksonomies have already been proposed in literature [6, 9, 14]. All these works lay on the model proposed by Mika [11] as an abstraction of the network of users, tags and resources generated by a collaborative tagging system.

2.1 The conceptual model

According to the abstraction provided in [11], a collaborative tagging system can be generally modeled as a tripartite 3-uniform hypergraph. Here, we first define a folksonomy in terms of a hypergraph structure and then, we present how to obtain from this model a personal folksonomy through a graph transformation process.

Definition 1. Given a collaborative tagging system or a folksonomy, where there are a set of registered users denoted with U , a set of applied tags denoted with T and a set of annotated resources denoted with R we can define $F=(N,E)$ as the tripartite 3-uniform hypergraph model representing the system. The set $N=U\cup T\cup R$ represents all the *entities* within the collaborative tagging system while $E=\{(u,t,r) \mid u\in U, t\in T, r\in R\}$ is the set representing all the *annotations* that compose the folksonomy.

Informally, a hypergraph is a generalization of a graph, in the sense that it extends the notion of graphs allowing edges to connect any number of nodes. While graph edges are pairs of nodes, hyperedges are arbitrary sets of nodes, therefore they contain an arbitrary number of nodes.

The graph in definition 1 is tripartite as the set of nodes is partitioned in three disjoint sets, namely $U=\{u_1, \dots, u_k\}$, $T=\{t_1, \dots, t_l\}$, $R=\{r_1, \dots, r_m\}$, representing the set of users, tags, and resources, respectively. Further, it is a 3-uniform hypergraph as each edge connects exactly 3 nodes, one from the each set U , T and R . Hence, each edge represents an annotation in the system performed by a particular user with a specific tag for a certain resource.

From the original hypergraph representing a folksonomy, we can obtain a bipartite graph (with no hyperedges), which represents all the annotations performed by a single user within the system. We can thus provide the following definition:

Definition 2. Given the model of a collaborative tagging system, $F=(N,E)$, as defined in definition 1, a *personal folksonomy* is defined as the bipartite graph $P^F=(N_u, E_u)$, where $N_u=T\cup R\subset N$ and $E_u=\{(t,r) \mid \exists u\in U \ni (u,t,r)\in E\}$. The set N_u denotes all the entities of the collaborative tagging system F related to the user u , whereas E_u is the set representing all the annotations of the user u within the system.

The expression “personal folksonomy” can appear as an oxymoron because it combines the terms personal (i.e. individual, private) and folks (e.g. collective,

shared). Nevertheless, it is useful to express in a concise way the personal view of a specific user on the collaborative tagging system. The personal folksonomy is thus a projection of the initial tripartite graph in a two-dimension scale where the user entity is fixed.

Furthermore, using a generic abstract model to represent any collaborative tagging system, regardless their distinguishing features, leads to the following opposite definition:

Definition 3. Given a generic collaborative tagging system as defined in definition 1, a *global folksonomy* is defined as an indexed family of hypergraphs $G^F = (F_i)_{i \in I}$, where I is the index set and $\forall_i \in I \exists F_i = (N_i, E_i)$. The set $N_i = U_i \cup T_i \cup R_i$ represents the entities of a collaborative tagging system modeled as F_i , while the set E_i depicts all the annotation in F_i .

2.2 The RDF implementation of the conceptual model

RDF metadata are encoded in statements defined as $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ triples. Given the definition of a folksonomy as a hypergraph, we propose to model such a structure using the RDF. Any RDF graph can be mapped to a simple ordered 3-uniform hypergraph, where every statement corresponds to a hypergraph edge, with the nodes being the subject, predicate and object in this order [7]. The RDF abstract model is thus well suited to represent a tagging system as defined in definition 1, since an RDF triple can naturally represent an annotation corresponding to a 3-node edge in the hypergraph. Furthermore, the order of nodes allows to distinguish the role of each entity of the system within a single annotation.

For each edge $e \in E$ in F , representing an annotation within the folksonomy, we can store an RDF statement $\langle s, p, o \rangle$, where $s \in U$, $p \in T$, $o \in R$. There are different possible syntaxes to store RDF statements in XML. TriX⁷ (Triples in XML) is a serialization for named graphs with the purpose to provide a highly normalized, consistent XML representation for RDF graphs, allowing the effective use of generic XML tools [3]. Figure 1 shows a hypergraph representing a folksonomy, while the following excerpt of an RDF file expressed in TriX syntax depicts how to map such a graph in RDF.

```
<?xml-stylesheet type="text/xml"
href="http://www.w3.org/2004/03/trix/all.xsl"?>
<TriX xmlns="http://www.w3.org/2004/03/trix/trix1/"
xmlns:u="http://example.com/userentity/"
xmlns:t="http://example.com/tagentity/"
xmlns:r="http://example.com/resourceentity/">
  <graph>
    <uri>http://example.org/folksonomy</uri>
    <triple>
      <qname>u:U1</qname>
      <qname>t:T2</qname>
      <qname>r:R3</qname>
```

