Authors vs. Readers

A Comparative Study of Document Metadata and Content in the WWW

by Michael G. Noll

HPI Hasso Plattner Institut

http://www.michael-noll.com/publications/

ACM Symposium on Document Engineering, Winnipeg, Canada, August 2007

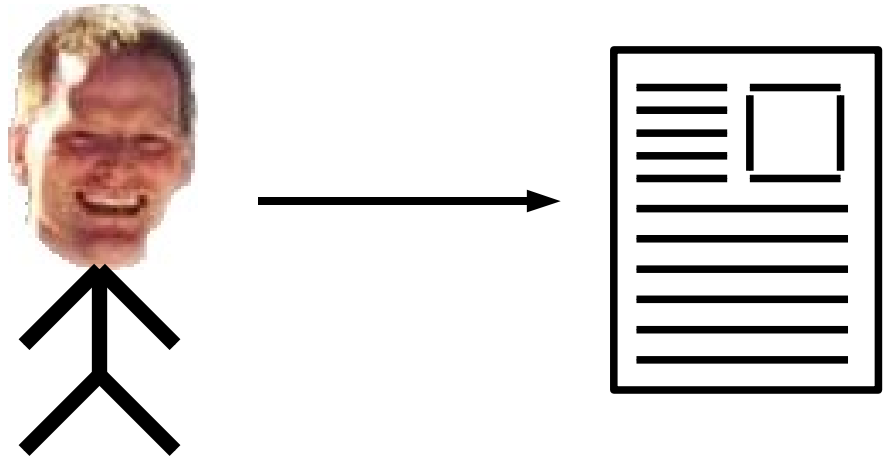
Overview

- Introduction
- Authors
- Readers
- Authors vs. Readers
- Summary & conclusion

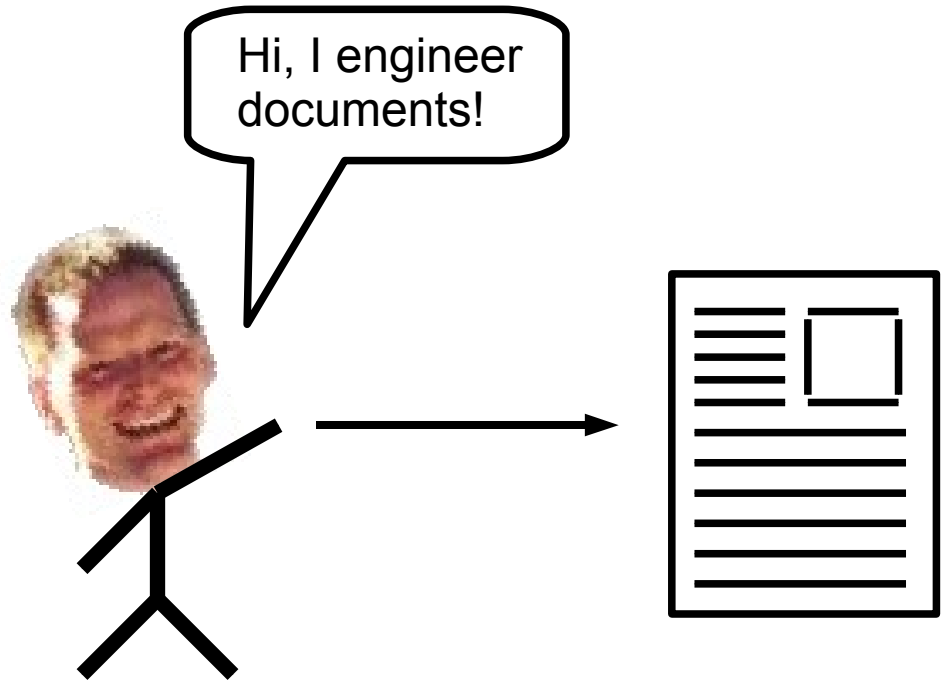
Introduction

- how can we benefit from things like “Web 2.0”? = what do end users bring to the table?
- how much and which kind of user-supplied (meta)data is out there?
- what can we expect to do with it?
- how does it compare to “traditional” metadata?

