

Relating Folksonomies with Dublin Core

Maria Elisabete Catarino
University of Minho, Portugal / Capes-
MEC-Brazil
ecatarino@dsi.uminho.pt

Ana Alice Baptista
University of Minho, Portugal
analice@dsi.uminho.pt

Abstract

Folksonomy is the result of describing Web resources with tags created by Web users. Although it has become a rich basis for the description of resources, in general terms it is not being conveniently integrated in metadata. However, if the appropriate metadata elements are identified, then further work may be done in order to automatically assign tags to these elements (RDF properties) and use them in Semantic Web applications. This article presents research carried out to continue the project Kinds of Tags, which intends to identify elements required for metadata originating from folksonomies and to propose an application profile for DC Social Tagging. It will provide information that may be used by software applications to assign tags to metadata elements and, therefore, means for tags to be conveniently gathered by metadata interoperability tools. Despite the unquestionably high value of DC and the significance of the already existing properties in DC Terms, the pilot study show revealed a significant number of tags for which no corresponding properties yet existed. A need for new properties, such as Action, Depth, Rate, and Utility was determined. Those potential new properties will have to be validated in a later stage by the DC Social Tagging Community.

Keywords: folksonomy; social tagging; metadata; Dublin Core.

1. Dublin Core and Folksonomies: the context

The highly active participation of users in the construction and organization of Internet contents arises from the evolution of the technologies used in the Web, the so-called Web 2.0. It is “the network as platform, spanning all connected devices; Web 2.0 applications are those that make the most of the intrinsic advantages of that platform: delivering software as a continually-updated service that gets better the more people use it, consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others, creating network effects through an ‘architecture of participation’, and going beyond the page metaphor of Web 1.0 to deliver rich user experiences”. (O’Reilly, 2005).

Among the new possibilities of the Web 2.0 folksonomy comes up as “the result of personal free tagging of information and objects (anything with an URL) for one's own retrieval. The tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information” (Wal, 2006). The tags which make up a folksonomy would be key-words, categories or metadata (Guy; Tonkin, 2006). In this brief definition of tag, it can be noticed that tags can play different roles.

Folksonomies describe the Web resources and as such it may be expectable that they are intelligible by machines and thus used by Semantic Web applications. To do so, properties (also known as “RDF links”) are needed in order to clarify and express how given tags relate to the resource they describe. The DC Terms properties (from now on only referred as DC properties) are of high value to be used as a basis for interoperability and their wide acceptability is a good measure of this value. However, they are oriented to describing resources from the classical standpoints of authors and libraries, whereas in Web 2.0, resources are described from the highly diverse perspective of users.

The project Kinds of Tags (KoT) focuses its attention “on the analysis of tags that are in common use in the practice of social tagging, with the aim of discovering how easily tags can be ‘normalised’ for interoperability with standard metadata environments such as the DC Metadata Terms” (Baptista et al., 2007). Within KoT it was observed that there are some tags to which none of the existing DC properties could be adequately assigned. This indicates that other metadata elements might need to be identified. Preliminary results from this project were presented in DC-2007 and NKOS-2007 describing some probable new elements: Action_Towards_Resource, To_Be_Used_In, Rate and Depth (Baptista et al., 2007 and Tonkin et al., 2007).

In order to continue this analysis a deeper and more detailed research is underway and it aims to answer the following questions:

- do the DC properties have the necessary semantics to clarify and express how given tags relate to the resource they describe?
- if not, which other properties that hold this semantics can be identified to complement DC and to be used in social tagging applications?

This research uses the same data set that was used in KoT and begun with a detailed pilot study regarding the tags of the first five resources of the data set. This article presents the results of the pilot study and also refers some preliminary results of the final study. These indicate that some new properties may be needed for social tagging applications, which implies the possible construction of an application profile to be proposed to the Social Tagging community.

2. The research Project: an in-depth study following up KoT preliminary results

The dataset used in this project is the same of KoT: it is composed of 50 records of resources which were tagged in two systems of social bookmarking: Connotea and Delicious. Each record is composed has information distributed in two groups of data: a) data related to the resource as a whole: URL, number of users, research date; and b) data related to the tags assigned to the resource: social bookmarking system, user’s nickname, bookmarked date and the tags.