⁷ <http://www.w3.org/2004/03/trix/>

```

</triple>
<triple>
  <qname>u:U2</qname>
  <qname>t:T3</qname>
  <qname>r:R2</qname>
</triple>
<triple>
  <qname>u:U3</qname>
  <qname>t:T1</qname>
  <qname>r:R1</qname>
</triple>
</graph>
</TriX>

```

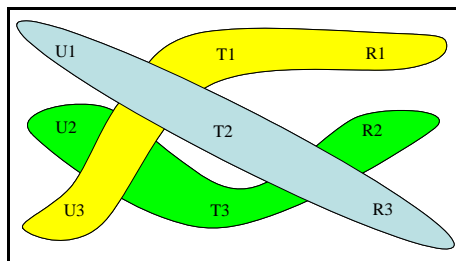


Figure 1. Hypergraph representing a folksonomy

When inserting and searching large amounts of data, proper tools to store the RDF statements are needed because of the low performance. Sesame [2] is a framework for storage and querying of RDF information which offers parsers and writers supporting the TriX syntax. Sesame can be used on different platforms, as it is written in Java and connects to common products like Tomcat and MySQL. Other projects have already successfully used Sesame to support distributed Semantic Web applications [10].

3 An illustrative scenario

As an illustrative context for our approach, we consider the digital library of the Association for Computing Machinery (ACM). The interaction process of a user with the digital library can be characterized as a three-step iteration [4].

1. **Selection.** It involves discovering and choosing a specific citation in the whole repository. This step is already available in a common digital library.
2. **Organization.** It involves creating and structuring a personal information space according to individual interests. This step goes beyond current opportunities because it allows not only to store collections of citations of interest but also to group them using the desired metadata and structure.
3. **Sharing.** It involves making public some selected collections and corresponding metadata in order to support a community knowledge evolution.

In the following, we focus on the second step, depicting a simple scenario that shows how the experience of a single user can reflect on the previously proposed abstract model.

Alice is an ACM member with a web account on the online portal of the library. She has just performed a query on the web portal using as keywords the sentence *collaborative tagging* and then, she has selected a citation of interest from the list of results, e.g. the article “*Usage patterns of collaborative tagging systems*” referred in this paper as [5]. Alice now has the opportunity to save the selected citation into her own personal information space using the “Save this Article to a Binder” feature provided by the ACM.

Saving an article into a virtual personal space is a sign of a real interest for the citation, hence we can assume that Alice is wishful to provide the metadata she considers most appropriate for annotating the selected citation. However, to avoid burdening Alice’s experience, authoring metadata has to remain as simple as in existing collaborative tagging systems. The task assigned to Alice is just to browse a space of suggested metadata, pointing out the most favorites and eventually proposing new ones. Through the DOI assigned to every citation, the system is able to univocally identify the selected citation, and a large set of metadata related to that publication can be retrieved from different systems freely available on the web. For example for the citation selected by Alice the system could retrieve tags from services like CiteULike, Bibsonomy⁸ and Connotea⁹. In terms of the model presented, assuming that *res* is the identifier for the citation selected by Alice, we can consider all the tags associated to *res*, T_{res} , retrieved from a finite set of collaborative tagging systems, where:

$$T_{res} = \{t \in T_i \mid (u, t, res) \in E_i\}, \forall F_i = (N_i, E_i) \in G^F$$

In order to create such a space of metadata, a web service is needed to retrieve this kind of information from different operational systems. This web service can use the model presented in section 2 to uniformly represent in RDF the metadata collected around the Web. Given in input the identifier of a specific resource (e.g. a DOI for a paper) the web service will return as output a subset of T_{res} . Whether available, the web service will invoke the APIs that some systems already offer. Conversely, for systems that do not make their tags available via APIs, our web service will invoke screen scrapers to extract relevant information from web pages and restructure it into RDF. The scraping strategy together with the use of RDF promotes the sharing across a wide variety of tagging sources, regardless of the APIs they provide, and, at the same time, supports users in their personal organizational activity.

