Wisdom of the Crowds

Mathias Lux

mlux@itec.uni-klu.ac.at
http://www.itec.uni-klu.ac.at/~mlux

The Long Tail
Folksonomies
Wikis
Tagging
Web 2.0
Ontologies
Users

Web 2.0

- **The Long Tail**
  - Small groups make up the big mass (Amazon makes no fortune with chartbreakers ...)
- **Data is the Next Intel Inside**
  - Applcations heavily depend on underlying data (e.g. Amazon and Barnes&Noble: Amazon started with the same data, but is now provider)
- **Users Add Value**
  - Implicit & Explicit integration of user generated data (social bookmarking, see later ...)
- **Network Effects by Default**
  - Metcalfe's Law: The value of a communication network

Web 2.0

- **Some Rights Reserved**
  - E.g. Creative Commons allow creative re-use of original work (mashups, aggregation, remixability)
- **The Perpetual Beta**
  - Continous development and rollout instead of monolithic releases.
- **Cooperate, Don‘t Control**
  - Offering services and APIs, Feeds and Data
- **Software above the Level of the Single Device**
  - Pervasive computing, mobile devices, „the web as platform“ ...

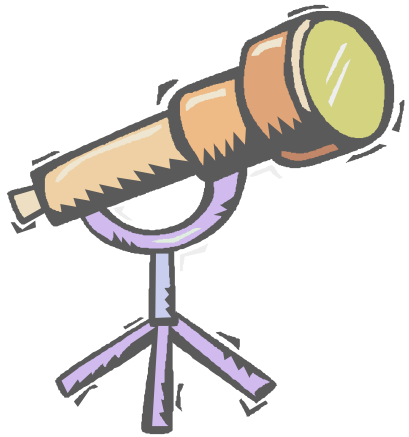
Contents

- Wisdom of the Crowds: Examples
- Folksonomies
- Folksonomy Analysis (an example)
- Summary & Conclusions

Examples for WOTC

In Web 2.0 lots of examples exist ...

- Social Bookmarking
- Collaborative Annotations
- Social Media Sharing
- Collaborative Content Creation
- Blogging

