Margaret Kipp <margaret.kipp@gmail.com>
Faculty of Information and Media Studies
University of Western Ontario

Information Organisation Practices on the Web:

Tagging and the Social Organisation of Information

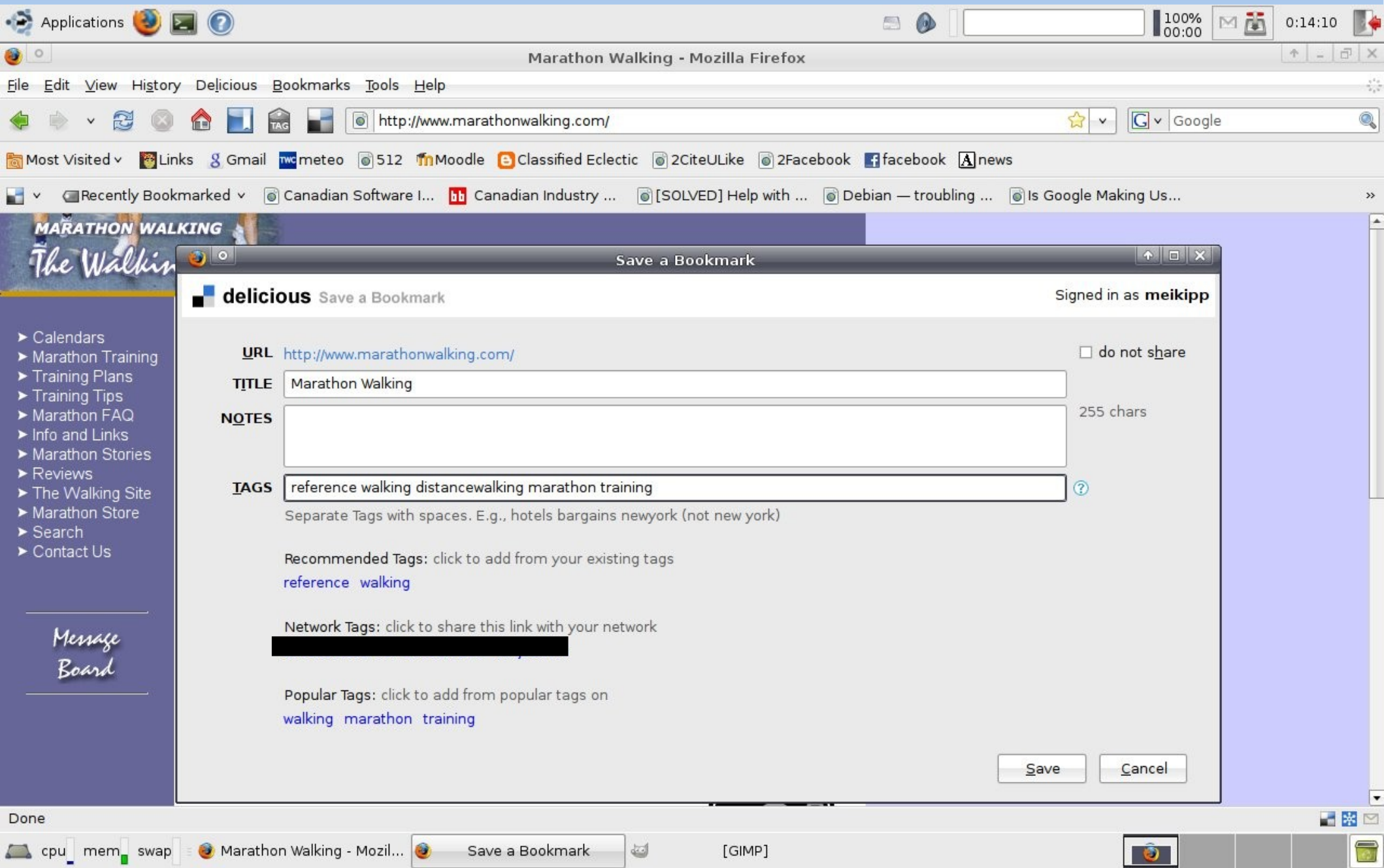
Pratiques d'organisation de l'information en ligne:

Le tagging et l'organisation social de l'information

Background

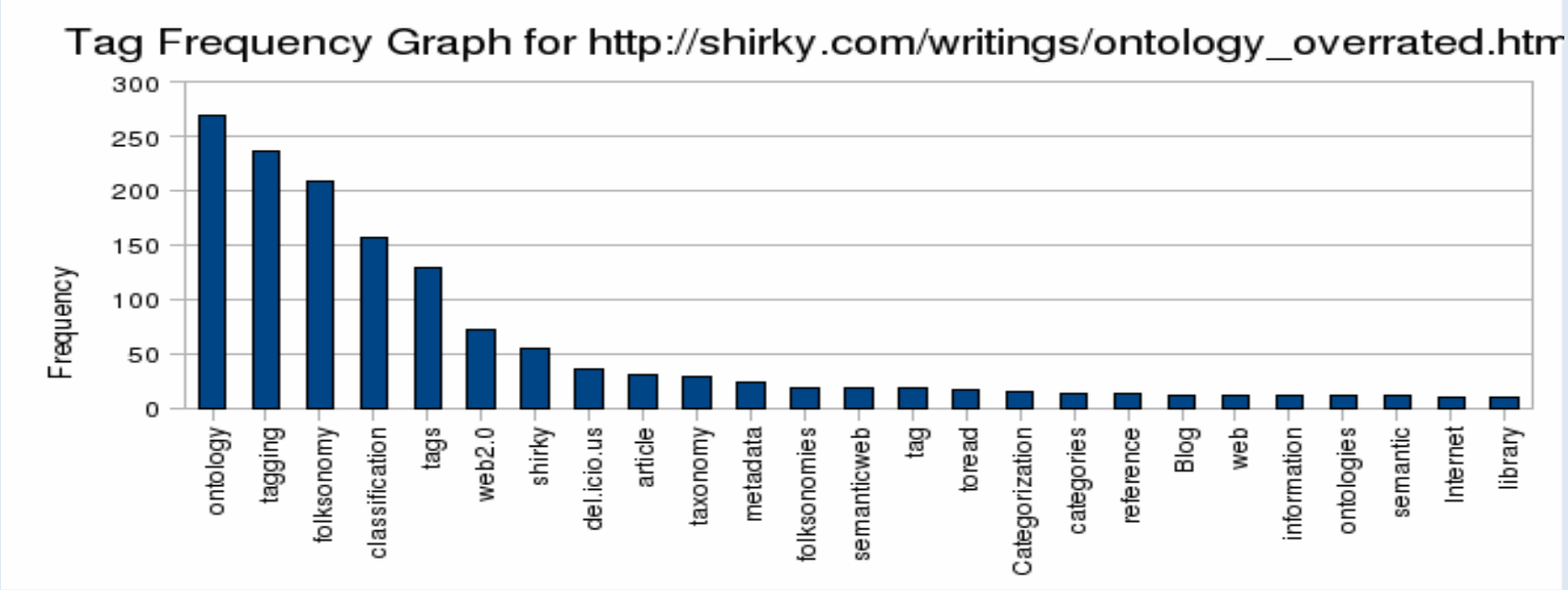
- My research examines:
  - how people organise things on the web and how this compares to traditional library classification and indexing techniques
- Specific points of interest:
  - structures and the creation of structures in social tagging systems
  - relationship between user, author and library indexing
  - relationship between personal information management and indexing

Social Tagging Process



Tagging as Collaborative Indexing

- Tagging is increasingly examined as a form of collaborative indexing
- Multiple studies have examined the consensus shown by frequency graphs



Tagging as User Indexing

- Creation of a folksonomy or user based taxonomy (Shirky 2005; Hammond et al. 2005)
- Examination of consensus in tagging (Golder and Huberman 2006; Kipp and Campbell ASIST2006, etc.)
- Comparisons of tags and subject headings (Kipp CAIS2006, CAIS2007; Smith SIGCR2007)

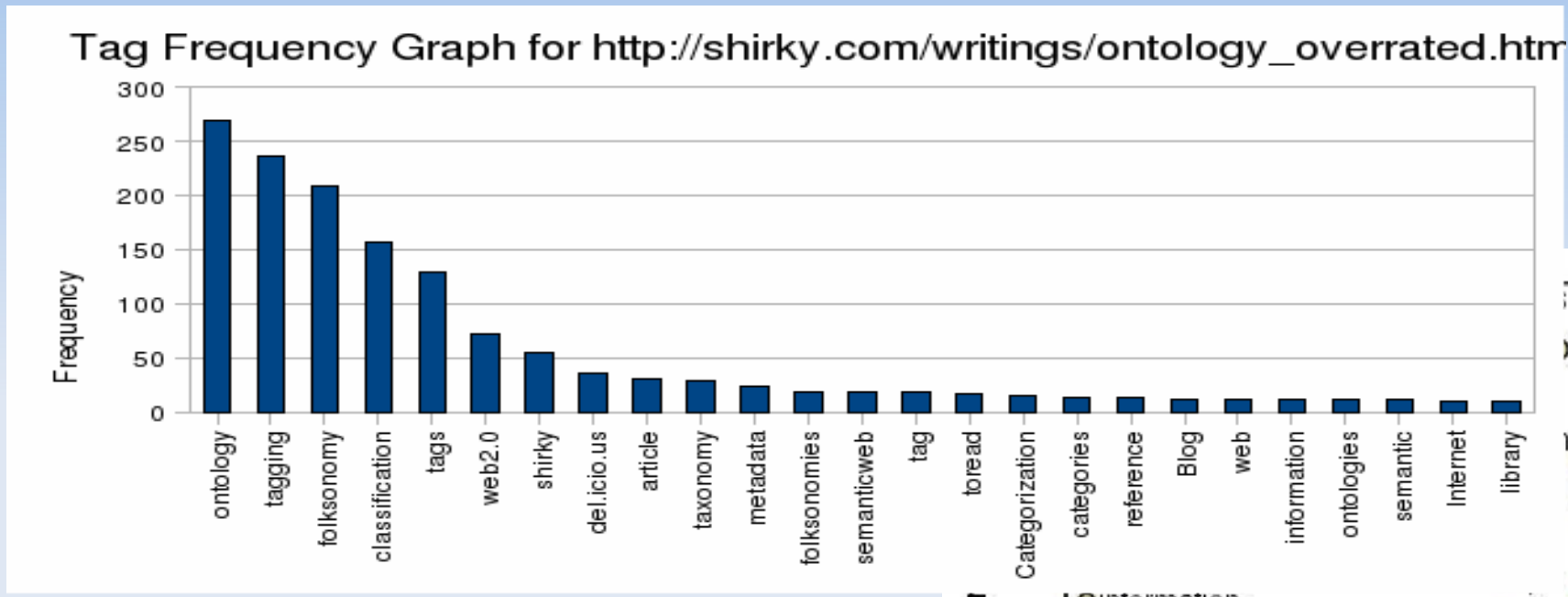
Tagging as Personal Information Management

- Tagging and PIM (see Shirky 2005, Kipp and Campbell CAIS2006; Kipp IASummit2007, etc.)
- PIM: paper based (Malone 1983, etc.) and electronic studies (Jones et al. 2005, etc.)
- Examine how users organise or arrange their personal/work documents for use
- Also examines differences between current and archival document organisation