This space of metadata can be then normalized, using a filtering process to discard useless tags (like those occurring isolated) and to group those very similar to each other (e.g. singular and plural). As a result, Alice is presented a space of metadata similar to the one depicted in Figure 2. She, then, selects the term *classification* to annotate the previously selected citation. Using a lexical resource, such as Wordnet¹⁰, a search for possible multiple senses associated to the selected term can be performed. Four senses are retrieved from Wordnet for the noun *classification* and Alice

⁸ <http://www.bibsonomy.org>

⁹ <http://www.connotea.org>

¹⁰ <http://wordnet.princeton.edu>

disambiguates them selecting the sense one (Figure 3). Furthermore, Wordnet can provide synonyms, hypernyms and hyponyms related to the selected sense. The system can thus map the term chosen by Alice to a corresponding concept including relationships with other related concepts.



Figure 2. An example of the space of metadata

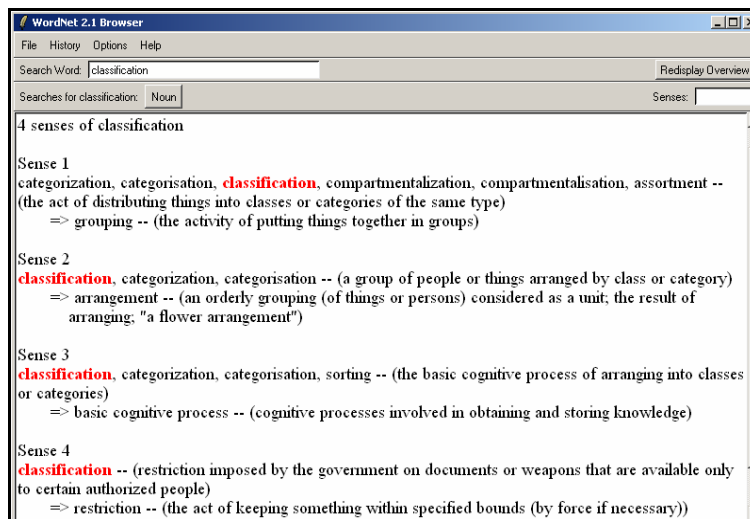


Figure 3. Senses for the term classification

4 Conclusion and future work

We have presented an abstract model for representing a generic collaborative tagging system. Using the RDF as the underlying technology to store the metadata created by different online communities, we depicted a scenario in the domain of a scientific digital library.

We are developing a tool to support the user interaction process, from the selection of information to the sharing of their personal knowledge. We intend to develop a software agent which is able to monitor users' interactions with the system and learn about users' interests. The agent will gain access to metadata in users' personal information spaces to discover topics of interest. The agent will benefit from

processing metadata expressed as RDF statements, rather than simple keywords expressed in natural language.

Although we have depicted a scenario for a research community, our approach applies to online communities in general. We are currently involved in a project aiming to build a user community around a large archive of ancient and contemporary literature, owned by a national book publisher.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* (2001).
2. Broekstra, J., Kampman, A., Harmelen, F. v.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In *Proceedings of the 1st International Semantic Web Conference, LNCS, Vol. 2342. Springer-Verlag, (2002) 54-68.*
3. Carroll, J. J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web, (2005).*
4. Gendarmi D., Abbattista F., Lanubile F.: Fostering knowledge evolution through community-based participation. In *Proceedings of the 1st Workshop on Social and Collaborative Construction of Structured Knowledge at WWW'07, (2007), to appear.*
5. Golder, S., Huberman, B.: Usage patterns of collaborative tagging systems, *Journal of Information Science, 32, 2 (2006), 198-208.*
6. Halpin, H., Robu, V., Shepard, H.: The Dynamics and Semantics of Collaborative Tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop at ISWC'06, (2006).*
7. Hayes, J., Gutierrez, C.: Bipartite Graphs as Intermediate Model for RDF. In *Proceedings of the 3rd International Semantic Web Conference, LNCS, Vol. 3298. Springer-Verlag, (2004) 47-61.*
8. Hepp, M.: Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies. *IEEE Internet Computing, 11, 1, (2007), 90-96.*
9. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G: BibSonomy: A social bookmark and publication sharing system. In *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th Int. Conf. on Conceptual Structures, (2006).*
10. Huynh, D., Mazzocchi, S., Karger, D.: Piggy Bank: Experience the Semantic Web Inside Your Web Browser. In *Proceedings of the 4th International Semantic Web Conference, LNCS, Vol. 3729. Springer-Verlag, (2005) 413-430.*
11. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference, LNCS, Vol. 3729. Springer-Verlag, (2005) 522-536.*
12. Udell, J.: Collaborative knowledge gardening. *InfoWorld, August 2004.*
13. Vander Wal, T.: Explaining and Showing Broad and Narrow Folksonomies, *Personal InfoCloud, February 2005.*
14. Zhdanova, A.V., Predoiu, L., Pellegrini, T., Fensel, D.: A Social Networking Model of a Web Community. In *Proceedings of the 10th International Symposium on Social Communication, (2007).*