Workshop on
Wikipedia Research

Jakob Voss

WikiSym 2006, Odense, Denmark
23 August 2006

Workshop on

Wikipedia Research

[[/Intro]]

{{ **Definition** }}

* WikiPedia
* WikiMedia
* MediaWiki

{{ About [[user:nichtich|me]] }}

* Active in German Wikipedia since 2002
* Founding board member
  of Wikimedia Germany
* M.A. Computer Science
  and Library & Informtion Science
* System librarian in Göttingen

> I bring order out of chaos, I shine light into the dark
> because power comes from knowledge just like fire from a spark
> and like Gutenberg and Luther with press and pen in hand
> I take the message to the masses in a form they understand
>
> Jonathan Rundman: *I'm a librarian*

{{ **Agenda** }}

* [[**/Intro**]]
  * You are here
* [[**/Topics**]] & Results
  * What?
* [[**/Methods**]]
  * How? Where? Who?

{{ **Wikipedia Research?** }}

* New discipline
  * Wikipedia is a singular phenomenon
  * Very heterogenious
* General meta-questions
  * Research vs. developement?
  * Inside or outside Wikipedia?
  * What science?

{{ Wikipedia Research? }}

“Understanding and fostering the growth of the World Wide Web, both in engineering and societal terms, will require the development of a new interdisciplinary field.”

Berners-Lee, Hall, Hendler, Shadbolt, Weitzner:
Creating a Science of the Web.
In: *Science* 11 Aug 2006. *v* 313, n 5788, p 769-771
http://dx.doi.org/10.1126/science.1126902

Workshop on
Wikipedia Research

Jakob Voss

WikiSym 2006, Odense, Denmark
23 August 2006

Workshop on

Wikipedia Research

[[/Topics]]

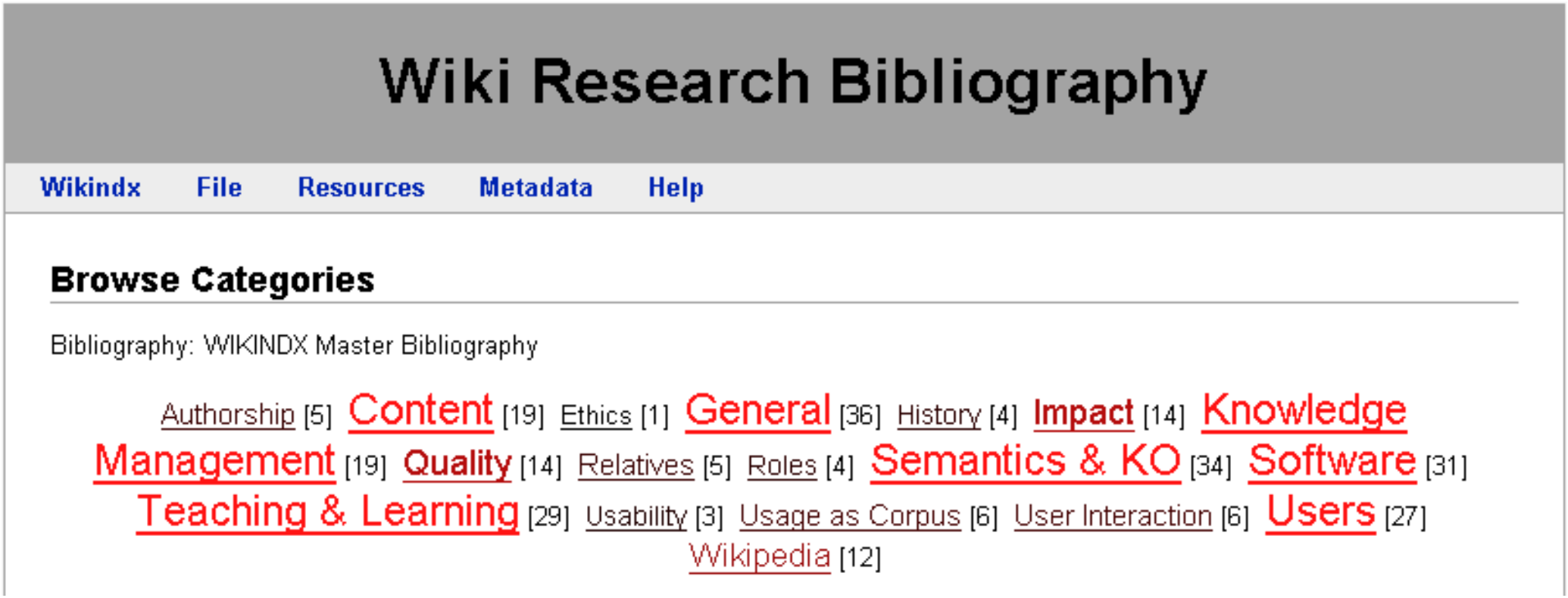
(&Results)

{{ The real object of science }}

Please apologize if I haven't mentioned your publication!

Tell me if you paper is missing in the bibliography!

{{ Infrastructure : Bibliography }}



* http://bibliography.wikimedia.de/

* 230 resources (also about wikis in general)

Jakob Voss : **Workshop on Wikipedia Research.** *WikiSym 2006-08-23*, Odense, Denmark

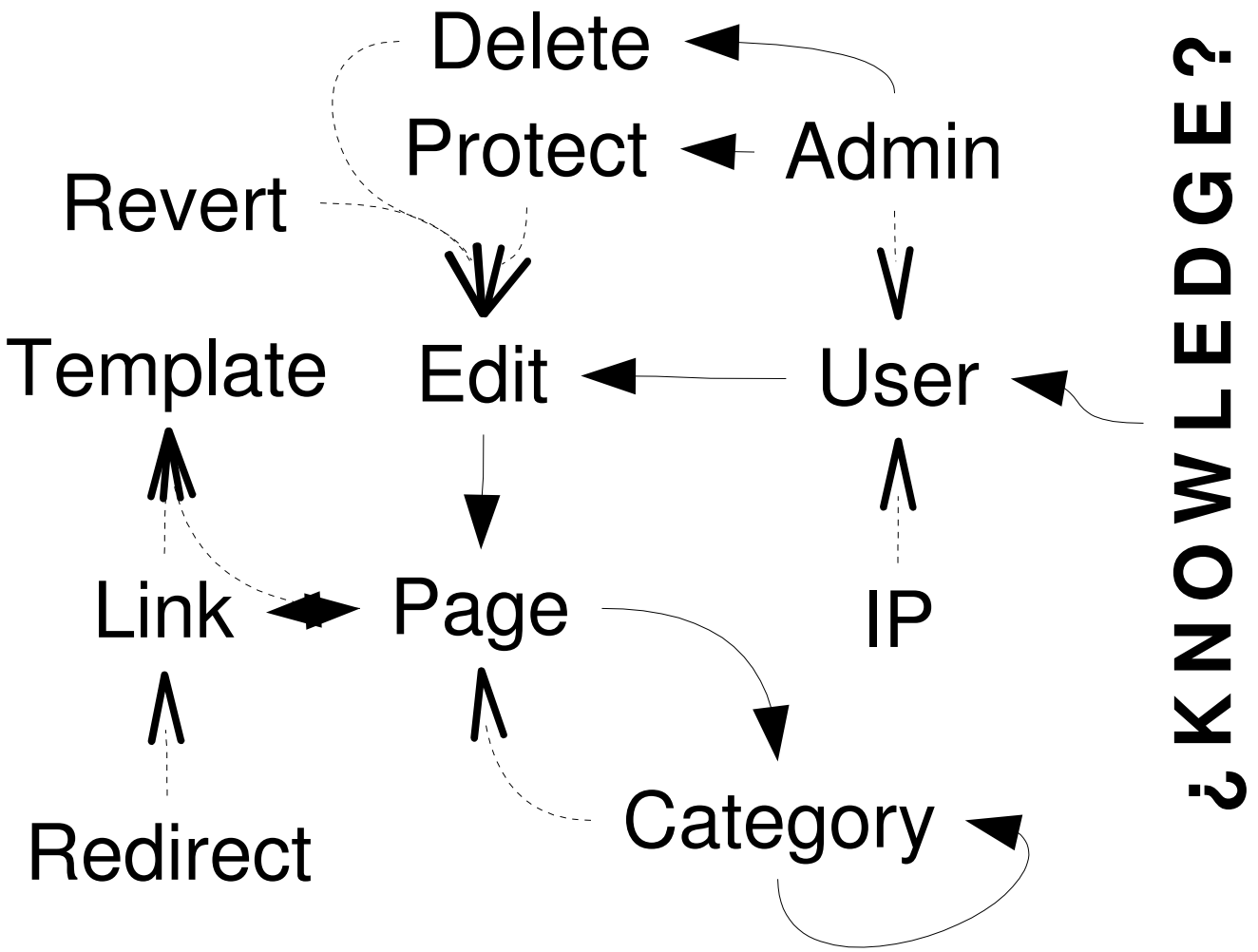
{{ **Topics** }}

* Content*
  * Structure, Semantics, Quality...
* Users
  * Motivation, Roles, Interaction...
* Impact
  * Society, Recipation...

