Kinds of Tags - Progress Report for the DC-Social Tagging Community

This presentation is available in Open Access at http://hdl.handle.net/1822/6881

Ana Alice Baptista - Universidade do Minho
Emma L. Tonkin - UKOLN
Andrea Resmini - Università di Bologna
Seth Van Hooland - Université Libre de Bruxelles
Susana Pinheiro - Universidade do Minho
Eva Mendéz - Universidad Carlos III Madrid
Liddy Nevile - La Trobe University

http://creativecommons.org/licenses/by-nc-sa/2.5
What is KoT and how it began

How we did it

The first indications we found and what we are willing to find

How to get involved
How it Began

- Liddy’s post on DC-Social Tagging mailing list;
- Preparation of a proposal and posting it to the mailing list;
- Receiving expressions of interest from people from the UK, Spain, France, Belgium, Italy and USA;
- The authors of this presentation are working on it, but newcomers are always welcome.
Conditions / Restrictions

- it is a **bottom-up project**: it was born inside the community;
- it is **completely Internet-based** as:
  - it was born in the electronic environment;
  - most of the participants don’t know each other personally: all communication was Internet-based (Google docs was of extreme help) and, *note*, mostly asynchronous;
- there was **no financial support** and it was all developed based on a common interest of the participants.
The questions

It is focused 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.

We are starting to have some indications to provide (still foggy) answers to the following questions, for this particular set of documents:

- Into which DC elements can tags be mapped?
- What is the relative weight of each of the DC elements?
- What other elements come up from the analysis of the tags?
- Do tags correspond to atomic values?
The Process of Data Collection

- **Fifty** scholarly documents were chosen, with the constraints that:
  - each should exist both in Connotea and Del.icio.us; and
  - each should be noted by at least five users.

- A corpus of information including user information, tags used, temporal and incidental metadata was gathered for each document by an automated process;

- This was then stored as a set of spreadsheets containing both local and global views.
The Data Set

- 4964 different tags corresponding to 50 resources (documents): repetitions were removed;
- no normalisation of tags was done at this stage;
- all work was performed at the global view: easier to work with;
Assignation of DC elements

- Each of the 4964 tags in the main dataset was analyzed in order to manually assign one or more DC elements;
- In certain cases in which it was not possible to assign a DC element and where a pattern was found, other elements were assigned;
- Thus, four new elements have been "added" (indications to the question: What other elements come up from the analysis of the tags?):
  - "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).
Assignation of DC elements (2)

- **Multiple alternative elements** were assigned in the event where:
  - meaning could not be completely inferred (additional contextual information would help in some cases);
  - tags had more than one value (e.g., dlib-sb-tools - elements: publisher and subject).
  - When there were enough doubts a question mark (?) was placed after the element (e.g., subject?)
Assignation of DC elements (3)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Non DC element</th>
<th>Number of Non-DC elements</th>
<th>DC element</th>
<th>DC element</th>
<th>DC element</th>
<th>DC element</th>
</tr>
</thead>
<tbody>
<tr>
<td>#great-idea</td>
<td>Rate</td>
<td>1</td>
<td>Audience?</td>
<td>Description?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#it_administrator</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Subject</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#toread</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>situ_web2.0</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>articulo</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta).url</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>delicious</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*best</td>
<td>Rate</td>
<td>1</td>
<td>Subject?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>clippings</em></td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*essay</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*read</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*to_read</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overview</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>paralar</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>work</td>
<td>To Be Used In</td>
<td>1</td>
<td>Subject?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rss</td>
<td>Action Towards Resource?</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>article</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blogging</td>
<td>Action Towards Resource?</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>oreilly</td>
<td>Action Towards Resource?</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0, business,internet,social,2.0,we</td>
<td>Action Towards Resource?</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4doctors</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4lee</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aan-June2006</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>academia</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>academicis</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>article_archive</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>article_read</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>article_resource_06.03.%233</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>article_titles</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>articleis</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>articles</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>articles:web2.0</td>
<td>Action Towards Resource</td>
<td>0</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>articles_i_should_read</td>
<td>Action Towards Resource</td>
<td>1</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some Indications
(Work in Progress)

• Users apply tags not only to describe the resource, but also to
describe their relationship with the resource (e.g. to read, to
print,...)

• Do tags correspond to atomic values? Many of the tags have
more than one value, which potentially 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.
Some Indications
(Work in Progress) (2)

• What is the relative weight of each of the DC elements?
  • It was possible to allocate metadata elements to 3406 out of the total number of 4964 tags (meaning was inferred somehow);
  • 3111 out of these 3406 were assigned with one or more DC elements - (no contextual information).
  • The Subject element was the most commonly assigned (2328), and was applied to under 50% of the total number of tags.
Conclusions

• A revision of all assigned elements was made; however, normalised markup of such a large corpus is an enormous task.

• The indications we show here are not preliminary findings. This work is in an initial phase. Further work (that may invalidate these indications partially or totally) has to be done, preferably by the whole community.

• Assigning metadata elements to tags is a difficult task even for a human - Contextual information may ease it, but we still don’t know at what extent (because we didn’t do it).
Questions for the Future

• Current question: **how easily can tags be ‘normalised’ for interoperability** with standard metadata environments such as the DC Metadata Terms?

• Future:

  • Should we have a more structured interface for motivated users to tag? Would that be used? Would that be useful?

  • Will we be able to infer meaning from tags? To what extent? Is it really needed?
How to Get Involved

• Emma Tonkin is leading new developments;

• Contact Emma (e.tonkin@ukoln.ac.uk) or any of the authors;

• Share your ideas and say how you are willing to help.
Thanks!!!

Ana Alice Baptista - analice@dsi.uminho.pt
Emma L. Tonkin - e.tonkin@ukoln.ac.uk
Andrea Resmini - root@resmini.net
Seth Van Hooland - svhoolan@ulb.ac.be
Eva Mendéz - emendez@bib.uc3m.es
Liddy Neville - liddy@sunriseresearch.org