# Relating folksonomies with Dublin Core

Maria Elisabete Catarino*

Departamento de Ciência da Informação,
Universidade Estadual de Londrina,
Campus Universitário, Caixa Postal 6001,
Londrina, Paraná, Brazil
E-mail: beteca@uel.br
*Corresponding author

Ana Alice Baptista

Departamento de Sistemas de Informação,
Universidade do Minho,
Campus de Azurém, Guimarães, Portugal
E-mail: analice@dsi.uminho.pt
Website: http://www.dsi.uminho.pt/~analice

**Abstract:** This article presents a research carried out to continue the project Kinds of Tags, which intends to identify elements required for metadata originating from folksonomies. It will provide information that may be used by intelligent applications to assign tags to metadata elements. Despite the unquestionably high value of DC and DC Terms, the pilot study revealed a significant number of tags for which no corresponding properties yet existed. A need for new properties was determined. This article presents the problem, motivation and methodology of the underlying research. It further presents and discusses the findings from the pilot study.

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

**Reference** to this paper should be made as follows: Catarino, M.E. and Baptista, A.A. (2010) ‘Relating folksonomies with Dublin Core’, *Int. J. Metadata, Semantics and Ontologies*, Vol. 5, No. 4, pp.285–295.

**Biographical notes:** Maria Elisabete Catarino is a Professor of Information Technology applied to the Organization of Information and Web Information Management at the Universidade Estadual de Londrina, Brazil. She has experience in the area of information science, with emphasis on information technology, working primarily in the following topics: Web, Ontologies, Folksonomies, Dublin Core and Information Technologies.

Ana Alice Baptista is a Researcher and a Professor in the Information Systems Department at the Universidade do Minho, Portugal. She belongs to the Directive Commission of the Doctoral Programme in Information Systems (University of Minho). She is a member of the DCMI Advisory Board. She participated in several R&D projects and has acted as evaluator of FP7 project proposals. She is the head of the Odisseia Research Group. Her main areas of interest are Scholarly Communication, Digital Libraries and the Semantic Web. She is also interested in the social aspects of the internet. More information is available at http://www.dsi.uminho.pt/~analice.

## 1 Dublin Core and folksonomies

The highly active participation of users in the construction and organisation 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)

The Web 2.0 is growing, more and more social networks are being created, and the current social networks are gaining popularity. Some statistics about Web 2.0 show the real effects of the ‘architecture of participation’ on the social networks. The following are a few numbers (Gulati, 2009; Schroeder, 2009). Facebook, a social network service, has more than 150 million active users who have uploaded over

10 billion images. Twitter, a free social networking and microblogging service, has about 1 million active users, and it will have nearly 100 million visitors some time in this year. Youtube is serving 75 billion video streams to 375 million unique visitors. Wikipedia has 11,461,663 registered users.

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)

Folksonomies are relatively recent, but perfectly justified for the organisation of web resources. Studies have been conducted to analyse the folksonomies in the context of information organisation. Some recent papers concerning the analysis of tags in this context can be cited (Spiteri, 2007; Thomas et al., 2009).

Thomas et al. (2009) in a study, whose proposal was to provide a quantitative analysis of the extent to which folksonomies replicate the Library of Congress Subject Headings (LCSH) and see if folksonomies would successfully complement cataloguer-supplied subject headings in library catalogues, concluded that "social tagging does indeed augment the LCSH providing additional access to resource".

Spiteri (2007) considered that "Folksonomies have the potential to add much value to public library catalogues by enabling clients to: store, maintain and organise items of interest in the catalogue using their own tags". To understand this context, a research was developed with the purpose of examining how the tags that constitute folksonomies are structured. Spiteri concluded that "… folksonomies could serve as a very powerful and flexible tool for increasing the user-friendliness and interactivity of public library catalogues …".

According to Spiteri,

> "traditionally, such indexing is performed either by an authority, such as a librarian or a professional indexer, or else is derived from the authors of the documents; in contrast, folksonomies allow anyone to freely attach keywords or tags to content." (Spiteri, 2007)

Tags allow users to represent resources according to the way they perceive them, i.e., it is a form of representing a personal understanding or point of view one user has towards the resource (Mathes, 2004; Quintarelli, 2005; Feinberg, 2006). It results from the attribution of tags that may represent either the physical or the thematic description of a resource, as well as other aspects related to its functionalities, or its relation with the user that tags it (from now on referred as the tagger). Then, according to Guy and Tonkin (2006), we could say that the tags are keywords, category names, or metadata, because they represent either the physical or the thematic description.