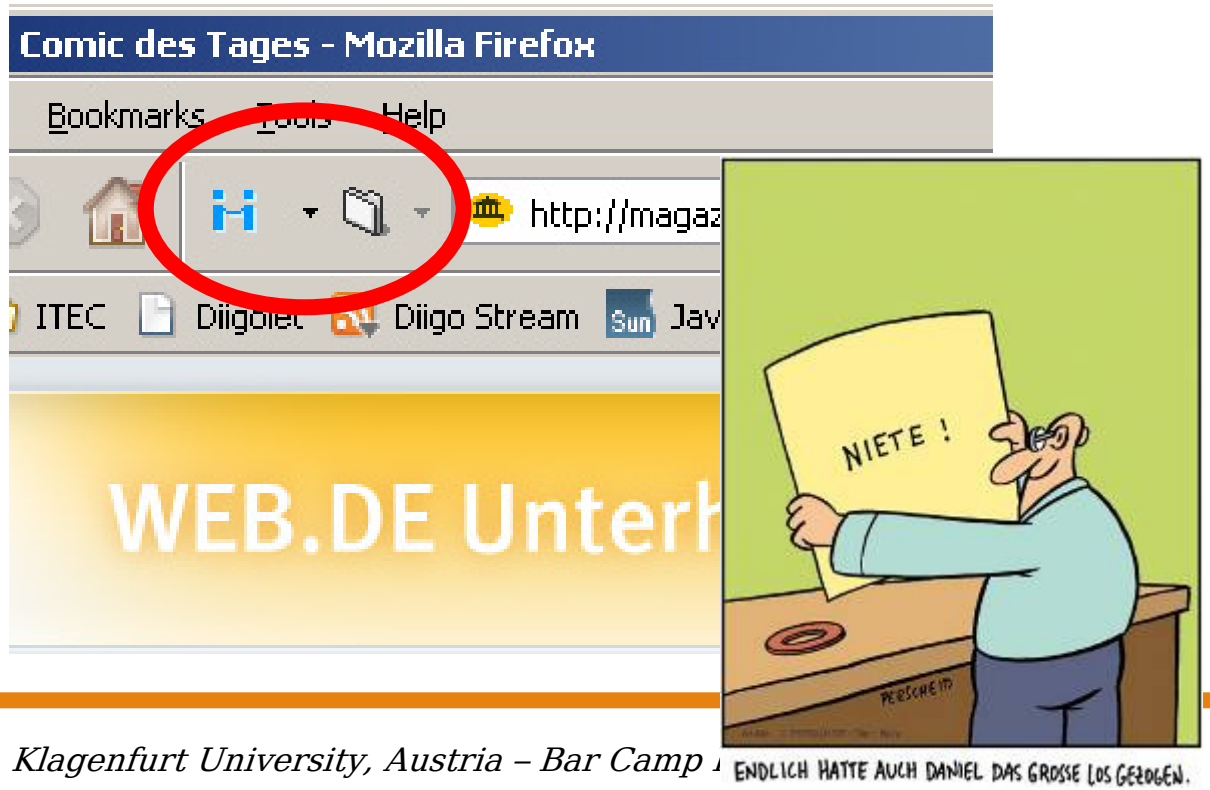


Obvious and Hidden

- Some approaches are very obvious

Obvious and Hidden

- Some are more subtle
  - Plug-Ins with client-server communication

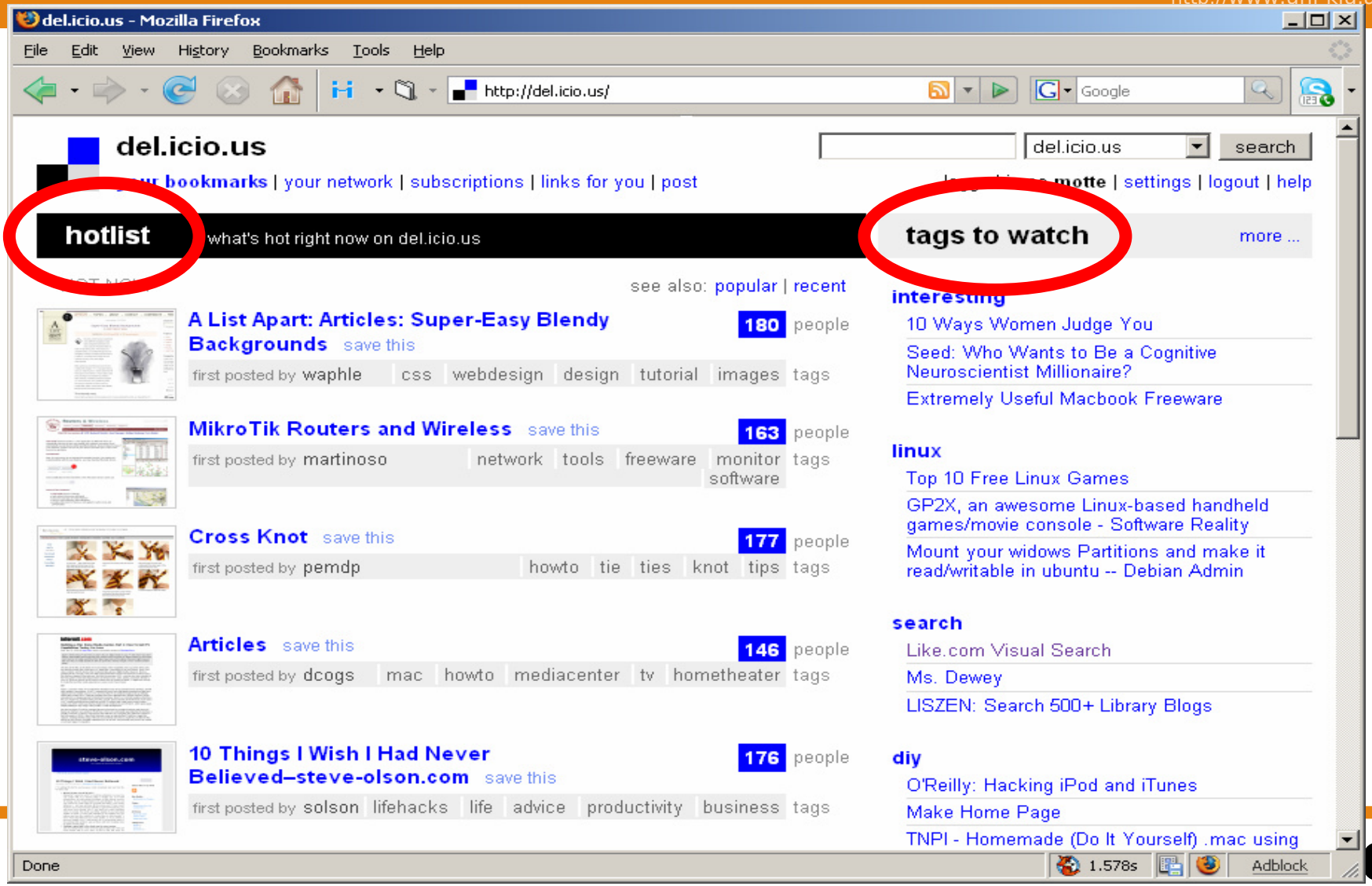


Examples: Social Bookmarking

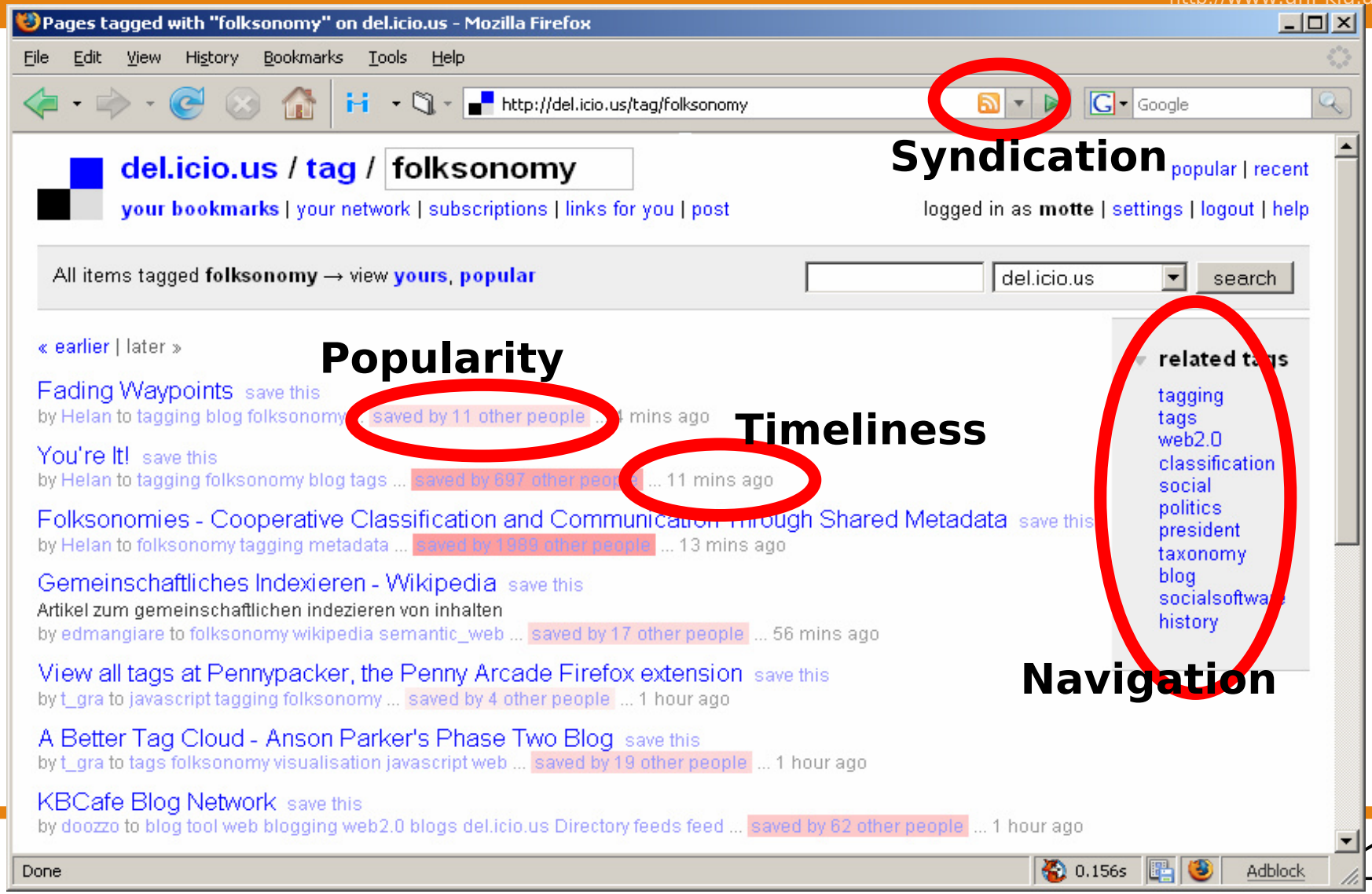
Social Bookmarking defined:

- Bookmarking Resources
- Providing a „stream of bookmarks“
- Eventually additional support for
  - Tagging (keywords)
  - Caching (Saving the state of the bookmark)
  - Organization & Collaboration (Groups)

Example: del.icio.us



Example: del.icio.us



Example: del.icio.us

- User Interface
  - Clean and easy2use
  - Powerful tools (bookmarklets & plugins)
- Additional Features
  - Thumbnails
  - Social Networking

del.icio.us

- User intentions are unclear:
  - Self-organization or group organization
  - Participation / Being part of it
- Explicitly Generated
  - Bookmarking & Tagging
  - Tag Bundles
- Implicitly Generated
  - Time, Interestingness, The „Seen Web“
  - User Profile, Social Network

Example: Web Clippings & Sticky Notes

- Several Applications exist:
  - Google Notebook
  - Clipmarks.com
  - etc.
- Our example:
  - Annotate parts of web pages ...
  - Done using Diigo

image from http://versaphile.com/fanwork/sg/post-it.html

Demo: Diigo

- Video ...

diigo

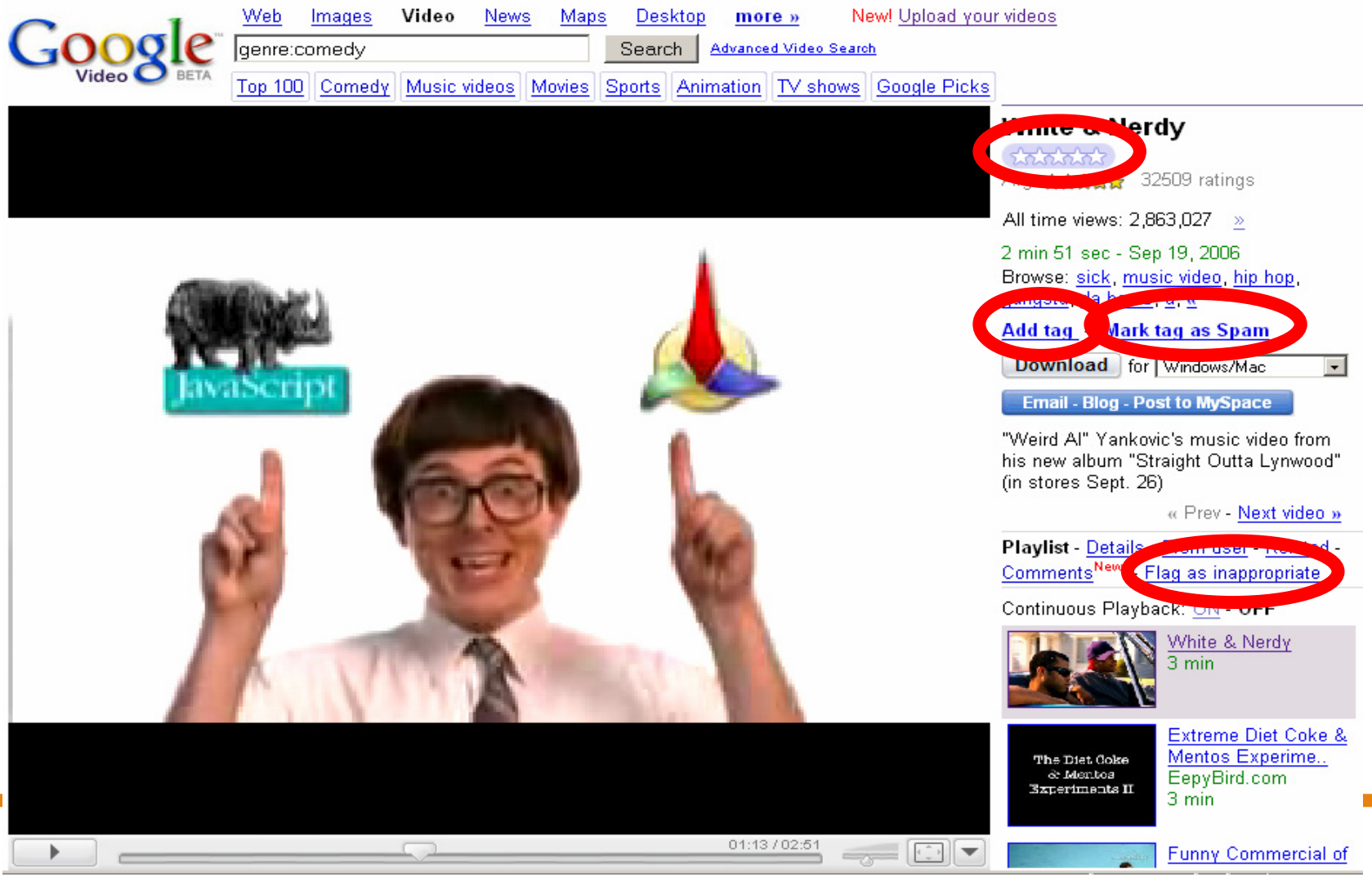
Example: Diigo

- User intentions are unclear:
  - For himself (later reading / work)
  - For a group of people / coworkers
- Explicitly Generated
  - Highlighting & bookmarking
  - Tagging & Description, Sticky Note
- Implicitly Generated
  - Time, Interestingness, The „Seen Web“
  - User Profile

Examples: Social Media Sharing

- Flickr.com, Bubbleshare.com, Zooomr.com, ...
  - Sharing images & annotations
- YouTube.com, Google Video, VideoEgg.com. ...
  - Sharing videos & annotations
- Pandora, Last.fm
  - Sharing music & flavors

Example: Google Video



Google Video

- Explicitly Generated
  - Ratings, Flags, Tags, Spam Filtering
- Implicitly Generated
  - Interestingness (Charts, etc.)
  - Usage (Client reports events like start, pause, stop, ...)
  - From the HTTP-Request: GET http://video.google....&reportevent=pause...

