Web 2.0:
User Generated Metadata

Mathias Lux
Klagenfurt University, Austria
Dept. for Information Technology
mlux@itec.uni-klu.ac.at

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.0 License. See http://creativecommons.org/licenses/by-nc-sa/2.0/at/

Agenda

- Introduction
- Examples for User Generated Metadata
- What is User Generated Metadata?
- Folksonomies
- Emergent Semantics
- Summary

Introduction

- Why should we care about metadata?
  - Why don't just analyze the data?
- Why should we care about users generating something?
  - Isn't it always error prone?
  - Isn't it always unreliable?

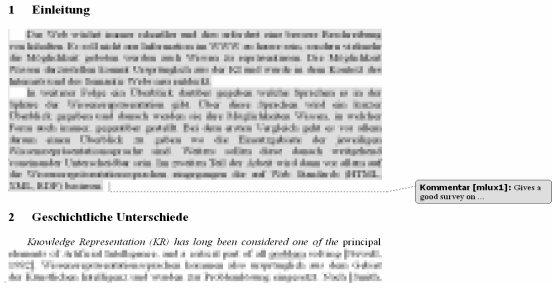
Why metadata?

- Understanding of data is a key issue
- Interpretation often not possible
  - e.g. Abstract Semantics in Images like „Friendship"
  - Cannot „infere" metadata

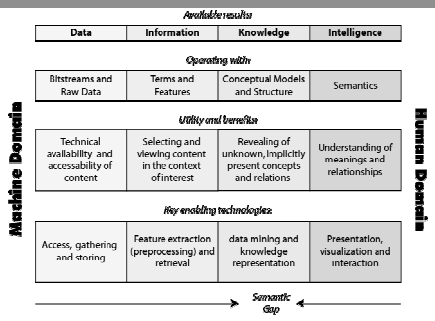
Why metadata?

- Compare to Information Search & Retrieval:
  - e.g. User need: some „good survey or state of the art"
  - Often available in „related work" sections of „related" articles
  - Might be possible with metadata.

Example for Annotations



Semantic Gap



Why should we care about users generating something?

- Users are the driving force
  - Consumers/Sellers in e-business
  - Reader & Ad-Clickers on Content Portals
- So we've got to give them „something"
  - Give them what they need (see Google Mail)
  - Give them what they like (see YouTube)
- And take a added value for us
  - e.g. money (in case of ads) or generated content (in case of mails)

Isn't it always error-prone?

- We are not talking about
  - Life Support Systems
  - Ontological Knowledge
- We have to „live" with errors, e.g. in
  - Information Retrieval: Incomplete sources and queries
  - Emails: Spams and Phishing Mails

Isn't it unreliable?

- In many cases UGM (and UGC) is unreliable.
- But there can be ways to create added value with a **certain** quality level
  - Compare to Internet Search Engines
  - Compare to Data Mining issues (Outliers, Data Cleaning, etc.)

So what can UGM do for us?

- Provide additional data & structure
- Add abstract (not obvious) annotations
- At certain quality levels
  - Depending on domain and application

Examples for UGM

In Web 2.0 lots of examples exist ...

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

Obvious and Hidden

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

- Some approaches are very obvious

Obvious and Hidden

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

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

Examples: Social Bookmarking

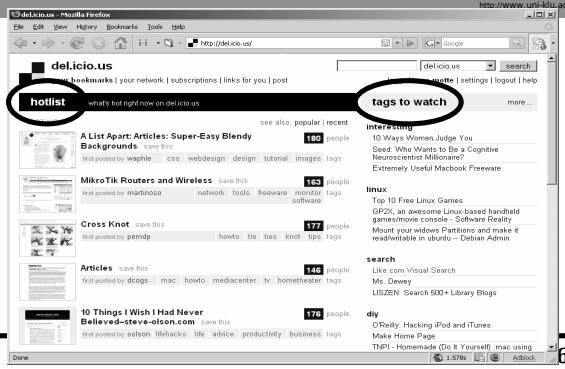
http://www.uni-klu.ac.at

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)