{{ **Content** }}

* Structure
  * Network analysis, Visualization...
* Semantics and Knowledge Organization
  * Categories, Semantic Wikis...
* Quality
* Other usage (as corpus)

{{ Content : Structure }}



{{ Content : Structure }}

* Network analysis
  * Zlatic et al. 2006: *Wikipedia as complex networks.*
  * Copacci et al 2006: *Preferential attachment in the growth of social networks: the case of Wikipedia.*
* Visualization
  * Hollaway et al 2005: *Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors.*
  * Viégas, Wattenberg, Dave 2004: *Studying Cooperation and Conflict between Authors with History Flow Visualizations*

{{ Semantics & KO }}

* **Semantics:**
  computers understand what people meant
* **Knowledge Organization:**
  structure the creation of meaning

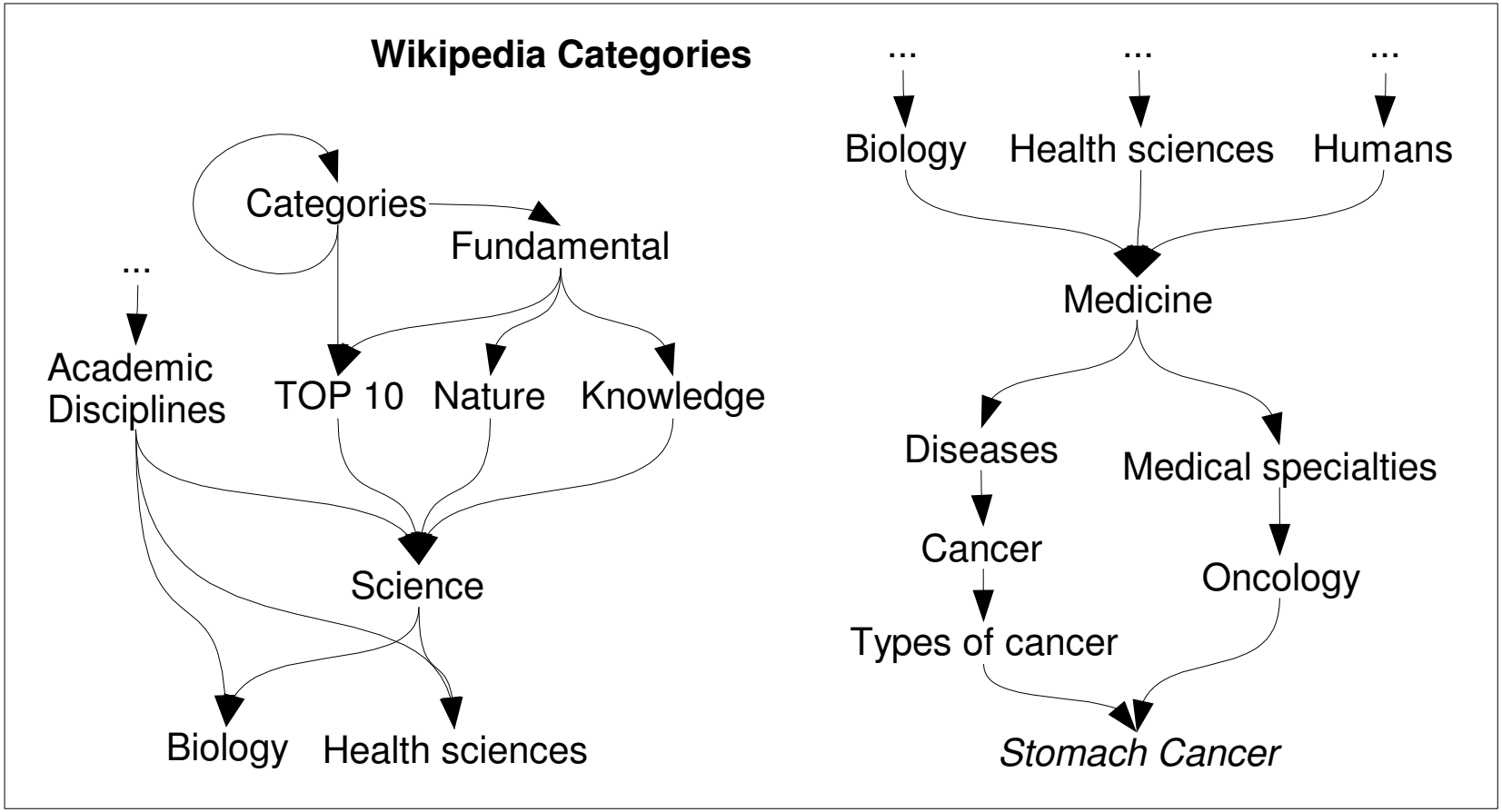
{{ Semantics & KO }}

* **Semantics:**
  computers understand what people meant
* **Knowledge Organization:**
  structure the creation of meaning
* **Knowledge Management:**
  people understand what people meant
* **Teaching & Learning:**
  people understand the creation of meaning

{{ Semantics & KO }}

* **Semantics:**
  computers understand what people meant
* **Knowledge Organization:**
  structure the creation of meaning
* **Knowledge Management:**
  people understand what people meant
* **Teaching & Learning:**
  people understand the creation of meaning
* **The fable of Strong AI:**
  computers understand the creation of meaning

{{ Semantics : Categories }}



[http://arxiv.org/cs/0604036 **Collaborative thesaurus tagging the Wikipedia way**]

{{ Semantics : Categories }}

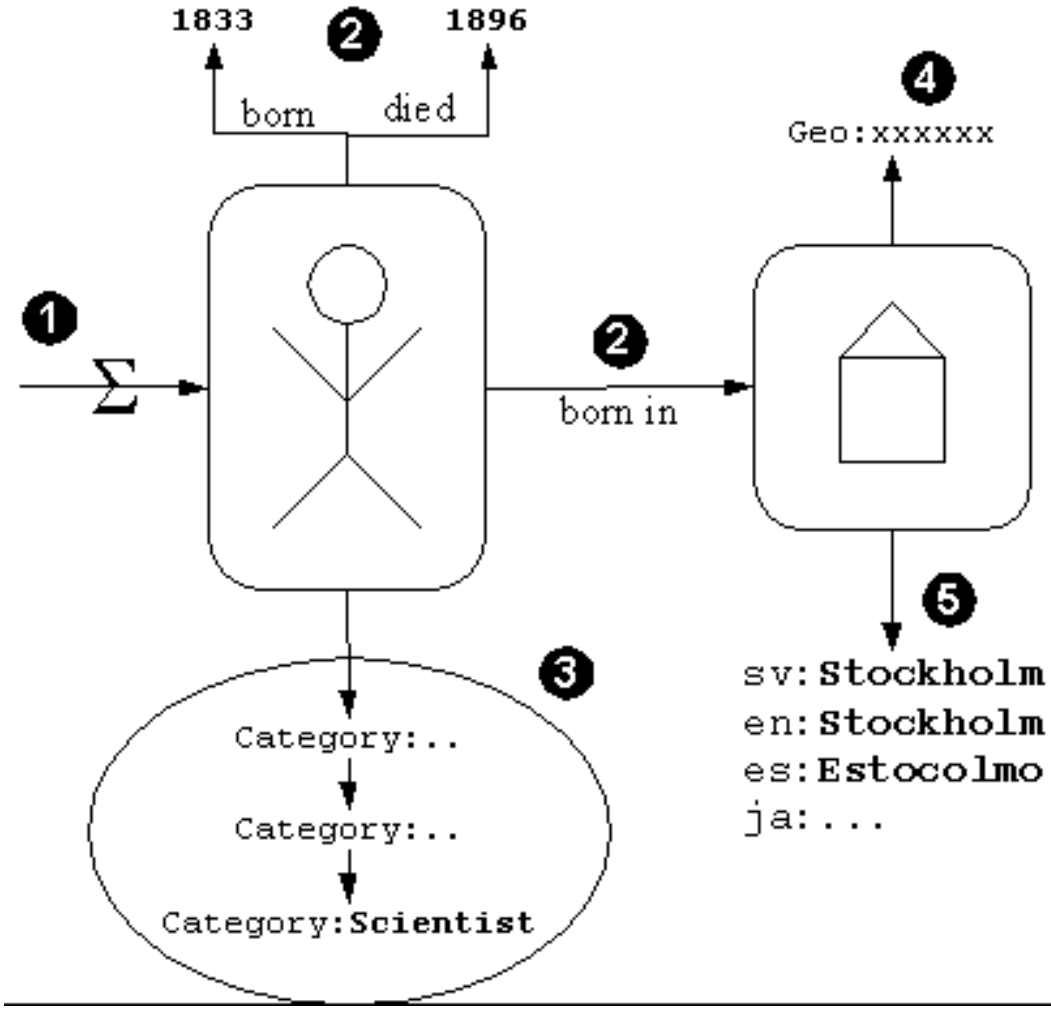
* Chernov et al. 2006:
  Extracting Semantics Relationships between Wikipedia Categories.
* Holloway et al. 2006:
  Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors.
* Voss 2006:
  Collaborative thesaurus tagging the Wikipedia way.

{{ **Semantics** }}

* Krötzsch, Vrandecic & Völkel 2005: *Wikipedia and the Semantic Web – The missing Links.*
* Voss 2005: *Metadata with Personendaten and beyond.*
* Ruiz-Casado et al. 2005: *Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets.*
* Hepp et al. 2006: *Harvesting Wiki Consensus: Using Wikipedia Entries as Ontology Elements.*
* Strube & Ponzetto 2006: *WikiRelate! Computing Semantic Relatedness Using Wikipedia.*
* *...*