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 to clarify and express how given tags relate to the resource they describe.

The Dublin Core Metadata Element Set, aka Dublin Core or just DC, is a vocabulary of 15 properties for use in resource description that have been endorsed in several standards, including ISO Standard 15836-2003 (DCMI, 2008). It arouse after some workshops intended to discuss issues regarding the description of web resources. One of those issues was the need of creating a pattern of metadata, addressing the interoperability of data and the recovery of information.

The DCMI Metadata Terms (aka DCMI Terms) is a set of all metadata terms maintained by the DCMI. It includes the DC 15 elements and other properties (DCMI Usage Board, 2008). This subset of DCMI Terms that is composed by all the properties and sub-properties maintained by DCMI will be referred to as DC properties in the context of this paper.

This set was created and is maintained by the DCMI. The DCMI is "an open organisation engaged in the development of interoperable metadata standards that support a broad range of purposes and business models" (DCMI, 2010). DCMI is highly committed to promoting interoperability at several levels, as can be demonstrated by a series of recommendations on this topic (e.g., The Singapore Framework for Dublin Core Application Profiles (Nilsson et al., 2008), Guidelines for Dublin Core Application Profiles (Coyle and Baker, 2008), the Interoperability Levels for Dublin Core Metadata (Nilsson et al., 2009)).

The DC is used in the scope of various projects and tools across the web (see http://dublincore.org/community-and-events/ for more information). The fact that it is endorsed by formal standards and the fact that it is widely used makes DC properties a good basis for interoperability.

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 describe 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 current project follows the project KoT, a small non-funded project that resulted from a challenge made by Ana Alice Baptista to the DC Social Tagging mailing list. As a response to this challenge, several people joined 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). As a non-funded project, it could not go beyond its first results. These were, nevertheless, relevant enough to open room for more research on the same subject. This paper presents the first results of that more in-depth research that followed KoT.

This study is, however, much more detailed than the one in KoT, which generated some indicative results:

- "Users apply tags to describe not only the resource, but also their relationship with them (e.g., to read, to print, …)"
- "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"
- "Into which DC elements can tags be mapped? 14 out of the 16 DC elements, including Audience, have been allocated" (Baptista et al., 2007).

Within KoT, it was observed that there are some tags with which none of the existing DC properties could be adequately related. This indicates that other metadata elements might need to be identified. Preliminary results from this project were presented in DC-2007 and NKOS-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`); Depth (e.g., `overview`) (Baptista et al., 2007; Tonkin et al., 2007).

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. It began with a detailed pilot study regarding the tags of the first five resources of the data set. This pilot study was performed to enhance and refine the project's methodology. As stated by Yin (1989), a pilot study helps the researcher refine the procedures for collection and data recording and gives him the opportunity to test the established procedures.

This paper presents the problem, motivation and methodology of the underlying research. It further presents and discusses the findings from the pilot study, which indicate that some new properties may be needed for social-tagging applications. In Section 2, we will describe the methodological procedures used in the research project. In Section 3, we will present the rules used to properly analyse tags, establish Key-tags and relate DC properties with them. In Section 4, the analysis of the tags is described. In Section 5, we will present the results of the pilot study. Lastly, we conclude with a discussion of future work in Section 6.

## 2 The research project: an in-depth study following up KoT preliminary results

In this research, we intended to do a more in-depth analysis of the same data set used in KoT. Our intention was to validate its results and provide more accurate information about the KoT that was being used by regular social bookmarking users. In KoT, two social bookmarking systems were selected: Delicious and Connotea.

At the time KoT was started in 2007, Connotea favoured consistency between tags of the same user, i.e., for a given resource it used tags that the user had already input for other resources. On the other hand, Delicious favoured consistency between tags of different users for the same resource, i.e., it used tags that other users had input for that same resource. Although it was not a hypothesis to be verified in this study, we assumed that this difference might influence the final set of tags of a given resource. Therefore, systems implementing both perspectives should be included in the study.

Each record of the data set is composed of two groups of data:

a *Data related to the resource*: URI, number of users and research date

b *Data related to the tags assigned to the resource*: Social bookmarking system, user's nickname, bookmarked date and the tags themselves.

There is a total number of 5098 tags that correspond to a total of 79,146 tag occurrences. It is important to consider the total number of tag occurrences because a single tag might relate to different metadata elements, depending on the resource to which it was assigned, i.e., a tag could relate to Title in a resource and to Date in another. For instance, the tag `May` in the paper whose title is "As We May Think" could relate to Title and in another paper entitled "Social Bookmarking Tools" could relate to Date.

