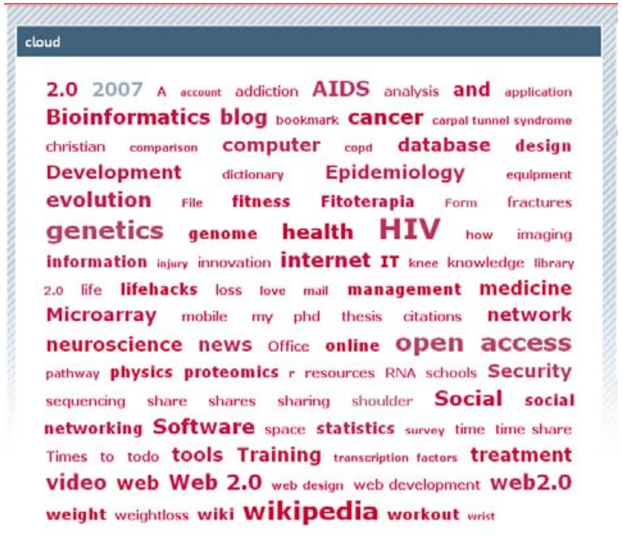
Tagging tagging.

Analysing user keywords in scientific bibliography management systems



cloud

2.0 2007 A account addiction AIDS analysis and application Bioinformatics blog bookmark cancer carpal tunnel syndrome christian comparison computer copd database design Development dictionary Epidemiology equipment evolution File fitness Fitoterapia Form fractures genetics genome health HIV how imaging information injury innovation internet IT knee knowledge library 2.0 life lifehacks loss love mail management medicine Microarray mobile my phd thesis citations network neuroscience news Office online open access pathway physics proteomics r resources RNA schools Security sequencing share shares sharing shoulder Social social networking Software space statistics survey time time share Times to todo tools Training transcription factors treatment video web Web 2.0 web design web development web2.0 weight weightloss wiki wikipedia workout wrist

Markus Heckner, Susanne Mühlbacher, Christian Wolff

University of Regensburg

Outline

1. Introduction - Research context and related work
2. Goals and Method
3. Tag Category Models (LTCM, FTCM, T2TCM)
4. Tags versus author keywords
5. Conclusion and Discussion

Background

- growing number of systems that use tagging (e.g. flickr, del.icio.us, citeulike, connotea, google video, youtube)
- user provided vocabulary for the annotation of resources
- tagging as a possible solution to the „vocabulary problem“ stated by Furnas (1987)
- tags can “identify qualities or characteristics” of resources (Kipp and Campbell 2006, Kipp 2007, Feinberg 2006, Kroski 2005)

Related work

- Empirical research rare and limited to...
  - Automatic statistical analyses (Golder and Hubermann 2006, Hammond 2005)
  - Systems from personal or private domain

- Still little research on functional and linguistic aspects of tags (especially in the context of scientific bibliography management systems)

Research questions

- Is it possible to **discover regular patterns in tag usage** and to establish a stable category model?
- To what degree are social **tags taken from** or findable in the **full text** of the tagged resource?
- How do social **tags differ from author keywords**?
- Does **tagging go beyond content description** and how?

