A Cloud-Based Infrastructure for Community-Driven Data Mining on Research Outputs

Robert Jäschke\textsuperscript{1,3} & Andreas Hotho\textsuperscript{2,3}

\textsuperscript{1}The University of Sheffield, UK
\textsuperscript{2}The University of Würzburg, Germany
\textsuperscript{3}L3S Research Center Hannover, Germany
Agenda

1. future (what we need)

2. past & present (what we have)
Agenda

1. future (what we need)

2. past & present (what we have)
Research Outputs are Big Data

<table>
<thead>
<tr>
<th>Volume</th>
<th>Velocity</th>
<th>Variety</th>
<th>Veracity</th>
</tr>
</thead>
</table>

Images representing the four 'V's.
Velocity

In every 24-hour period approximately 20,000,000 words of technical information are being recorded. A reader capable of reading 1,000 words per minute would require 1.5 months, reading 8 hours every day, to get through 1 day’s technical output, and at the end of that period, he would have fallen 5.5 years behind in his reading!
Velocity

In order to emphasize the problem, it is worthwhile to cite some figures. The rate at which technical documents are produced at the present time is estimated to be well over 500,000 per year. In every 24-hour period approximately 20,000,000 words of technical information are being recorded. A reader capable of reading 1,000 words per minute would require 1 1/2 months, reading 8 hours every day, to get through 1 day's technical output, and at the end of that period, he would have fallen 5 1/2 years behind in his reading! Even in attempting to read the portion of the literature in a single subject field such as chemistry, he would find himself falling behind an estimated 850,000 pages per year. This production rate of scientific information will undoubtedly increase as countries such as China and India begin to produce technical work commensurate with their size.
Velocity

REPORT NO. RSIC-510

METHODS FOR SATISFYING THE NEEDS
OF THE SCIENTIST AND THE ENGINEER
FOR SCIENTIFIC AND TECHNICAL INFORMATION

by

Hubert Murray, Jr.

January 1966
Volume

How much science is there?
Scientific publishing has been accelerating—a new paper is now published roughly every 20 seconds. Let’s imagine a bibliography listing every scholarly paper ever written. How long would it take?

- If we can fit 100 citations per page...
- 1,000 pages per book...
- ...and then we start stacking books...

...this is what the full list would look like:

All scholarly articles from before 1880 fit in just a few volumes.

How open is it?
Since the advent of the web, much of scientific publishing has been moving to open access. According to Science-Metric, open access reached a “tipping point” around 2011; more than 50% of new research is now available free online.

Open-access papers
As journals move to open access and digitize their archives, old papers from every period move up here...

...in addition to the flood of new papers being published here directly.

Traditional publication

OPEN-ACCESS PAPERS

25% of open-access papers are freely available on publication. The rest become free within 12 months on journal websites or other repositories.

By Randall Munroe • Design by Joe Cador and David Phamott

(R. Munroe)
Some scientists claim it takes less time to do an experiment than to find out whether or not it has been done before.

(J. Naisbitt, 1982)
Veracity

REFERENCES

16. NEED CITATION FROM TELECOM-CITIES


20. Institute for the Future. NEED CITATION FROM TEN YEAR FORECAST?
State of the Art

How do we tackle this issue?

Well, ...

(R. Munroe, XKCD)
Vision

make the state of the art accessible
Proposal

Combine the power of these systems ... 
... into a federated, open platform ... 
... as part of the European Open Science Cloud.
Agenda

1. future (what we need)

2. past & present (what we have)
The Social Bookmark and Publication Management System BibSonomy

Dominik Benz, Andreas Hohto, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme.
The VLDB Journal 19(6):849--875 (December 2010)

Abstract

Social resource sharing systems are central elements of the Web 2.0 and use the same kind of lightweight knowledge representation, called folksonomy. Their large user communities and ever-growing networks of user-generated content have made them an attractive object of investigation for researchers from different disciplines like Social Network Analysis...

Links and resources

DOI: 10.1007/s00778-010-0208-4
URL: http://www.kde.cs.uni-kassel.de/pub/pdf/ benz2010social.pdf

community

In collection of:
- @dmir
- @kde-alumni
- @kde-alumni
- @bibsonomy-ruby
- @michael-t-weber
- @tioni
- @asx0312
- @schmidt2
- @schmitz
- @kibaranov
- @ls
- @olnya
- @folke
- @tioni
- @noebrain
- @victorinostar
- @juese-biel

Meta data

Last update a year and 3 months ago
Created 4 years and 10 months ago

edit tags

2010 bibsonomy bookmark edit...
History
Bibsonomy is a system for sharing bookmarks and lists of literature. When discovering a bookmark or a publication on the web, you can store it on our server. You can add tags to your entry to retrieve it more easily. This is very similar to the bookmarks/favorites that you store within your browser. The advantage of Bibsonomy is that you can access your data from wherever you are. Furthermore, you can discover more bookmarks and publications from your friends and other people.