The whole study was made manually to be as precise as possible regarding the meaning of the tags. It was divided in five stages:

1 database implementation

2 analysis of tags

3 identification of complementary properties

4 validation of the proposal

5 formalisation of the new properties in an ontology-like representation.

On the first stage, a relational database was set up with information about the DCMI Metadata Terms and the KoT data set that was imported from its original files.

The second stage involves an analysis of all tags contained in the data set. 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`.

In this stage, it was necessary to use lexical resources to help identify the meanings and translations of terms where necessary. In some cases, there was also the need to perform searches on the web using search engines, to identify the meaning of a tag. The most used lexical resources were WordNet, Infopedia and Webster.

WordNet is a database of lexical data in English. It is composed of nouns, verbs, adjectives and adverbs grouped into sets of cognitive synonyms (synsets), each expressing distinct concepts (Wordnet, 2008). WordNet was an extremely important instrument once the visualisation of the several synonym sets allowed the analysis of tags with a dubious meaning.

Infopedia is a Portuguese database of reference contents that covers all areas of knowledge. It includes a broad set of material nature encyclopaedic, linguistic and graphic (dictionaries, encyclopaedias, atlas and graphic resources) (Infopedia, 2008).

Webster is an online multilingual dictionary. The dictionary will soon consist of over 400 modern languages and 10 ancestral languages with some 30 million individual entries across languages (including expressions, technical terminologies and words) (Webster's Online Dictionary, 2008). It allowed for a more efficient translation of tags written in idioms other than English.

In some situations, neither these lexical resources nor the web search engines were effective in the translation or identification of the meanings of tags. In such cases, if the taggers' e-mail was available in some way, an e-mail was sent requesting a clarification. Generally, the contacts of Delicious' users were the only ones available. The Information that was collected through e-mail was very elucidative. Nevertheless, despite all the efforts, there were many tags left whose meanings were impossible to identify.

There were other situations that made it hard to identify the meanings of the tags. There were tags composed by signals and symbols, numbers, representations, abbreviations, mnemonic and mixed formulas, as well as tags with spelling errors. To identify the meaning of signs and symbols is a hard task.

For example, the tag `*****` could be interpreted as meaning 'five stars'. However, what does the tag 'five stars' mean to the user? It could be inferred that it had the purpose of rating the resource on its quality, but this is a conclusion hard to prove.

There were tags made of nothing but numbers (e.g., `11072006`). Identifying the meanings of numbers is not an easy task. Exceptions are dates (about publishing or tagging), or numbers that were explicitly related to the resource (title, topic, editor, etc.).

Some tags were made of representations (`bmj`), abbreviations (`tech`), and others appeared in mnemonic formulas (`2bread`). Representations and mnemonic formulas also make the interpretation hard. It was frequently needed to go to for searches on web search engines.

There were cases where more than one term was found in a single tag (e.g., `boeingreadinglist`). In such cases, the meaning was, generally, easy to interpret.

Mixed tags, i.e., tags composed of words and graphic signs, punctuation signs, symbols, or numbers (e.g., `005 – lagoze`), were analysed as words.

There were also tags with spelling errors (e.g., `buisness`). In such cases, tags were grouped to Key-tags that were correctly spelled.

In the most difficult cases, it was often possible to interpret the meanings of tags through the analysis of the tagger's set of tags. This was done either by analysing his or her tags for that specific resource or for all of his or her resources. As a last try, in the cases where the e-mail of the tagger was available, an e-mail was sent requesting help.

The third stage aims at proposing complementary properties to the ones already existing in the DCMI Metadata Terms (DCMI Usage Board, 2008). Key-tags to which no DC property was assigned in stage 1 will now be subject to further analysis 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 fourth stage intends to validate the proposal of new properties through online questionnaires sent to the community DC.

Finally, the last stage comprises the adaptation of an already-existing DC ontology-like representation of the DC elements and their semantics.

A pilot study was conducted for the first three stages with the first five resources of the data set. It allowed for refining the proposed methodology and, in the second stage, to verify whether the proposed variants for grouping and analysing tags was adequate. In the third 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, this inevitably leads to a lack of homogeneity in their form. Therefore, it was necessary to establish some rules to properly analyse 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 such as Greek, Cyrillic, Chinese and Japanese.

Another rule is related to language. The data set is composed of 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. Table 1 shows the distribution of tags per language, for tags whose meaning was identified.