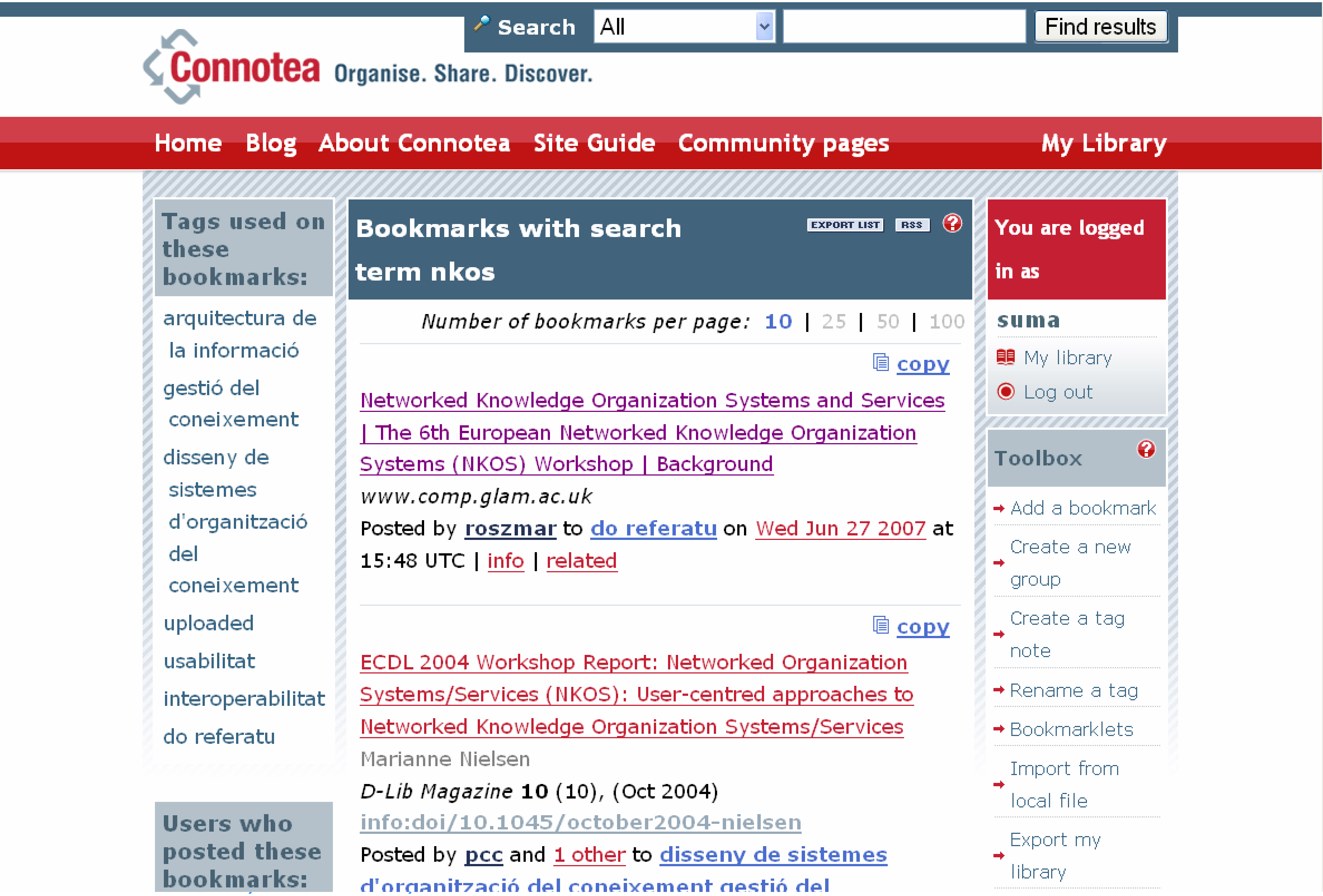
Method

Dataset and model

- (Step 1) *Explorative creation of a category model*
    - *Random sample from connotea.org (Web API)*
    - *Creation of individual classes by information scientists*
    - *Consolidation to preliminary model*
- *(Step 2) Explanatory case study: Applying and verifying the category model*
    - Second sample (500 ICT related articles, 1191 tags)
    - Assign to preliminary model

  ➔Evolution of stable category model

Connotea (search for “NKOS“)



Tags used on these bookmarks:

arquitectura de la informació

gestió del coneixement

disseny de sistemes d'organització del coneixement

uploaded

usabilitat

interoperabilitat

do referatu

Users who posted these bookmarks:

Bookmarks with search term nkos

Number of bookmarks per page: 10 | 25 | 50 | 100

copy

Networked Knowledge Organization Systems and Services | The 6th European Networked Knowledge Organization Systems (NKOS) Workshop | Background

www.comp.glam.ac.uk

Posted by roszmar to do referatu on Wed Jun 27 2007 at 15:48 UTC | info | related

copy

ECDL 2004 Workshop Report: Networked Organization Systems/Services (NKOS): User-centred approaches to Networked Knowledge Organization Systems/Services

Marianne Nielsen

D-Lib Magazine **10** (10), (Oct 2004)

info:doi/10.1045/october2004-nielsen

Posted by pcc and 1 other to disseny de sistemes d'organització del coneixement gestió del

You are logged in as

suma

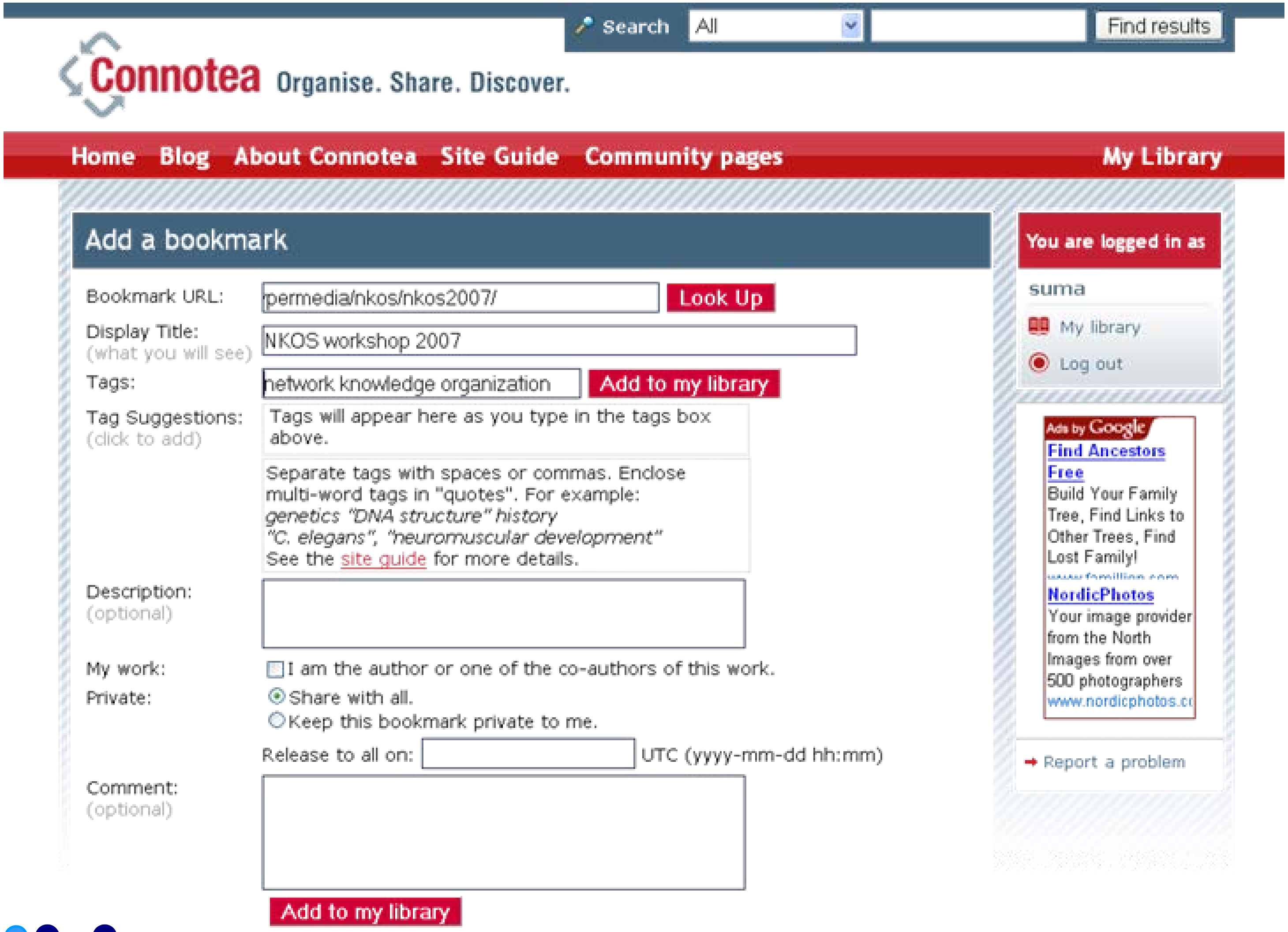
My library

Log out

Toolbox

- Add a bookmark
- Create a new group
- Create a tag note
- Rename a tag
- Bookmarklets
- Import from local file
- Export my library

Connnotea: Tagger's view (tagging NKOS 2007)



Add a bookmark

Bookmark URL: permedia/nkos/nkos2007/ Look Up

Display Title: (what you will see) NKOS workshop 2007

Tags: network knowledge organization Add to my library

Tag Suggestions: (click to add) Tags will appear here as you type in the tags box above.