Example: Wikipedia

Wikipedia is

- A user driven encyclopedia
- Self-organizing & self-directed

General issues:

- Reliability and truth
- Completeness
  (e.g. Computer Science)

Example: Wikipedia: Heinrich Rudolf Hertz

Sign in / create account

article | discussion | edit this page | history

Your continued donations keep Wikipedia running!

Heinrich Rudolf Hertz

From Wikipedia, the free encyclopedia

Heinrich Rudolf Hertz (February 22, 1857 - January 1, 1894) was the German physicist and mechanician for whom the hertz, an SI unit, is named. In 1888, he was the first to demonstrate the existence of electromagnetic radiation by building an apparatus to produce UHF radio waves.

Contents [hide]

1 Biography
- 1.1 Early years
- 1.2 Meteorology
- 1.3 Electromagnetic research
- 1.4 Death and afterwards

2 See also

3 Further reading

4 References, external links, and other articles

Biography [edit]

Early years [edit]

Hertz was born in Hamburg, Germany, to Gustav Ferdinand Hertz, whose father converted from Judaism to Lutheranism and married into a Lutheran family, and Anna Elisabeth Pfefferkorn, herself a

Heinrich Rudolf Hertz

"I do not think that the wireless waves I have discovered will have any practical application."

Born	February 22, 1857 Hamburg, Germany
Died	January 1, 1894 Bonn, Germany
Residence	Germany
Nationality	German

Example: Wikipedia: Heinrich Rudolf Hertz

A „standard“ Wikipedia article having:

- A wiki name: Heinrich_Rudolf_Hertz
- Lots of Wiki-Code
  - Several Outlinks
  - Several Inlinks
  - And the InfoBox ...

Example: Wikipedia: Heinrich Rudolf Hertz

```
{{Infobox_Scientist
|name = Heinrich Rudolf Hertz
|image = Heinrich Rudolf Hertz.jpg
|caption = <div style="font-size:
   90%">'''I do not think that the
   [[wireless]] waves I have discovered
   will have any [[radio|practical
   application]].'''</div>
|birth_date = [[February 22]], [[1857]]
|birth_place = [[Hamburg, Germany]]
|residence = [[Germany]]
   [[Image:Flag_of_Germany.svg|20px|]]
|nationality = [[Germany|German]]
   [[Image:Flag_of_Germany.svg|20px|]]
|death_date = [[January 1]], [[1894]]
|death_place = [[Bonn, Germany]]
```

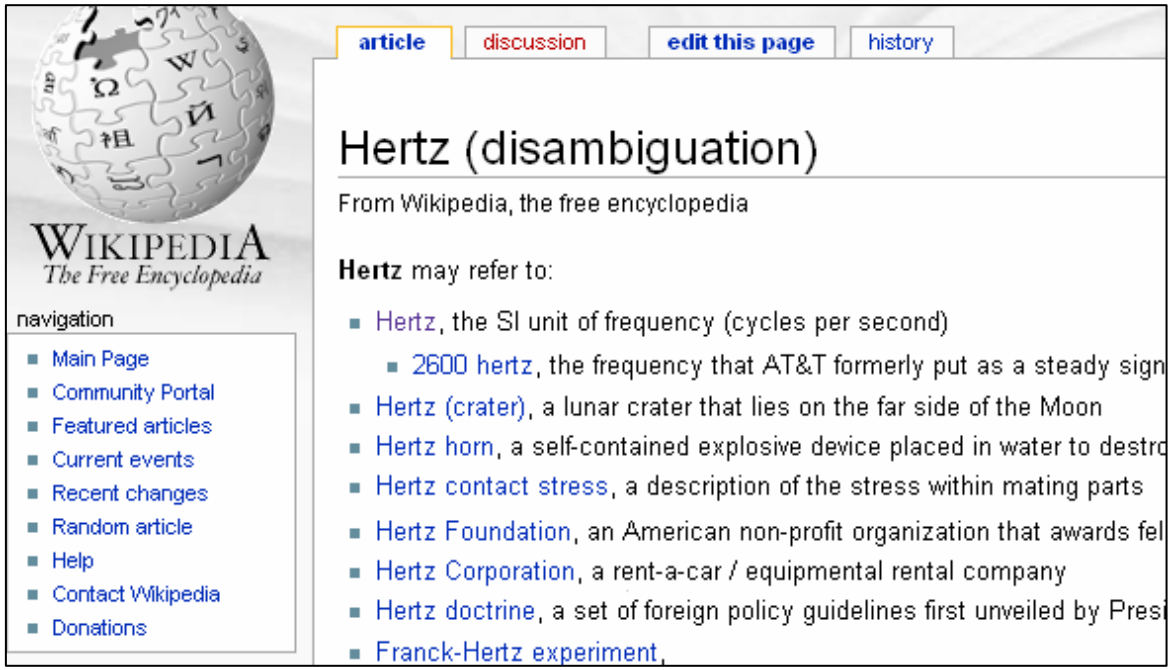
Example: Wikipedia: Heinrich Rudolf Hertz

http://www.uni-klu.ac.at

- Explicitly Generated Metadata
  - Embedded in the data
  - Based on key = concept (identified by Wiki word)
- Implicitly Generated Metadata
  - Popularity (length of the article, browsing behavior)
  - Structure & Links (partially through bots)
  - Introduction of disambiguated concepts (Wiki names)

Wikipedia: Disambiguation

- Disambiguation through Wiki Words
  - Hertz, Hertz_(crater), Heinrich_Rudolf_Hertz, Arne_Hertz



Example: Blogs

- Most prominent concept of Web 2.0
- Explicitly Generated
  - Strong typing of structure (order, header, categories, body text, etc.)
- Implicitly Generated
  - Structure & Links (Trackbacks, bidirectional)
  - Introduction of categories (tags)

What is Wisdom of the Crowds in Web 2.0?

- Bottom up
  - In contrast to controlled vocabularies
  - In contrast to quality ensured content creation processes
- Superimposed structure
  - Instead of using predefined hierarchies
  - Through heavy use of linking
- Huge and fuzzy
  - Unimaginable mass of links & tags
  - Lots of redundant information
- Spammed
  - Just starting ...