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

| *ISO 639 acronym* | *Language* | *No. of tags* |
|---|---|---|
| CA | Catalan | 43 |
| CS | Czech | 3 |
| DA | Danish | 3 |
| DE | German | 51 |
| ES | Spanish | 47 |
| ET | Estonian | 2 |
| EU | Basque | 1 |
| FI | Finnish | 9 |
| FR | French | 68 |
| HU | Hungarian | 9 |
| IT | Italian | 16 |
| MUL | Multiple languages* | 57 |
| NL | Dutch | 16 |
| NO | Norwegian | 9 |
| PL | Polish | 2 |
| PT | Portuguese | 77 |
| RO | Romanian | 4 |
| SV | Swedish | 8 |
| TR | Turkish | 1 |

*Tags that have the same spelling in several languages.

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 needed to be established based on a thesaurus structure and in its syntactical relations: simple or compound, singular or plural. 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`)
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 of a focus (or head) and a modifier (International Standards Organization, 1986). The focus is the noun component that identifies the general class of concepts to which the term as a whole refers, and the modifier refers to one or more components that serve to specify the extension of the focus; in the above-mentioned example: `Institutional` is the modifier and `Repositories` is the 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 example, there is not a relation of focus/difference between the Key-tags.

## 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 input in the social bookmarking sites: the way tags are input may 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 input containing only the space as separator, the system considers two tags: `Social` and `Tagging`. To be input 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 inputs `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.

As an example, a Delicious user when assigning tags to the resource 'The Semantic Web', written by Tim Berners-Lee, input 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-coordinated[1] 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 analysed as a compound term for the assignment of the DC Property. This analysis could only be carried out in relation to only one resource's user at a time and never to a group, since it can mischaracterise the assignment of properties.

The second rule concerns tags that correspond to more than one DC Property. It is taken into account 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 may 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 `april2002/Weibel` corresponds to two Key-tags, `April 2002` and `Stuart L. Weibel`, each one of them corresponding to a different property: Date and Creator (respectively). There may also be cases of compound tags that represent two different values for the same property, as in `folksonomiestagging`, which was split 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; i.e., `Folksonomies`, `Web 2.0`. Another example is the case of the resource entitled '`Social Bookmarking Tools`', where the tags Social, Bookmarking and Tools were assigned by the same user, and, thus, are post-coordinated.

## 4 Tag analysis

As stated earlier, this stage comprises an analysis of all tags contained in the data set. At this stage, all tags assigned to the resources are analysed, 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, 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 e-mail was necessary to understand the exact meaning of a given tag.

The first step of tag analysis comprises grouping tag variants:

- language
- simple/compound
- abbreviations and acronyms
- singular/plural
- 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, meta-data, metadata, metadata/, métadonnées, metadata.tags`; Key-tag: `METADATA`
- *Tags*: `informationscience, information science, information.science, Ciències de la informació, is`; Key-tag: `INFORMATION SCIENCE.`

The above-mentioned Key-tags show a variation in:

- *Spelling*: `Information science, informationscience, information.science and is`
- form (Singular/Plural): `metadata, metadados, métadonnées`
- Language: `information science (EN), ciènces de la informació (CA); metadados (PT), metadata (EN) and métadonnées (FR).`

The above-mentioned examples also show the two kinds of compound tags. Compound Tags focus/modifier like `information science` are assigned to only one Key-tag. Tags composed of two focus components like `metadata.tags` are assigned to two distinct Key-tags: `Metadata` and `Tags`.

After the definition of Key-tags, 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, data related to the first five resources of the data set was analysed. This implied the analysis of a total of 311 tags with 1141 occurrences and assigned by 355 users.

The results from the current pilot study confirm those of KoT that identified the need for new metadata elements for Social-Tagging applications. However, it points out the need for 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 analysed in the pilot study, 212 Key-tags were created.

The majority of Key-tags (55.3%) were composed of a single tag, i.e., tags that did not have variant forms. However, there were Key-tags with groups of 2–12 tags. As shown in Table 2, there is a large concentration of tags (80%) grouped into Key-tags of up to 4 tags and 20% grouped in Key-tags of 5–12 tags.

From this amount, 159 Key-tags (75%) corresponded to the following DC properties: Creator, Date, Format, Is Part Of, Publisher, Subject, Title and Type. From these, 90.6% correspond 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); Format 0.6% (see Table 3).