Separate tags with spaces or commas. Enclose multi-word tags in "quotes". For example:
genetics "DNA structure" history
"C. elegans", "neuromuscular development"
See the site guide for more details.

Description: (optional)

My work: □ I am the author or one of the co-authors of this work.

Private: ● Share with all.
○ Keep this bookmark private to me.

Release to all on: UTC (yyyy-mm-dd hh:mm)

Comment: (optional)

Add to my library

You are logged in as

suma

My library

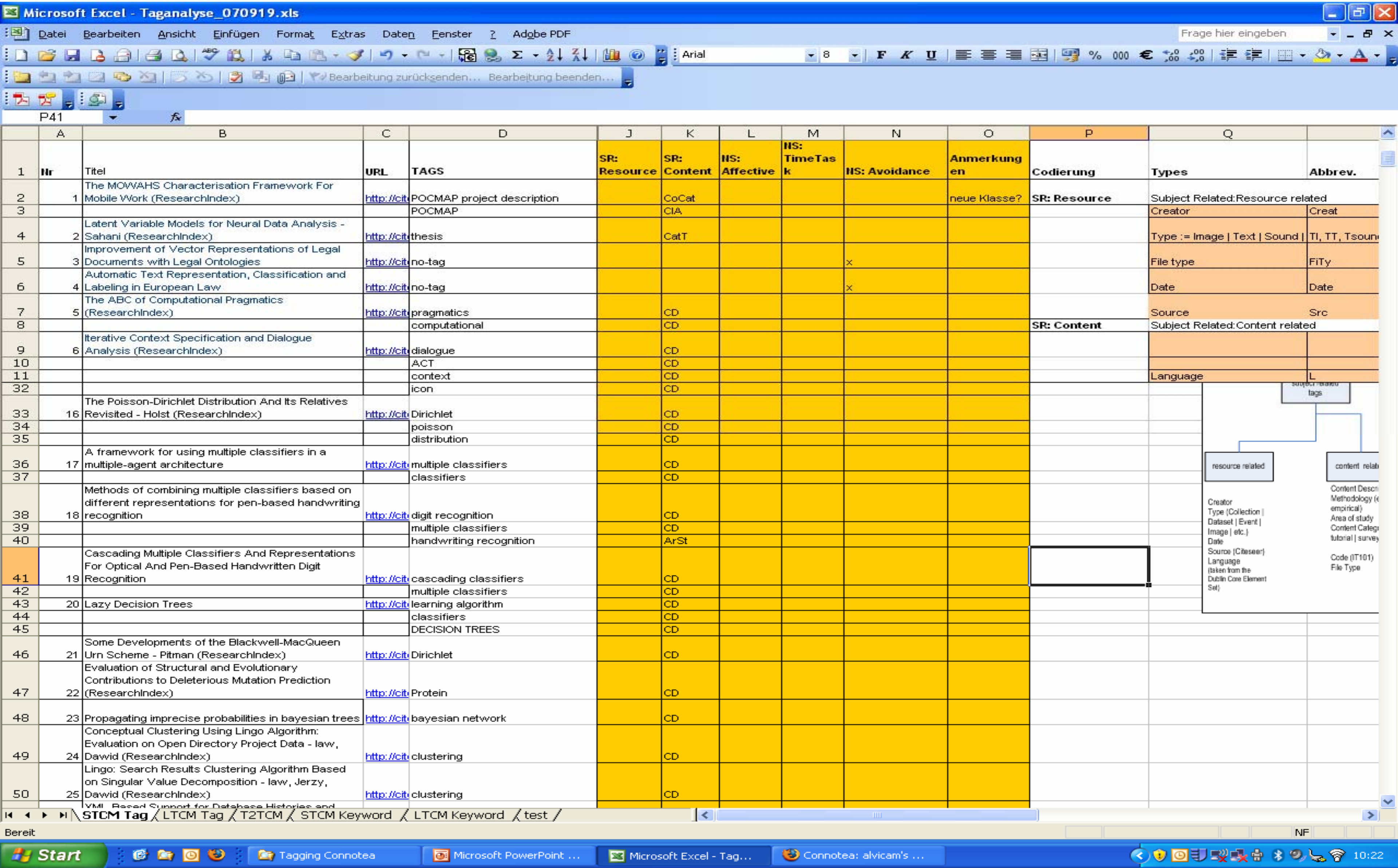
Log out

Ads by Google
Find Ancestors Free
Build Your Family Tree, Find Links to Other Trees, Find Lost Family!
www.familliop.com

NordicPhotos
Your image provider from the North Images from over 500 photographers
www.nordicphotos.c