Folksonomies

http://www.uni-klu.ac.at

- Definition & Description
- Why do tagging systems work?

Folksonomies

Network of Tags, Users and URLs

- Users describe resources
- By using (multiple) tags

Examples:

- Social bookmarking, media sharing, etc.

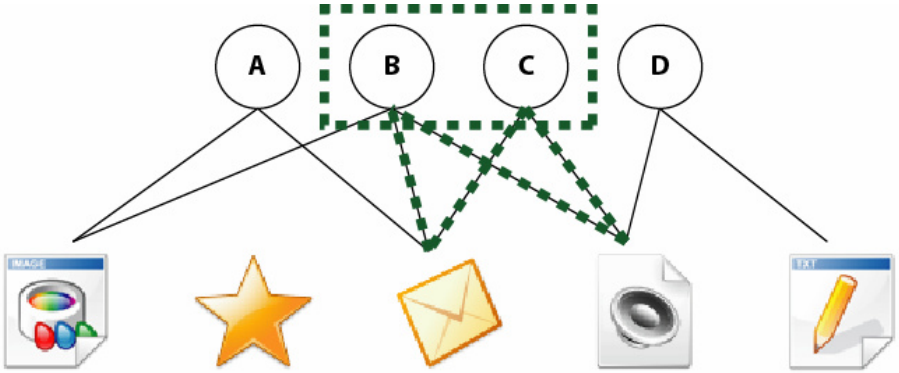
Folksonomies: The Structure

User *tags* resource (URL)

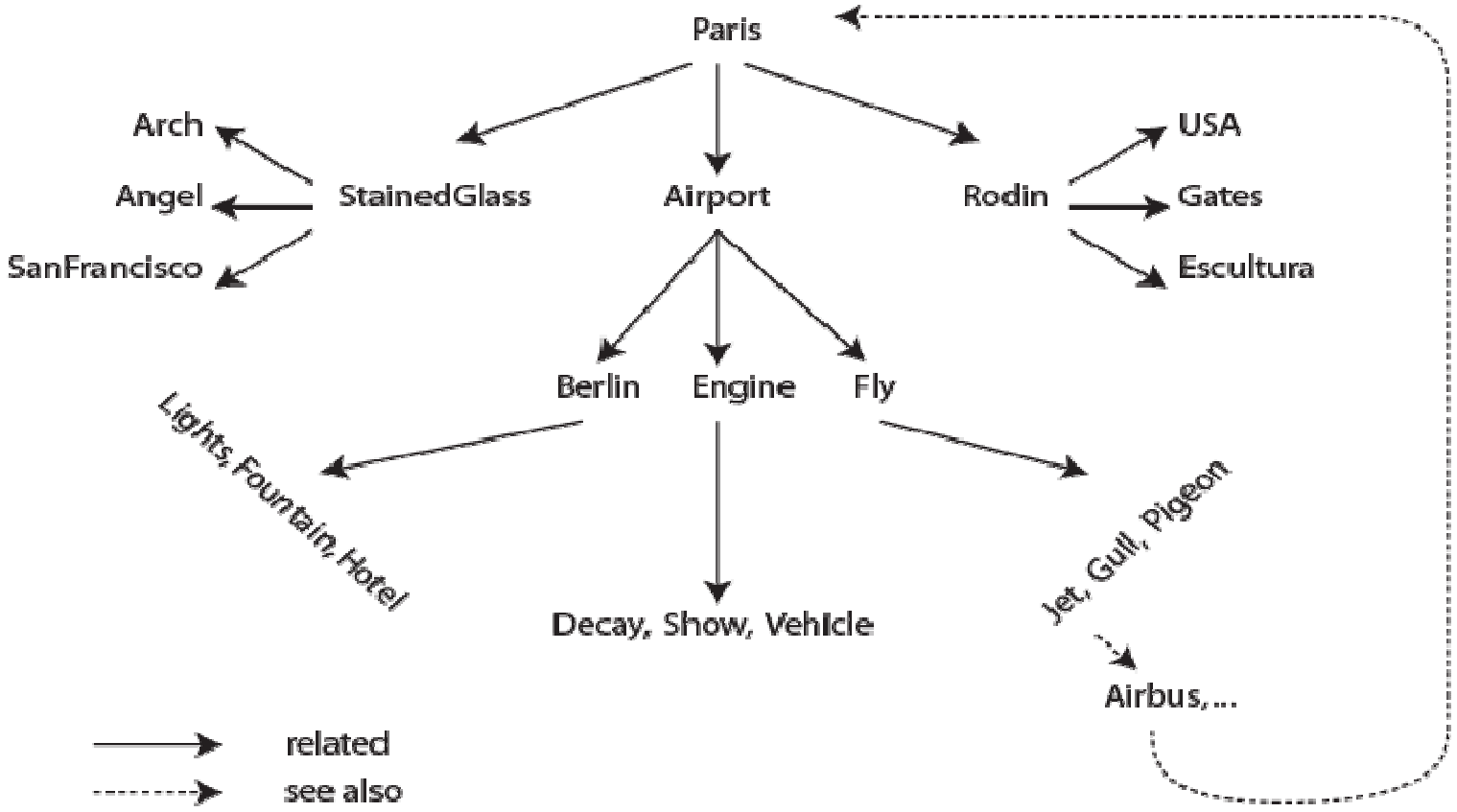
- 1+ words or phrases (bonn, „mathias lux“)
- No controlled vocabulary, taxonomy
- No quality control
- No constraints (language, length, number)