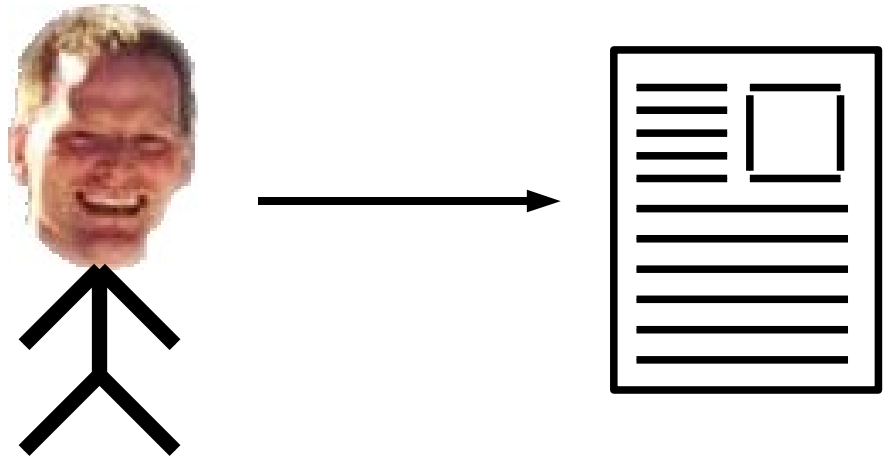
Introduction



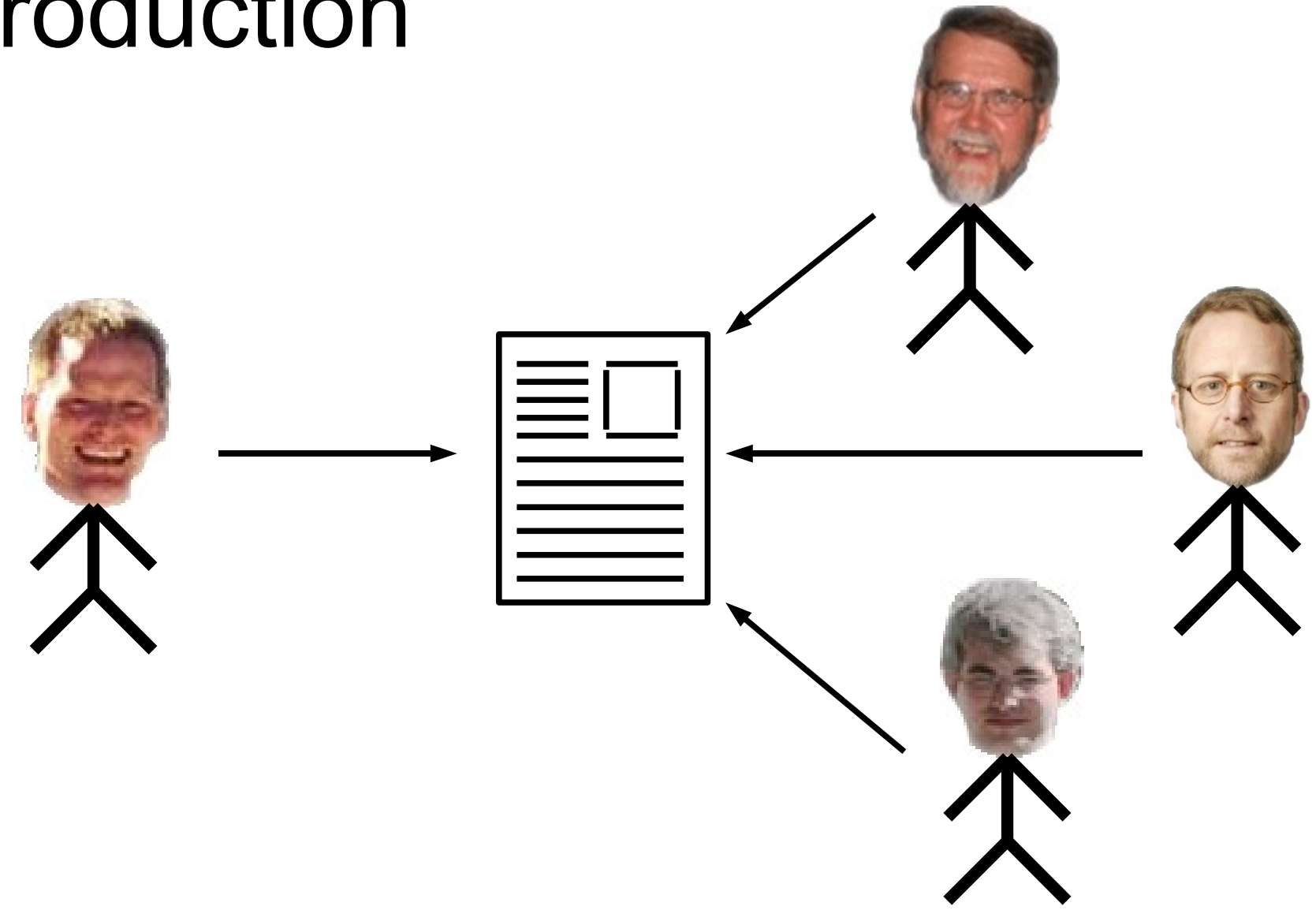
Introduction



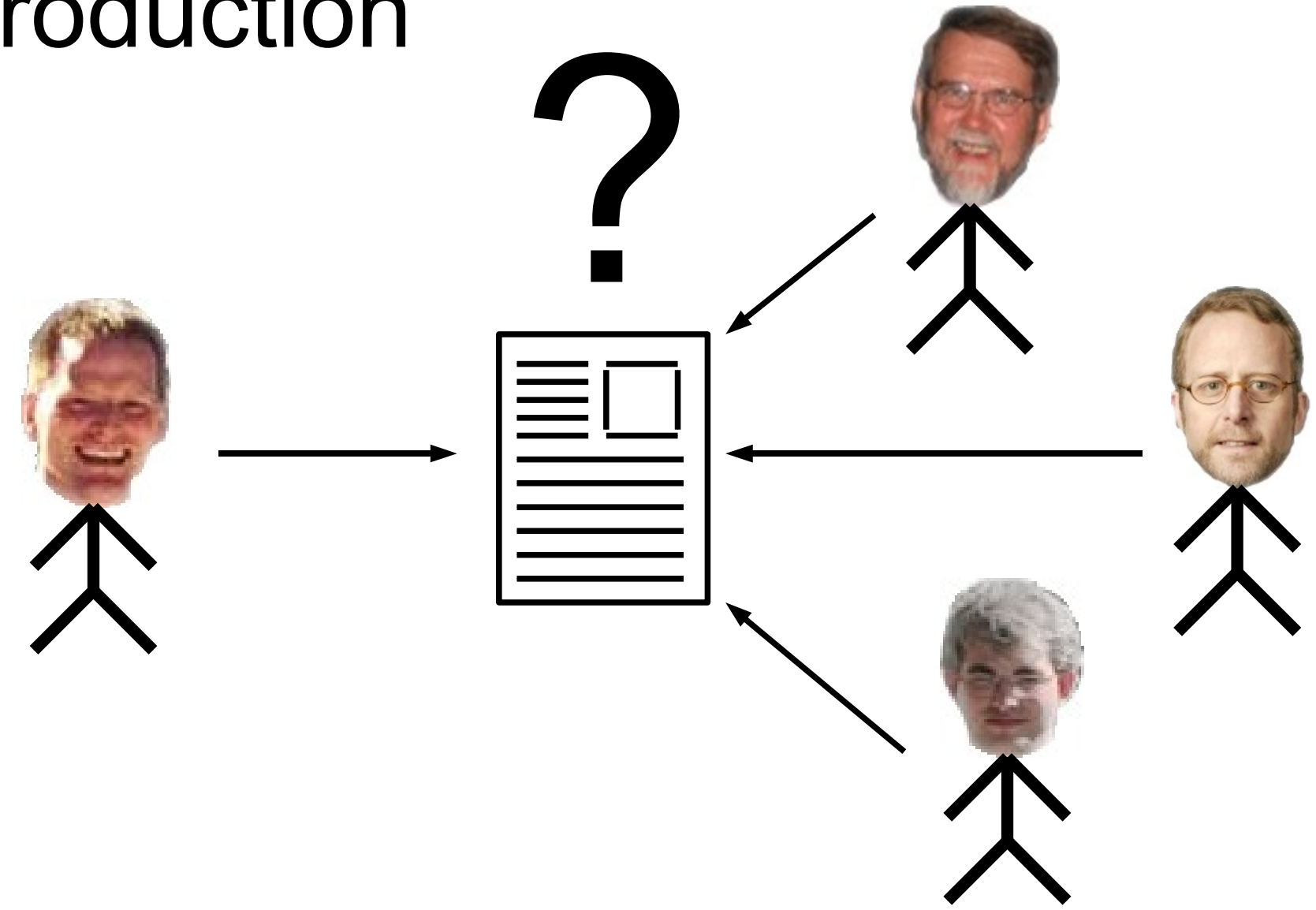
Introduction



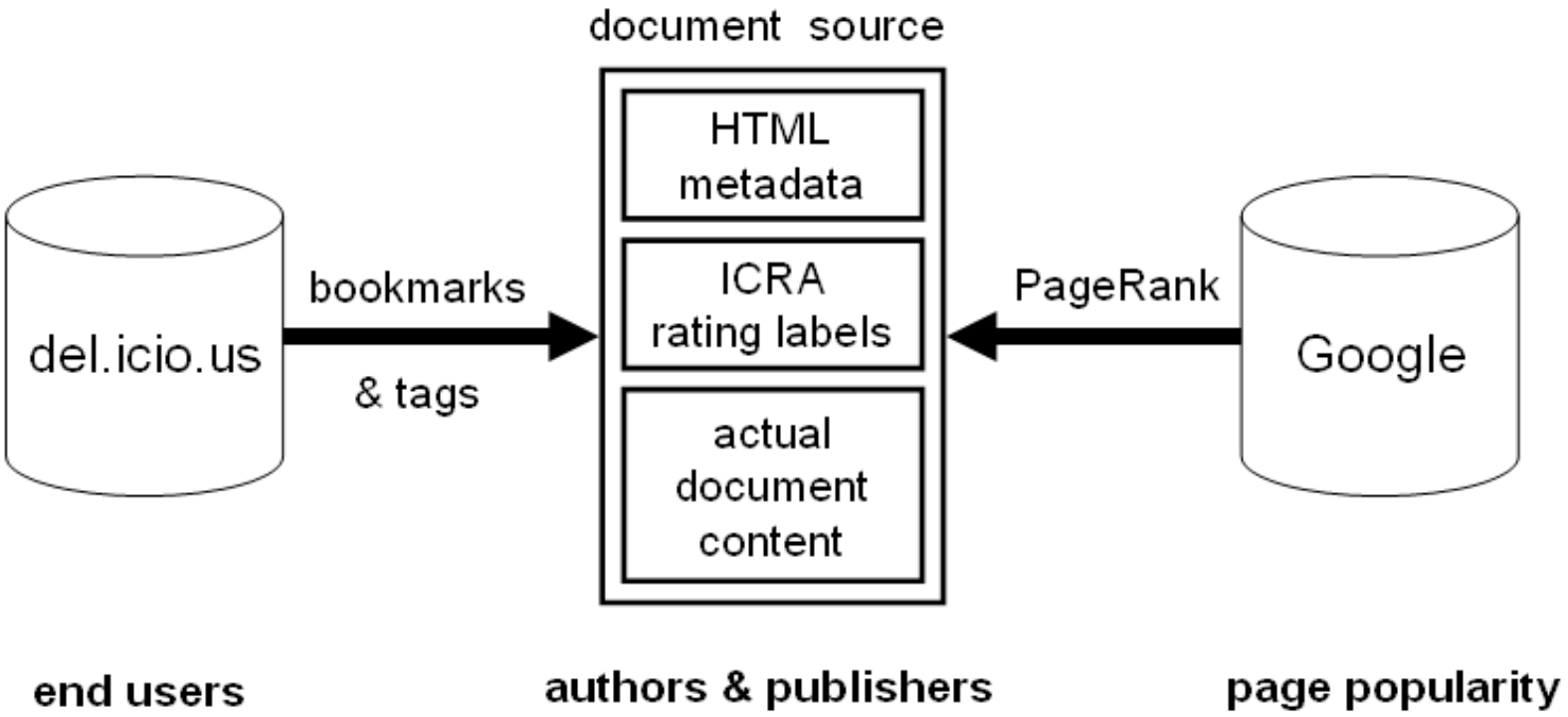
Introduction



Introduction



Information sources



DMOZ100k06

- data set created from these information sources