4 Studies

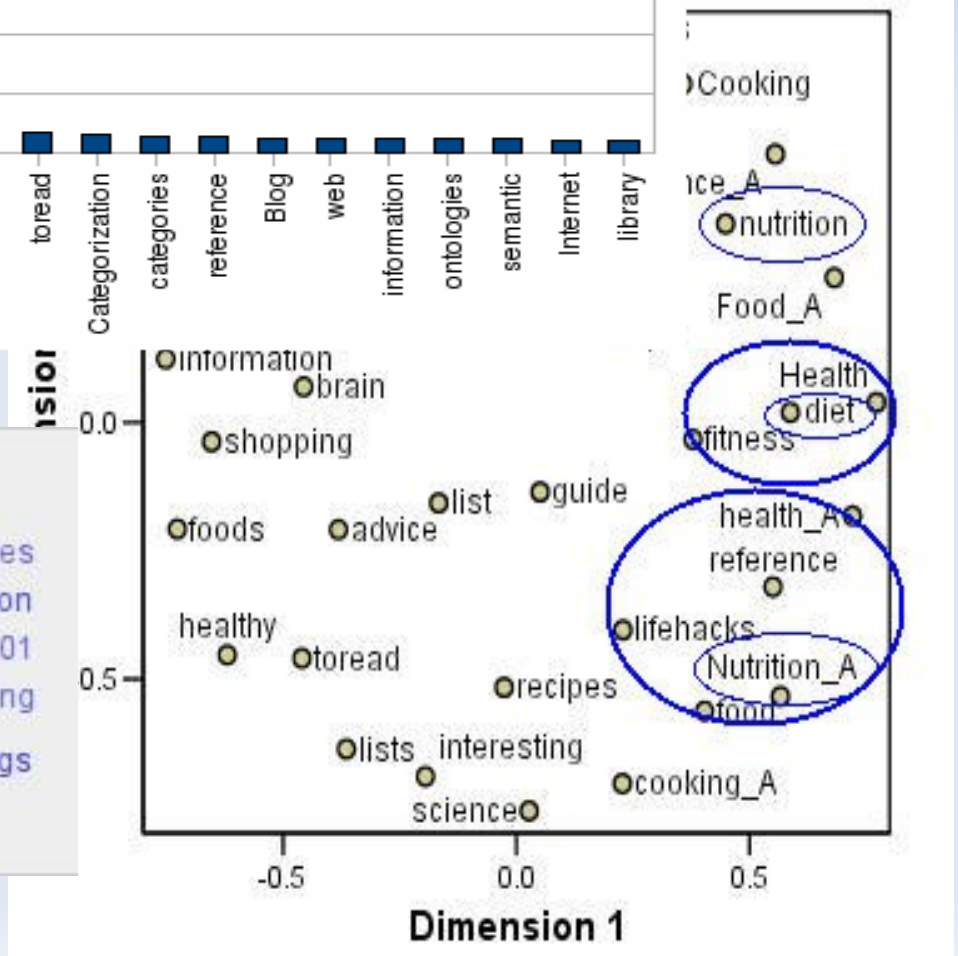
1. Tagging as indexing
    1. convergence and divergence in tagging
2. Tagging as user indexing
    1. are there differences in context and use between user, author and intermediary indexing?
3. Tagging as associative user indexing
    1. affective, subject and associative tagging
4. Tagging for information retrieval
    1. user use of subject access (e.g. indexing, tagging)

Study 1: Tagging as Indexing



common tags cloud | list

Information Social Web2.0 academic article articles bookmarking cataloging classification collaboration del.icio.us folksonomies folksonomy indexing kcb201 network paper research socialbookmarking socialnetworking socialtagging **tagging** tags taxonomy tonkin



Methodology

- all posts for 63 popular URLs collected from del.icio.us (http://del.icio.us/popular) in 2006 and 2008
- examine user tags and compare to traditional indexing methods
  - examine similarities and differences
  - analyse relationships, related tags and structures
  - examine frequency charts for a single URL
  - coword analysis of tags for a single URL

Research Questions

1. What patterns of consistent user tagging activity emerge through analyses of tagging frequency and co-word analysis?
2. To what extent do these patterns of tagging support and enhance some of the other traditional ways of indexing documents?
3. To what extent do these patterns defy these traditional methods, suggesting viable and promising alternatives to traditional subject access tools?