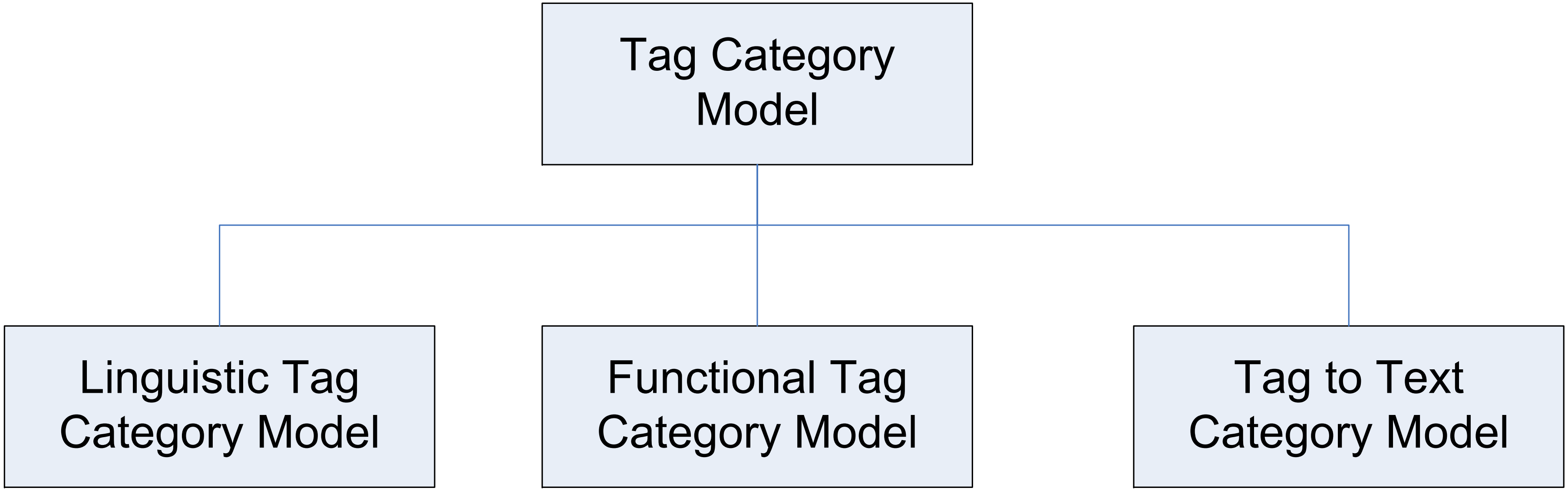
Report a problem

Data Analysis in Excel



	A	B	C	D	J	K	L	M	N	O	P	Q	
1	Nr	Titel	URL	TAGS	SR: Resource	SR: Content	NS: Affective	NS: TimeTask	NS: Avoidance	Anmerkungen	Codierung	Types	Abbrev.
2	1	The MOWAHS Characterisation Framework For Mobile Work (ResearchIndex)	http://cit	POCMAP project description		CoCat				neue Klasse?	SR: Resource	Subject Related:Resource related	
3				POCMAP		CIA						Creator	Creat
4	2	Latent Variable Models for Neural Data Analysis - Sahani (ResearchIndex)	http://cit	thesis		CatT						Type := Image | Text | Sound |	TI, TT, Tsoun
5	3	Improvement of Vector Representations of Legal Documents with Legal Ontologies	http://cit	no-tag					x			File type	FiTy
6	4	Automatic Text Representation, Classification and Labeling in European Law	http://cit	no-tag					x			Date	Date
7	5	The ABC of Computational Pragmatics (ResearchIndex)	http://cit	pragmatics		CD						Source	Src
8				computational		CD					SR: Content	Subject Related:Content related	
9	6	Iterative Context Specification and Dialogue Analysis (ResearchIndex)	http://cit	dialogue		CD							
10				ACT		CD							
11				context		CD						Language	L
32				icon		CD							
33	16	The Poisson-Dirichlet Distribution And Its Relatives Revisited - Holst (ResearchIndex)	http://cit	Dirichlet		CD							
34				poisson		CD							
35				distribution		CD							
36	17	A framework for using multiple classifiers in a multiple-agent architecture	http://cit	multiple classifiers		CD							
37				classifiers		CD							
38	18	Methods of combining multiple classifiers based on different representations for pen-based handwriting recognition	http://cit	digit recognition		CD							
39				multiple classifiers		CD							
40				handwriting recognition		ArSt							
41	19	Cascading Multiple Classifiers And Representations For Optical And Pen-Based Handwritten Digit Recognition	http://cit	cascading classifiers		CD							
42				multiple classifiers		CD							
43	20	Lazy Decision Trees	http://cit	learning algorithm		CD							
44				classifiers		CD							
45				DECISION TREES		CD							
46	21	Some Developments of the Blackwell-MacQueen Urn Scheme - Pitman (ResearchIndex)	http://cit	Dirichlet		CD							
47	22	Evaluation of Structural and Evolutionary Contributions to Deleterious Mutation Prediction (ResearchIndex)	http://cit	Protein		CD							
48	23	Propagating imprecise probabilities in bayesian trees	http://cit	bayesian network		CD							
49	24	Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data - law, Dawid (ResearchIndex)	http://cit	clustering		CD							
50	25	Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition - law, Jerzy, Dawid (ResearchIndex)	http://cit	clustering		CD							
		XML Based Support for Database Histories and											