**Table 2** Quantity of tags in the Key-tags

| *Tags grouped* | *Key-tags* | *Total tags* | % | *Cumulate* (%) |
|---|---|---|---|---|
| 1 | 172 | 172 | 55.3 | 55.3 |
| 2 | 24 | 48 | 15.4 | 70.7 |
| 3 | 3 | 9 | 2.9 | 73.6 |
| 4 | 5 | 20 | 6.4 | 80.0 |

**Table 2** Quantity of tags in the Key-tags (continued)

| *Tags grouped* | *Key-tags* | *Total tags* | % | *Cumulate* (%) |
|---|---|---|---|---|
| 5 | 1 | 5 | 1.6 | 81.6 |
| 6 | 3 | 18 | 5.8 | 87.4 |
| 7 | 1 | 7 | 2.3 | 89.4 |
| 8 | 0 | 0 | 0 | 89.4 |
| 9 | 1 | 9 | 2.9 | 92.6 |
| 10 | 0 | 0 | 0 | 92.6 |
| 11 | 1 | 11 | 3.5 | 96.1 |
| 12 | 1 | 12 | 3.9 | 100.0 |
| *Total* | *212* | *311* | *100* | *100* |

**Table 3** Dublin Core properties Key-tags

| *DC property* | *Key-tag* (*N* = 159) | % |
|---|---|---|
| Creator | 5 | 3.1 |
| Date | 2 | 1.3 |
| Format | 1 | 0.6 |
| Is part of | 5 | 3.1 |
| Publisher | 2 | 1.3 |
| Subject | 144 | 90.6 |
| Title | 5 | 3.1 |
| Type | 8 | 5.0 |

The second and third columns did not result in 159 Key-tags and 100%, respectively, because some Key-tags corresponded to more than one DC property.

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, Note, Rate, User Name and Utility.

From these eight possible new properties, four had already been suggested in 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 proposed properties were identified: 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 13 Key-tags (24.5%) where 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 (see Table 4).

**Table 4** No DC properties Key-tags

| *No. DC properties* | *Key-tag* (*N* = 53) | % |
|---|---|---|
| Action | 8 | 15.1 |
| Category | 6 | 11.3 |
| Depth | 5 | 9.4 |
| Note | 4 | 7.5 |

**Table 4** No DC properties Key-tags (continued)

| *No. DC properties* | *Key-tag* (*N* = 53) | % |
|---|---|---|
| Rate | 8 | 15.1 |
| User name | 1 | 1.9 |
| Utility | 8 | 15.1 |
| No. assigned | 13 | 24.5 |

## 5.1 Action

There is a group of Key-tags that represents the action of the user in relation to the tagged resource (see Table 5).

**Table 5** Description of the action property

| *Label* | *Action* |
|---|---|
| Definition | An action that a tagger intends to take or suggests to take regarding the resources |
| Comment | Action may be used to describe the action taken by the tagger on the resource |
| Example | `Check it, Delete, Look at, to do, to evaluate, to listen, _toread, a_lire` |

It is a kind of tag that can be easily identified. As an example, the tags, which represent the action To Read, were input by 6 users from Delicious: `_toread`, `a_lire`, `toread`.

This property does not describe the resource itself. Instead, it indicates which action the user executed or intends to execute. It is useful mainly for who inputs the tag. Anyhow, these tags are able to signal subjectively a quality evaluation, or at least an expectancy of quality, of the resource for that tagger. Another example is the Key-tag `HighLight`, which indicates the tagger highlighted or intends to highlight the resource in any manner.

It was observed that the infinitive form of a verb in English would be used in most of the action tags, as for example `!tobechecked`, `*tostudy`, `.todo`, `_toblog`, `_to-read`, `2try`, `articlestoevaluate`, `is:toread, library_to_read`.

## 5.2 Category

This property relates to tags whose function is to group resources into categories. This property does not group resources into subject, but other categories (see the description of this property in Table 6). For instance, the tag `Faq` was assigned several resources that contained answers to frequent questions.

Another example for the Category property is the tag `DC Tagged`. During the analysis of the Key-tag `DC Tagged`, it was noticed that the corresponding resources also had 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 of 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.

**Table 6** Description of the category property

| *Label* | *Category* |
|---|---|
| Definition | Category of a group of resources |
| Comment | Category may be used to classify a set of resources, according to classifications other than theme or subject |
| Example | `How To, Faq` |

### 5.3 Depth

This type of tag represents the degree of intellectual depth to the tagged resource (see Table 7). According to Webster's dictionary, 'depth' is synonymous of profundity. In this paper, it is used in the following meaning: "Degree of intellectual depth" or "The intellectual ability to penetrate deeply into ideas" (Webster's Online Dictionary, 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, semanticweb. overview, semwebintro, overview. These tags mean that users are describing a resource whose content is thought as a schematic or a summarised explanation, introductory and general.

