

Tag-Based Metadata Management for Information Integration

Miika Nurminen

Department of Mathematical Information Technology
University of Jyväskylä, Finland



Outline

- Background & Concepts
- Research goals
- Implementation ideas
- Conclusion



About my graduate studies

- Graduate studies so far...
 - M.Sc (software engineering) in 2005
 - Various R&D projects, teaching, thesis supervision, and other coding/tech. support activities
 - Research topics with a shifting focus: text mining, business process management, content management, requirements engineering...
 - 7 conference or workshop papers
 - Required courses for graduate studies completed in 2008 => ABD
 - Supervisors: prof. Tommi Kärkkäinen, Anne Honkaranta (-2008), Anneli Heimbürger (2009-)
- 2010: **refocused** research plan
 - *Tag-Based Metadata Management for Information Integration*
 - Format: monograph (with supporting papers from earlier studies)
 - Est. completion time: 2012



Central concepts

•Metadata

- "Data about data"
- May be implicit (e.g. structural data in xml), explicit, external (e.g. rdf annotations), embedded (e.g. mp3 id3), centralized, or distributed.
- The division between data and metadata depends on context (e.g. are comments to a blog post data or metadata?)

•Tags

- User-defined, free-form keywords
- Popularized by del.icio.us (renamed as delicious), started in sep. 2003

•Triple tags (machine tags)

- Machine tags are tags that use a special syntax to define extra information about a tag (e.g. time, location, standard metadata fields).
- Introduced in Flickr: <http://www.flickr.com/groups/api/discuss/72157594497877875/>
- Highly elaborate use of machine tags: <http://www.delicious.com/cldwalker> and http://tagaholic.me/blog.html#post:*

•Hierarchical tags

- Allow adding **lightweight** semantic information to tags
- Enables conceptual search
- See: <http://www.bibsonomy.org/user/msn>

•Semantic tags

- Tags integrated to an explicit knowledge structure (e.g. ontology)



Central concepts

•Context

- Many definitions (usually involving location, time, or environment)
- In this context, we are interested in the context as *the situation/task at hand*.

•Ontology

- “An ontology is a specification of a conceptualization” -Gruber
- In practice, ontology often describes the concepts and vocabulary used in a domain in a way that allows automatic inferences
- Essential component semantic web – used by rdf metadata – can be interpreted as *non-constraining schemas* for metadata
- The formality and specificity of different ontologies vary greatly – see [http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-\(with-citation\).htm](http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm)