Emerging models



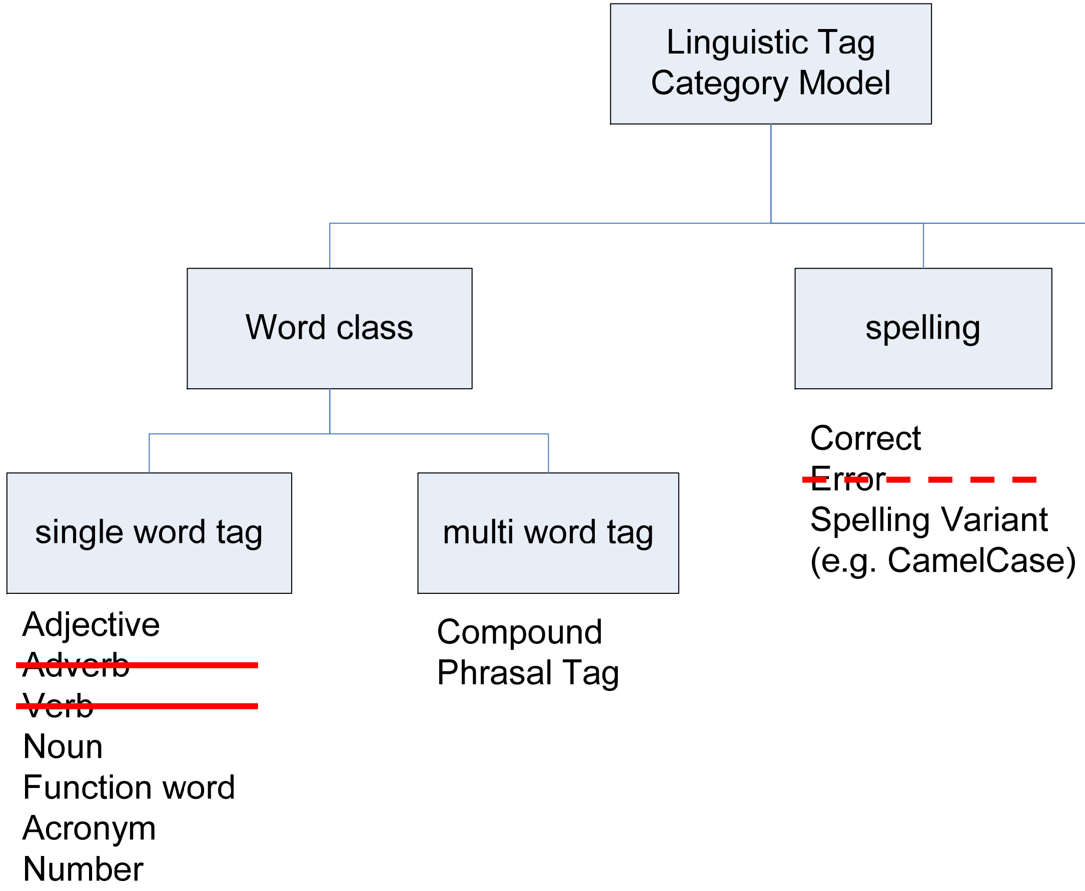
Tag Category Model

Linguistic Tag Category Model

Functional Tag Category Model

Tag to Text Category Model

Linguistic model (morphosyntax, lexicon, orthography)



Linguistic Tag Category Model

Word class

spelling

neologisms

language

single word tag

multi word tag

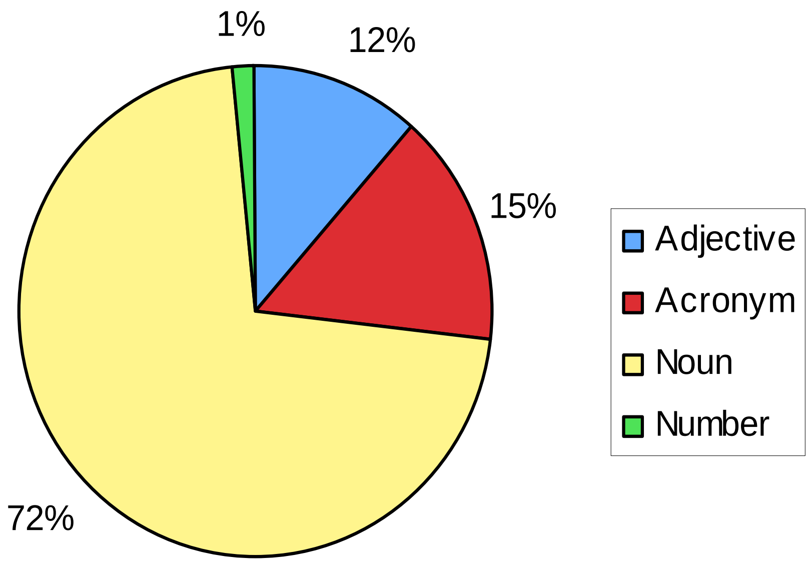
Adjective
~~Adverb~~
~~Verb~~
Noun
Function word
Acronym
Number

Compound
Phrasal Tag

Correct
~~Error~~
Spelling Variant
(e.g. CamelCase)

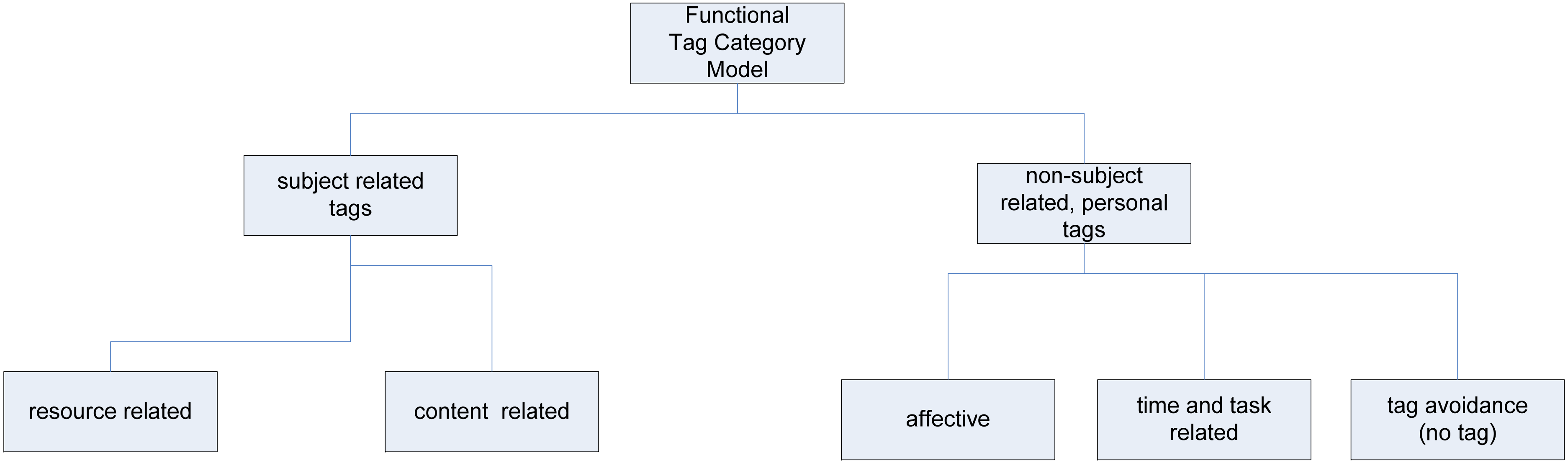
Linguistic Model

- Users do not tag with verbs or adverbs
- Acronyms and Adjectives rather common



Single word tags

Functional / Semantic model



Functional Tag Category Model

subject related tags

non-subject related, personal tags

resource related

content related

affective

time and task related

tag avoidance (no tag)