- based on random sample of 100,000 web documents (2.1%) from the Open Directory
- idea: help researchers and allow comparison of results
- freely available:

  http://www.michael-noll.com/dmoz100k06/

DMOZ100k06

- overall statistics

Total documents	97,578	
Total bookmarks	180,246	
Total (common) tags	25,311	6,090 unique
Bookmarked documents	13,771	14.1%
Tagged documents	4,992	5.1%

DMOZ100k06

- overall statistics

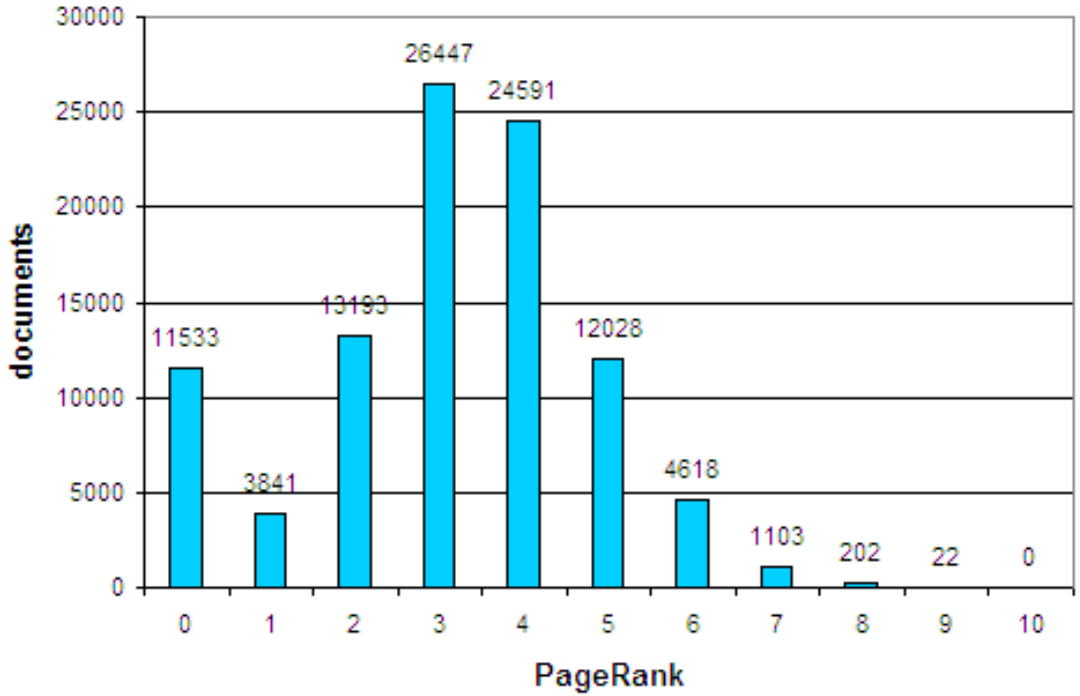
Total documents	97,578	
Total bookmarks	180,246	
Total (common) tags	25,311	6,090 unique
Bookmarked documents	13,771	14.1%
Tagged documents	4,992	5.1%