A relational database was set up with the DCMI Metadata Terms and the KoT data set that was imported from its original files. The following tables were created: Tags, Users, Documents, Key-tags and Metadata.

There is a total number of 5098 tags (Connotea: 901; Delicious: 4819). The total number of users amounts to 15.381 (Connotea: 509; Delicious: 14.872). Considering that different users in different resources repeatedly assigned a tag, there is a total of 75.429 tag occurrences (Connotea: 3.698; Delicious: 71.731). It is important to consider the total number of tag occurrences, since a tag could correspond to different metadata elements depending on the resource to which it was assigned (for instance, a tag could correspond to Subject in a resource and to Title or Description in another).

The whole study is made manually in order to be as precise as possible regarding the meaning of the tags. It was divided in four stages: 1 – Analysis of tags; 2 – Identification of complementary properties; 3 – Formalization of the new properties in an ontology-like representation; 4 – Validation by the community and release of the first version of the proposal.

The first stage consists of an analysis of all tags contained in the dataset. At this stage all tags assigned to the resources are analysed, grouped in what we call key-tags and then DC properties are assigned to them when possible. A Key-tag is a normalised tag that represents a group of similar tags. For instance, the key-tag `Library Science` stands for tags `library.science`, `library_science` or `library-science`.

Once that the meaning of tags is not always clear, it is necessary to dispel doubts by complementarily turning to lexical resources (dictionaries, encyclopedias, Word Net, Wikipedia,

etc), and analyzing other tags of the same users. Contacting the users may be a last alternative to try to find out the meaning of a given tag.

The second stage aims at proposing complementary properties to the ones already existing in the DCMI Metadata Terms (DCMI Usage Board, 2008). Key-tags to which none DC property was assigned in stage one will now be subject to further analysis in order to identify new properties specific to Social Tagging applications. This analysis takes into account all DC standards and recommendations, including the DCAM model, the ISO Standard 15836-2003 and the NISO Standard Z39.85-2007.

The next stage comprises the adaptation of an already existing DC ontology-like representation of the DC elements and their semantics. This will make use of Protégé, an ontology editor developed at Stanford University. The ontology will be encoded in OWL, a language endorsed by the W3C.

Finally, the fourth stage intends to submit a proposal for a DC Social Tagging application profile to the DC social tagging community for comments and feedback via online questionnaires. After this phase, a first final version of the proposal will be submitted to the community.

A pilot study was conducted for the first two stages with the first five resources of the data set. It allowed to refine the proposed methodology and, in the first stage, to verify whether the proposed variants for grouping and analyzing tags are adequate. In the second stage, the pilot study allowed to have a preliminary overview of the percentage of tags to which DC properties could be assigned and, complementarily, the percentage of tags that would fit in new properties. As it was impossible to determine the meaning of some tags, there is a high percentage of non-assigned tags.

An important concern regarding tag analysis is the fact that as tags are assigned by the resources' users, that inevitably leads to a lack of homogeneity in their form. Therefore, it was necessary to establish some rules in order to properly analyze tags, establish key-tags and relate DC properties with them.

3. Rules for the first two stages

3.1. Rules for the first Stage

The first rule to be observed concerns the alphabet. In this project, only tags written in Latin alphabet were considered. Further studies should involve the analysis of tags written in different alphabets. For example: Greek/Ελληνική, cyrillic/Кирилица, Chinese/中國, Japanese/日本語, etc.

Another rule is related to language. The dataset comprises tags written in different languages. It was possible to identify and translate 425 tags written in languages other than English, which corresponded to 8,3% of the total number of tags as shown in Table 1.

Table 1 - Number of identified and translated tags in languages other than English

ISO 639 acronym	Language	No. of tags	ISO 639 acronym	Language	No. of tags
CA	Catalan	43	HU	Hungarian	9
CS	Czech	3	IT	Italian	16
DA	Danish	3	MUL	Multiple Languages ¹	57
DE	German	51	NL	Dutch	16
ES	Spanish	47	NO	Norwegian	9
ET	Estonian	2	PL	Polish	2
EU	Basque	1	PT	Portuguese	77

¹ Tags that have the same spelling in several languages.