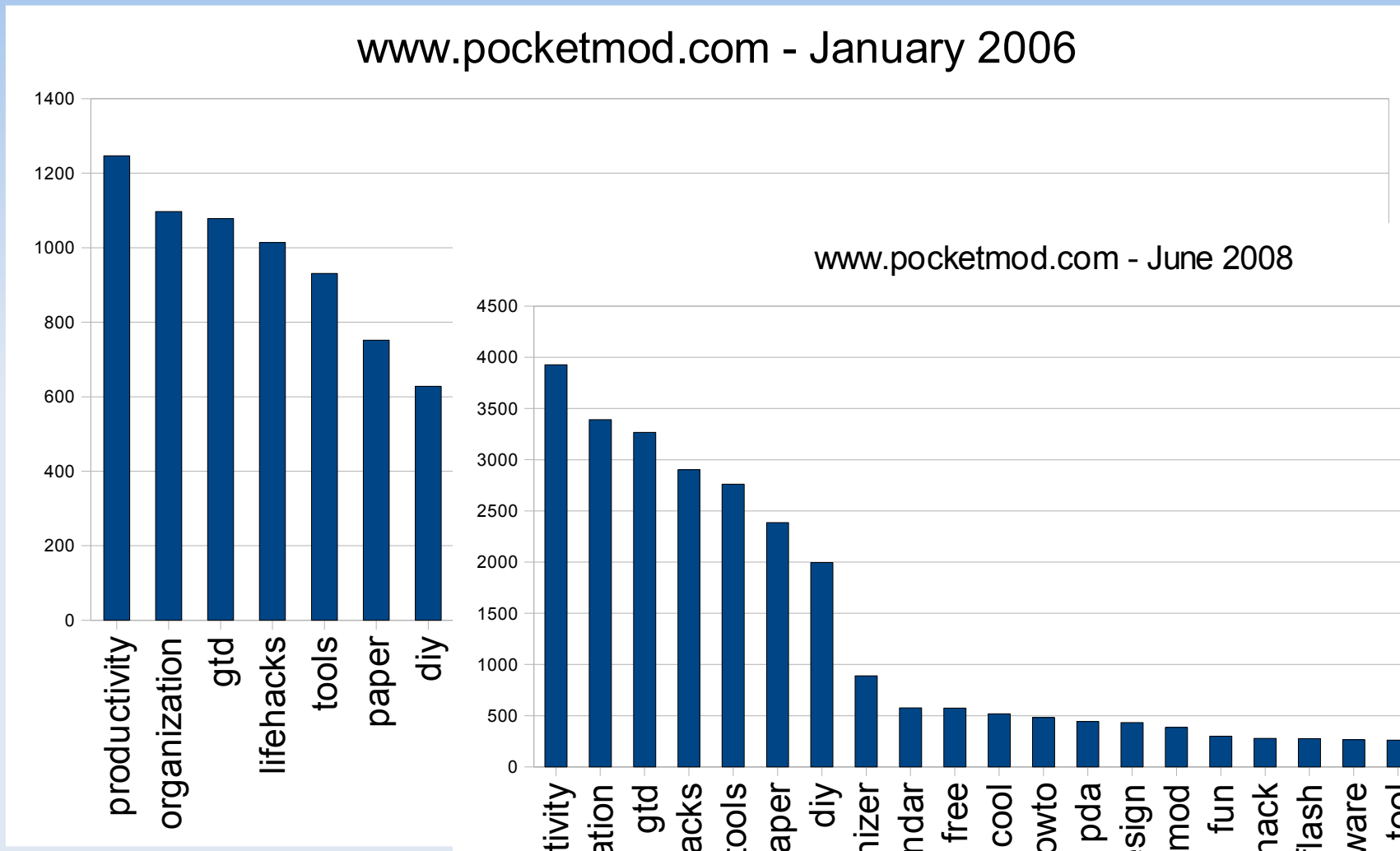
Descriptive Statistics

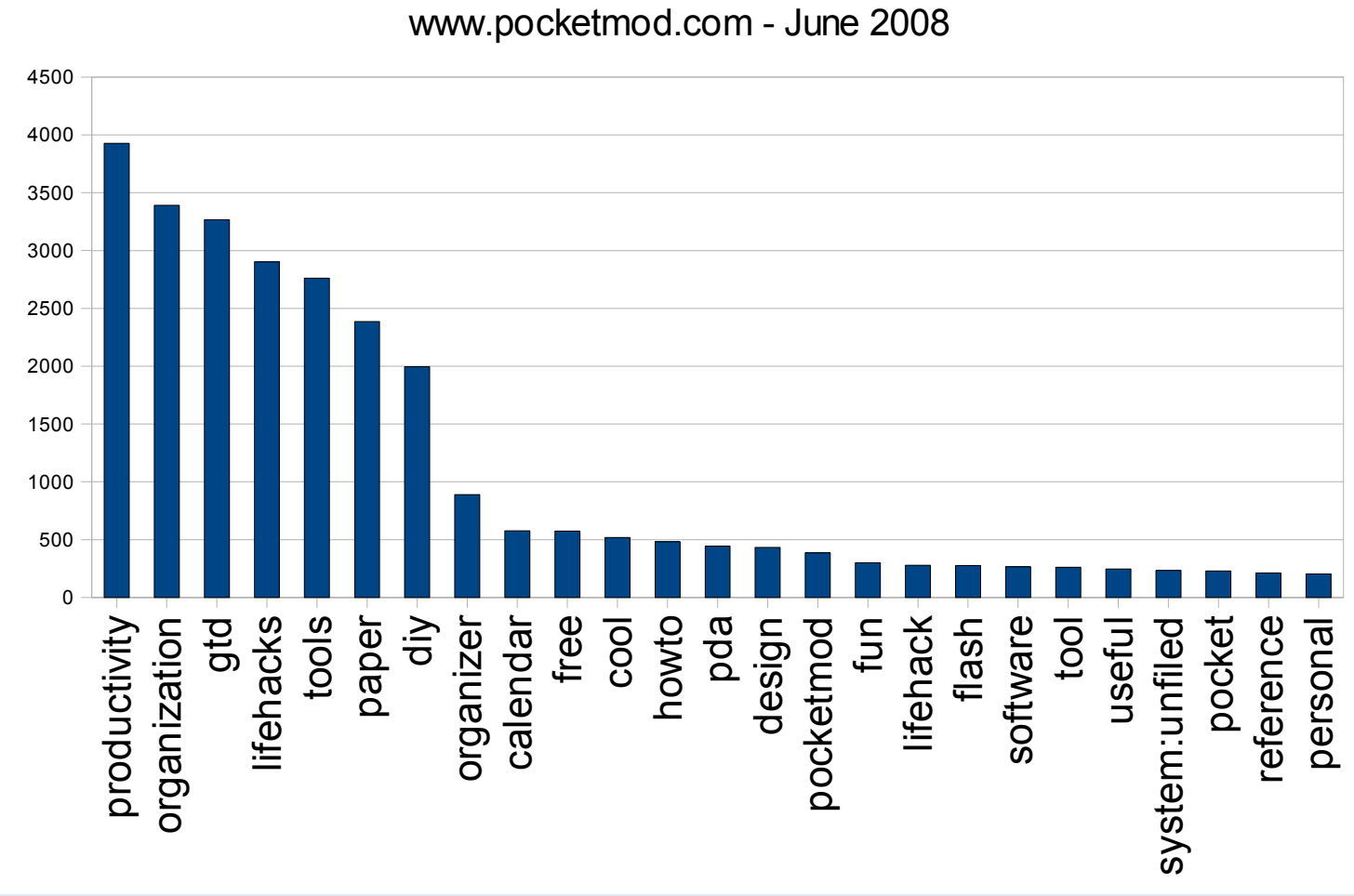
- number of users: 58728
  - average users per URL: 917 (max: 5172, min: 53)
- number of tags: 165831
  - average tags per URL: 295 (max: 13809, min: 49)
- users who did not tag: 6%
- users who used 1-3 tags: 65%