{{ Semantics : Extract }}

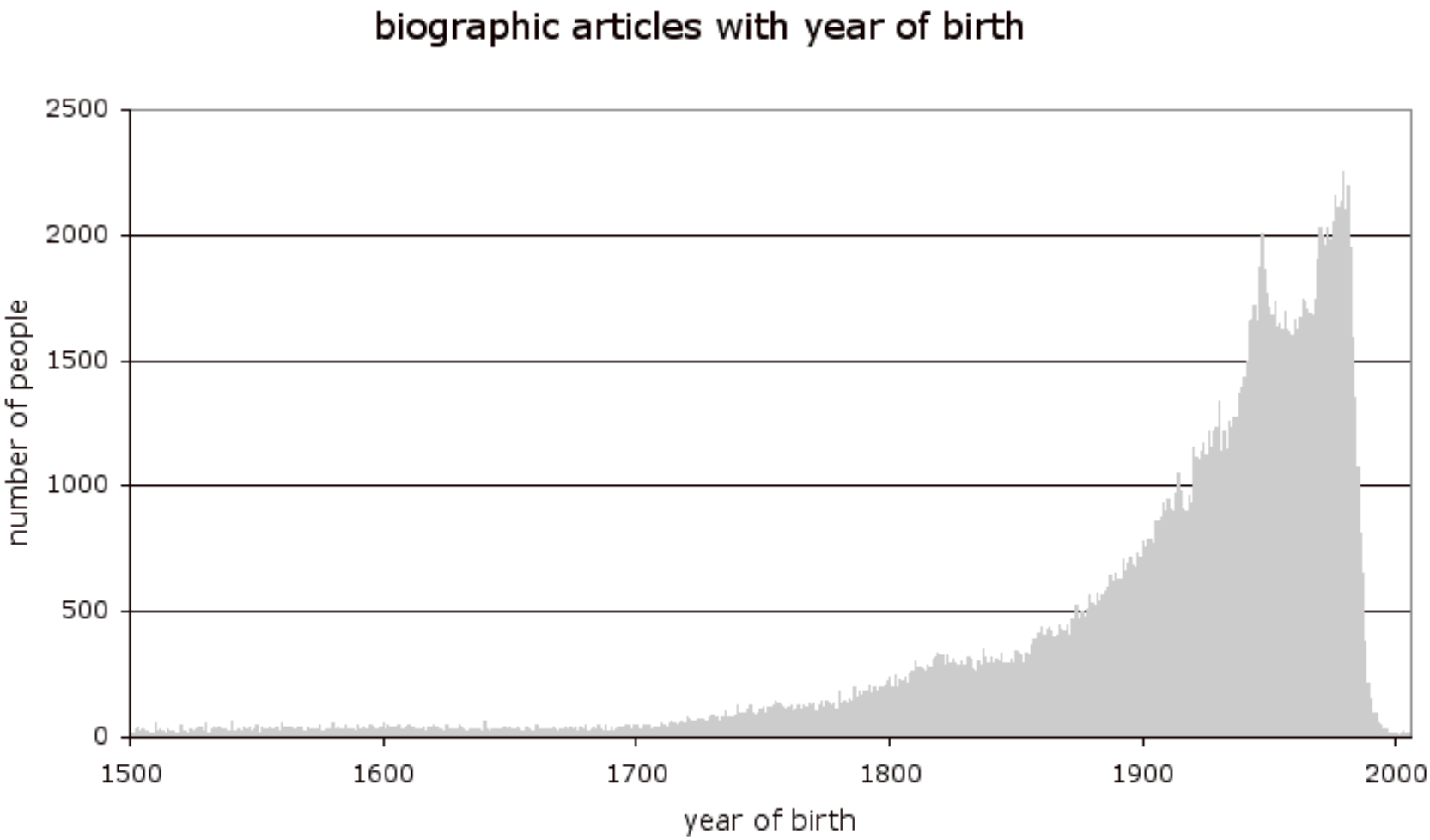
„Famous 19th century scientists of this town?“



Semantic Information in Wikipedia

1. *famous*: sum of **inlinks**
2. *19th century*: **personendaten** ~ 1780 < born < 1880
3. *scientist*: **categories**
4. *this town*: GEO...
4. *town*: **Interwiki**-links

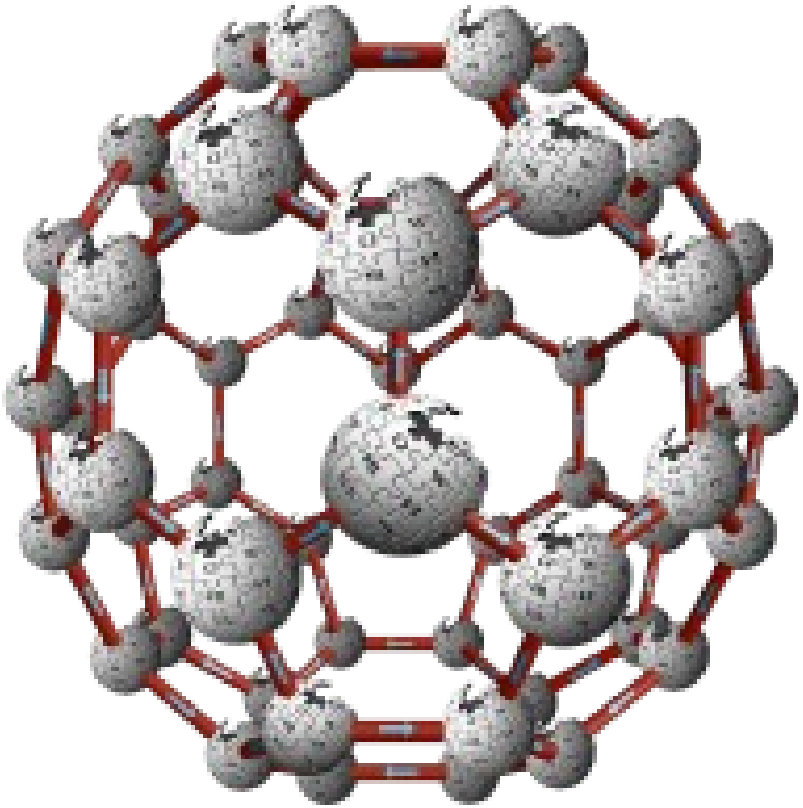
{{ Semantics : Extract }}



[http://en.wikipedia.org/wiki/User:Dsp13 **Biographical coverage in Wikipedia**]

Jakob Voss : **Workshop on Wikipedia Research.** *WikiSym 2006-08-23*, Odense, Denmark

{{ Semantics : Build }}



Relations
[[located in::Fyn]]

Attributes
[[has population:=186 595]]

it's [[Odense]]

[http://wiki.ontoworld.org/wiki/Semantic_MediaWiki]

{{ Quality }}

* „*Good Wikipedia, Bad Wikipedia*“

* What **is** Quality, Neutrality, Usefulness...?

{{ **Quality** }}

* Lots of articles in the press & blogosphere
  * Seigenthaler '05, House of Representatives '06, etc.
  * http://en.wikipedia.org/wiki/Criticism_of_Wikipedia
* Study of Science Journal Nature
  * Experts reviewed 42 selected articles of Encyclopædia Britannica and Wikipedia
  * EB had 123 errors (average 2.9)
  * Wikipedia had 162 errors (averages 3.9)
* Better *studies* on information quality???

{{ **Quality** }}

* Sanger 2004:
  Why Wikipedia Must Jettison Its Anti-Elitism.
* Brändle 2005:
  Zu wenige Köche verderben den Brei.
* Anthony et al. 2005:
  Explaining Quality in Internet Collective Goods.
* Stvilia et al. 2005:
  Information Quality: Discussions in Wikipedia.
* Reagle 2005: *Is the Wikipedia Neutral?*
* Lanier 2006: *Digital Maoism:*
  The Hazards of the New Online Collectivism.

{{ Content : Usage as corpus }}

* INEX Test corpus for XML Retrieval
  http://inex.is.informatik.uni-duisburg.de/2006/
* TREC-13, 2004
  http://staff.science.uva.nl/~gilad/pubs/uams-trec-2004-final-qa.pdf
* WiQA at CLEF 2006
  http://ilps.science.uva.nl/WiQA/
* Human Knowledge Compression Contest
  http://prize.hutter1.net/

linguistic (in contrast to logic) approach to AI

{{ **Users** }}

* Authorship
* Roles
  * Powerusers, Fluctuation
* Interaction, Roles, Motivation...

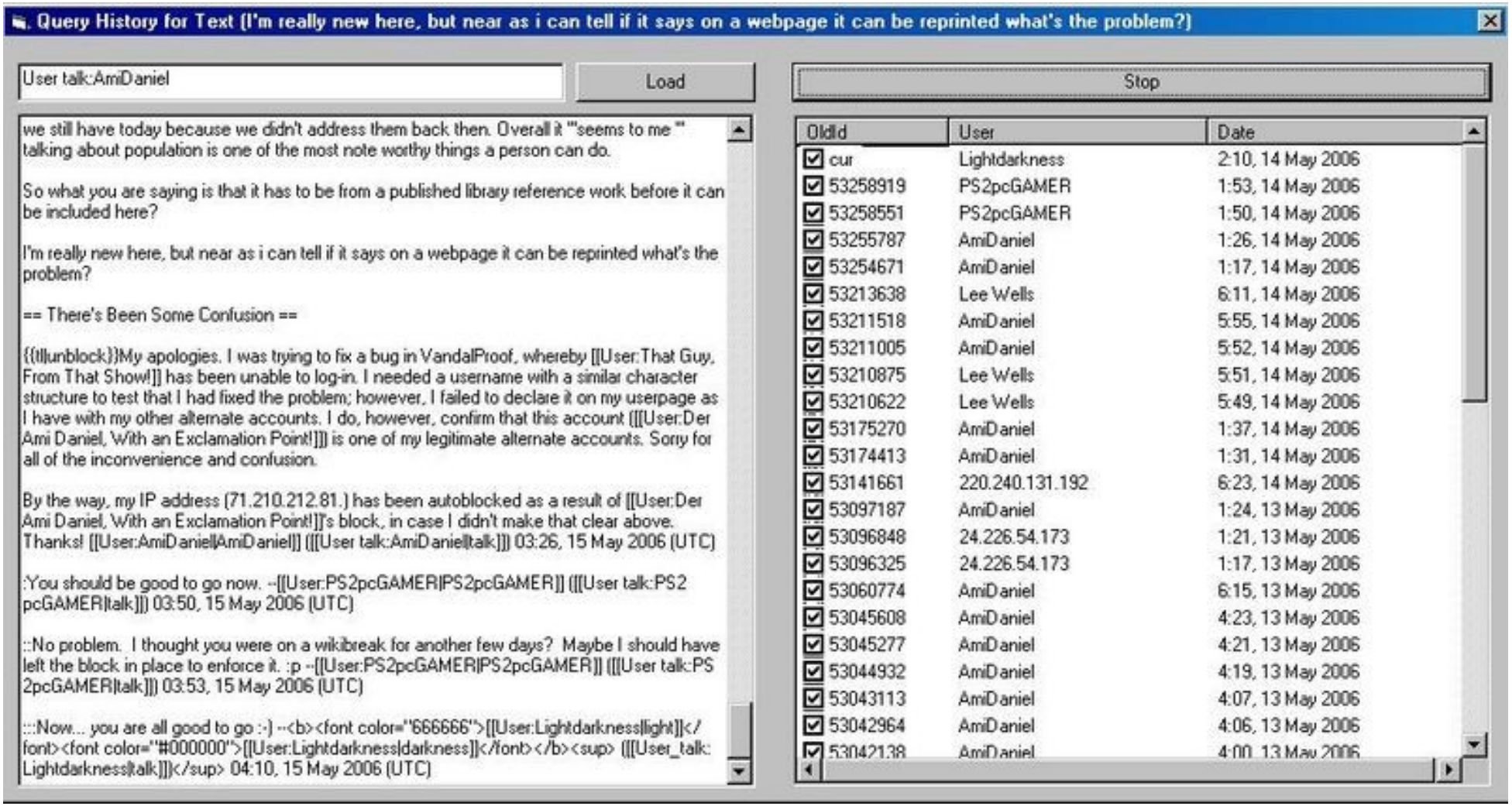
This is social science – Please help me!

