Kinds of Tags

Emma L. Tonkin – UKOLN

Ana Alice Baptista - Universidade do Minho

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

UKOLN is supported by:

JISC

Supported by

MLA

Museums, Libraries and Archives Council

UNIVERSITY OF BATH

www.bath.ac.uk

Social tagging

- “A type of distributed classification system”
- Tags typically created by resource users
- Free-text terms – keywords in camouflage…
- Cheap to create & costly to use
- Familiar problems, like intra/inter-indexer consistency

Characteristics of tags

- Depend greatly on:
  - Interface
  - Use case
  - User population
  - User intent: by whom is the annotation intended to be understood?

Perspectives on the problem

- Each participant has very different motivations:
  - Ana: applying informal communication as a means for sharing perception and knowledge – as part of scholarly communication
  - Andrea: enabling faceted tagging interfaces
  - Seth: evolution to a hybrid situation where professional and user-generated metadata can be searched through a single interface
  - Emma: where sociolinguistics meets classification? “Speaking the user's language” - language-in-use and metadata

What's in a tag?

Reviewing Marshall's dimensions of annotation:

Formal	Informal
Explicit	Implicit
Writing	Reading
Extensive	Intensive
Permanent	Transient
Published	*Private*
Institutional	*Individual*

'computationally tractable & interoperable, but expensive'

'descriptive, but not necessarily computationally tractable'

– "To reduce the overhead of description, we may use methods of extracting more formal description from informal annotations."

The Future of Annotation in a Digital (Paper) World, Catherine C Marshall

Hence:

- At least part of a given tag corpus is ‘language-in-use’:
  - Informal
  - Transient
  - Intended for a limited audience
  - Implicit
- Also note 'Active properties'

Dourish P. (2003). The Appropriation of Interactive Technologies: Some Lessons from Placeless Documents. Computer-Supported Cooperative Work: Special Issue on Evolving Use of Groupware, 12, 465-490

Consistency

- *Inter/intra-indexer consistency*
- *Definitions:*
  - *Level of consistency between two indexers' chosen terms*
  - *Level of consistency between one indexer's terms at different occasions*
- *Why is there inconsistency and what does it mean? Is it noise or data?*

Context

- Language as mediator - of?
- Extraneous encoded information: *informal, infinite, dynamic*

  Coping with Unconsidered Context of Formalized Knowledge, Mandl & Ludwig, Context '07
- *How does one handle unconsidered context?*
- *Could it ever consist of useful information?*

A primary aim in tag systems

- To improve the signal-to-noise ratio:
  - Moving toward the left side of each dimension
- Cost of analysis vs. cost of terms
- Can be a lossy process - many tags may be discarded
- Systems with fewer users are likely to prefer the cost of analysis than the loss of some of the terms

Analysis of language-in-use?

- Something of a linguistics problem
- You might start by:
  - Establishing a dataset
  - Identifying a number of research questions
  - Investigation via analysis of your data
  - Some forms of investigation might require markup of your data

Approaches to annotation

- Corpora are often annotated, eg:
  - Part-of-speech and sense tagging
  - Syntactic analysis
- Previous approaches used tag types defined according to investigation outcomes
- A sample tag corpus annotated with DC entity - to investigate the links between (simple) DC and the tag

Related Work

- Kipp & Campbell – patterns of consistent user activity; how can these support traditional approaches; how do they defy them? Specific approach: Co-word graphing. Concluded: Predictable relations of synonymy; emerging terms somewhat consistent. Also note 'toread' 'energetic' tags
- Golder and Huberman – analysed in terms of 'functions' tags perform:
  What is it about? What is it? Who owns it? Refinement to category. Identifying qualities or characteristics. Self-reference. Task organisation.

What
KoT
is
about

KoT

What is KoT and how it began

How we did it

The first indications we found
and what we hope to find

How It Began

- Liddy Nevile'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, the USA and most recently, Singapore

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 see some indications that 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)

33 34	Tag	Non DC element	Non DC element	Number of Non-DC elements	DC element	DC element	DC elemen
145	#great-idea	Rate		1			
146	#it_administrator			0	Audience?	Description?	
147	#toread	Action Towards Resource		1			
148	$itu_web2.0			0	Subject		
172	(artículo)			0	Type		
173	(beta).url			0			
174	(delicious)			0	Description?		
184	*best	Rate		1			
185	*clippings*			0	Subject		
186	*essay			0	Type		
190	*read	Action Towards Resource		1			
191	*to_read	Action Towards Resource		1			
219	.overview	Depth		1			
220	.paraler	Action Towards Resource		1			
243	.work	To Be Used In		1	Subject?		
244	/rss			0	Subject		
245	:article			0	Type		
246	:blogging	Action Towards Resource?		1	Subject?	Type?	
253	:oreilly			0	Publisher?	Creator?	
320	2.0,business,internet,social,2.0,we			0	Subject	Subject	Subject
353	4doctors			0	Audience?		
354	4lee			0	Audience?		
381	aan-june2006			0	Date		
388	academia			0	Subject?	Audience?	
389	academialis			0	Subject?	Audience?	
542	article_archive	Action Towards Resource		1	Type		
543	article_read	Action Towards Resource		1	Type		
544	article_resource_06.03.%233			0	Type	Date	
545	article_titles			0	Type		
546	articlelis			0	Type	Subject	
547	articles			0	Type		

Some Indications

(Work in Progress)

(Work in Progress)

- Users are seen to 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

- 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.

Working towards automated annotation?

- Approaches:
  - Heuristic
  - Collaborative filtering
  - Corpus based calculation
- Eventual aim: to create lexicon of possibilities, to disambiguate where there is more than one possible interpretation

Conclusions

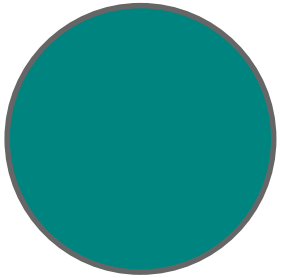
- A revision of all assigned elements was made; however, normalised markup of such a large corpus is an enormous task.
- The indications we show here are not true preliminary findings. This work is in an initial phase. Further work (that may invalidate these indications partially or totally) has to be 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 yet 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 neded?

Criticisms

- Is Simple DC a 'natural' annotation (good fit) for a real-world tag corpus?
  - (If not, then what?)
- Does anybody really want a faceted interface? Indications are: this easily becomes confusing and unusable.
  - (If not, then how else do we apply this information to improve the user experience?)



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 Nevile - liddy@sunriseresearch.org