This page shows you the latest updates of Bibsonomy. Why don't you just try it yourself? After a free registration, you can organise your own bookmarks and publications, and discover related entries.

bookmarks

BibTeX

Comparison of visualizations in formal concept analysis and correspondence analysis

K. E. Wolff and S. Gaerler. Visualization of categorical data. (to appear)
ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
ELECTRICAL ENGINEERING & COMPUTER SCIENCE, UNIVERSITY OF KASSEL

The University of Sheffield.

Information School.

Forschungszentrum · Research Center

L3S
History
Features

tag
user
group
author
concept
BibTeX key
search
search:jaeschke
in group
- 20dc13
- bibsonomy_dev
- iccs
- kde
- l3s
- nepomuk
- stair

go by @jaeschke
You shall not pass: detecting malicious users at registration time

Authors: Christian Kast & Robert Jäschke

Published in:
Proceedings of the 1st International Workshop on Online Safety, Trust and Fraud Prevention
ACM, New York, NY, USA © 2016
ISBN: 978-1-4503-4364-0
doi:10.1145/2915388.2915370

Bibliometrics:
- Downloads (6 Weeks): 8
- Downloads (12 Months): 30
- Citations: 0
Features

You shall not pass: Detecting malicious users at registration time

Authors: Christian Kater, Robert Jäschke

Published in: Proceedings of the International Workshop on Trust and Privacy in Online Social Networks (TrustonSocial '16)

ACM, New York, NY, USA

general information

entry type: proceedings

title: You Shall Not Pass: Detecting Malicious Users at Registration Time

author(s): Christian Kater, Robert Jäschke

I am an author

tags - describe the post

tags: 2016, bibsonomy, fraud, myown, spam

recommendation: web, social, bibsonomy, blog, bibsonomynews
You Shall Not Pass: Detecting Malicious Users at Registration Time


Abstract

Spam is a widespread problem for many online services. The use case in this paper is the social bookmarking system BibSonomy, which received over 150 times more registrations from spam users than from normal users over the last ten years. A common approach to fight spam is to use machine learning to classify the users into good or malicious users. … (more)

10.1145/2915368.2915370
http://doi.acm.org/10.1145/2915368.2915370

There is no review or comment yet. You can write one!

Tags
fraud spam bibsonomy sys.relevantforu1a

edit your publication post

general information

home myBibSonomy add post groups popular genealogy admin

You shall now

Features


2016


thinking about hidden features
removing unused features is sometimes necessary
there is never a shortage of ideas for good features
explaining (new) features is important
having an HCI expert is important
Integration with other Websites and Services

Bibsonomy integrates well with other websites, content management systems, and reference managers. Furthermore, you can connect Bibsonomy to your own software using one of our API (application programming interface) clients. The following list describes ways to integrate Bibsonomy with various systems and technologies - ranging from code snippets to complete and powerful, easy-to-install plugins or API clients.

- **Integration with other Websites and Services**
  - Browser
    - Chrome / Chromium
    - Firefox
      - Add on
      - Search
      - Zotero
    - Safari
  - Programs
    - Citavi
    - Emacs
    - JabRef
      - Bibsonomy Plugin
      - Extract JabRef layouts from Jar files
    - KBibTex
    - Sublime Text
    - Textile
  - Websites
    - Confluence
    - GoogleDocs
    - Google Scholar
    - Moodle
  - Your own website
    - Bookmark Links
    - Tag Clouds
    - Post Lists
    - JSON feed
    - TYPO3
    - Wordpress
    - XWiki
    - Zope
  - Libraries
    - Digital Libraries
    - Discovery Service
    - OpenURL resolver
  - Applications
    - Bookmark Exporter
    - Link checker
    - RSS
  - REST API
    - Java
    - PHP
    - Python
  - Others
    - Android
    - OAuth
    - Search Engine

---

**Add-ons**

**Bibsonomy Buttons**

Three useful Bibsonomy button: myBibsonomy as well as post bookmark and post publication.