•Folksonomy

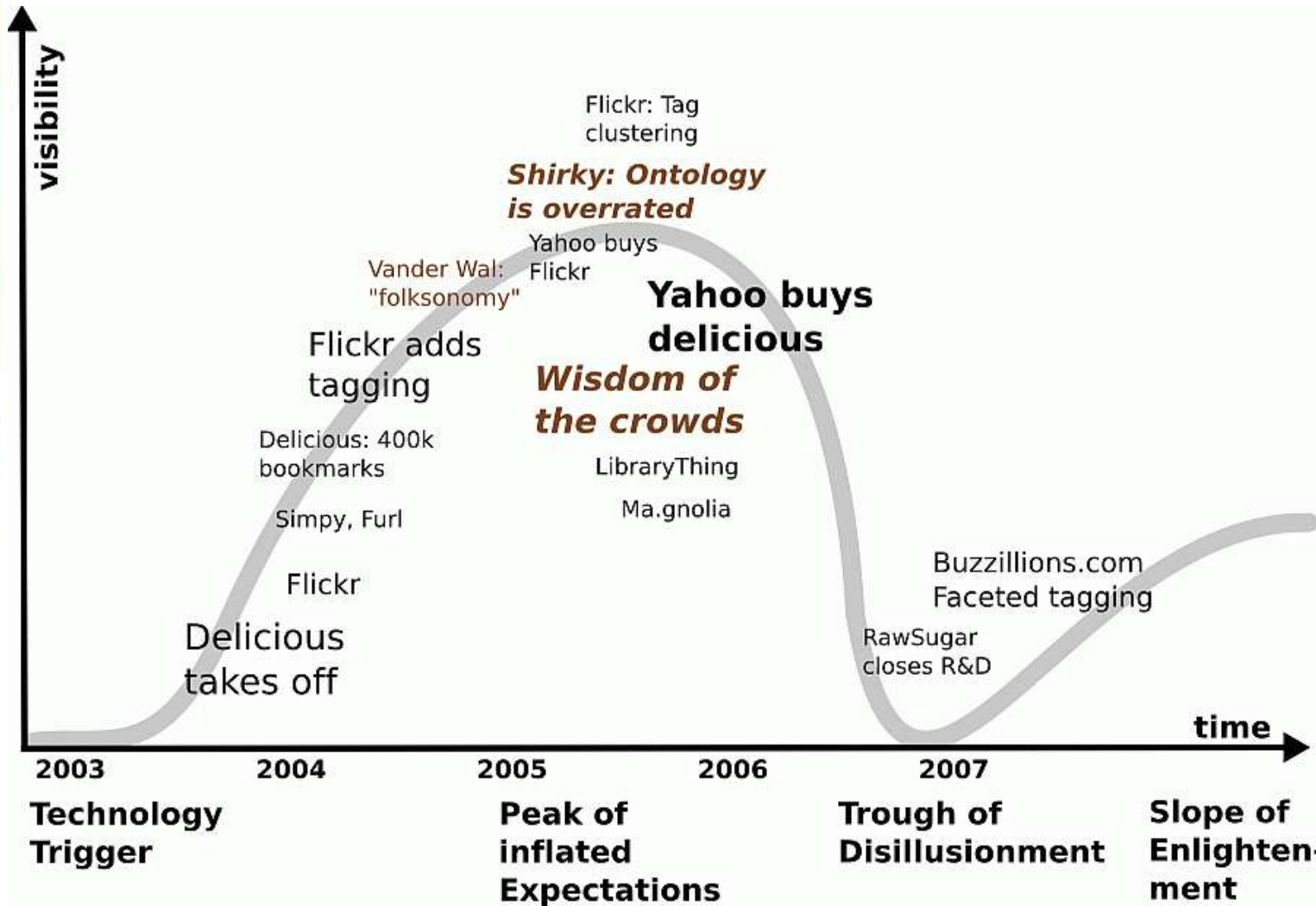
- An informal knowledge structure that emerges from collaborative tagging
- Can be regarded as a lightweight ontology
- Term coined by T. Vander Wal, popularized (among others) by A. Mathes (<http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>)

•Faceted classification

- Allows multiple classifications to on object (cf. traditional library classification)
- Effective information representation and search technique for both tagged and semantic web data - see: <http://www.museosuomi.fi/>
- Hierarchical tagging can be interpreted as faceted classification



Some history...



Tags vs ontologies

•Tags...

- Have become a common component to many web applications (blogs, wikis, cataloging services, time management, e-commerce...)
- Highly effective and flexible personal information management technique
- "ubiquitous" – because of the simplicity tags can be potentially shared and integrated with relatively little effort
- Help to "tie things together" in an ad-hoc way, in the spirit of Memex
- With hierarchical tags, serve as a superior generic organization metaphor (both for files and abstract structures) compared to traditional directories or categories

•Ontologies and Semantic web...

