Kinds of Tags: a collaborative research study on tag usage and structure

Emma Tonkin[1], Ana Alice Baptista[2], Seth van Hooland[3],
Andrea Resmini[4], Eva Mendéz[5] and Liddy Neville[6]
1 UKOLN, University of Bath e.tonkin@ukoln.ac.uk
2 Universidade do Minho, Portugal, analice@dsi.uminho.pt
3 Université Libre de Bruxelles, Belgium, svhoolan@ulb.ac.be
4 CIRSFID, University of Bologna, resmini@cirsfid.unibo.it
5 University Carlos III, Madrid, Spain, emendez@bib.uc3m.es
6 Sunrise Research, Australia, liddy@sunriseresearch.org

KoT (Kinds of Tags) is an ongoing joint collaborative research effort with many participants worldwide, including the University of Minho, UKOLN, the University of Bologna, the Université Libre de Bruxelles and La Universidad Carlos III de Madrid. 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.

The initial phase of research entailed a preliminary study of tag use, data gathering and analysis. For this phase, fifty scholarly documents on the topic of tagging were chosen, with the constraint that each should exist both in Connotea and Del.icio.us and that 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.

Three stages of analysis were then applied:

1. Raw tag frequency count
2. On a small, manually gathered subset, we performed preliminary normalisation: stemming, gender and number (singular/plural), and normalised tag frequency count. No synonyms or corrections were made on the primary dataset.
3. Group tags into clusters of KoT (e.g. subject, depth, audience...) and include a column with the number of tags per cluster.

Each of the 4964 different 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 "invented": "Action Towards Resource" (e.g., to read, to print...), "To Be Used In" (e.g. work, class material), "Rate" (e.g., very good, great idea) and "Depth" (e.g. overview). Some tags were assigned metadata elements tentatively, or marked with multiple alternative elements in the event that meaning could not be inferred without additional contextual information. In the event that tags had more than one value, two or more elements were assigned simultaneously (e.g., dlib-sb-tools - elements: publisher and subject).

A revision of all assigned elements was made; however, normalised markup of such a large corpus is an enormous task. For this reason, further revision is planned via the DCMI Social Tagging group prior to an envisaged release of this corpus as a research tool.

Preliminary findings:

- Users apply tags not only to describe the resource, but also to describe their relationship with the resource (e.g. to read, to print,...)
- Many of the tags have more than one value, which potentially results in more than one metadata element assigned. 473 tags have more than one possible value.
- From a total number of 4964 tags, 3406 have metadata elements allocated to them (meaning

was inferred somehow), from which 3111 have one or more DC elements allocated.

- 14 of the 16 DC elements, including audience, have been allocated.
- The Subject element was the most commonly allocated (2328), and was applied to under 50% of the total number of tags.
- Assigning metadata elements to tags without context information is a difficult task even for a human.

Further work is expected to focus on ensuring the quality of assigned elements, on application of this corpus to investigate related research questions, and on the possibilities offered by automated tag normalization and markup.