FI	Finnish	9	RO	Romanian	4
FR	French	68	SV	Swedish	8
HR	Croatian	1	TR	Turkish	1

Most of the tags were, however, written in English. Thus, English was the chosen language to represent Key-tags.

Depending on the Key-tags, certain criteria concerning the classification of words need to be established: simple or compound, singular or plural, based on a thesaurus structure in its syntactical relations. In these cases, the rules to establish thesauri structure were followed as indicated by ISO 2788-1986 Standard.

It was still necessary to create rules to deal with compound tags, as they contain more than one word. There are two kinds of compound tags: (1) the ones that are related to only one concept and therefore originate only one key-tag (e.g. `Institutional Repositories`); and (2) the ones that are related to two or more concepts and therefore originate two or more key-tags (e.g. `digital-libraries:dublincore`).

In the first kind, compound tags are composed by a focus (or head) and a modifier (International Standards Organization, 1986). The focus, i.e. the noun component which identifies the general class of concepts to which the term as a whole refers, and the modifier, i.e. one or more components which serve to specify the extension of the focus; in the example above: `Institutional` (modifier) `Repositories` (focus). It is a compound term that comprises a main component or focus and a modifier that specifies it.

In the second kind, compound tags are related to two or more distinct Key-tags, as for example: `digital-libraries:dublincore`, which would be part of the group of two distinct Key-tags: `Digital Libraries` and `Dublin Core`. In this second segment there is not a relation of focus/difference between the components as they are totally independent.

3.2. Rules for the second stage

In the occurrence of Simple tags there is a peculiarity to be noticed that relates to the way tags are inserted in the social bookmarking sites: the way tags are inserted can interfere with the system's indexation. In Delicious the only separator is the space character and everything that is typed separated by spaces will be considered distinct tags. For example, if the compound term `Social Tagging` is inserted containing only the space as separator, the system will consider two tags: `Social` and `Tagging`. In order to be inserted as a compound tag it is necessary to use special characters such as underscore, dashes and colons. Some examples of such kind of compound tags are: `social+tagging`, `social_tagging`, `social-tagging`.

In Connotea tags are also separated by a space or a comma. However, Connotea suggests to users to type compound tags between inverted commas. For example, if the user inserts `Controlled Vocabularies` without placing the words between inverted commas, the words will be considered two distinct tags; however, if they are typed between inverted commas ("`Controlled Vocabularies`") the system will generate only one compound tag. This simple, yet important issue, has a high implication on the system's indexation of the tags.

To exemplify what is said above there is an example of a Delicious user who, when assigning tags to the resource "The Semantic Web", written by Tim Berners-Lee, inserted the following tags: `the`, `semantic`, `web`, `article`, `by`, `tim`, `berners-lee`, without using the characters of word combination (`_`; `-` etc). The system generated seven simple tags. However, it is clear that these tags can be post-coordinated² to have a meaning such as Title, Creator and Subject.

Thus, as a first rule, in the cases when simple tags could clearly be post-coordinated, they were analyzed as a compound term for the assignment of the DC Property. However, this analysis

² Post-coordination is the principle by which the relationship between concepts is established at the moment of outlining a search strategy (Angulo Marcial, 1996 apud Menezes; Cunha; Heemann, 2004).

could only be carried out in relation to only one resource's user at a time and never to a group, since it can mischaracterize the assignment of properties.

The second rule concerns tags that correspond to more than one DC Property. It is considered two different situations: simple and compound tags. The easiest case is the one of simple tags. If simple tags occur to which two or more properties can be assigned, then all the properties are assigned to the tag. For example in the resource entitled "An Architecture for Information", the properties "Title" and "Subject" are assigned to the Key-tag Architecture.

As explained earlier, compound tags, however, can correspond to two or more key-tags. Thus the relationship with DC properties is made through the key-tags. These are treated as simple tags in the way they are related to DC properties. For example the tag doi:10.1045/april2002-weibel, corresponds to three Key-tags, doi:10.1045, april 2002 and Stuart L. Weibel, each one of them corresponding to a different property: Identifier, Date and Creator (respectively). There may also be cases of compound tags that represent two different values for the same property, as in folksonomiestagging, that was splitted into two Key-tags: Folksonomy and Tagging, to which both the subject property was assigned.