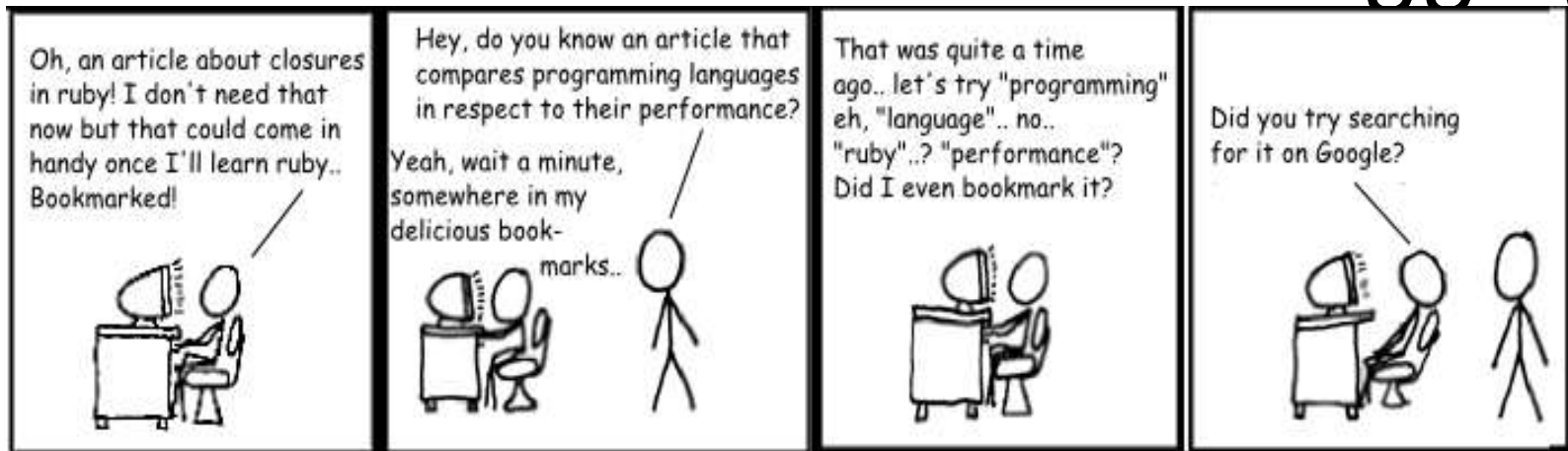
- Semantic web has been more or less around since 2003 but has yet changed the web little compared to individual services, applications, web search, and... tagging.
- Semantic web standards are complex to implement especially integrate – developing generally accepted ontologies is highly problematic
- Ontologies in more limited scale (e.g. in one organization or service) work better – also, an rdf-based database might be good back-end solution even for tagging-based system (semantic tagging)

•Recently: tagging and semantic web have been seen as complementary, rather than competitive approaches

- E.g. extract ontological data from Wikipedia: <http://dbpedia.org/>
- Linked data, Semantic Web 2.0...



Problems with conventional tagging



<http://www.pui.ch/phred/archives/2007/10/remembering-on-the-web-5-reasons-why-social-bookmarking-doesnt-work.html>

- The tagging process itself can be laborious and time-consuming, especially if a large number of tags are used
- Tags from different services do not integrate well (attempt: technorati)

However, tags can still be highly effective for

- *Contextual* information (e.g. work processes, projects) that have nothing to do with original linked content
- Automatically generated metadata (e.g. temporal or spatial information)
- Representing an information retrieval model where conventional metadata fields are exposed as tags
- Annotating audiovisual content (cf. ESP game, steve.museum)
- Bridging social and semantic web with *semantic* tags



Research goals

- Defining a **generalized information retrieval model for tagging** systems that allows **hierarchical tags** (in a similar way to faceted classification systems, producing a directed acyclic graph or a lattice), optional **simple datatypes for tagging** (e.g. explicit search mechanism for tags like "publicationYear:2010", or other spatiotemporal information), and **explicit contextualization** (e.g. definition of synonyms and matching between different "tag schemas" based on tags used in different services or by different persons)
- Developing techniques to **integrate tags and other metadata from different kinds of data sources** (web services, databases, documents in local filesystems) such that a single search interface can be used to retrieve data, manage tags (e.g. one interface to define tag hierarchies, rename tags, find duplicates, suggest new tags, etc) and **present it in different contexts** (faceted search and profiled multichannel publishing).
- Utilizing computational methods (e.g. conceptual clustering, object consolidation, retrieval fusion) to **merge tags with close similarity** (based on the metadata or content) from different contexts (e.g. services or users) and **match tags with external sources**, such as rdf metadata to enable semantic search.



Ideas from personal information management

- Utilizing personal blogwikis (blikis) for published (and private) notes, ideally with file-based storage mechanism to ease integration with different editing tools – organized with hierarchical tags.
 - Vanilla (<http://www.langreiter.com/>)
 - Dokuwiki (<http://www.dokuwiki.org/>)
- Tagging helps to integrate and utilize to do –lists, meeting minutes, lecture notes and other time-sensitive data that should be browsed from different viewpoints
 - Evernote (<http://www.evernote.com/>)
 - ”unification” of tags and ”folders” in Evernote 2 with auto-assignment
 - Ecco Pro (http://en.wikipedia.org/wiki/Ecco_Pro)
 - The earliest known system to implement typed, hierarchical tags in the desktop (1993!) +outlines +auto-assignment functionality +scripting
 - The Holy Grail of outliners.
 - InfoQube (<http://www.infoqube.biz/>)
 - More recent attempt for typed outlines
 - currently in beta, draws ideas from Ecco Pro and adds even more, such as allowing multiple parents in content nodes.