- per document

	mean	**std.dev.**
Bookmarks	1.85	47.68
Tags (common)	0.26	1.80
PageRank	3.13	1.66

DMOZ100k06

- PageRank distribution of docs in the data set

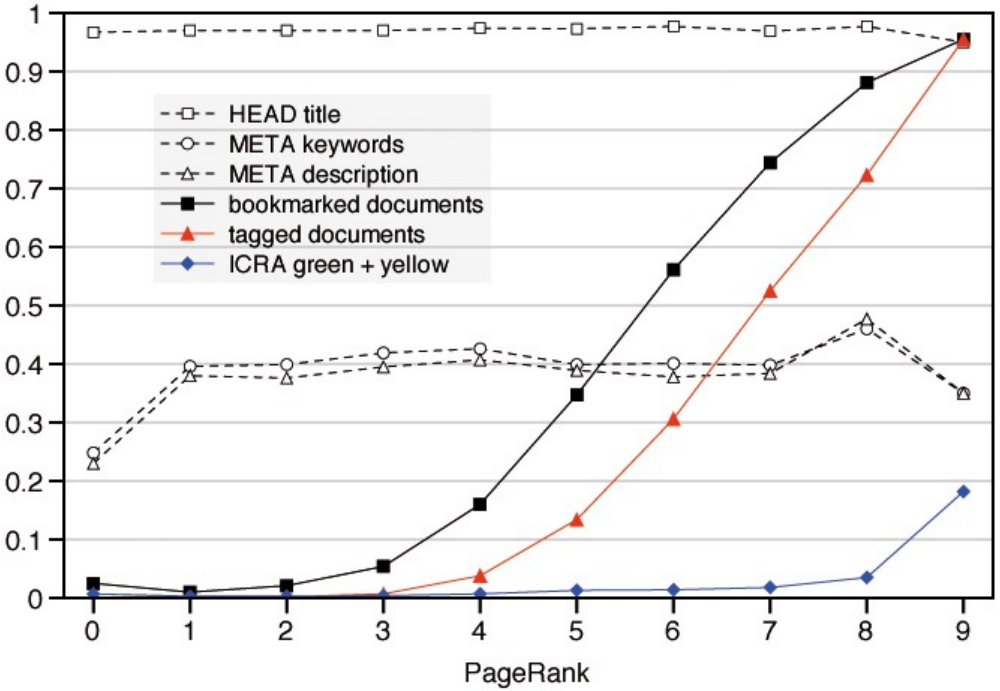


Authors

Authors: available metadata

Results

- document title >> everything else
- keywords > description
- forget about ICRA
- cf. Google study [Dec'05]