**Table 7** Description of the depth property

| *Label* | *Depth* |
|---|---|
| Definition | Degree of intellectual depth of the resource |
| Comment | Depth may be used to represent the degree of intellectual profundity of the resource, as estimated by the tagger |
| Example | `Introductory, Synthesis, General, Rhizome` |

Another example is the tag State of the Art "the highest level of development at a particular time" (WordNet, 2008). This tag means that the resource content is in a State of the Art form and, thus, represents the grade of intellectual depth of the topic.

### 5.4 Note

This element may be proposed to represent the tags that are used as a note or reminder (see Table 8). As WordNet, a note is "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 categorisation (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).

**Table 8** Description of the note property

| *Label* | *Note* |
|---|---|
| Definition | A note or annotation |
| Comment | Note may be used to express a comment or observation with the objective of reminding somebody about something or registering an observation, comment or explanation related to the resource |
| Example | `729 week 01, Via Popular, Hey e OR2007` |

To identify the Note property, it was necessary to analyse the whole set of tags of a given user, taking into account a specific resource. All tags from all resources tagged by that user were analysed. This simple fact made the task hard to perform. From this analysis, it was inferred that those tags were assigned to make a note as some kind of observation, comment or explanation.

From the five analysed resources, the following tags considered as 'Note' were identified: Hey, Ingenta, OR2007, PCB Journal Club. As an example, 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 resource.[2] The second tag makes reference to the Open Repositories 2007, an event where Tony Hey made a Keynote speech. However, interestingly enough, the tagged resource does not have any direct relation either with that event or with Tony Hey.[3]

### 5.5 Rate

Rate, meaning pattern, category, class or quality, is related to tags that are evaluating the tagged resource. Thus, the user categorises the resource according to its quality when using this type of tag (see Table 9).

**Table 9** Description of the rate property

| *Label* | *Rate* |
|---|---|
| Definition | The quality of the tagged resource |
| Comment | Rate may be used to express a qualitative evaluation that a tagger assigns to a resource |
| Example | `Bad, Good, Great, Important, Boring, Brilliant, Best of, Excellent` |

Golder and Huberman (2006a, 2006b) had already identified this kind of tag, which would have the function of "identifying qualities or characteristics" in the resources. Authors consider them tags of the adjective type, such as "scary, funny, stupid, inspirational", which express an opinion of the resources' users.

It has to do with the evaluation results of the users themselves of resources, which makes it an important property, as it allows the opinion of those who accessed the resource to be known. This property is not related to the resource description itself, but it represents a perception (subjective) the users have of it.

The following tags were related to the Rate 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 lead to misunderstandings, and it may become necessary to analyse them in relation to the whole set of tags of a given tagger. For instance, the tag `Vision` could have several meanings, but after an analysis of the collection of resources, it may be concluded that it is classifying the quality of the resource.

Tags composed of symbols (e.g., `****`) or mixed (e.g., `5 stars rating`) were identified as Rate after the analysis of the whole set of tags of a given tagger, where bundles such as ranking, *!rated*, *z!rated* were observed. Such tags mean a grade or rate given or assigned to resources.

### *5.6 User name*

The Tag ‘User Name’ labels the resource with the name of a user. The analysed resource had the name of the user of the tagged resource (see Table 10).

**Table 10** Description of the user name property

| *Label* | *User name* |
|---|---|
| Definition | Name or Nickname of the tagger |
| Comment | User Name may be used to register the names of taggers of the resources. User name may be the tagger’s nickname or his own name in full, in initials or in an abbreviated form |
| Example | `Alttabilib, Bokardo, Jwelles, SYP, MFL` |

User Name is the property with which one can relate the name or nickname of a resource tagger himself. Tags that are identical to the nickname of the tagger were found. An example is `bokardo` (tag) and bokardo (nickname of the tagger). There are other tags that are not identical to the nickname but that refer to the user who is tagging the resource. For example: `mlf` (tag = user’s name initials) and morgaine (nickname of the tagger).

### *5.7 Utility*

This property relates to the tags that represent the utility of the resource for the user.

It represents a specific categorisation of the tags, so that the user may recognise which resources are useful to him/her regarding certain tasks and utilities (see Table 11).