{{ Users : Authorship }}

* Default diff function
  * Not very intelligent (newlines, moves etc.)
* Main authors:
  * http://de.wikipedia.org/wiki/Wikipedia:Hauptautoren/S
* Better diff feature:
  * http://217.147.83.36:9000/history::171=169

* ***What is*** authorship in Wikipedia?

{{ Authorship : WhodunitQuery }}



[http://en.wikipedia.org/wiki/User:AmiDaniel/WhodunitQuery **WhodunitQuery**]

Jakob Voss : **Workshop on Wikipedia Research.** *WikiSym 2006-08-23*, Odense, Denmark

{{ Powerusers }}

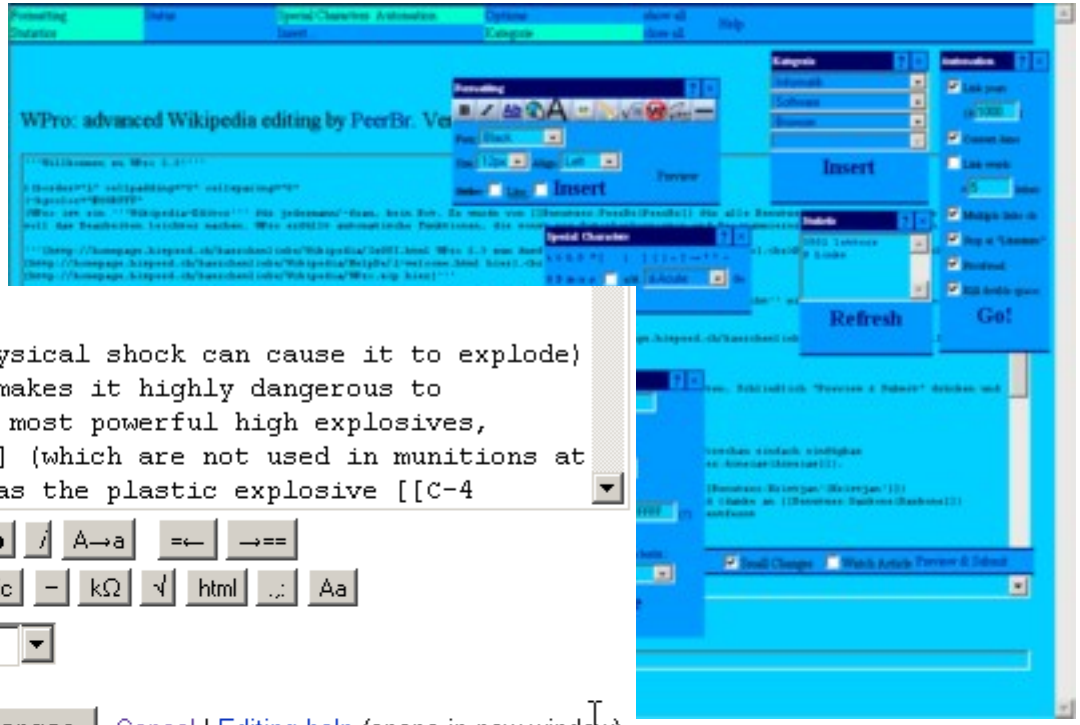
* Hardcore Wikipedians / Core community
  * Definition ?
  * Elite?
  * Influence?
  * Addiction?
* may have other working environments, for instance **user scripts**
  * [http://en.wikipedia.org/wiki/Wikipedia:Scripts]
* may have other focus
  for instance arbitration, VfD, Wikimedia...

{{ Powerusers : Definition }}

* Above average activity
  * What average?
  * What activity?
* Straight Power-law (Lotka's law)
  * <10% accounts do >90% edits
  * >50% accounts do < 10 edits

{{ Powerusers : Editors }}

[http://de.wikipedia.org/wiki/Benutzer:PeerBr/WPro **WPro**]




```
== Instability and desensitization ==
In its pure form, it is a [[contact explosive]] (i.e., physical shock can cause it to explode)
and degrades over time to even more unstable forms. This makes it highly dangerous to
transport or use. In this undiluted form it is one of the most powerful high explosives,
comparable to the military explosives [[RDX]] and [[PETN]] (which are not used in munitions at
full concentration because of their sensitivity) as well as the plastic explosive [[C-4
```

Get | ←Find | (\w*)glycer(ine?|ol) | Find→ | ↑↓ | ← | → | Undo all | b | i | A→a | =← | →==

All | ←Repl. | $1glycerol | Repl.→ | Case | Regexp | Fix: | Basic | − | kΩ | √ | html | .:. | Aa

Edit summary: Copyedit

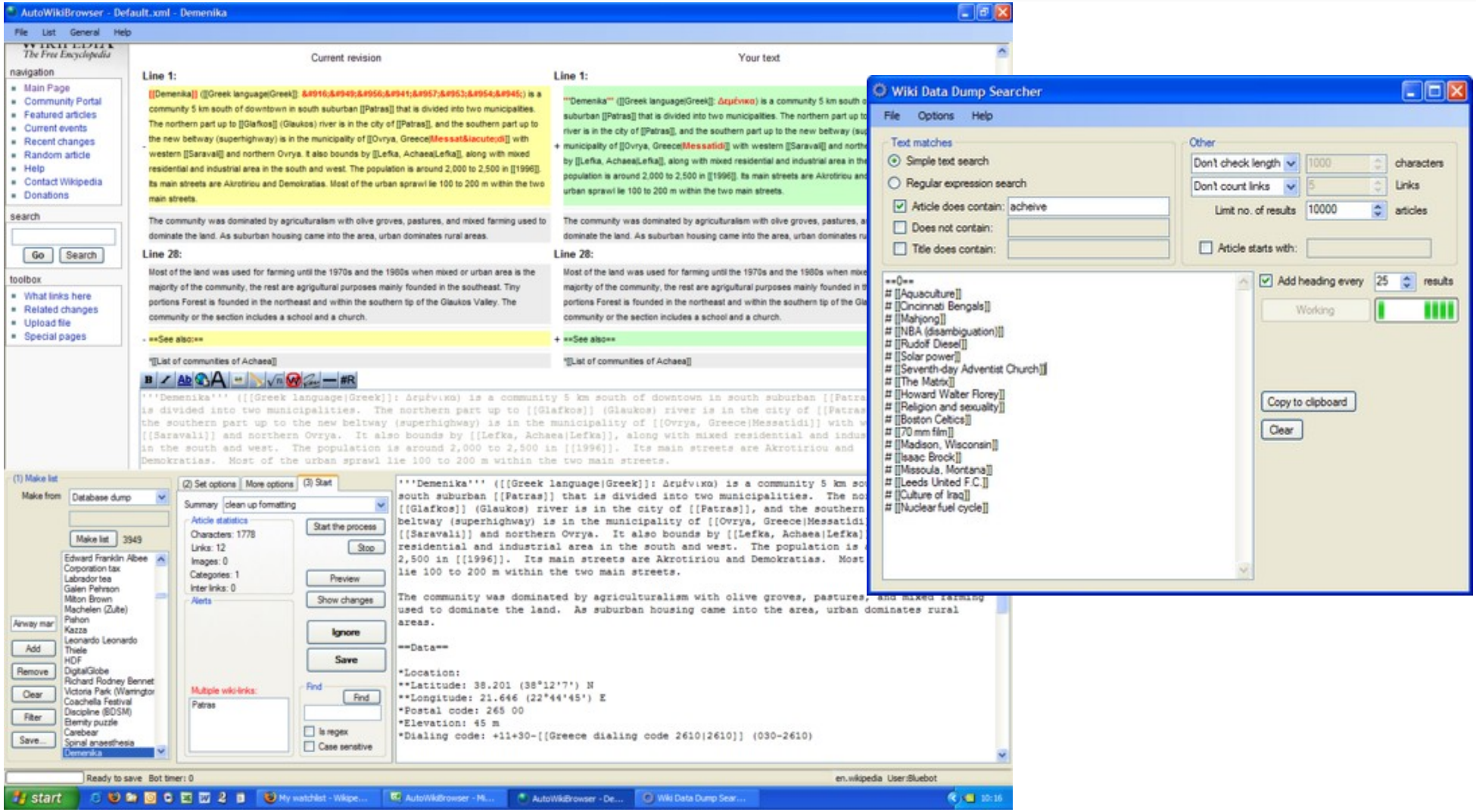
This is a minor edit (?) Watch this page

Save page | Instant: Preview | Changes | Clear | Server: Preview | Changes | Cancel | Editing help (opens in new window)

'''Nitroglycerin''', also known as '''nitroglycerine''', '''trinitroglycerin''', and '''glyceryl trinitrate''', is a [[chemical compound]]. It is an extremely dangerous heavy, colorless, poisonous, oily, explosive liquid obtained by [[nitration | nitrating]] [[glycerol]]. ›
Nitroglycerin is used in the manufacture of [[explosive]]s, specifically [[dynamite]], and as such is employed in the [[construction]] and [[demolition]] industries. It is also used medically as a [[vasodilator]] to treat [[heart]] conditions. It is colored yellow when it is decomposing due to acidic [[pH]].