Interesting

- drop at PR0, peak at PR8

Readers

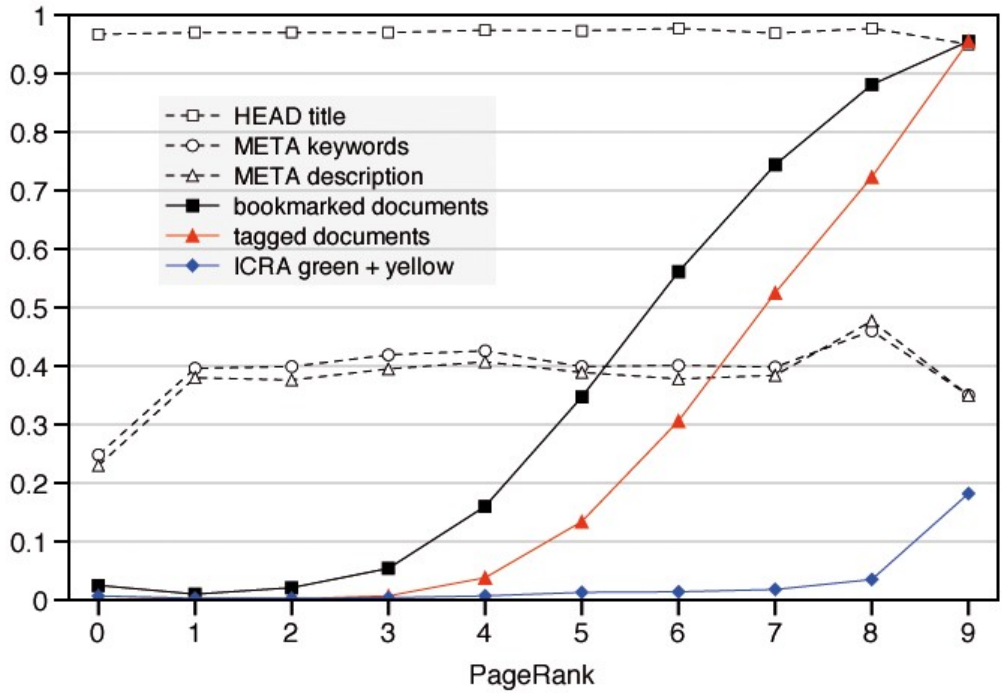
Readers: available metadata

Results

- bookmarks :≥ tags due to experiment setup
- popularity is king

Interesting

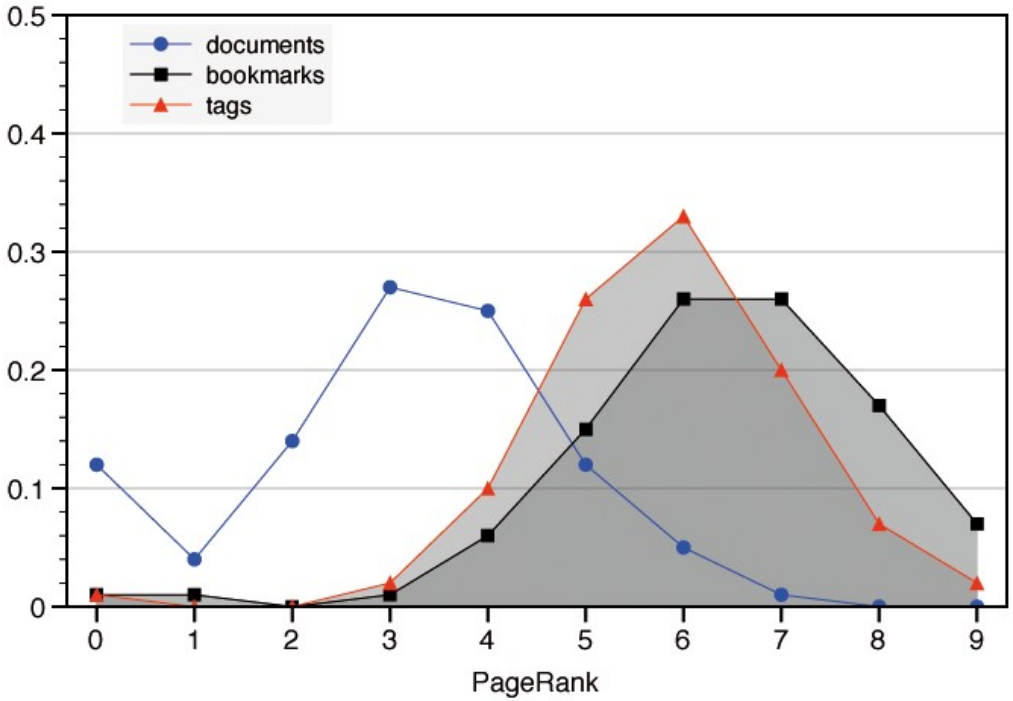
- search engines meet readers' taste, or “follow the mass” effect?



Readers: available metadata

Looking only at bookmarked/tagged documents

- main tagging window between PR5 and PR7
- tagging is shifted towards lower PR compared to bookmarking



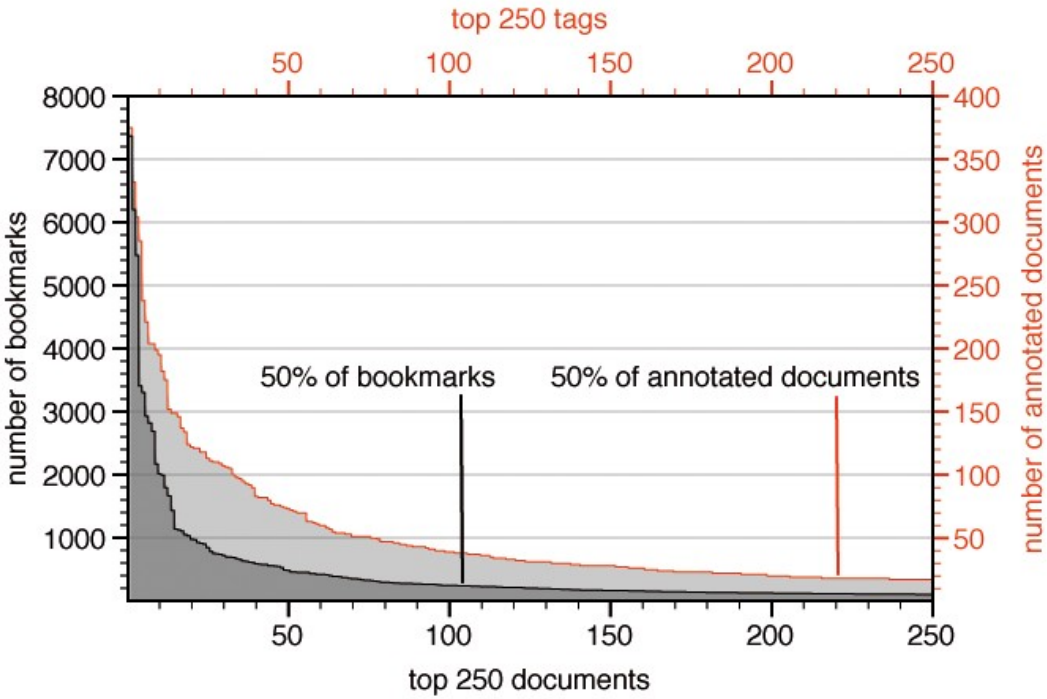
Interesting

- no “long tail”