Vocabulary Use

- spelling variations
  - British versus American spelling
  - singular or plural
  - conjugated verb versus stem
  - caps versus lower case (no longer an issue with del.icio.us but some systems still make distinctions)
- synonyms or related terms
  - diet, nutrition, health, food, eating
- non subject tags
  - toread, cool
- acronyms
  - IR for information retrieval, KM for knowledge management

Tag Frequency Graphs



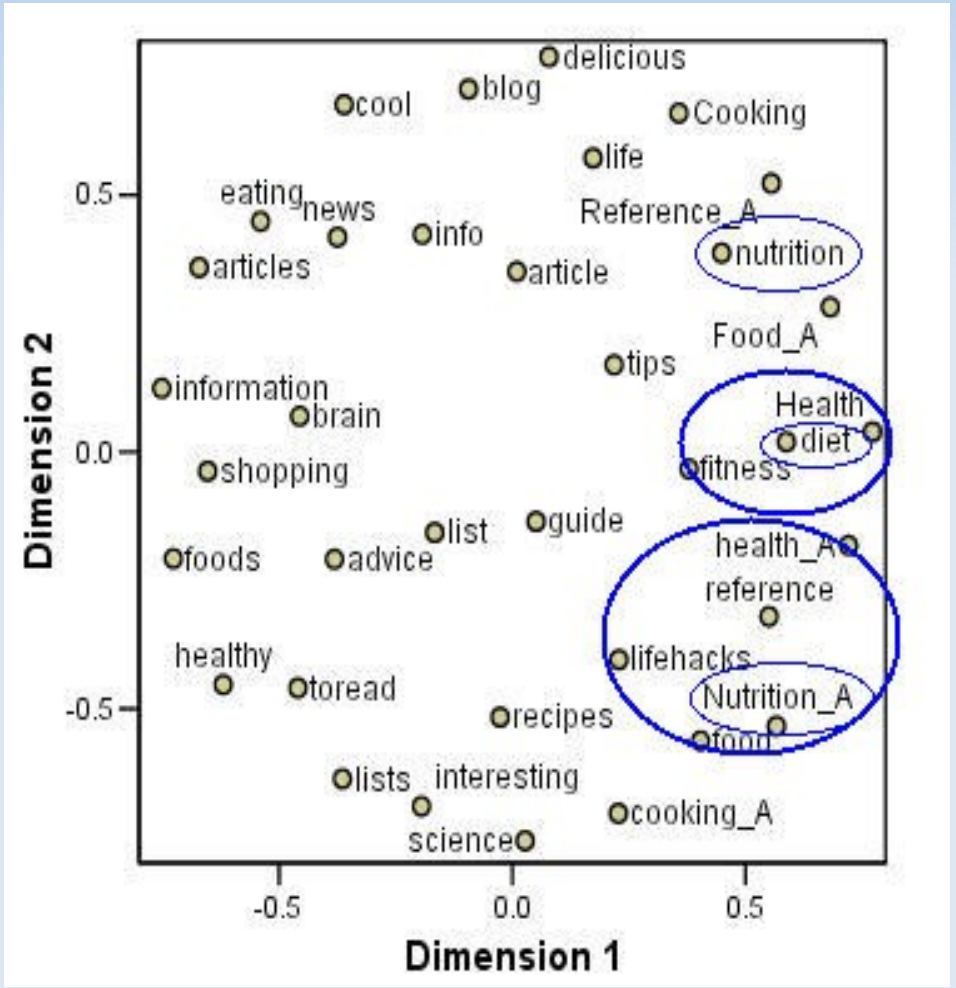


Tag Coword Analysis

- examine frequency of occurrence of pairs of tags (cotag analysis)
- if users A, B, and C have all tagged the same URL with tags X and Y, then X and Y co-occur
- clusters show relationships between items
- cotag analysis shows similar tags may be far apart

Cotag Graphs

- nutrition and diet do not cluster together
- neither nutrition nor Nutrition cluster with diet (Nutrition does cluster with food)
- perhaps evidence of different user groups in the tag clusters



Cotag graph www.bellybytes.com

General Results

- closely-related terms are not necessarily revealed through co-occurrence
- users employ many conventions in constructing tags, but apply them inconsistently
- users use both narrower and broader terms than is common in indexing
- users want to represent more than just the aboutness of a document

Link to Study 2 and 3

- continuities between tagging and indexing suggest the two may be complementary and that a combination would enrich both (study 2)
- use of time and task or affective tags shows that tagging expresses a dynamic relationship between users and documents, suggesting possible new ways of modelling information access (study 3)

Study 2: Tagging as User Indexing

- RoMEO studies 1: the impact of copyright ownership on academic author self-archiving Journal of Documentation 59(3): 243-277
  - tags: **copyright**, openaccess, romeo
  - author keywords: universities, publishing, academic staff, **copyright**, ownership, document management
  - descriptors: copy protection, **copyright**, electronic publishing, meta data, scholarly publishing, Great Britain, surveys

Academic Tagging: 2 Part Study

- compare user tags, author keywords and intermediary indexing terms
- Part 1: Kipp CAIS2006: study of LIS related articles tagged on CiteULike
  - users do use words from thesaurus as tags, but often use similar or related terms from other fields
- Part 2: Kipp CAIS2007, NASKO2007: examines articles from medicine/biology using similar methodologies

Research Questions

- To what extent do term usage patterns of user tags, author keywords and intermediary descriptors suggest a similar (or differing) context between users and indexers?
- How do tags assigned to academic articles reveal clues to the information context of the taggers?