**Activities: PUMA/Bibsonomy Module**

This PB is the PUMA/Bibsonomy plugin for Moodle. It helps you to publish publication lists from the Publication Management System http://www.academic-puma.de or http://www.bibsonomy.org.

**TYPO3**

Add a list of publications and tag clouds from Bibsonomy or PUMA to your TYPO3 pages. Style your publication lists using CSS stylesheets (http://citationstyles.org).
- abundance of options for integration
- a plugin/addon is quickly written … but maintenance is challenging
- time consuming but very important
• invest time in proper architecture early on
• migration is challenging
• many invisible improvements
• do not fear change
BibSonomy is a system for sharing bookmarks and lists of literature. When discovering a bookmark or a publication on the web, you can store it on our server. You can add tags to your entry to retrieve it more easily. This is very similar to the bookmarks/favorites that you store within your browser. The advantage of BibSonomy is that you can access your data from wherever you are. Furthermore, you can discover more bookmarks and publications from your friends and other people.

This page shows you the latest updates of BibSonomy. Why don't you just try it yourself? After a free registration, you can organise your own bookmarks and publications, and discover related entries.

bookmarks
previous | next
NICHTLUSTIG.DE
to humor cartoon lang:de as public by schmitz and 1 other person on 2005-12-18 23:13:12.0

BibTeX
previous | next
Comparison of visualizations in formal concept analysis and correspondence analysis
K. E. Wolff and S. Gable: Visualization of categorical data, (to appear)
### bookmarks (697)

<table>
<thead>
<tr>
<th>Title</th>
<th>Date</th>
<th>Time</th>
<th>Edit</th>
<th>Delete</th>
<th>Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>LaTeX Beamer Theme Previews « Jinto Reedwine's Blog</td>
<td>Jun 28, 2010</td>
<td>8:33 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PUMA - erste stabile Version jetzt online verfügbar</td>
<td>Jun 28, 2010</td>
<td>2:49 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature of the week: New help pages</td>
<td>Jun 28, 2010</td>
<td>2:44 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coranac » Off on a tangent : a look at the tangent implementation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### publications (392)

<table>
<thead>
<tr>
<th>Title</th>
<th>Date</th>
<th>Time</th>
<th>Edit</th>
<th>Pick</th>
<th>Unpick</th>
<th>Gold-standard</th>
<th>URL</th>
<th>BibTeX</th>
<th>openURL</th>
<th>Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Logs as Folksonomies</td>
<td>Jun 14, 2010</td>
<td>10:11 AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning optimal ranking with tensor factorization for tag</td>
<td>Jun 14, 2010</td>
<td>10:09 AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A fast string searching algorithm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bibsonomy denotes a social bookmarking system as well as the underlying technical platform. The system enables users to share and annotate bookmarks and publications. This page is the entry point for developers interested in the software platform underlying Bibsonomy and PUMA.

The web application is based on Java, backed-up by a MySQL database and Elasticsearch for fulltext search. The Bibsonomy platform is developed by the KDE group of the University of Kassel, the DMIR group of the University of Würzburg, the LS Research Center, and HethoData.

http://dev.bibsonomy.org/
- plan for times with limited resources
- invest time in community building
- do not underestimate complexity of development
Dear Webmaster,

I am the webmaster of www.__________au. It would be greatly appreciated if you remove links to my website. My Links can be found on these following pages:

http://www.bibsonomy.org/user/lawcostseo11?lang=ru
http://www.bibsonomy.org/user/lawcostseo11?resourceType=publication
http://www.bibsonomy.org/user/lawcostseo11?resourceType=bookmark

Referring_Page_URL...Please could you let me know when you have completed this removal or if you are unable to meet this request and the reasons for this.

Thank you in advance for your co-operation in this matter.

Kind Regards,
Webmaster
www.__________au
● be professional early on
● distribute responsibilities
● do not underestimate effort for user interaction
● be prepared for weird user behavior ;-)
bookmarks | publications
resources: 625k | 938k
posts: 727k | 1045k
documents: - | 51k
users: 9k | 10k

+ more than 5 million posts from DBLP & DNB
+ 2.5 million users that are flagged as spammers
- many more users would require more resources
- measures against spam crucial
- optimizing database schema + queries pays off
• unique opportunities if you run your own system
• can do A/B testing like the „big guys“
• research → production: a long and winding road
• need enthusiastic and persistent PhD students
FUTURE
● ORCID integration
● web archiving
● extended support for groups of users
● federated search, fulltext document search
● …
A Cloud-Based Infrastructure for Community-Driven Data Mining on Research Outputs

jaeschke@L3S.de