[http://en.wikipedia.org/wiki/User:Cacycle/editor **WikEd**]

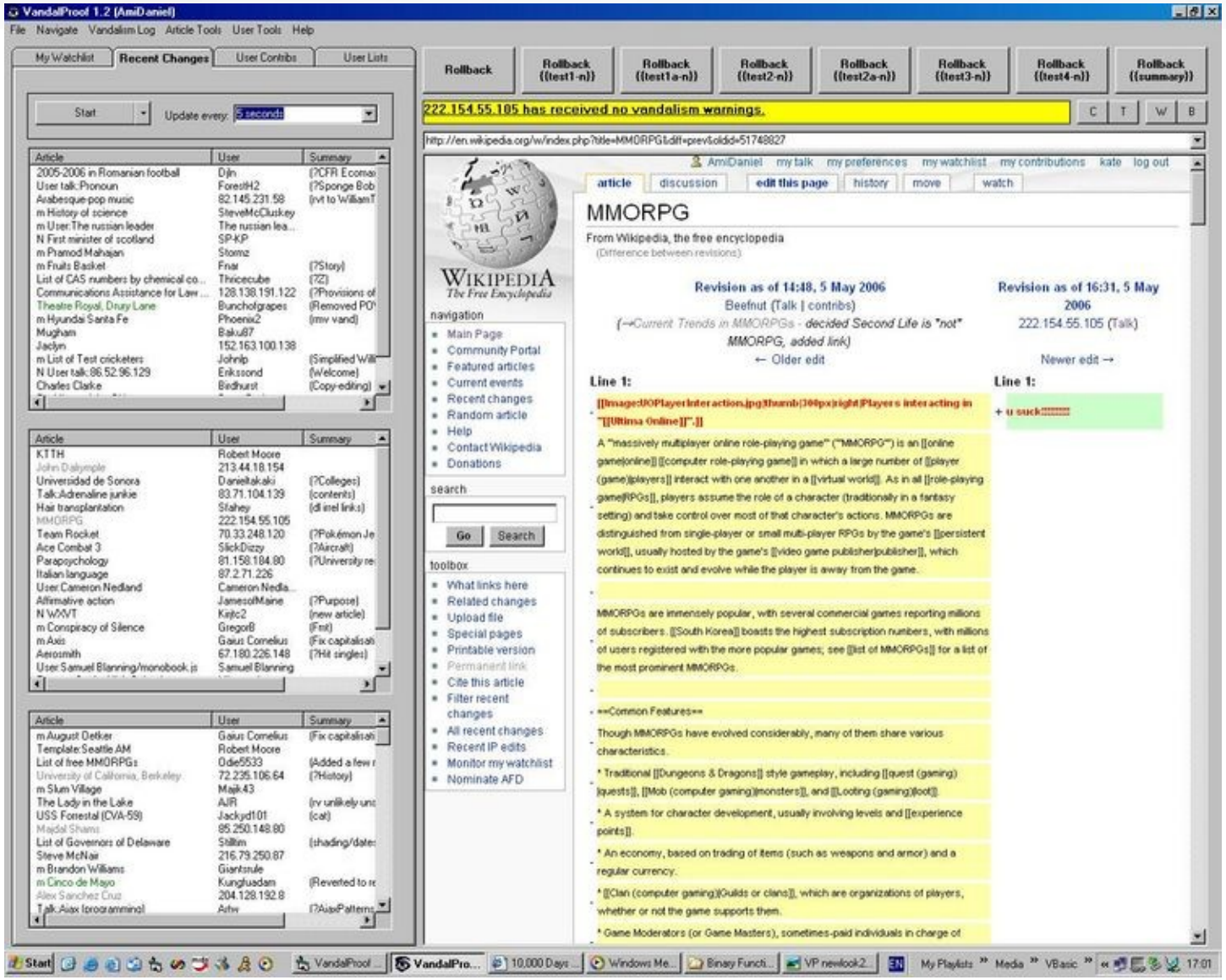
{{ Powerusers : AWB }}



[http://en.wikipedia.org/wiki/Wikipedia:AutoWikiBrowser **AutoWikiBrowser**]

Jakob Voss : **Workshop on Wikipedia Research.** *WikiSym 2006-08-23*, Odense, Denmark

{{ Powerusers : VandalProof }}

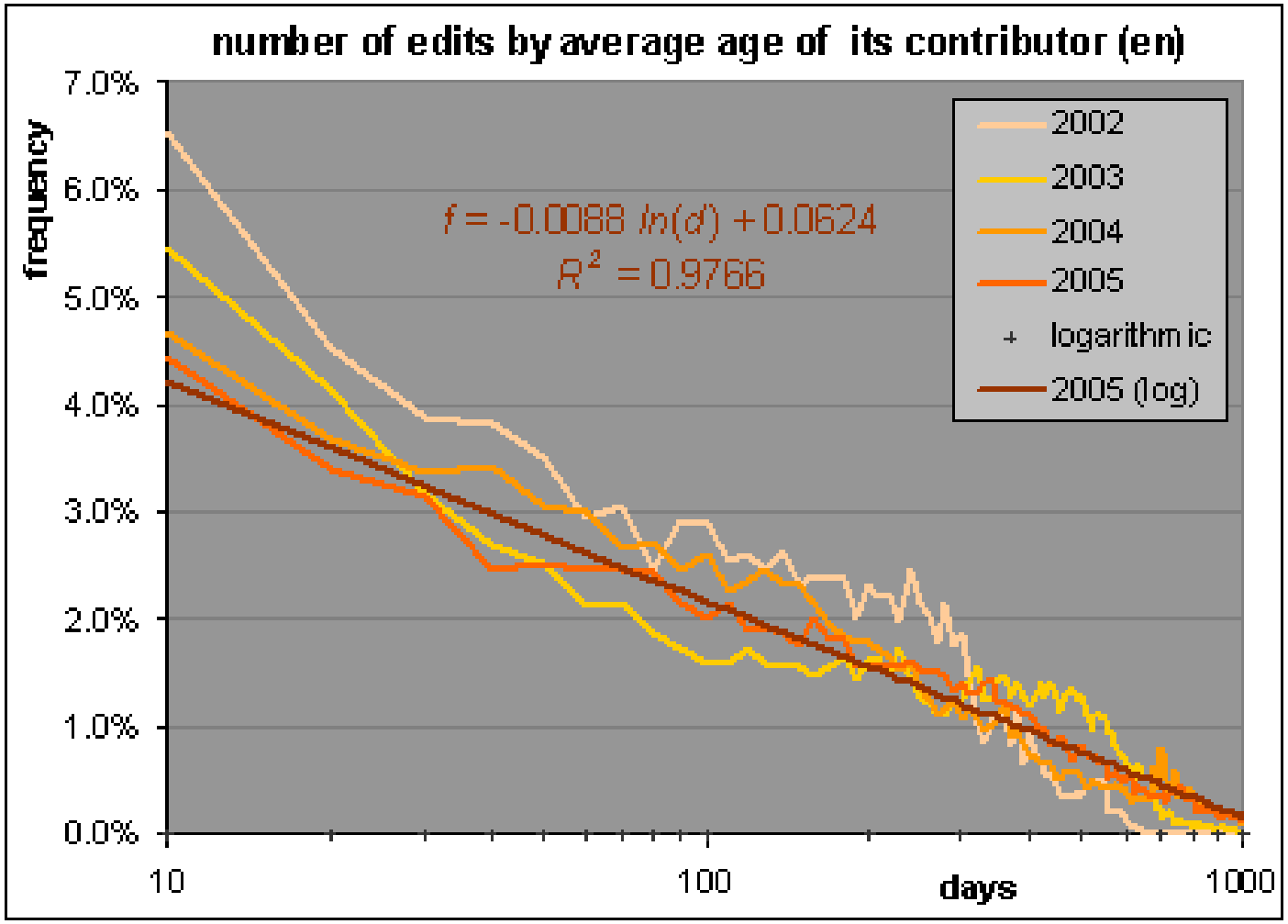


[http://en.wikipedia.org/wiki/User:AmiDaniel/VandalProof **VandalProof**]

Jakob Voss : **Workshop on Wikipedia Research.** *WikiSym 2006-08-23*, Odense, Denmark

{{ Users : Fluctuation }}

{{ Users : Newbies and edits }}



~25% of all edits are by "newbies" that have done their first edit less than 100 days ago

[http://wm.sieheauch.de/?p=44 **Days since first edit**]
[http://meta.wikimedia.org/wiki/Days_since_first_edit **Days since first edit**]

Jakob Voss : **Workshop on Wikipedia Research.** *WikiSym 2006-08-23*, Odense, Denmark

{{ **Users** }}

* Motivation
  * Frost 2005: *Zivilgesellschaftliches Engagement in virtuellen Gemeinschaften*
  * Forte et al 2005: *Why do people write for Wikipedia?*
* Roles
  * Ciffolilli 2003: *Phantom authority, self–selective recruitment and retention of members in virtual communities*
  * Reagle 2006: *Do as I do: leadership in the Wikipedia*
* Collaboration, Discussion
  * Matei et al 2006: *Ambiguity and conflict in the Wikipedian knowledge production system*
  * Pentzold et al 2006: *Focault at Wiki.*

{{ Users : More }}

* More roles
* Usage patterns
* …
* BUT: Privacy!