Readers: top tags and top bookmarks

Results

- most metadata concentrated on a small set of docs
- power law graph
- Zipf's law for tags, starting at #100
- distribution for docs (us) similar to findings for users (others)



Authors vs. Readers

Authors vs. Readers

Matching metadata of authors and readers

- authors: title, keywords, description, body
  + <combined>
- readers: tags

Authors vs. Readers

Matching metadata of authors and readers

- authors: title, keywords, description, body
  + <combined>
- readers: tags

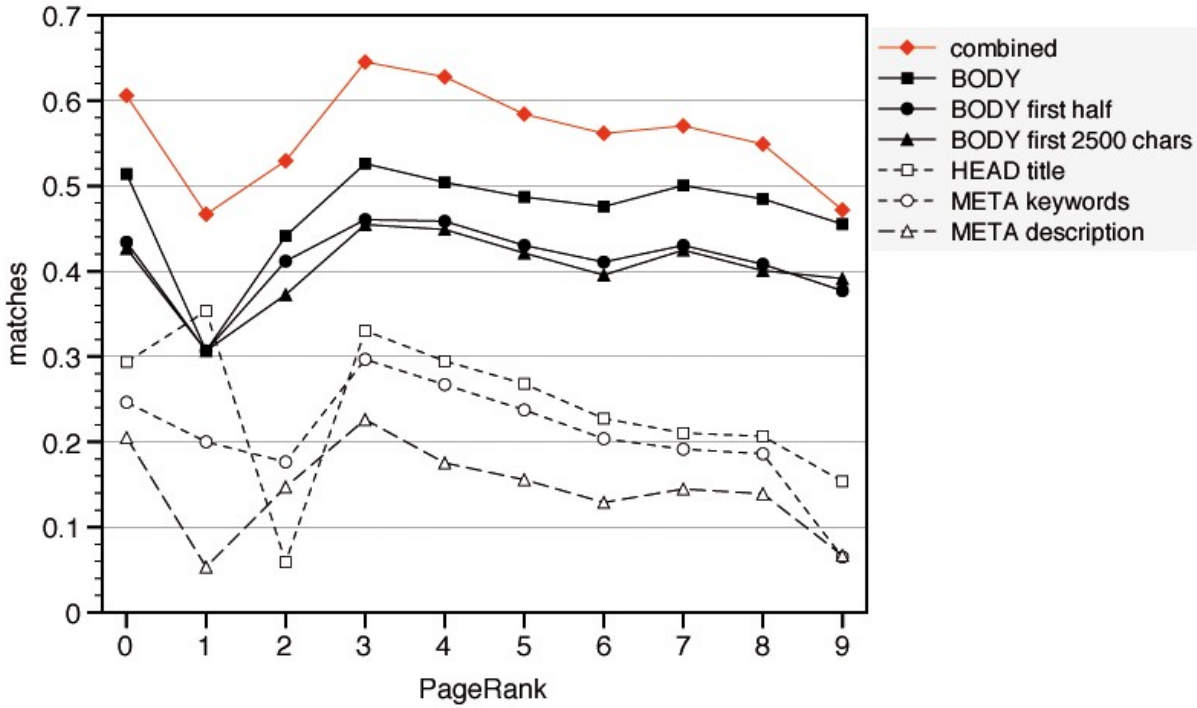
Preprocessing improved performance

- e.g. 46% → 58% matches for <combined>

Authors vs. Readers

Results

- body >> HTML metadata
- parts of body can be sufficient
- relatively stable for body, decreasing for metadata with higher PR