Ideas from personal information management

Mail & RSS feeds

- Gmail & Google reader are great by allowing tags to mail messages & rss entries. Gmail even supports automatic filters for tags
- However... Mail tags are separate from rss tags (and potential tags used by the author that produced the rss are omitted), and the concept of "folder" is separated from tags in Reader.
- More integration is still needed!

Outlines

- Hierarchical *tags* alone are not enough!
- The granularity of the information should be accounted in the data model as well
- Cf. Bookmarking to a book (esp. Proceedings) vs bookmarking to specific article in the book, *still retaining the info about the original book*
- When composing large outlines, this is an essential feature
- Freemind (<http://freemind.sourceforge.net/>) might have the best UI for composing outlines, but without tagging functionality, outlines don't "scale" well for complex domains.
 - Ecco Pro comes to rescue (again)
 - Microformats help as well, e.g. dokuwiki way:
{{tag>research.tagging, presentation, course:TIEJ601}}



Ideas from personal information management

- Reference management
 - BibTeX is an ideal format for storing references (esp. with JabRef+Bibsonomy, <http://www.bibsonomy.org/help/doc/jabref-plugin/>) ... Once you know that you are going to need the reference
 - Storing (and at worst, writing it manually) BibTeX data "just in case" is even more tedious than conventional bookmarking with tags (even if a bookmarklet is used for saving it)
 - Solution: store only BibTeX id (=filename) with as little contextual information (e.g. current project/paper) as possible
 - Some tools (e.g. Mendeley, SciPlore) allow extracting metadata automatically from PDF files as well
 - New metadata fields should be added in an incremental way as they are needed and as the reading process goes on
 - E.g. add "author" tag when going through papers by certain author, switch to "conference" or "journal" tag when doing a systematic review
 - Tags could provide a partial view to BibTeX dataset
 - Using the BibTeX id as a tag, notes from different sources (e.g. outlines, wikis, other note-taking tools) could be weaved together along with the original file



Implementation

- ❏ Systematic literature review based on current research on tagging systems
- ❏ Critical evaluation of essential tagging-like services and software, including examination of tags and content used in these services
- ❏ Construction and evaluation of a new prototype service that utilizes the proposed tagging model and integration capabilities with a focus in personal information management.



Other potential applications

- Collaborative annotation of complex structured data (e.g. museum database)
- Improving cataloging and recommendation systems (cf. LibraryThing, RateYourMusic)
- Cleaning up noisy databases with manually marked text fields
 - Museum databases (again), last.fm...
 - Object consolidation – essentially the same problem as unifying tags with close similarity
- Profiled multichannel publishing based on tagged content items
 - Learning objects, study guides, etc

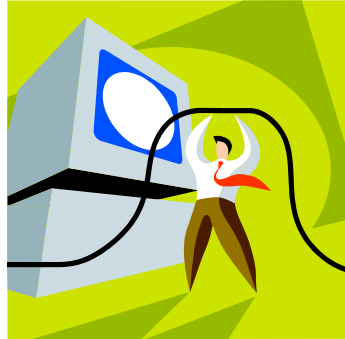


Conclusion

- ❏ People tend to use highly individualized tagging structures (even by same person in different contexts) - makes implementing tag recommendation systems challenging
- ❏ Bookmarking is **not** the killer application with tagging. More promising approach: marking the information **in context** and organize it as needed.
- ❏ Tag structure produced by a crowd can be more detailed and accurate than traditional metadata marked up by an information professional (not to mention ontologies)
- ❏ More expressive - *commonly accepted* retrieval model (still as low entry-level as possible) for tagging is needed to effectively integrate different data sources and make tags *a little more* semantic.



Thank You!



minurmin@jyu.fi

<http://users.jyu.fi/~minurmin/>

<http://www.mendeley.com/profiles/miika-nurminen/>

 Questions?



JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