{{ Semantics & KO }}

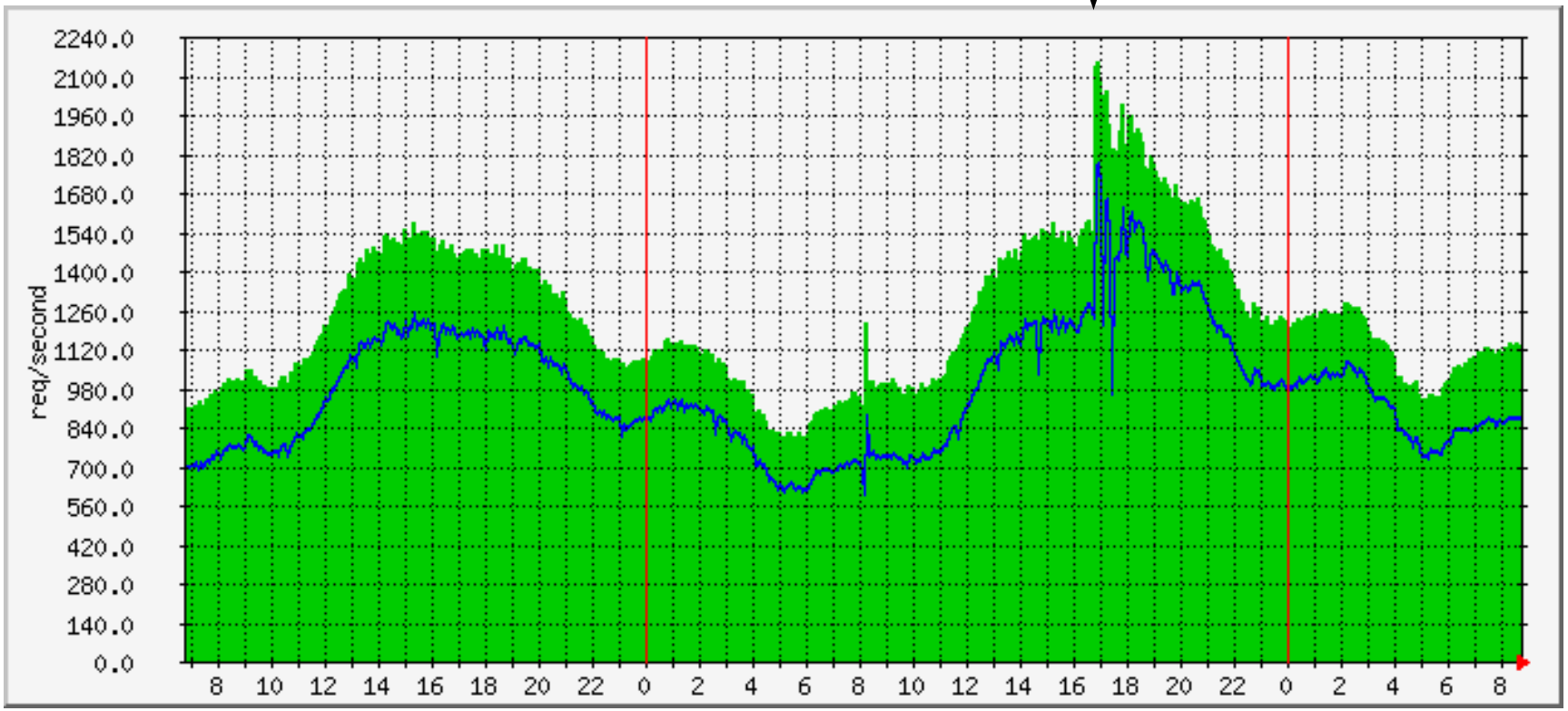
* **Semantics:**
  computers understand what people meant
* **Knowledge Organization:**
  structure the creation of meaning
* **Knowledge Management:**
  people understand what people meant
* **Teaching & Learning:**
  people understand the creation of meaning
* **The fable of Strong AI:**
  computers understand the creation of meaning

{{ **Impact** }}

* Attention
* Open Content, open society?
* History of Wikipedia and Encyclopaedias
* Usability
* Ethics
* ...

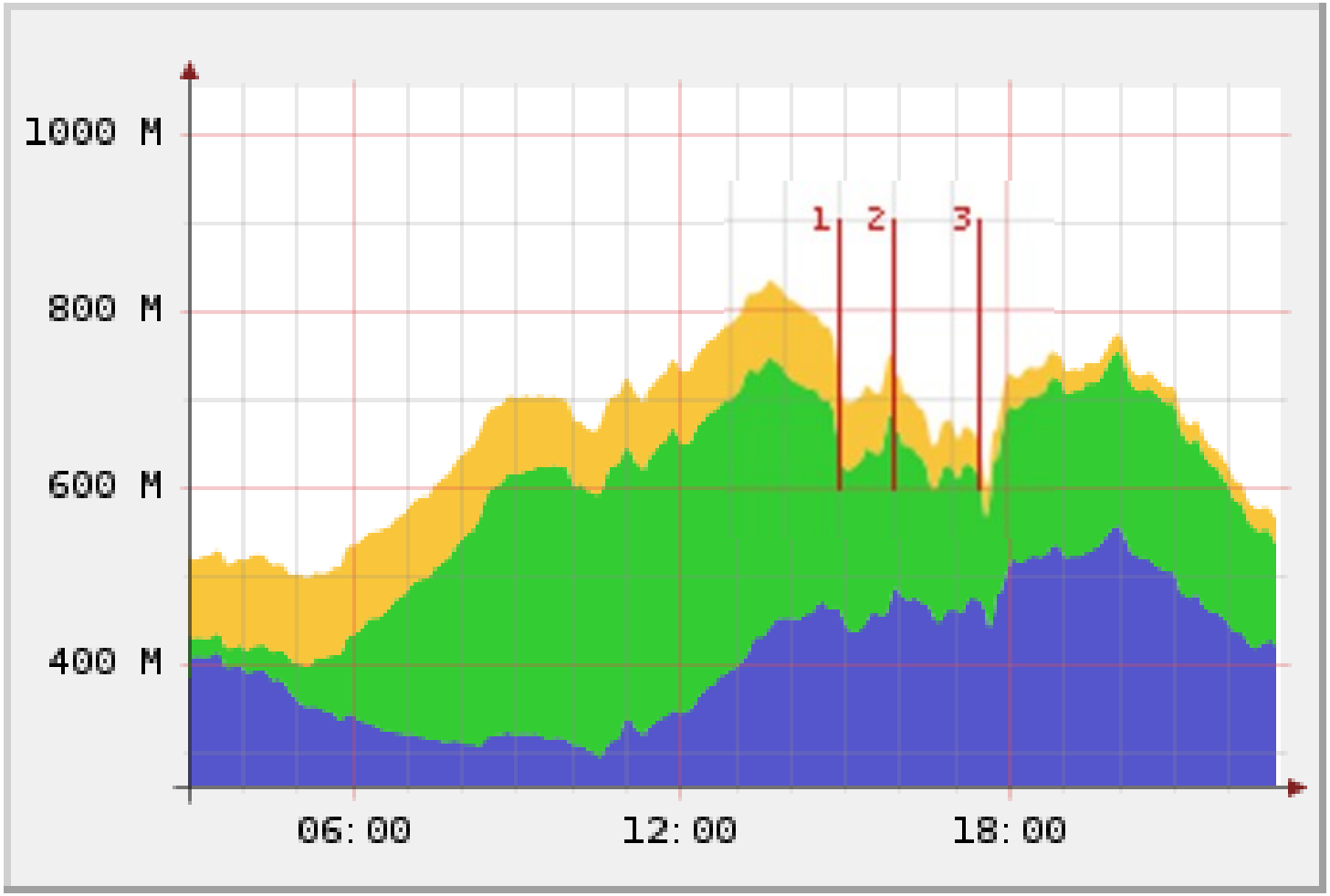
{{ Impact : Attention }}

2005-05-19: *habemus papam*



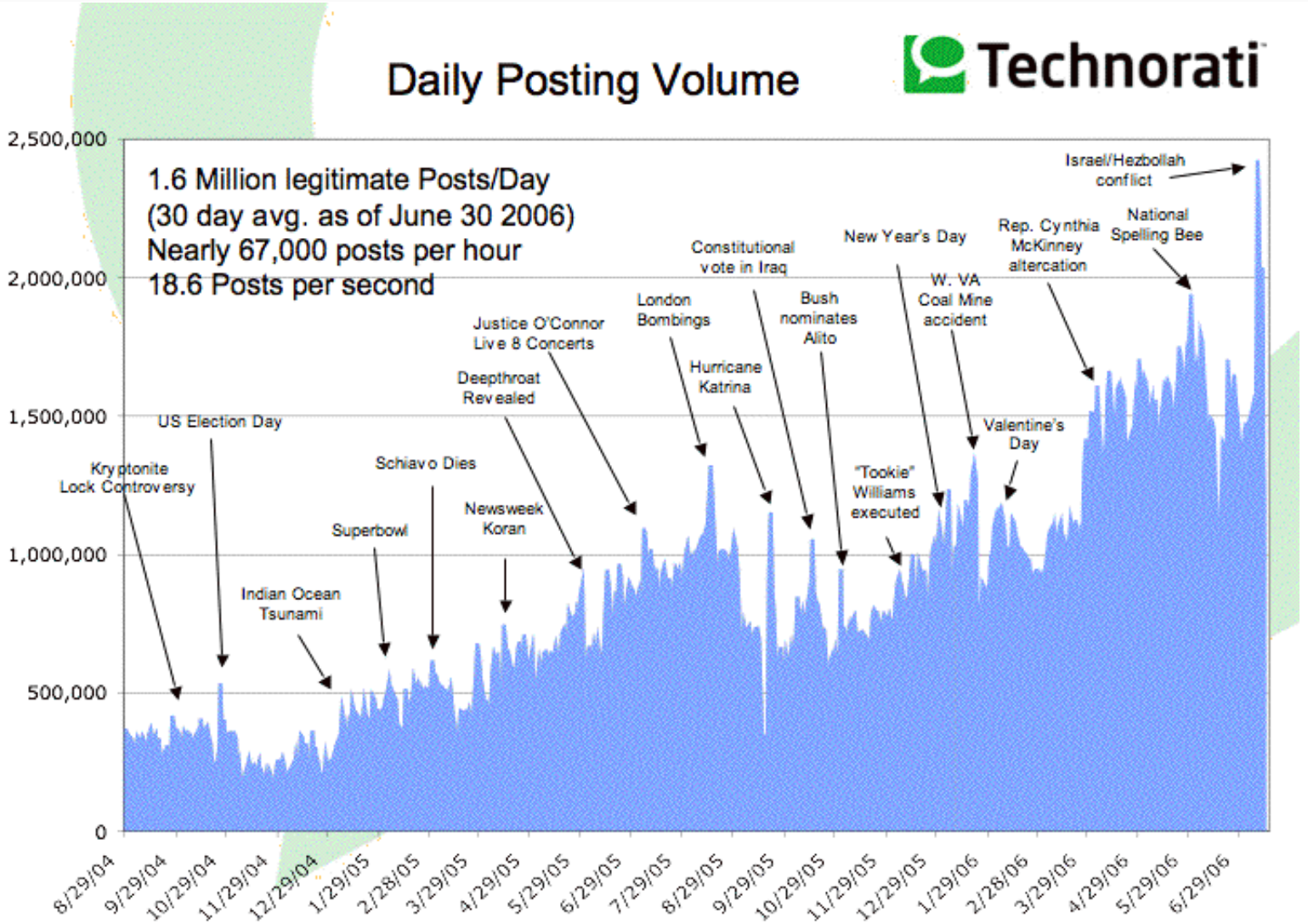
{{ Impact : Attention }}

2006-06-30: Germany-Argentina (start, half, penalty)



[http://wikipedistik.de/2006/07/01/enzyklopaedie-fussball-01/ **Tim Bartel**]

{{ Impact : Attention }}



[http://www.sifry.com/alerts/archives/000436.html **State of the Blogosphere**]

Jakob Voss : **Workshop on Wikipedia Research.** *WikiSym 2006-08-23*, Odense, Denmark

{{ **Research Needed** }}

* Comparision between different languages (intercultural studies)
* Impact of gender to Wikipedia and knowledge (gender studies)
* Ethical aspects (responsability, privacy...)