Another rule is related to tags whose value corresponds to the property Title. Tags will be related to the element "Title" when they are composed by terms found in the main title of the resource. For example, Folksonomies, WEb2.0. Another example is the case of the resource entitled "Social Bookmarking Tools", where the tags Social, Bookmarking, Tools, that were assigned by the same user, and thus, are post-coordinated.

4. Tag Analysis

As stated earlier, this stage consists of an analysis of all tags contained in the dataset. At this stage all tags assigned to the resources are analyzed, grouped in key-tags and then DC properties are assigned to them when possible. In this stage it was necessary to use lexical resources (dictionaries, WordNet, Infopedia, etc) and other online services, such as online translators, in order to fully understand the meaning of tags. In some cases further research and analysis of other tags of a given user, or even a direct contact with this user by email was necessary in order to understand the exact meaning of a given tag.

The first step of tag analysis comprises grouping tag variants: a) language; b) simple/compound; c) abbreviations and acronyms; d) singular/plural; e) capital letter/small letter. Then a Key-tag is assigned to each of these groups according to the rules presented in section 3. Following, there are two examples of tags and their assigned key-tags:

- Tags: metadados, metadata, meta-data, metadata/, métadonnées, metadata.tags; Key-tag: METADATA;
- Tags: informationscience, information science, information.science, Ciències de la informació, is; Key-tag: INFORMATION SCIENCE;

The above key-tags show a variation in :

- spelling: information science, informationscience, information.science and is;
- form (Singular/Plural): metadata, metadados, métadonnées;
- language: information science (EN), ciències de la informació (CA); metadados (PT), metadata (EN) and métadonnées (FR).

The examples above also show the two kinds of compound tags. Compound Tags focus/modifier like information science are assigned to only Key-tag. Tags composed of

two focus components like metadata . tags are assigned to two distinct Key-tags: Metadata and Tags.

After Key-tags definition, an analysis to verify which DC Properties correspond to these tags is carried out. This analysis becomes more complex as the DCMI Terms definitions are purposely general enough so that the description of the electronic documents with a small, though sufficient, number of metadata is possible.

5. Complementary Properties - Results from the Pilot study

In the pilot study it was analyzed data related to the first five resources of the data set. This implied the analysis of a total of 311 tags with 1141 occurrences and assigned by 355 users.

The accomplishment of the pilot study was also important in order to compare its results with the results of KoT. This study, is, however, much more detailed than the one in KoT which generated some indicative results: 1) "Users apply tags not only to describe the resource, but also to describe their relationship with them (e.g. to read, to print,...)"; 2) "Do tags correspond to atomic values? Many of the tags have more than one value, with potential results in more than one metadata element assigned"; 3) "Into which DC elements can tags be mapped? 14 out of the 16 DC elements, including Audience, have been allocated" (Baptista et al., 2007).

The results from KoT indicated that the following new elements could be added to the DC Social Tagging Application Profile: Action Towards Resource (e.g., to read, to print...); To Be Used In (e.g. work, class); Rate (e.g., very good, great idea) and Depth (e.g. overview).

The preliminary results from the current pilot study confirm the need for the proposal of new metadata elements for Social Tagging applications. However, it points out for some more elements than KoT did. The results of this study are presented in the following sections and, when pertinent, they will be compared with the results of KoT.

From the 311 tags analyzed in the pilot study, 212 Key-tags were created. From this amount, 159 Key-tags (75%) of which corresponded to the following DC properties: Creator, Date, Description, Format, Is Part Of, Publisher, Subject, Title and Type. From these, 90,5% corresponds to Subject. The other properties present the following percentages of allocation: Type 5%; Creator, Is Part Of and Title 3,1% each; Date and Publisher 1,3% each and Format 0,6%.

No DC properties could be assigned to the other 53 Key-tags (25%). New complementary properties were defined and their definition is still in process. The following properties that were identified in the pilot study will be described: Action, Category, Depth, Rate, User Name, Utility and Notes.