Folksonomies: Structure

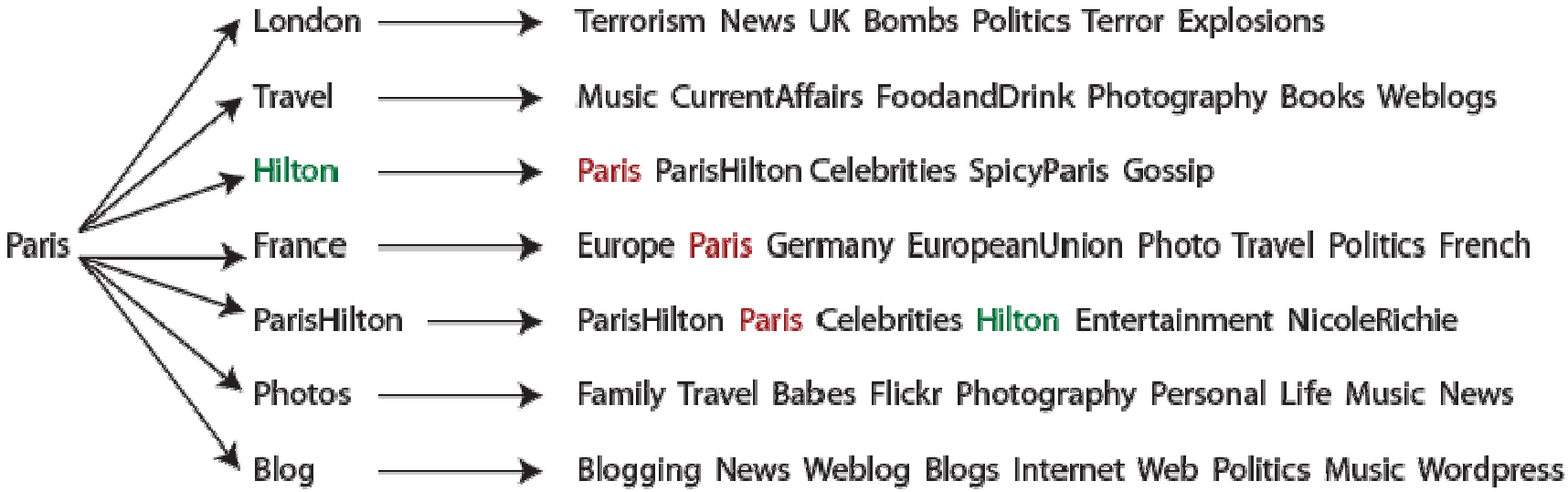
- Tag to URL is a n:m relation
- Superimposed structure through bidirectional links
- Structure is called „folksonomy“



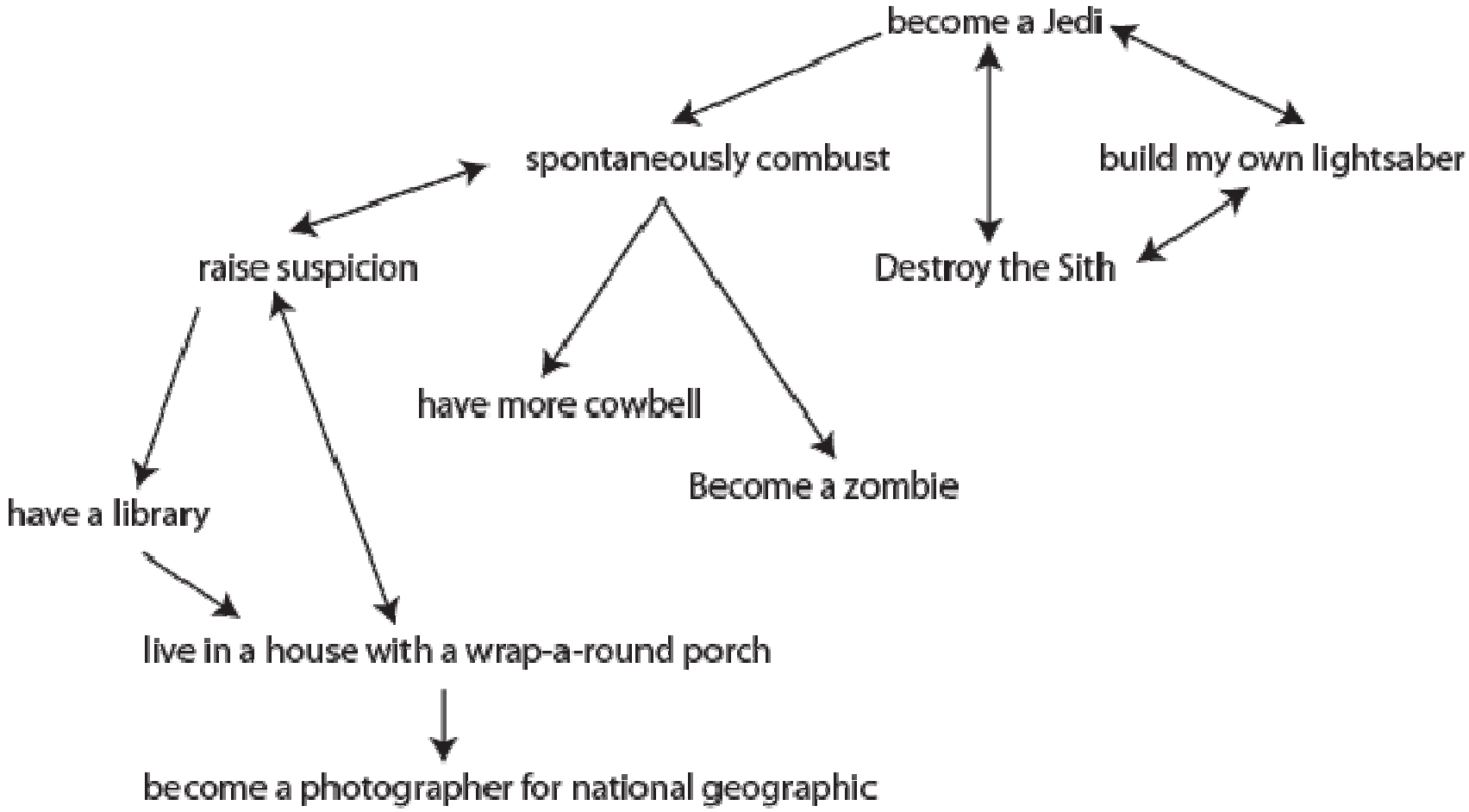
Folksonomy Example: Flickr



Folksonomy Example: Technorati



Folksonomy Example: 43things



Why do tagging systems work?