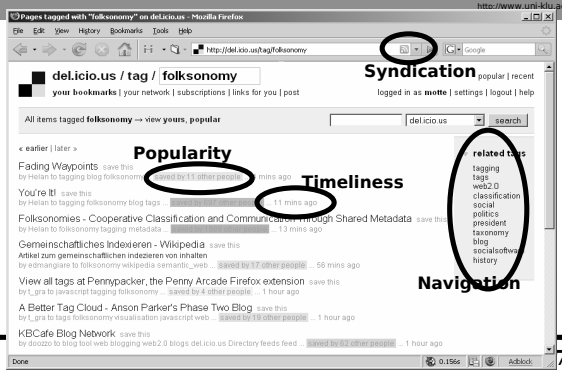
Example: del.icio.us

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



Example: del.icio.us

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



Example: del.icio.us

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

- 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 Metadata
  - Bookmarking & Tagging
  - Tag Bundles
- Implicitly Generated Metadata
  - 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 a part of a web page ...
  - Done using Diigo

Demo: Diigo

- Video or
- Live Demo ...

Example: Diigo

- User intentions are unclear:
  - For himself (later reading / work)
  - For a group of people / coworkers
- Explicitly Generated Metadata
  - Highlighting & bookmarking
  - Tagging & Description, Sticky Note
- Implicitly Generated Metadata
  - 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 Metadata
  - Ratings, Flags, Tags, Spam Filtering
- Implicitly Generated Metadata
  - 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)

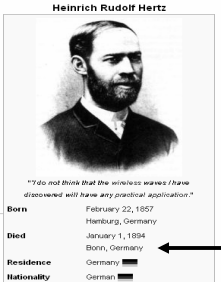
Example: Wikipedia: Heinrich Rudolf Hertz