From these eight possible new properties, four had already been suggested in the KoT. Nonetheless, until the end of the full study, others may be added, or even, some of the ones proposed here may be withdrawn, depending on the evolution of the study.

In the group of the 53 Key-tags the following percentages for the properties proposed were observed: Action, Rate and Utility (15,1% each), Category (11,3%), Depth (9,4%), Notes (7,5%) and User Name (1,9%). There is also a group of Key-tags (24,5%) to which it was not possible to assign or propose any property as their meaning in relation to the resources and users was not possible to identify.

5.1. Action

There is a group of Key-tags that represents the action of the user in relation to the tagged resource. It is a kind of tag that can be easily identified since the action is expressed in the very term itself when tagging the resource. As example the tags which represent the action To Read, attributed to 6 users, all from Delicious: `_toread`, `a_lire`, `toread`.

5.2. Category

This property includes Tags whose function is to group the resources into categories, that is, to classify the resources. The classification is not determined by subjects or theme of the resource, since, in these cases, the key-tags could correspond to the Subject property.

This property is not easy to identify, since it is necessary to analyze the given tag in the context of the totality of tags that user has inserted, independently of the resource under analysis. In some cases it may become necessary to analyze the whole group of resources the user has tagged with the tag that is object of analysis.

For instance, during the analysis of the Key-tag DC Tagged it was noticed that the corresponding resources had also other tags with the prefix dc: (e.g.: dc:contributor, dc:creator, dc:Publisher, dc:language or dc:identifier, among others). It was concluded that the tag "DC Tagged" could be applied to group all the resources that were tagged by tags that were prefixed by dc:. Therefore it was considered a "Category" since it is not a classification of subjects or a description of the content of the resource.

5.3. Depth

This type of tag confers the degree of intellectual depth to the tagged resource. As Word Net, Depth "degree of psychological or intellectual profundity" (WorNet, 2008). A resource was tagged by six users who assigned the following tags to represent the degree of profundity of the resource: diagram, doc/intro, overview, semanticweb.overview, semwebintro. These tags mean that users are describing a resource which content is thought as a schematic or a summarized explanation, introductory and general.

5.4. Notes

This element may be proposed to represent the tags that are used as a note or reminder. As WordNet, "a brief written record" that has the objective of registering some observations concerning the resource, but that does not refer to its content and does not intend to be used as its classification or categorization (WordNet, 2008). A note should be understood as: an annotation to remind something; observation, comment or explanation inserted in a document to clarify a word or a certain part of the text (Infopedia, 2008).

From the five analyzed resources, the following tags considered as "Notes" were identified: Hey, Ingenta, OR2007, PCB Journal Club. For instance, there is a resource that received the tags Hey and OR2007. The first tag, Hey, refers to Tony Hey, a well-known researcher who made a debate on important issues that were related to the tagged resource³.

The second tag makes reference to the Open Repositories 2007, event where Tony Hey mentioned above made a Keynote speech. However, interestingly enough, the tagged resource does not have any direct relation neither with that event nor with Tony Hey⁴.

5.5. Rate

Rate, meaning pattern, category, class or quality is important to include tags that are evaluating the tagged resource. Thus, the user categorizes the resource according to its quality when using this type of tag.

The following tags were related to the property: academic, critical, important, old, great, good and vision. These are generally easily identified as Rate in each one of the terms. In other cases, the tags may be doubtful and it becomes necessary to analyze them in

³ This information was given by the user who assigned the tags.

⁴ This information confirmed by the author of the resource himself (the creator).

relation to the tags assigned by the user to the resource under analysis as well as to the whole collection of resources tagged by that user. For instance, the tag `vision` could have several meanings, but, after an analysis to the collection of resources, it may be concluded that it is classifying the quality of the resource.

5.6. User Name

The Tag “User Name” labels the resource with the name of a user. The analyzed resource had the name of the user of the tagged resource.

Only one tag of this type was identified in the pilot study. Despite the preliminary results presented here, it is assumed that here may be other occurrences.

5.7. Utility

This property would gather the tags that registered the utility of the resource for the user.

It represents a specific categorization of the tags, so that the user may recognize which resources are useful to him in relation to certain tasks and utilities.