This was topic of a panel at CHI 2006, following conclusions were drawn:

- Tagging has a benefit for the user
  - Similar to bookmarking, integrated apps
  - Benefit of accessibility from everywhere in the internet
- Tagging allows social interaction
  - Connecting a user to a community trough tags
  - People can subscribe your stream

Why do tagging systems work? (2)

- Tags are useful for retrieval
  - Synonyms and typos vanish in the mass of tags
  - Communities can retrieve “their” stuff (e.g. by special tag)
- Tagging Systems have a low participation barrier
  - Apps are easy to use, intuitive, responsive
  - Free text is used to do the tagging
  - Requires no previous considerations & training

Folksonomy Analysis

- Some mathematical background ...
- .. or geeky stuff ;)

image from http://www.squaredot.com/geek.html

Unified Model for Social Networks & Semantics

Mika P. (2004) *"Ontologies are us: A unified model of social networks and semantics"*

- Ontologies contain instances *I* and concepts *C*
- Ontologies are formal specifications
  - Which are stripped from their original social context of creation
  - Which are static and may get outdated

Where do semantics emerge from?

A third set besides *C* and *I* is needed

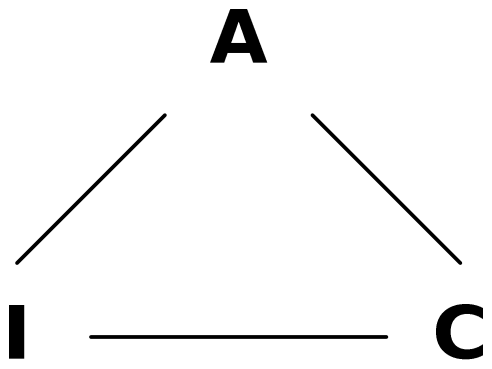
- Agents *A* are those who specify
- Agent defines
  - which Concept *C* is
  - assigned to Instance *I*

⇒ A **tripartite model** can be identified

A tripartite model

- 3 partitions: A, C & I
- Hyperedges connect exactly one $a \in A$ with one $c \in C$ and $i \in I$
- One edge denotes that a user assigns a concept to a resource.

But tripartite graphs are rather hard to understand and to work with!



Simplifying the tripartite Model

Similar to the introduced structure of folksonomies:

- An instance is connected to a concept
  - like a tag to a resource
- The edge is labeled by the user or
- Weighted by the number of assignments

A bipartite Model ...

A graph connecting

- Instances *i* to
- Concepts *c*

We call this **IC-Graph**

The weights can be expressed in an association matrix

	c1	**c2**	**c3**	**...**
i1	1	5	0	...
i2	0	3	0	...
i3	4	2	2	...
...	...	...	...	...

The Association Matrix

- This matrix connects two different sets
- Folding allows to transform the Matrix to a one mode network
- Just like the co-occurence matrix in text retrieval:

$$M_c = M_{IC} \cdot M'_{IC}$$

- Result is a matrix connecting concepts to concepts

Beispiel: Konzepte

	computer	pda	cellphone	wlan	network
i1	7	5	0	6	1
i2	7	1	1	1	2
i3	0	4	5	0	0
i4	8	0	0	0	6
i5	3	3	0	4	0

	computer	pda	cellphone	wlan	network
computer	111	62	20	62	60
pda	62	56	9	68	28
cellphone	20	9	41	0	12
wlan	62	68	0	100	24
network	60	28	12	24	34

The Association Matrix

- Also instance based co-occurrence can be calculated

$$M_I = M'_{IC} \cdot M_{IC}$$

- Based on the co-occurrence clustering algorithms can be applied:
  - Instance Clustering
  - Concept Clustering

Other Association Matrices

- Based on the AC-Graph
  - Bipartite agent2concept graph
  - Instances are used as weights
- Based on the AI-Graph
  - Bipartite agent2instance Graph
  - concepts are used as weights
- Based on A[C|I]-Graph the social network between agents can be analyzed

Application to Folksonomies

http://www.uni-klu.ac.at

- Concepts, agents and instances in Folksonomies:
  - Tags are concepts
  - Agents are users
  - Resources are instances
- Tags are error prone, but semantics can eventually emerge (see P. Mika for the example del.icio.us)