Methodology (Part 2)

- Articles from Proteins and Journal of Molecular Biology tagged in CiteULike
- Author Keywords and Pubmed Descriptors from journal sites and Pubmed respectively
- 1083 articles (1588 posts, 239 users)
- Informetric analysis
- Thesaural analysis (Voorbij 1998, Kipp 2006)
  - comparison of terms using Pubmed thesaurus (range from SAME, SYN, NT, BT, RT, related and Not related)

Tags, Keywords and Descriptors

	Tags	Keywords	Descriptors
Unique	1136	3181	2746
Total	3788	4866	12473

- ratio of unique terms to total terms highest for author keywords
- supports findings from previous study in which author keywords were found to be more diverse than tags or descriptors

Popular Tags, Keywords and Descriptors

Tags	Frequency
protein_structure	140
no-tag	114
protein	114
structure	103
docking	97

Author Keywords	Frequency
protein folding	58
protein structure	49
molecular dynamics	46
protein structure predictio	38
docking	31

Descriptors	Frequency
Models, Molecular	649
Protein Conformation	511
Proteins	388
Amino Acid Sequence	306
Binding Sites	280

- 645 tags were used only once in the data set
- 2548 keywords were used only once
- 731 descriptors were used only once

Tags, Keywords and Descriptors by Article

By Article	Max	Min	Median
Tags	29	1	2
Keywords	13	1	5
Descriptors	36	2	11

- journal policies often require around 6 keywords, database like Pubmed have similar indexing policies
- Note: article with 29 tags was tagged by 14 users (most still use 1-3 tags per article)

User Vocabulary Length

User	Max tag list	Min tag list	Articles posted
3109	7	2	15
3063	6	1	73
4068	15	2	9

- number of unique terms used by a user
- largest user vocabulary length was 62 (min. 1, median 2)
- most users use a small number of tags

Thesaural Analysis

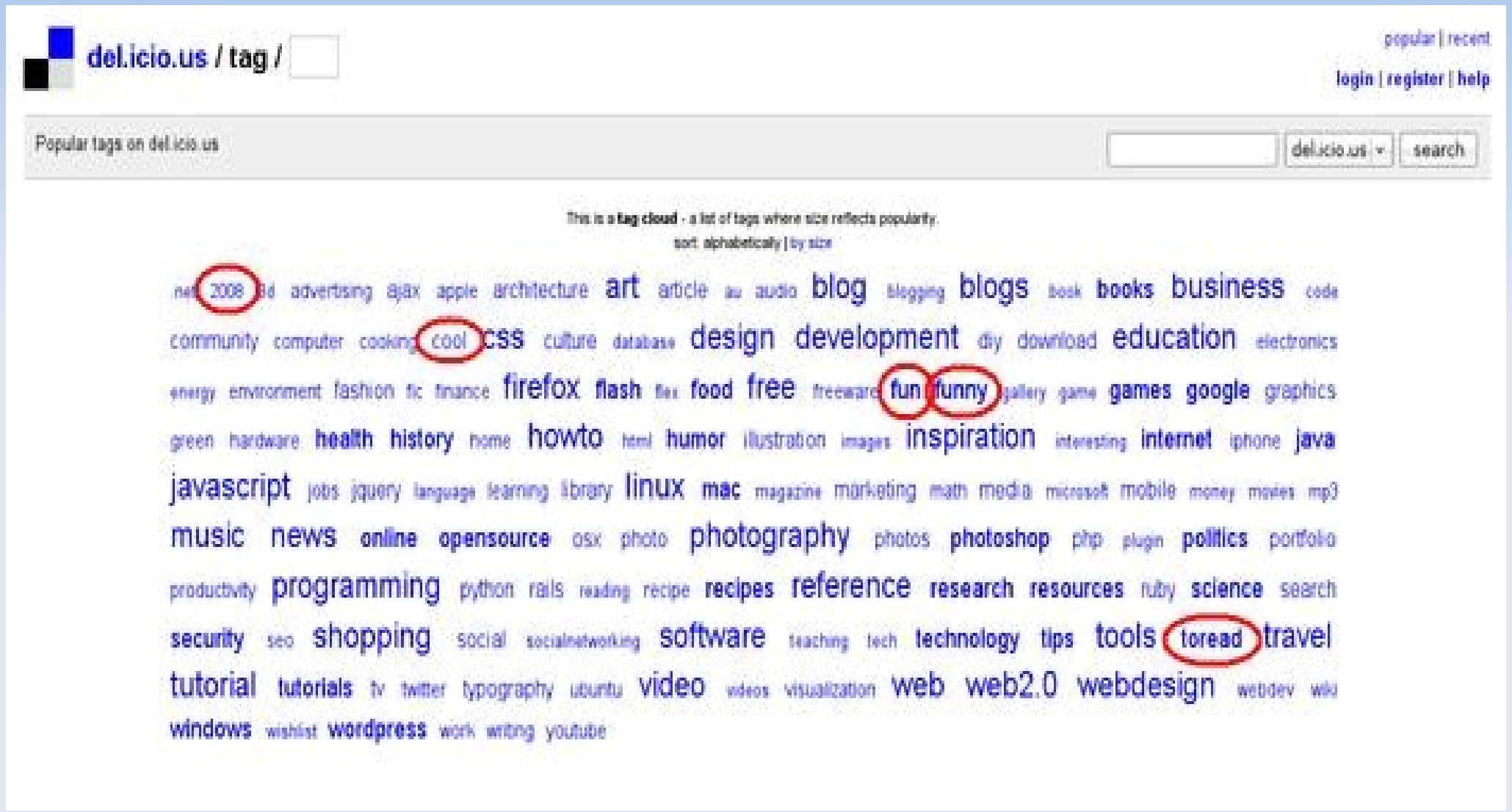
Tags	Keywords	Descriptors
3d	**16 S RNA**	Base Sequence
algorithms	ribosome	Computer Simulation
prediction	computer modeling	Cross-Linking Reagents
rna	distance geometry	Escherichia coli
16s		Models, Molecular
distance_geometry		Molecular Sequence Data
bioinformatics		Nucleic Acid Conformation
structure		**RNA, Ribosomal, 16S**