`Maass` is a tag that was bundled in “Study”. The term represents the name of a teacher, information found in the user’s notes in two resources tagged with `Maass`: “Forschung von Prof. Maass an der Fakultat Digitale Medien an der HFU”; and “Unterlagen für Thema ‘Folksonomies’ für die Veranstaltung “Semantic Web” bei Prof. Maass”.

6. Final Considerations

In the pilot study 212 key-tags were generated. DC properties could be assigned to 159 (75%) of those. The identified new properties were assigned to 40 key-tags (18,9%) and 13 key-tags (6,1%) were left without assignment because it was not possible to identify their meaning. As this data shows, DC properties can be assigned to a great part of the tags analyzed in the pilot study. However, still, 25% of them are left out.

The final study has already been finalized and although it is not yet possible to show the final results, it is possible to say that the percentage tags unassigned to DC elements is higher and it will probably range between 35% and 45% (39,5% is the provisory number, but some further analysis will still be done). It is not possible to assign properties to a great number of those tags because their meaning could not be identified. However, new properties could be assigned to most of them (the provisory number for tags assigned with new properties is 26,5%, while the provisory number for tags left unassigned is 13%).

DC plays a fundamental role as a foundation for metadata interoperability. From this study it is evident that DC keeps this role even in the presence of a paradigm shift, as with Web 2.0 and the social tagging applications. However, as in these applications the user is in the centre of the description process, there is a significant number of new kinds of values (terms/tags) not previously foreseen in the scope of DC and to which current DC properties cannot be assigned.

This research aims at discovering if the DC properties have the necessary semantics to hold tags and, if not, it aims at finding which other properties that hold the lacking semantics can be coined to complement DC and to be used in social tagging applications. This application profile will allow rich descriptive tags to be handled by metadata interoperability protocols and consequently, to enrich the semantic Web.

This work begun with a pilot study for the first five resources of the KoT data set in order to refine the methodology and have a preliminary overview of the possible new properties that could be identified, if any. This article presents the results from the pilot study and already gives some lights on the final study. The final research results will then be submitted to the DC community for evaluation and validation purposes.

References

- BAPTISTA, A. A. et al. (2007). Kinds of Tags: progress report for the DC-Social tagging community. *Presented in DC-2007, International Conference on Dublin Core and Metadata Applications, 2007*. Retrieved September 4, 2007, from <http://hdl.handle.net/1822/6881>.
- CURRÁS, Emília. (2005). *Ontologías, taxonomía y tesauros: manual de construcción y uso*. 3.ed. act. y ampl. Madrid: Treas.
- DCMI Usage Board. (2008). *DCMI Metadata Terms*. Retrieved March 10, 2008, from <http://dublincore.org/documents/dcmi-terms/>.
- GUY, Marieke; TONKIN, Emma. (2006, January). Folksonomies: tidying up tags?. *D-Lib Magazine*, (12,1). Retrieved December 12, 2006, from <http://www.dlib.org/dlib/january06/guy/01guy.html>.
- INFOPEDIA. (2008). Retrieved March 10,2008, from <http://www.infopedia.pt>.
- MENEZES, E. M.; CUNHA, M. V.; HEEMANN, V. M. (2004). *Glossário de análise documental*. São Paulo: ABECIN. (Teoria e Crítica, 01).
- O'REILLY, T. (2005). *Web 2.0: Compact definition?* O'Reilly Radar Blog, 1 October 2005. Retrieved November 6, 2006, from http://radar.oreilly.com/archives/2005/10/web_20_compact_definition.html.
- TONKIN, E. et al. (2007). Kinds of tags: a collaborative research study on tag usage and structure (Presentation). *Presented in European Networked Knowledge Organization Systems (NKOS), 2007*. Retrieved December 10, 2007, from <http://www.us.bris.ac.uk/Publications/Papers/2000724.pdf>.
- WAL, Thomas Vander. (2006). *Folksonomy definition and wikipedia*. Retrieval November 22, 2006, from <http://www.vanderwal.net/random/entrysel.php?blog=1750>.
- WORDNET. (2008). Retrieval November 22, 2006, from <http://wordnet.princeton.edu/>.