* *...what are you interested in?...*

Workshop on
Wikipedia Research

Jakob Voss

WikiSym 2006, Odense, Denmark
23 August 2006

Workshop on

Wikipedia Research

[[/Methods]]

{{ **Agenda** }}

* Scientists
* Methods
  * statistics, surveys, interviews
  * bots, API, database
* Infrastructure
  * Mailing lists, blogs, conferences, bibliography

{{ **Scientists** }}

* Community
* Students (Diploma/Master theses)
* Journalists
* PhD theses
* University projects
* Papers at journals and conferences

{{ Scientists : Disciplines }}

* Media Studies, Journalism
* Psychology, Sociology
* Library & Information Science
* Computer Science
* Pedagogy, Knowledge Management

* Law, Economics, Gender studies ....

{{ **Methods** }}

* Comparisions
* Statistics
* Surveys
* Interviews
* Literature review

* ...depending on the discipline...
* *Don't use tools if you don't know how they work!*

{{ **Philosophy** }}

* Focault @ wiki
* Habermas's Theory of Communicative Action
* Heidegger's Hammer

{{ **Statistics** }}

* Statistics of Wikipedia's size and growth
  http://stats.wikimedia.org/
* Statistics of communication on mailing lists:
  http://www.infodisiac.com/Wikipedia/ScanMail/

Erik's the best! :-)

* Many simple / smaller statistics in the projects
* Don't even trust your own statistics!

{{ Statistics : Tools }}

* User edit counter
  http://tools.wikimedia.de/~interiot/cgi-bin/
  PRIVACY!
* CoCat
  http://tools.wikimedia.de/~voj/cgi-bin/coca
  slow!
* Build in MediaWiki methods

{{ Statistics : Usage }}

* Too much
* No stats, no manipulation
* WikiCharts
  http://tools.wikimedia.de/~leon/stats/wikicharts/
* Main page, Current news (Hisbollah, Harry Potter...), Countries, Sexuality

{{ **Surveys** }}

* Difficult
  * Validity
  * Acceptance
* Motivation of Contributors (2005)
  http://wy2x05.psychologie.uni-wuerzburg.de/ao/research/wikipedia.php?lang=en
  => 88% men, 50% singles
* General (Official) User Survey
  http://meta.wikimedia.org/wiki/General_User_Survey

{{ **Bots** }}

* (semi)automatic processes interacting with Wikipedia – navigation and editing
* get a special account flag
* Correct spelling, add categories, change links, detect vandalism etc.

{{ Bot Frameworks }}

* [http://pywikipediabot.sourceforge.net/ **PyWikipediaBot**] (Python)
* [http://sourceforge.net/projects/dotnetwikibot **DotNetWikiBot Framework**] (.NET, C#)
* **Query API**

{{ **Query API** }}

* Query data directly from MediaWiki
* [http://en.wikipedia.org/w/query.php en]
* Returns many, many properties
* Multiple output formats (XML, JSON, TXT...)

{{ API : Example }}

#**Question:** Main authors of an article?

#**Idea:** Count number of edits

#**Implementation:** evaluate revision

{{ API : Implementation }}

* http://en.wikipedia.org/w/query.php
  ?**what**=revisions &**format**=json (for testing: jsonfm)
  &**rvlimit**=200 &**titles**=*article's title*

*
```
{ "pages":
        "3173243": {
            "title": "article's title"
            "revisions": [
                {
                    "timestamp": "2006-08-13T07:53:48Z",
                    "user": "user's name"
                },{
                    "timestamp": "2006-08-13T07:41:02Z",
                    "anon": ""
                    "user": "ip number"
                }
            ]
        }
}
```

{{ API : Implementation }}

```
...
eval("var queryResult=" + data);
var page = anyChild(queryResult.pages);
var revisions = page.revisions;
...
for(var i=0; i<revisions.length; i++) {
  var rev = revisions[i];

  if (rev.minor != null) minorCount++;
  if (rev.anon != null) anonCount++;
  else
    if (!editorCounts[rev.user]) {
      editorCounts[rev.user] = 1;
    } else {
      editorCounts[rev.user]++;
    }
}
...
```

{{ API : Example }}

#**Question:** Main authors of an article?

#**Idea:** Count number of edits

#**Implementation:** evaluate revision

#**Conclusion:** edit-counting is inaccurate

{{ **Database** }}

* Toolserver
  http://meta.wikipedia.org/wiki/Toolserver
* (almost) direct access (read-only)
* MediaWiki database layout
  http://meta.wikimedia.org/wiki/Database_layout

{{ Database : Dump }}

* Database Dumps (XML)
  http://download.wikimedia.org

{{ Database : Example }}

* **Question:** Sizes of categories?
* **Idea:** table categorylinks

```
cl_from
cl_to
cl_sortkey
cl_timestamp
```

* **Method 1:** MySQL statement

```
SELECT cl_to, COUNT(cl_from) FROM
categorylinks GROUP BY cl_to
```

* ...and wait

{{ Database : Example }}

* **Method 2:** Short script

#Extract categorylinks

```
echo "SELECT cl_from, cl_to FROM
categorylinks" | mysql enwiki_p > catlinks
```

#Skip 1st line & 2nd column, count, sort

```
awk "NR!=1 {print \$2;}" catlinks |
perl count.pl |
sort -r -n -k 2 > catsizes
```

#Show result

```
head catsizes
```

{{ Database : Example }}

```
12  NPOV_disputes
12  Articles_which_may_contain_original_research
12  Forms_of_government
12  Anarchism
12  Political_ideologies
12  Political_philosophies
12  Social_philosophy
```

Categories: NPOV disputes | Articles which may contain original research | Forms of government | Anarchism | Political ideologies | Political philosophies | Social philosophy

{{ Database : Example }}

```
#!/usr/bin/perl
# count.pl – create histogram

my %counter = ();

while (<STDIN>) {
  $_ =~ s/\n$//;  # remove linebreak
  $counter{$_}++;
}
foreach $key (keys %counter) {
  print "$key\t" . $counter{$key} . "\n";
}
```

{{ Database : Example : Results }}

GFDL_images.....102445
Living_people.....97298
Disambiguation.....55204
Redirects_from_US_postal_abbreviation.....40890
User-created_public_domain_images.....36300
Album_covers.....34007
Public_domain_images.....29220
Logos.....27666
Free_use_images.....20310
Promotional_images.....20013
User_en.....17015
Screenshots_of_television.....15628
Screenshots_of_films.....13862
Open_proxies_blocked_on_Wikipedia.....13152
Unprintworthy_redirects.....12755
1911_Britannica.....11991
Screenshots_of_computer_and_video_games....11648
Articles_with_unsourced_statements.....11553
Articles_lacking_sources.....10857