There is another property that was identified and previously described, which also aims at the personal organisation of tasks: Action. Thus, it should be made clear that Action differs from Utility. Action represents actions that the user intends to execute with the resource, e.g., print (`To print`), whereas Utility represents contexts or situations in which the resource shall or might be useful for the user. For example, the tag `Chapter8` identifies the resource as a useful resource for the writing of chapter 8 of a book (clarification provided by the tagger).

**Table 11** Description of the utility property

| *Label* | *Utility* |
|---|---|
| Definition | Represents the tagger’s intended use for the resource |
| Comment | Utility may be used to express the category of the resources according to the utility for the tagger |
| Example | `Book-Project, Dissertation, for work` |

The following tags that point out a specific activity for which the resource shall or might be useful were identified: `Bachelor Thesis`, `Chapter 2`, `Class Paper`, `Dissertation`, `IMT530`, `J-Hosp_Lib_Bib`, `Maass`, `Research`, `Search` and `Thesis`. It was not hard to identify most of them as being related with Utility. Others, though, demanded a grouped analysis of tags, resources and users. Some examples will be described here.

`Class Paper` is a tag, which is bundled to ‘1schoolwork’, and was assigned to three resources. By analysing the set of related resources and tags, it was verified that it refers to resources that would be or were used for a certain action: ‘school work’.

Some tags corresponded to subjects’, courses’ graduation or post-graduation codes. As, for instance, the tag `IMT530` that was bundled to Master of Science in Information (MSIM), and was referring to the course IMT530 – Organisation of Information and Resources. This tag was related to the Utility property since it is meant to identify the useful resources for a post-graduation programme and its courses.

`J-Hosp_Lib_Bib` is a tag that was assigned to identify the useful resources for the production of a paper for the *Journal of Hospital Librarianship*. This information appears in an explanatory note given by the tagger: This serves as the bibliography, list of tools, list of examples discussed and list of additional resources (tools, examples and papers) for the *Journal of Hospital Librarianship* paper, “Social Software for Libraries and Librarians”, by Melissa Rethlefsen and others, due for publication in late 2006.

`Maass` is a tag that was bundled in ‘Study’. The term represents the name of a teacher as the 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”; “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 analysed in the pilot study. However, 25% of them are still left behind.

The final study has already been completed and although it is not yet possible to show the results, it is possible to say that the percentage of 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%, whereas 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 will allow rich descriptive tags to be handled by metadata interoperability protocols and, consequently, to enrich the semantic web.

This work began with a pilot study for the first five resources of the KoT data set to refine the methodology and have a preliminary overview of the possible new properties that could be identified, if any. This paper presents the results from the pilot study and already gives some lights on the final study.

The results of this study are relevant for the projects that aim to automatically infer meaning from tags. They do not necessarily imply the creation of an application profile for social-tagging applications, either in or outside the scope of DC. They are useful as a reference to what can be expected from users' tags on text resources.

The study we did present only a small part of the overall picture of the KoT users place on social tools. Further studies need to be done for other types of resources (e.g., images, audio or video) and in other contexts (e.g., architecture, classic music sites or performing arts).

## References

Baptista, A.A., Tonkin, E.L., Remini, A., Hooland, S., van Pinheiro, S., Mendéz, E. and Nevile, L. (2007) 'Kinds of Tags: progress report for the DC-Social tagging community', Paper presented at *the DC-2007, International Conference on Dublin Core and Metadata Applications*, Singapore, Obtained through the internet: http://hdl.handle.net/1822/6881, Accessed 04/09/2007.

Coyle, K. and Baker, T. (2008) *Guidelines for Dublin Core Application Profiles*, Obtained through the internet: http://dublincore.org/documents/2008/11/03/profile-guidelines/, Accessed 12/02/2010.

DCMI (2008) *Dublin Core Metadata Element Set: Version 1.1*, Obtained through the internet: http://dublincore.org/documents/2008/01/14/dces/, Accessed 10/03/2008.

DCMI (2010) *Dublin Core Metadata Initiative Home Page*, Obtained through the internet: http://dublincore.org/, Accessed 12/02/2010.

DCMI Usage Board (2008) *DCMI Metadata Terms*, Obtained through the internet: http://dublincore.org/documents/dcmi-terms/, Accessed 10/03/2008.

Feinberg, M. (2006) *An Examination of Authority in Social Classification Systems*, Papers of 17th SIG/CR Classification Research Workshop, 17, Austin, USA, Obtained through internet: http://www.slais.ubc.ca/users/sigcr/sigcr-06feinberg.pdf, Accessed 06/11/2006.