Creator
~~Resource Type~~
(Software | Sound | Text | Image | etc.)
~~File Type~~
Date
Source (Citeseer)
~~Language~~
(partially taken from the Dublin Core Element Set)

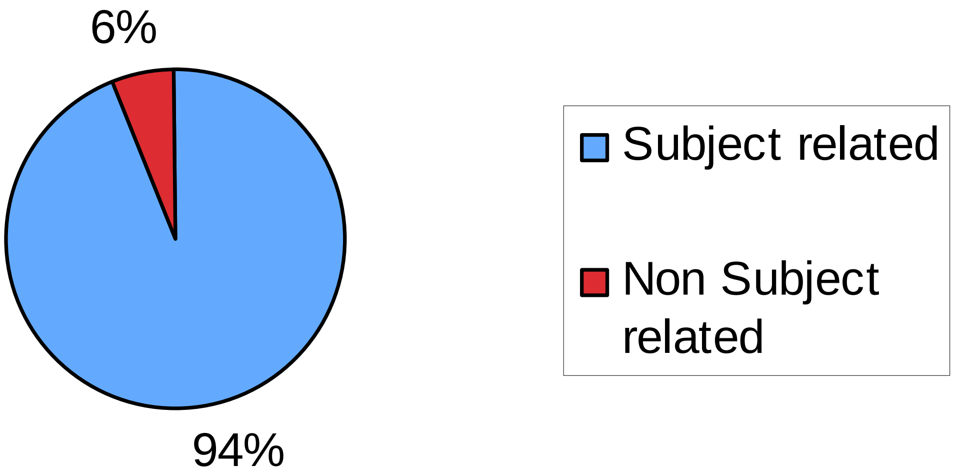
Content Description
Area of study
Classification Attempt

Content Category (review | tutorial | survey | manual)
Methodology (e.g. empirical)
Code (IT101)

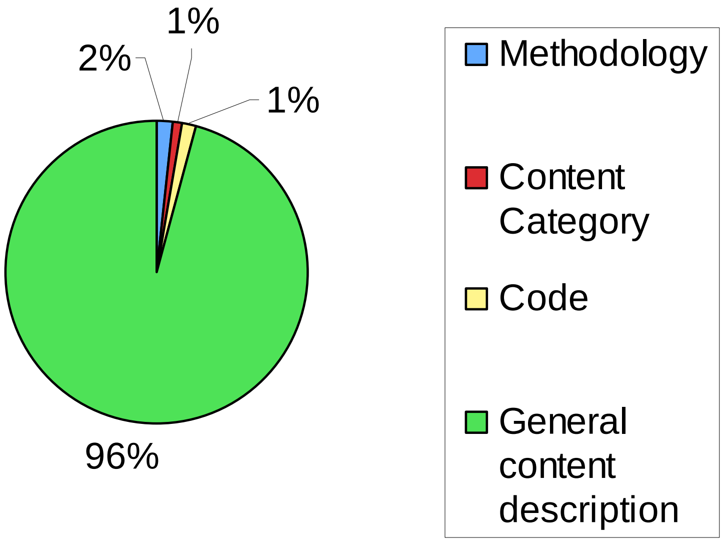
~~positive~~
~~negative~~

Action oriented
Context Workflow related

Functional / Semantic Model



Subject related vs. non-subject related tags



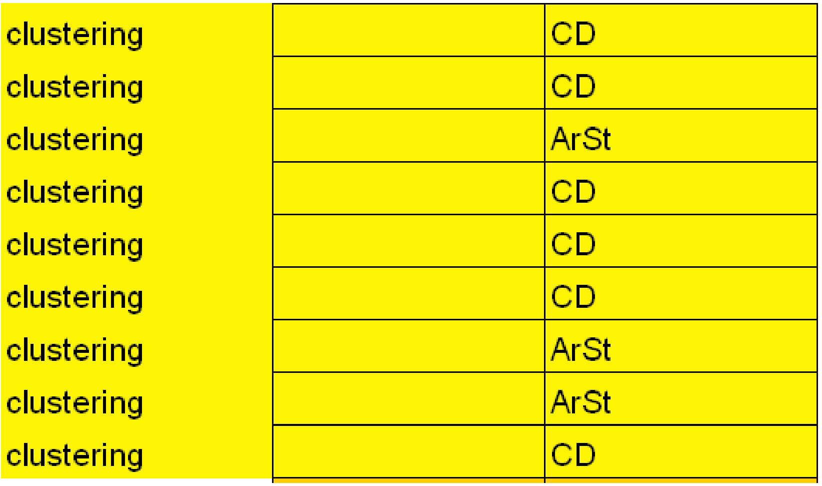
What do content related tags describe?

➔contrary to previous studies 16% non-subject related tags
Kipp and Campbell (2006)

What form of content description?

What kind of tag is “clustering”?

- Representation of content (CD, mental copy & paste)
- Description of the area of study (ArSt)
- Classification of content (ClA)



clustering		CD
clustering		CD
clustering		ArSt
clustering		CD
clustering		CD
clustering		CD
clustering		ArSt
clustering		ArSt
clustering		CD

➔ Tough decision, never independent of document content

Content description or more?!

- Tags ***exclusive*** to one user
- ***labeling*** function?