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 ]]], [[1894]]
|death_place = [[Bonn, Germany]]
...
```


Example: Wikipedia: Heinrich Rudolf Hertz

- 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

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

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

Example: Blogs

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

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

What is User Generated Metadata in Web 2.0?

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

- 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 ...

Summary: User Generated Metadata

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

- We have seen examples on UGM
- We have found some characteristics of UGM

Next:

- Getting more specific on folksonomies

Folksonomies

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

Network of Tags, Users and URLs

- Users describe resources
- By using (multiple) tags

Examples:

- Social bookmarking, media sharing, etc.

Folksonomies: The Structure

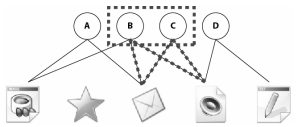
http://www.uni-klu.ac.at

User *tags* resource (URL)

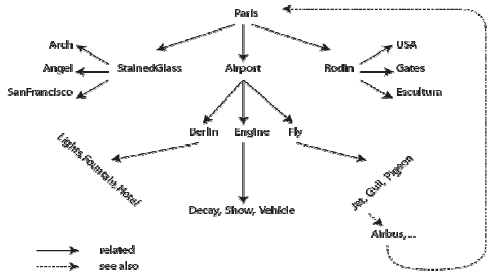
- 1+ words or phrases (bonn, „mathias lux“)
- No controlled vocabulary, taxonomy
- No quality control
- No constraints (language, length, number)

Folksonomies: Structure

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

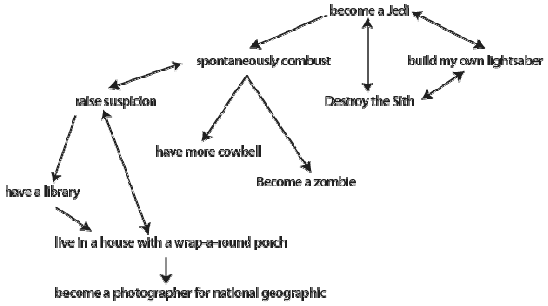


Folksonomy Example: Flickr



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

Summary: Folksonomies

- We have defined folksonomies
- We have talked about why they work.

Next:

- Emergent Semantics

Emergent Computation

Stephanie Forrester defines it in 1990:

- In context of parallel computing
- Global behavior emerges from local interactions
  - Mostly in systems without central authority
- It is called emergent computation if
  - the local interactions are computations
  - the global behavior is a computation

Emergent Semantics *(Steffen Staab)*

Emergent computation describes

- Creation of complex systems / organisms
- Without explicit purpose of creation

Emergent Semantics

- Arise from user interactions
- Without explicitly aiming at semantics

Emergent Semantics *(Steffen Staab)*

Semantics emerge through user interaction allow ...

- ... to cope with the problem of knowledge acquisition
- ... aggregation of semantics without additional efforts of the users
- ... the creation of a new basis for understanding (bottom up)

Emergent Semantics *(Simone Santini)*

Simone Santini – Visual Image Retrieval

- What is the meaning of an image?
  - Derived from interactions and relations
    - Behavior, queries, browsing, ...
  - Semantics of features should be formalized
    - Meaning of low level features like color, file size, ...
  - Relative to the social context of the user
    - Different people see different things in images

Emergent Semantics *(Simone Santini)*

Simone Santini – Visual Image Retrieval

- Semantics should be specified by the user
  - Selection through appropriate interfaces
    - Telling what is similar
    - Giving examples
    - Specifying constraints
  - Adaptation to user needs
    - relevance feedback
    - retrieval of metrics

Emergent Semantics *(Luc Steel)*

What are the ways to create semantics?

- Inductive
  - Starting from a reference set, “good example”
  - Learning from this reference set
  - Applying learned things by classifying, clustering ...
- Selective
  - Start with randomly generated data
  - Select the appropriate ()

Emergent Semantics *(Luc Steel)*

- Constructive
  - Example: Moving in a new home, organizing shoes ...
    - Inductive is not possible
    - Selective would be “a mess”
  - Includes cooperation & communication
- Emergent Semantics
  - Include cooperation & communication
  - Oppose the methods of induction & selection

A Unified Model for Emergent 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

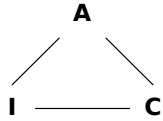
- Agents *A* are those who specify
- Agent defines
  - which Concept *C* is
  - assigned to Instance *I*

⇒ 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

- 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

Summary: Emergent Semantics

- We have seen what emergent semantics are
- We have learned about a model for semantics & social networks
- We have learned about a method for identifying emergent semantics

What can UGM do for us?

- Bridge the Semantic Gap?
  - (Multimedia) Information Retrieval
- Provide some additional input
- Do “Human Computing”
- Give us some headache
  - User rights
  - Mass of content & links

Outlook

- User Rights, DRM
  - Discussed on content but not on metadata
- Aggregation over different service providers
  - Standard formats & processes
- Openness & Accessibility
  - Sharing, licenses, availability, ...

Privacy issues

- Data is aggregated about users
- Eventual violations of privacy

- Something to think about ...
  - Video of the American Civil Liberty Union
  - http://www.aclu.org/pizza/

CfP: Workshop Multimedia Semantics - The Role of Metadata

The Web 2.0 & The Multimedia Web

Topics include: Multimedia Semantics, [...] Emergent Semantics, [...] Web 2.0

see http://www.multimedia-metadata.info/Events/WMS07

Thank you ...

... for your Attention!

Mathias Lux
mlux@itec.uni-klu.ac.at
Klagenfurt University
Austria

Armin Ulbrich
aulbrich@know-center.at
Know-Center Graz
Austria

References

Forrest, S. *(ed.)* Emergent Computation: Self-Organizing, Collective, and Cooperative Phenomena in Natural and Artificial Computing Networks *Emergent Computation, Proceedings of the Ninth Annual CLNS Conference, MIT Press,* **1990**, 1-11

Staab, S. Emergent Semantics *IEEE Intelligent Systems, IEEE Educational Activities Department,* **2002**, *17*, 78-86

Santini, S.; Gupta, A. & Jain, R. Emergent Semantics through Interaction in Image Databases IEEE Transactions on Knowledge and Data Engineering, IEEE Educational Activities Department, 2001, 13, 337-351

Golder, S.A. & Huberman, B.A. The Structure of Collaborative Tagging Systems Journal of Information Science, 2006, 32, 2

Mika, P. Ontologies Are Us: A Unified Model of Social Networks and Semantics International Semantic Web Conference, Springer, 2005, 522-536