Golder, S.A. and Huberman, B.A. (2006a) *The Structure of Collaborative Tagging Systems*, Obtained through internet: http://arxiv.org/abs/cs.DL/0508082, Accessed 14/11/2006.

Golder, S.A. and Huberman, B.A. (2006b) 'Usage patterns of collaborative tagging systems', *Journal of Information Science*, Vol. 32, No. 2, pp.198–208.

Gulati, A. (2009) *Statistics of Web 2.0*, Obtained through internet: http://gulati.info/2009/01/statistics-web-2, Accessed 14/01/2010.

Guy, M. and Tonkin, E. (2006) 'Folksonomies: tidying up tags?', *D-Lib Magazine*, Vol. 12, No. 1, Obtained through the internet: http://wwww.dlib.org/dlib/january06/guy/01guy.html, Accessed 12/12/2006.

Infopedia (2008) Obtained through internet: http://www.infopedia.pt, Accessed 10/03/2008.

International Standards Organization [ISO] (1986) *ISO 2788: Documentation: Guidelines for the Establishment and Development of Monolingual Thesauri*, [S.L.]: ISO.

Mathes, A. (2004) *Folksonomies – Cooperative Classification and Communication through Shared Metadata*, Computer Mediated Communication – LIS590CMC, Urbana, University of Illinois, Obtained through internet: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, Accessed 25/10/2006.

Menezes, E.M., Cunha, M.V. and Heemann, V.M. (2004) *Glossário de análise documentária*, ABECIN, São Paulo.

Nilsson, M., Baker, T. and Johnston, P. (2008) *The Singapore Framework for Dublin Core Application Profiles*, Obtained through internet: http://dublincore.org/documents/singapore-framework/, Accessed 12/02/2010.

Nilsson, M., Baker, T. and Johnston, P. (2009) *Interoperability Levels for Dublin Core Metadata*, Obtained through internet: http://dublincore.org/documents/interoperability-levels/, Accessed 12/02/2010.

O'Reilly, T. (2005) *Web 2.0: Compact definition? O'Reilly Radar Blog*, Obtained through internet: http://radar.oreilly.com/archives/2005/10/web_20_compact_definition.html, Accessed 06/11/2006.

Quintarelli, E. (2005) *Folksonomies: Power to the People*, Papers of Incontro ISKO Italia – UNIMIB, Milão, Obtained through internet: http://www.iskoi.org/doc/folksonomies.htm, Accessed 23/10/2006.

Schroeder, S. (2009) *The Web in Numbers: The Rise of Social Media*, Mashable the Social Media Guide, Obtained through internet: http://mashable.com/2009/04/17/web-in-numbers-social-media/, Accessed 14/01/2010.

Spiteri, L.F. (2007) *Structure and Form of Folksonomy Tags: The Road to the Public Library Catalogue*, *Webology*, Vol. 4, No. 2, Article 41, Obtained through internet: http://www.webology.ir/2007/v4n2/a41.html, Accessed 24/01/2010.

Thomas, M., Caudle, D.M. and Schmitz, C.M. (2009) 'To tag or not to tag?', *Library Hi Tech*, Vol. 27, No. 3, pp.411–434.

Tonkin, E., Baptista, A.A., Pinheiro, S., Hooland, S. van Resmini, A., Mendéz, E. and Neville, L. (2007) 'Kinds of tags: a collaborative research study on tag usage and structure (Presentation)', Paper presented at *the European Networked Knowledge Organization Systems (NKOS)*, Austin, Texas, Obtained through the internet: http://www.us.bris.ac.uk/Publications/Papers/2000724.pdf, Accessed 10/12/2007.

Wal, T.V. (2006) *Folksonomy Definition and Wikipedia*, Obtained through the internet: http://www.vanderwal.net/random/entrysel.php?blog=1750, Accessed 22/11/2006.

Webster's Online Dictionary (2008) With multilingual Thesaurus Translation, Obtained through internet: http://www.websters-online-dictionary.com, Accessed 05/01/2008.

WordNet (2008) Obtained through internet: http://wordnet.princeton.edu, Accessed 22/11/2006.

Yin, R.K. (1989) *Case Study Research: Design and Methods*, Thousand Oaks, USA.

## Notes

[1] According Ângulo Marcial (1996) Post-coordination is the principle by which the relationship between concepts is established at the moment of outlining a search strategy (as cited in Menezes *et al*., 2004).

[2] This information was given by the tagger.

[3] This information confirmed by the author of the resource (the creator).