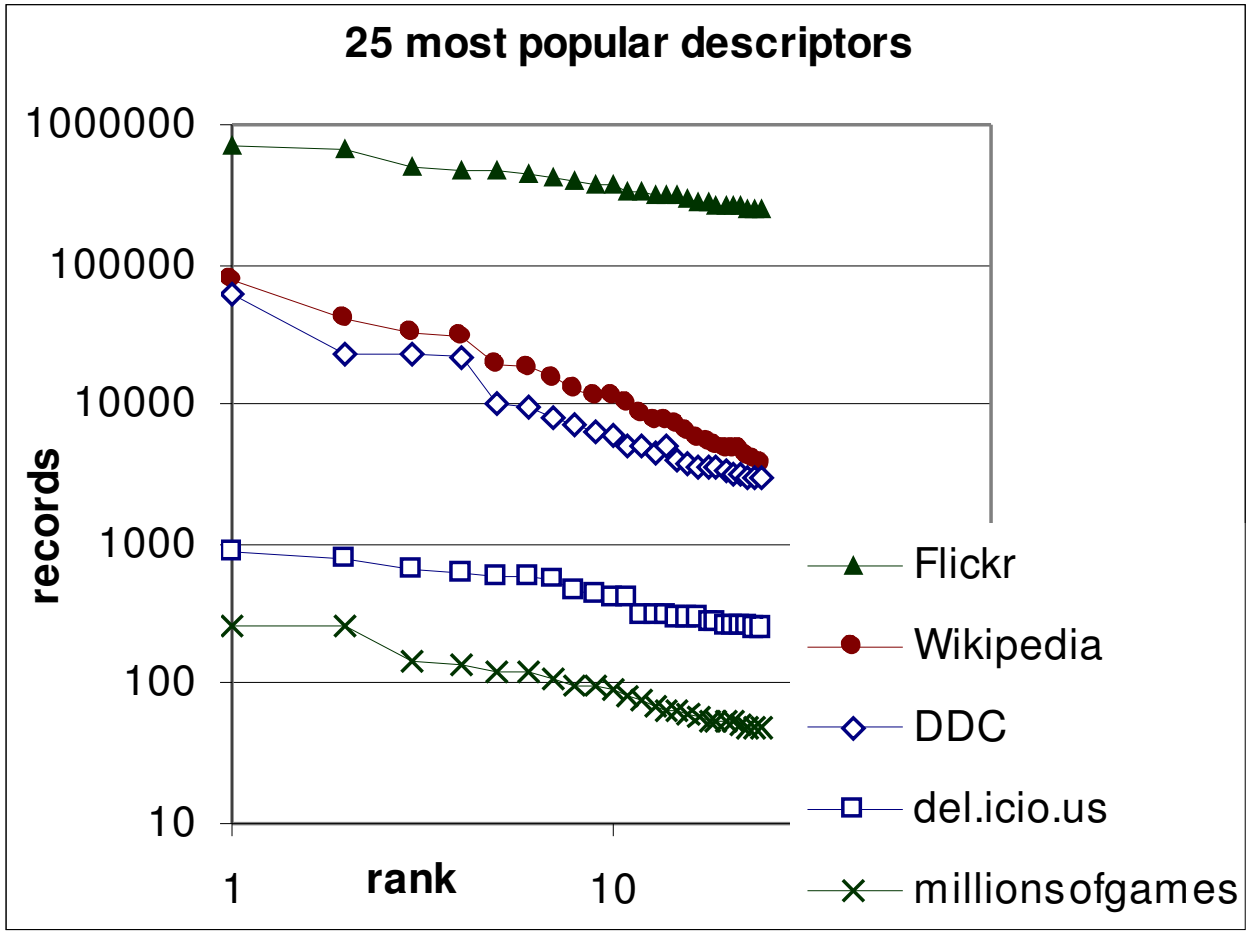
{{ Database : Example : Results }}

* Pages per Category

power law

Wikipedia $\lambda = 0.96$

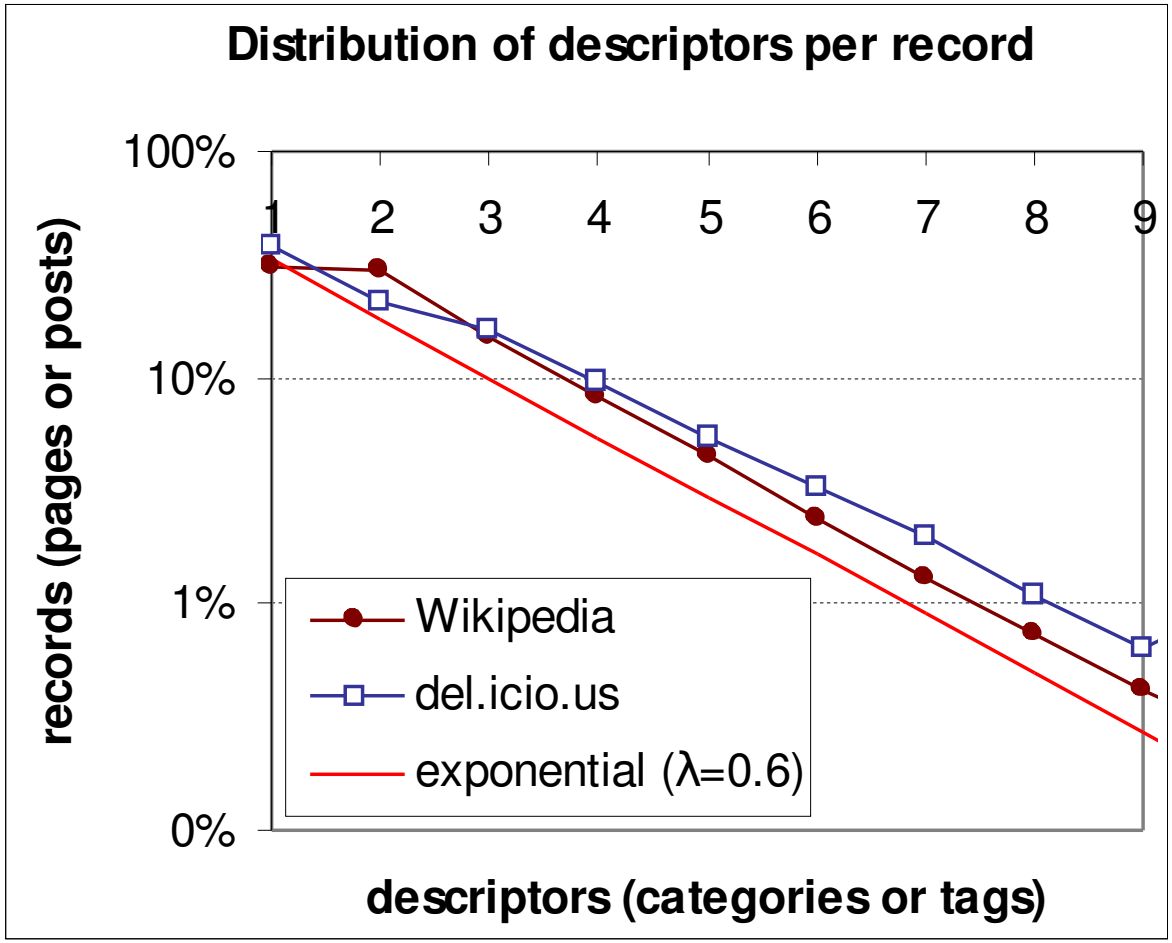
DDC $\lambda = 0.94$



{{ Database : Example : Results }}

* Categories per page:

  exponential distribution



{{ **Infrastructure** }}

* Mailinglists
* Weblogs
* Wikis
* Wiki Research Bibliography
* Conferences
* Journals
* Institutes ?

{{ Infrastructure : Mailing lists }}

* **wiki-research-l**
  http://mail.wikipedia.org/pipermail/wiki-research-l/
* **wiki-research**
  http://www.wikisym.org/cgi-bin/mailman/listinfo/wiki-research
* wikipedia-l, foundation-l...

{{ Infrastructure : Wikis }}

* [[meta:Research]]
* [[meta:Research/Social_Research_Collaborations]]
* [[en:Wikipedia:WikiProject Wikidemia]]
* [[de:Wikipedia:Wikipedistik]]

* Many distributed pages

{{ Infrastructure : Blogs }}

* Joseph Reagle
  http://reagle.org/joseph/blog/
* Wikimetrics (me)
  http://wm.sieheauch.de/
* Wikipedistik (Tim Bartel)
  http://www.wikipedistik.de/
* Andrew Lih
  http://www.andrewlih.com/blog/
* ...

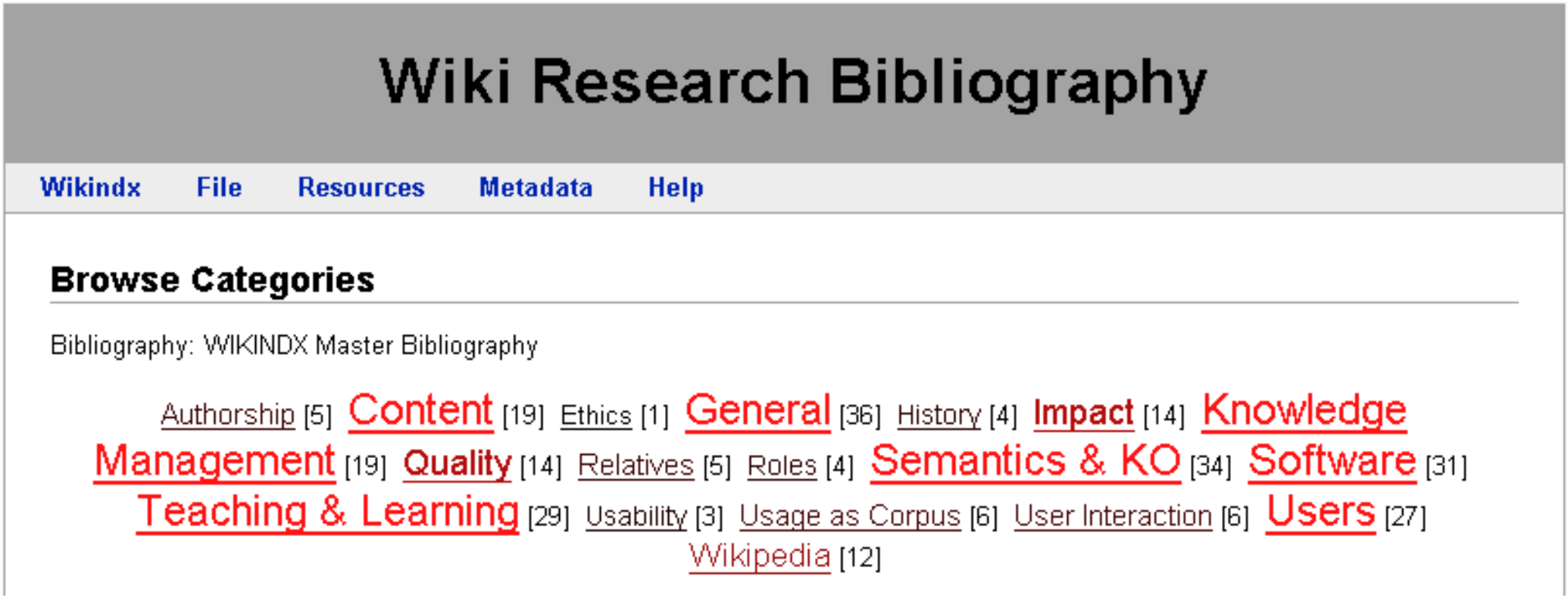
{{ Infrastructure : Conferences }}

WIKIMANIA 2006
The International Wikimedia Conference

August 4 - 6
Cambridge, Massachusetts, USA

The International Symposium on Wikis
* **WikiSym 2005**
* **WikiSym 2006**
* [[**WikiSym 2007**]]

{{ Infrastructure : Bibliography }}



* http://bibliography.wikimedia.de/

* 230 resources (also about wikis in general)

Jakob Voss : **Workshop on Wikipedia Research.** *WikiSym 2006-08-23*, Odense, Denmark