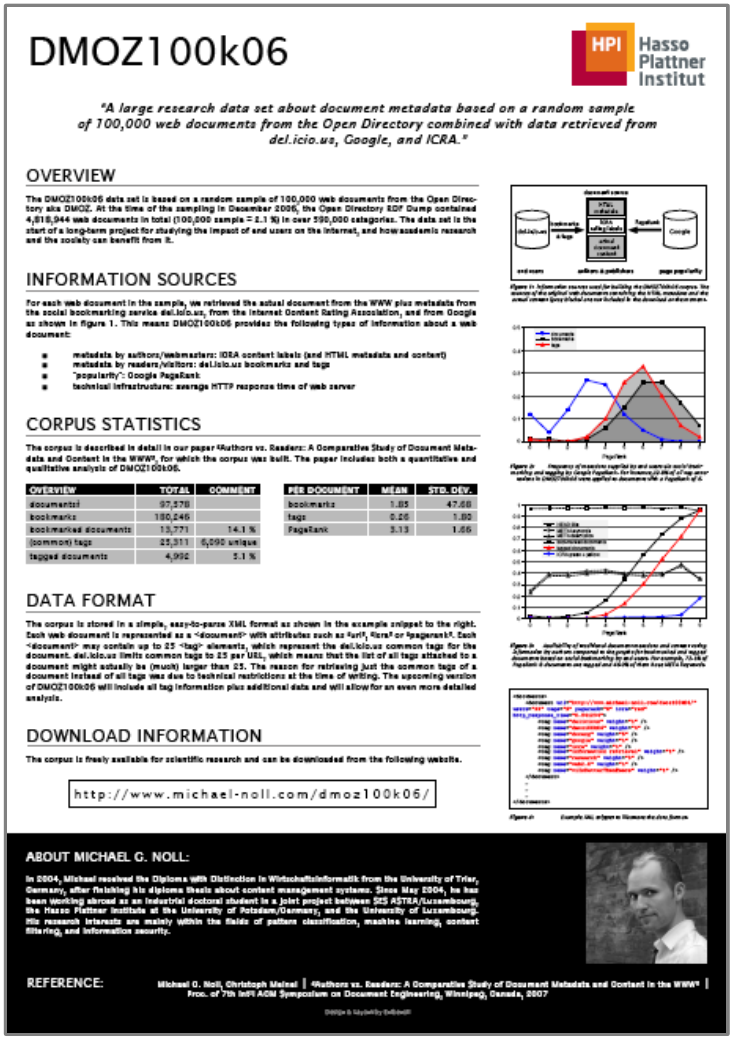
Authors vs. Readers

- the more popular a document, the less does its metadata reflect the perception of users
- user annotations provide additional information not available in a document itself (+ authors)
- keywords (23%) > description (15%), opposite of what search engines prefer

Summary and conclusion

- will not repeat results from previous slides :-)
- tags provide additional information which is not contained in a document itself
  - good: helps information retrieval and classification
  - bad: concentration on a relatively small subset of docs, but techniques such as PEBL can help
- upcoming DMOZ100k06 corpus will have much more information: *all* user annotations + more

DMOZ100k06



DMOZ100k06

HPI Hasso Plattner Institut

"A large research data set about document metadata based on a random sample of 100,000 web documents from the Open Directory combined with data retrieved from del.icio.us, Google, and ICRA."

OVERVIEW

The DMOZ100k06 data set is based on a random sample of 100,000 web documents from the Open Directory aka DMOZ. At the time of the sampling in December 2006, the Open Directory RDF Dump contained 4,818,944 web documents in total (100,000 sample ≈ 2.1 %) in over 590,000 categories. The data set is the start of a long-term project for studying the impact of end users on the Internet, and how academic research and the society can benefit from it.

INFORMATION SOURCES

For each web document in the sample, we retrieved the actual document from the WWW plus metadata from the social bookmarking service del.icio.us, from the Internet Content Rating Association, and from Google as shown in figure 1. This means DMOZ100k06 provides the following types of information about a web document:

- metadata by authors/webmasters: ICRA content labels (and HTML metadata and content)
- metadata by readers/visitors: del.icio.us bookmarks and tags
- "popularity": Google PageRank
- technical infrastructure: average HTTP response time of web server

CORPUS STATISTICS

The corpus is described in detail in our paper "Authors vs. Readers: A Comparative Study of Document Metadata and Content in the WWW", for which the corpus was built. The paper includes both a quantitative and qualitative analysis of DMOZ100k06.

OVERVIEW	TOTAL	COMMENT
documents	97,576	
bookmarks	180,246	
bookmarked documents	13,771	14.1 %
(common) tags	23,311	5,090 unique
tagged documents	4,992	5.1 %

PER DOCUMENT	MEAN	STD. DEV.
bookmarks	1.85	47.66
tags	0.26	1.80
PageRank	3.13	1.66

DATA FORMAT

The corpus is stored in a simple, easy-to-parse XML format as shown in the example snippet to the right. Each web document is represented as a "document" with attributes such as "url", "icra" or "pagerank". Each "document" may contain up to 25 "tag" elements, which represent the del.icio.us common tags for the document. del.icio.us limits common tags to 25 per URL, which means that the list of all tags attached to a document might actually be (much) larger than 25. The reason for retrieving just the common tags of a document instead of all tags was due to technical restrictions at the time of writing. The upcoming version of DMOZ100k06 will include all tag information plus additional data and will allow for an even more detailed analysis.

DOWNLOAD INFORMATION

The corpus is freely available for scientific research and can be downloaded from the following website.

http://www.michael-noll.com/dmoz100k06/

ABOUT MICHAEL G. NOLL:

In 2004, Michael received the Diploma with Distinction in Wirtschaftsinformatik from the University of Trier, Germany, after finishing his diploma thesis about content management systems. Since May 2004, he has been working abroad as an industrial doctoral student in a joint project between SES ASTRA/Luxembourg, the Hasso Plattner Institute at the University of Potsdam/Germany, and the University of Luxembourg. His research interests are mainly within the fields of pattern classification, machine learning, content filtering, and information security.

REFERENCE:

Michael G. Noll, Christoph Meinel | "Authors vs. Readers: A Comparative Study of Document Metadata and Content in the WWW" | Proc. of 7th Int'l ACM Symposium on Document Engineering, Winnipeg, Canada, 2007