User	Tag	used (# of docs)
linguini	958	19
fsyu2005	timetabling	6
mthomure	latent-semantic-analysis	7
mthomure	image-search	12
mreddington	HFSP-funded	87
radico	Trs	4
wyng	sensornet	18

- The “Super-label” / complex tags
- hierarchical structures in tags

data::gene perturbation

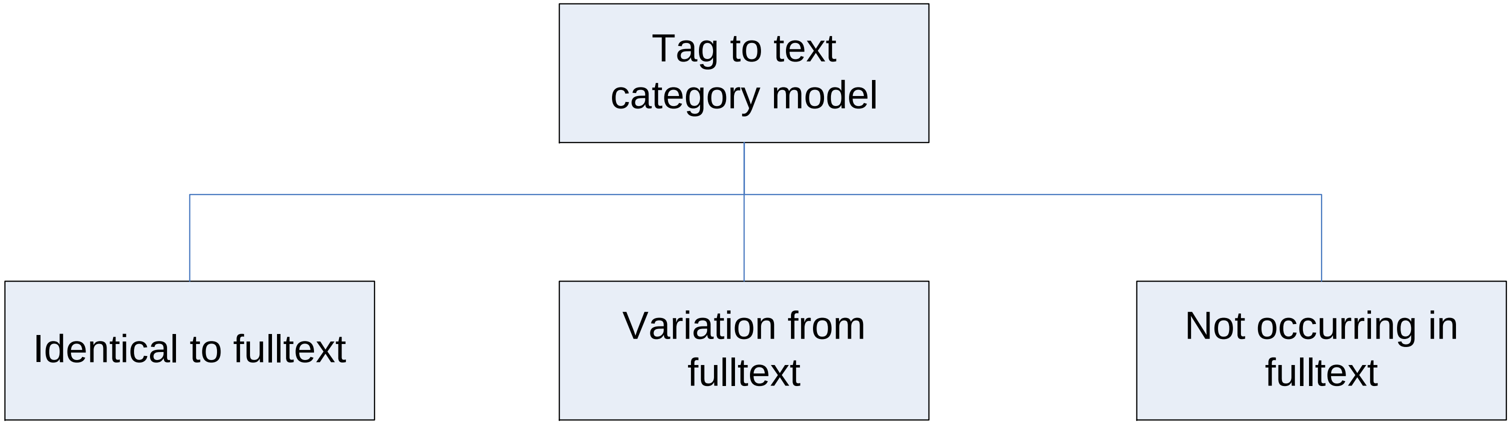
data::sequence

method::transitive reduction

➔Distinction between content description and labels used for workflow organisation is a difficult task!

Tag to text model

- relationship between tags and document (full) text – where are tags found in the text?



Tag to text category model

Identical to fulltext

Variation from fulltext

Not occurring in fulltext

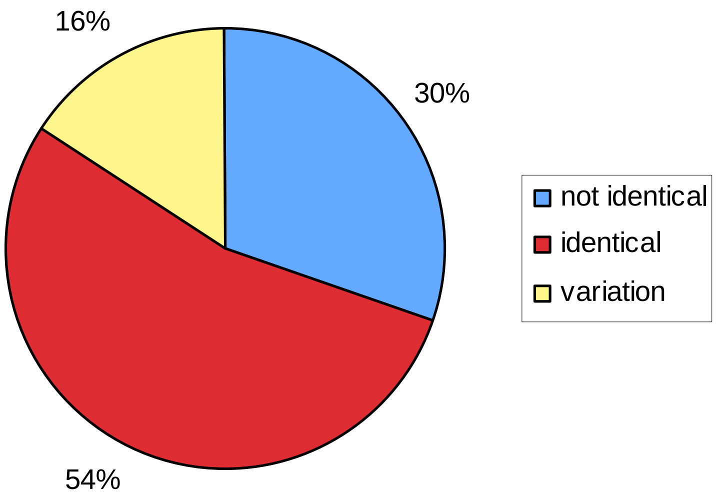
In title
In Abstract
In Fulltext
Same as keyword