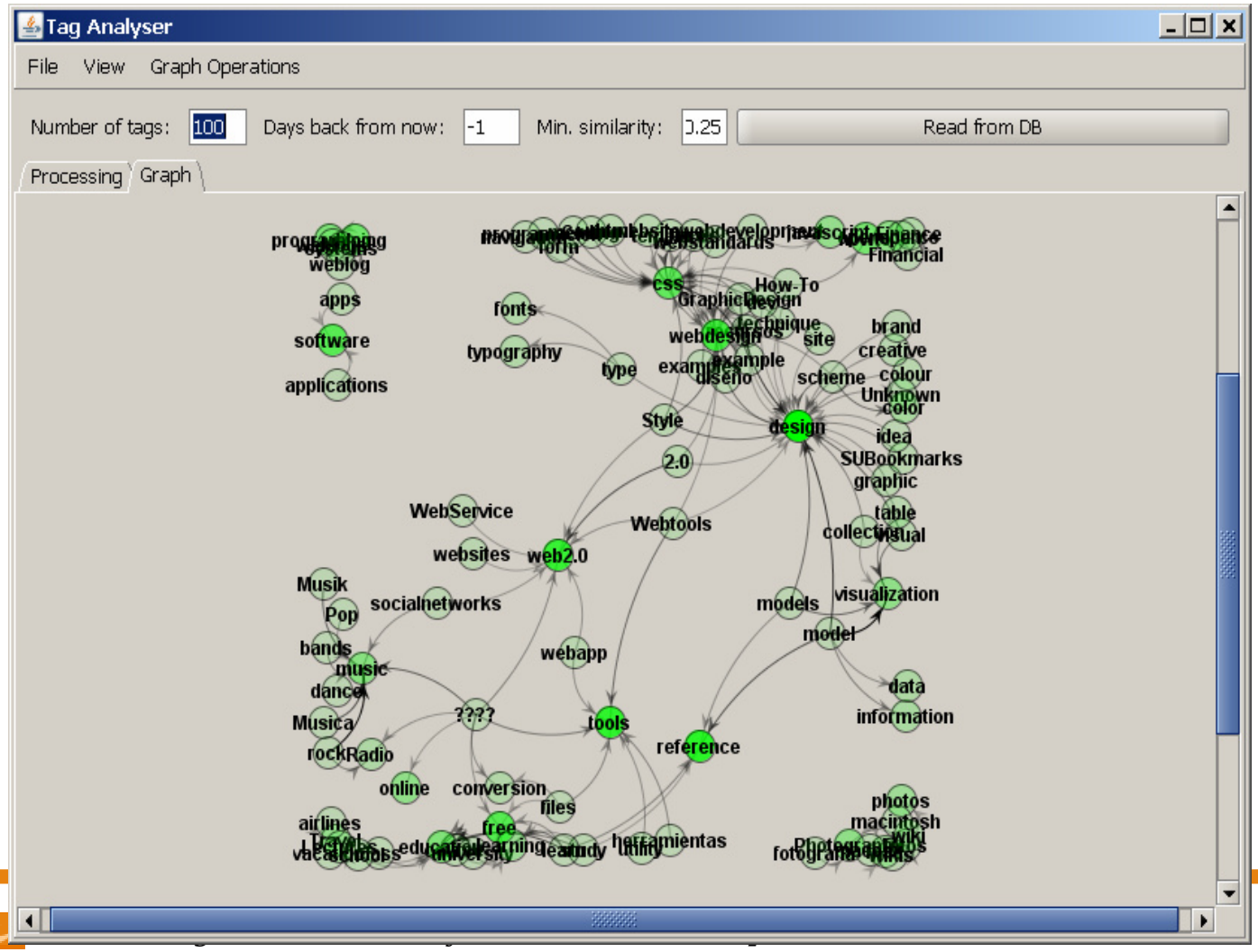
Problems of the approach

- Community based concepts & associations
- Tags have typos, synonyms
- Tags have different intentions
  - Abstract semantics (funny, sad, friendship)
  - Media description (pdf, online, word, image)
  - Rights and authors (persons names)
  - Organizational (2read, todo, marker)
  - etc.

Problems of the approach

- Computational problems
  - Big matrix multiplications are hard to compute
- Some folksonomies restrict tagging to the originating user:
  - Flickr tags can only be assigned by the uploader
  - YouTube has the same restriction

Folksonomy Analysis Example



Tag Gathering: del.icio.us

- Based on RSS feeds of del.icio.us
  - Read main feed
  - Get entries for each user
- Avoid spamming
  - Use entries on URIs with a min. of 2 users
- Write to relational database
  - In this case MySQL 5.1

Tag Database

http://www.uni-klu.ac.at

entries	
PK **PK,I1**	**usernameID** **urlID**
	title **creationdate** description

entry2tag	
PK **PK,I2,I1** **PK,I3,I1**	**usernameID** **urlID** **tagID**

tags	
PK	id
I1	**name**

urls	
PK	id
I1	**name**

usernames	
PK	id
I1	**name**

Tag database issues

- Group by & Having, Indexes
- Memory (temp tables)
- MySQL is just like Oracle:
  - tune it or leave it.
- Sample statement: Top tags:

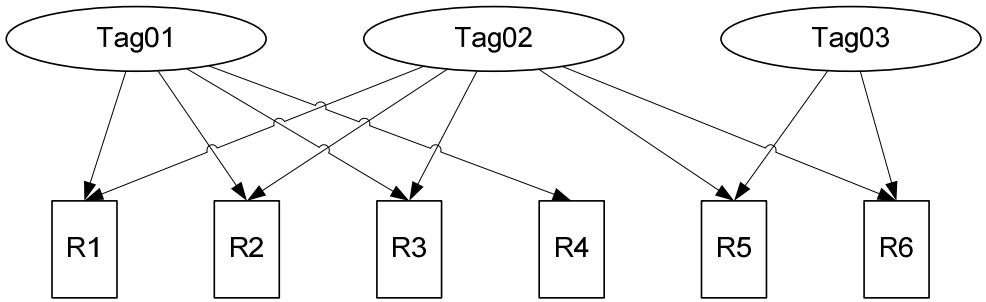
```
SELECT COUNT(e.tagid), t.name, t.id FROM
  entry2tag e, tags t WHERE t.id = e.tagid
  GROUP BY e.tagid ORDER BY 1 DESC
```

Tag similarity

- Tags are assigned to resources
- Tags describe same URIs-> Similarity
  - E.g. Javascript & Ajax
  - E.g. Windows & Software
  - E.g. Linux & Kernel
- Tags never describe same URIs-> Dissimilarity
  - E.g. Free & Shop
  - E.g. Usability & SAP

Tag similarity

- Mathematical background
  - Co-occurence Matrix C = M*M^t
  - M has resources in cols & tags in rows
  - Values are the number of assignments



	R1	R2	R3	R4	R5	R6
Tag01	1	1	1	1	0	0
Tag02	1	1	1	0	1	1
Tag03	0	0	0	0	1	1

Tag Merging: Objectives

- Main problems within del.icio.us (and possibly in many folksonomies due to their nature)
  - Synonyms
  - Basic level variation
- Encounter these problems by “merging” synonyms
  - Different spellings: e.g. eLearning & e-Learning
  - Typos & plurals

Tag Networks: Objectives

- What is the conceptual structure within a community?
- Which tags are similar / interconnected?
- Direction of the connection?
- Probability of transition for network edges?
- Network Analysis?
  - Hubs, central authorities
  - Clusters

Tag Centrality: Objectives

- Which are the most prominent nodes?
- Based on different measures?
  - In degree
  - In Betweenness
  - PageRank / HITS
- The removal of central nodes would hit the connectivity hard!

Tag Clustering: Objectives

- What are interesting conceptual clusters?
  - {design, webdesign, graphics}
  - {html, xhtml, css}
  - {ajax, javascript, prototype, script.aculo.us}
- What is a meaningful disambiguation of a topic / tag?

Clusters of tag programming

1. systems+unix (3,42)
2. developer+development (2,49)
3. webdevelopment+javascript+webdev (2,34)
4. ebook+books+book (2,19)
5. Coding (2,19)
6. programacao+ruby (2,14)
7. script+ajax (1,78)
8. DotNet+.NET (1,65)

Demo

- See TagAnalyzer / Folksonomist

Conclusions

- There is a whole lot of things one can do with tags
  - We can find basic patterns out of a mass of tags
  - We can visualized co-occurrence & networks
  - We can merge tags based on the network
  - We can cluster & disambiguate tags
- There is ongoing research in *Network Analysis* & *Social Software*

Conclusions

- We also encounter problems
  - Size of the database
  - Runtime of clustering algorithms
  - Matrix operations

Thanks ..

http://www.uni-klu.ac.at

... for you attention!

Contact me, I'm here for today :-)

mlux@itec.uni-klu.ac.at
http://www.itec.uni-klu.ac.at/~mlux