- Article: Computer modeling 16 S ribosomal RNA
  - bolded terms represent a thesaural match
  - bioinformatics versus computer modelling or simulation
- most common thesaural comparison was Related Term, followed by Equivalence

General Results

- users and authors use many terms that are, or are related, to thesaurus terms
- some user terminology is rare or completely absent from author keywords or descriptors
  - time and task tags, affective tags, project tags: @toread, cool, cais2006
  - faceted terms: 'protein' and 'structures' as separate tags instead of 'protein structures'
  - abbreviations: 'PDB' for 'Databases, Protein'

Study 3: Tagging as Associative User Indexing



Tagging Characteristics

- majority of tags used are subject related or form related -- exactly like traditional indexing (also geographic and date/time)
- a minority of tags are not subject related:
  - e.g. toread, cool, fun, cais2007
- 3 major categories:
  - affective (emotional) tags
  - time and task tags
  - project tags

Research Questions

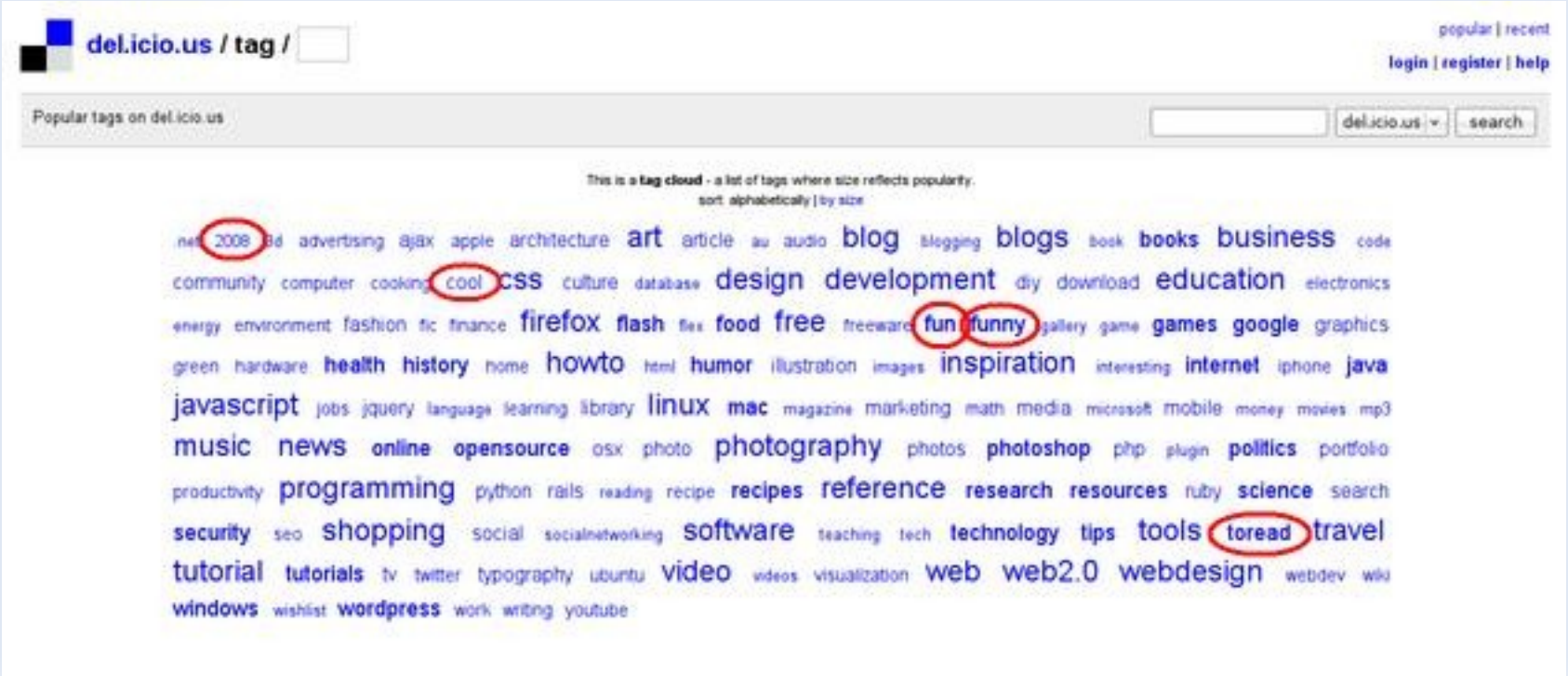
1. What patterns of user tagging activity emerge on examination of affective or time and task related tags?
2. How do users use time and task related tags or affective tags to indicate the value they see in a document?
3. What implications do the use of affective or time and task related tags have for the organisation of information?