Spelling error
Stemming / Inflection

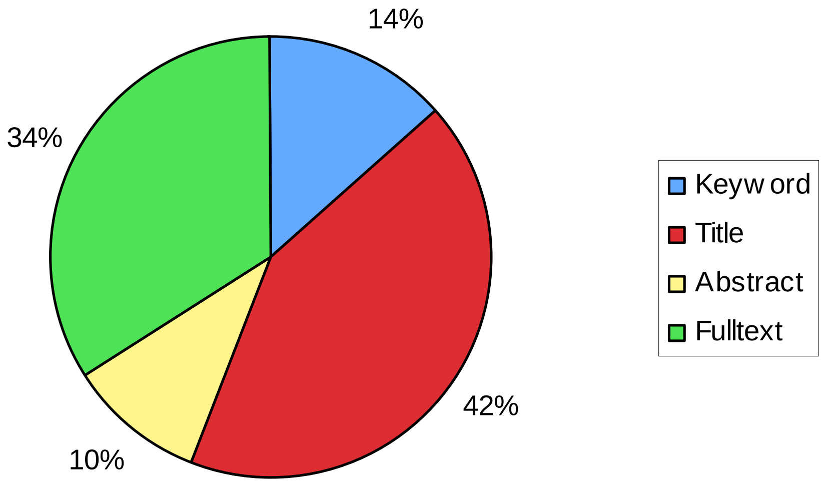
Synonym
Hyponym
Hyperonym

No relation at all

Tag to text category model

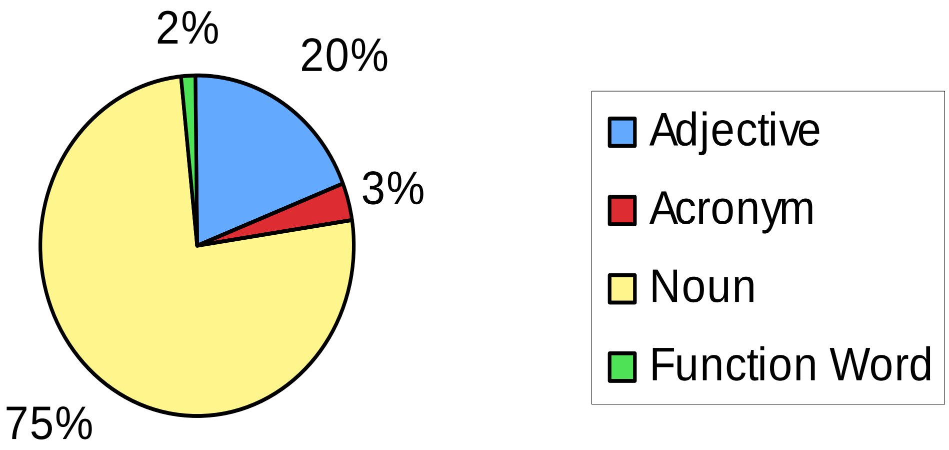


Relation of tag to full text

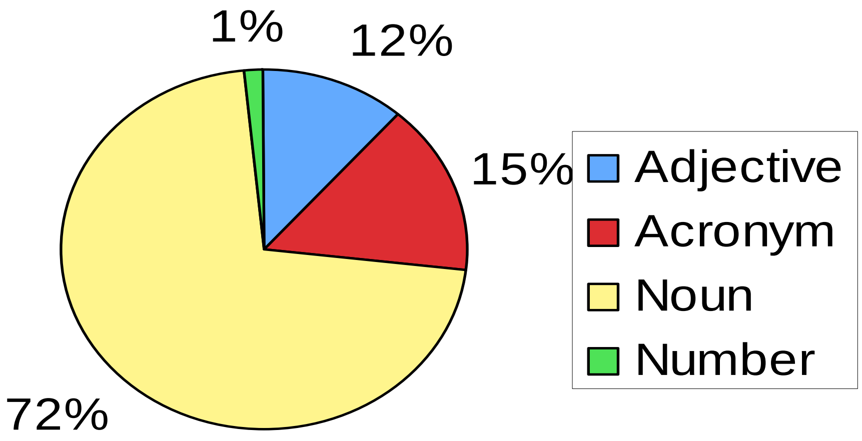


Position of tag in resource

Tags vs. author keywords – comparison of word classes



Author Keywords



Tags

Tags vs. author keywords (preliminary results)

- only documents where both are present were considered
- 1,3 words per tag vs. 1,8 words per keyword
- app. 2,2 tags/document vs. 5,6 keywords / document
- overlap:
  - identical or *near identical* concepts in tags and keywords
  - overlap bounded in almost all cases by the (lesser) number of tags
  - ca. 58% overlap in content
  - only 30% with respect to *all keywords*

Tags vs. author keywords: Relations

- typical relations between related tags and keywords:
  - more ***general*** tags (e.g. RNA (tag) vs. RNA secondary structures (keyword))
  - more ***specific*** tags (e.g. information visualization (tag) vs. visualization (keyword)
  - difference in number (e.g. wavelet (tag) vs. wavelets (keyword))
  - translation (recuperació de la informació (tag) vs. information retrieval (keyword))
  - different tags are part of multiword keywords (e.g. text, ..., input (tags) vs. text input (keyword)
- taggers tend to use less and more general concepts than authors

Words per Tag vs. Words per author keywords

Number of words per tag	**Occurrences**	**Percent total**
1	**844**	**70,87 %**
2	289	24,27 %
3	46	3,87 %
4	7	0,59 %
5	2	0,17 %
6	1	0,08 %
7	0	0
8	2	0,17 %
Overall	**1191**	**100 %**

Number of words per keyword	**Occurrences**	**Percent total**
1	331	34,4
2	**478**	**49,7**
3	128	13,3
4	19	1,98
5	4	0,42
6	1	0,20
Overall	**961**	**100 %**

Outlook

- further refinement of tag model and research method
- comparative studies concerning
  - the influence of system design on tagging strategies
  - comparison with ***expert*** keywords given by information professionals (e.g. in the INSPEC database)
- application of the model for different types of tagged content (videos, bookmarks, images)
- design hints for tagging systems
  - additional non-content-related tagging options (rating (content, readability, quality etc.), workflow)

References

Furnas, G. W.; Landauer, T. K.; Gomez, L. M. & Dumais, S. T. (1987), 'The vocabulary problem in human-system communication', Commun. ACM 30(11), 964--971.

Crystal, David (2006). Language and the Internet. Cambridge: Cambridge University Press.

Golder, S. & Huberman, B. A. (2006), 'The Structure of Collaborative Tagging Systems', Journal of Information Science 32, 198--208

Hammond, T., Hannay, T., Lund, B. and Scott, J. Social Bookmarking Tools – A General Overview. D-Lib Magazine11, 4 (April 2005)

Kipp, Margaret E. I. and Campbell, D. Grant (2006a) Patterns and Inconsistencies in Collaborative Tagging Systems : An Examination of Tagging Practices. In Proceedings Annual General Meeting of the ASIST, Austin, Texas (US).

Kipp, M. (2006). Complementary or Discrete Contexts in Online Indexing : A Comparison of User, Creator, and Intermediary Keywords., *Canadian Journal of Information and Library Science*.

Kipp, Margaret E. I. (2007). @toread and Cool: Tagging for Time, Task and Emotion. In: Proc. Information Architecture Summit 2007. Las Vegas. [Online: http://eprints.rclis.org/archive/00010445/]

Marlow, C.; Naaman, M.; Boyd, D. & Davis, M. (2006),HT06, tagging paper, taxonomy, Flickr, academic article, to read, in HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia', ACM Press, New York, NY, USA, pp. 31--40.

Sen, S.; Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M. & Riedl, J. (2006). Tagging, communities, vocabulary, evolution. *in* 'Proceedings of CSCW 2006.

Sinha, R. (2005). A cognitive analysis of tagging. [Online: http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html]. 2 August 2007.

Voß, J. (2007). Tagging, Folksonomy & Co - Renaissance of Manual Indexing? In: Osswald, A.; Stempfhuber, M.; Wolff, C. (Eds.): Open Innovation. Proc. 10th International Symposium for Information Science. Constance: UVK, 243-254.

Yew, J., Faison, G., Teasley, S. (2007). Learning by tagging: group knowledge formation in a self-organizing learning community. ICLS '06: Proceedings of the 7th international conference on Learning sciences.

Affiliations

Markus Heckner
markus.heckner@paedagogik.uni-regensburg.de
Media Educational Science

Susanne Mühlbacher
susanne1.muehlbacher@sprachlit.uni-regensburg.de
Information Science

Christian Wolff
christian.wolff@sprachlit.uni-regensburg.de
Media Computer Science

University of Regensburg