Data Collected

- all posts tagged with one of 78 specific tags
  - 48 tags were time and task related, 30 were affective tags
  - project tags were not collected due to the difficulty of locating these
- 203352 posts in total
  - 1831 from CiteULike
  - 2891 from Connotea
  - 198630 from Del.icio.us

Frequently Used Non Subject Tags

- Citeulike: fun, ToRead, todo, interesting, cool
- Connotea: fun, ToRead, important, unread, funny
- Del.icio.us: fun, ToRead, funny, cool, interesting



Time and Task Tags

- many time and task related tags are variations on toread:
  - @toread, @read, readlater, unread
- is the toread tag useful to other users?
  - Amazon's recommendation system relies on purchase data and view data as an indicator of interest
  - Search engines use hyperlinks as indicators of interest
- could a toread tag have a similar function?

Affective Tags

- affective tags represent an emotional reaction to an item
  - cool
  - fun
  - strange
- do not appear to add anything to the subject indexing of an item
- seem to be poor candidates for search terms for information retrieval... but they indicate interest

Non Subject Tags with Subject Tags

- non subject tags were frequently used with subject related tags
  - fun was found most often with physics and math, toread with biology
- Examples:
  - 320112 - ForensicsAge written in teeth by nuclear tests
    - carbon-dating, **cool**, forensics
  - 566928 - Action as language in a shared visual space
    - **519**, **519-week11**, action, co-location, **cscw**, shared-space

Associative Indexing

- non subject tags show that tagging expresses a dynamic relationship between users and documents
- PIM research show users classify by task and project as well as by subject
- What is the effect of personal and subjective terms such as cool, fun and toread in a social bookmarking system?
- How can these terms be used in aggregate?

Study 4: Tagging for Information Retrieval

"information 2N organization"
and "health information"
and "case stud*"

Are tags useful for finding or refinding?

- have users search traditional journal database (Pubmed) and a social bookmarking site (CiteULike) for academic articles
- 10 participants from LIS
  - all had prior search experience (online databases and the web)
  - very few had used social bookmarking tools
- participants were encouraged to discuss and compare their experiences searching each site

Research Questions

1. Do tags appear to enhance findability? Do users feel that they have found what they are looking for?
2. How do users find searching social bookmarking sites compared to searching more classically organised sites? Do users think that tags assigned by other users are more intuitive?
3. Do tagging structures facilitate information retrieval? How does this compare to traditional structures of supporting information retrieval?

Methodology

- brief intro to study and tools
- search Pubmed and CiteULike for information on a specific assigned topic
- screen capture using CamStudio and Xvidcap
- semi-structured interview after search

Search Topic

- "You are a reference librarian in a science library. A patron approaches the reference desk and asks for information about the application of knowledge management or information organisation techniques in the realm of health information. The patron is looking for 5 articles discussing health information management and is especially interested in case studies, but will accept more theoretical articles as well."

Study Timeline

- participants selected initial set of keywords
- searched for approx. 5 articles on each tool (order randomised)
- participants asked to make a second list of terms they would use if asked to search again
- semi structured interview covering:
  - usefulness of tags and subject headings
  - use of search terms
  - thoughts on the search process

Demographics

- 6 female, 4 male
- between 23-40 years
- 80% self identified as intermediate computer users with 6-22 years of computer experience
  - 20% had a website, 40% had a blog
  - all had experience with search but not social tagging
- majority with humanities/social sciences background
- generally worked in libraries/archives

Results

- Keyword Usage:
  - All participants used multi word keywords
  - most commonly used keyword was knowledge management
  - information organisation/organization also commonly used
- System Usage:
  - abstract considered most useful piece of information
  - list of related articles as useful as subject headings or tags

Desire for Structure

- "One of the things that kind of bothered me about [the tags] is that they weren't really grouped... you have care and health but you don't have health care together." **Participant 1**
- "I did find it useful in PubMed how they linked to related articles. That was useful." **Participant 1**
- "[I] wanted to be able to have subject headings [in PubMed] visible along with the abstract." **Participant 9**

Use of Social Tagging Features

- "[I thought] I wasn't using the tags, but I was actually using them to look at related articles" **Participant 10**
- "[A] lot of the keywords I used were already used as keywords in CiteULike, so I think they were good keywords. But because they list several keywords along the bottom, I can pick up new ones as I go." **Participant 5**
- "You can search by tags or you can search by people and it also shows the people who are interested in this idea... this search term that I put in." **Participant 7**

Implications for the Future Study of Tagging in Information Science

- libraries have begun to offer social networking and social bookmarking features in their OPACs
  - e.g. PennTags at the University of Pennsylvania, Bibliocommons at Oakville Public Library
- Study 1:
  - shows tag usage forms patterns of common "index" terms but lack of consensus continues
  - clustering is based on contextual relatedness not synonymy

Implications for IS 2

- Study 2: shows distinct differences in assumptions of taggers, authors and professional indexers
  - users violate indexing principle of co-extensivity but join indexers in grappling with issues of multi-word concepts and emerging fields
- Study 3: participants used tags as terms and as guides
- also found the notion of a person having gathered (through collection development) these items personally interesting

Implications for IS 3

- in tagging, term relationships grow as more people tag the same item with different terms
- higher number of unique terms in the 2008 data of study 1 suggests people are still using their own terms, despite evidence of convergence
- tagging raises questions about the nature of indexing structures and the relationship to PIM
- in the aggregate both subject and affective terms provide clues as to the relationships of other users to items

Tagging and Indexing

- thesis suggests that the convergence and divergence of terms present in and the personal dimension of tagging offer a way to combine the power of controlled vocabulary structures with the novelty of social tagging structures
- ultimately, tagging places entities in a helpful web of relationships based on personal and collective classification and indexing experiences

Margaret Kipp <margaret.kipp@gmail.com>
Faculty of Information and Media Studies
University of Western Ontario

Questions?

Thank you!
Merci!