Exploring the Academic Invisible Web

Das wissenschaftliche Invisible Web erkunden

Dr. Dirk Lewandowski
Heinrich-Heine-Universität Düsseldorf,
Information Science

Research done in collaboration with Philipp Mayr, Bonn

Agenda

1. Introduction
2. The (Academic) Invisible Web defined
3. The size of the (Academic) Invisible Web
4. AIW relevant to...
5. Opening the AIW – different models

1 Introduction

- Users expect their search services to be comprehensive and integrated.
- Up-to-dateness and completeness are important factors in research.

2 The Invisible Web defined

Definitions for Invisible/Deep Web

- "Text pages, files, or other often high-quality authoritative **information** available via the World Wide Web **that general-purpose search engines cannot**, due to technical limitations, or will not, due to deliberate choice, **add to their indices** of Web pages" (Sherman u. Price 2001).

- "The deep Web - those pages do not exist until they are created dynamically as the result of a specific search" (Bergman 2001).

Type of Invisible Web Content	Why It's Invisible
Disconnected page	No links for crawlers to find the page
Pages consisting primarily of images, audio, or video	Insufficient text for the search engine to "understand" what the page is about
Pages consisting primarily of PDF or Postscript, Flash, Shockwave, Executables (programs) or Compressed files (.zip, .tar, etc.)	Technically indexable, but usually ignored, primarily for business or policy reasons
Content in relational databases	Crawlers can't fill out required fields in interactive forms
Real-time content	Ephemeral data; huge quantities; rapidly changing information
Dynamically generated content	Customized content is irrelevant for most searchers; fear of "spider traps"

Sherman u. Price 2001

From the Invisible Web to the Academic Invisible Web

- Nowadays, the IW problem is mainly the problem with the contents of databases.
- For the academic sector, sources from the surface Web are relevant as well as sources from the Invisible Web.
- The Academic Invisible Web (AIW) consists of the databases relevant to academia.
- Or narrower: The AIW consists of the databases that libraries should index (using search engine technology).

3 The size of the Invisible Web

Bergman's calculation

- Average size of IW databases:
  - 5,43 million documents (mean)
  - 4.950 documents (median)
- Total size:

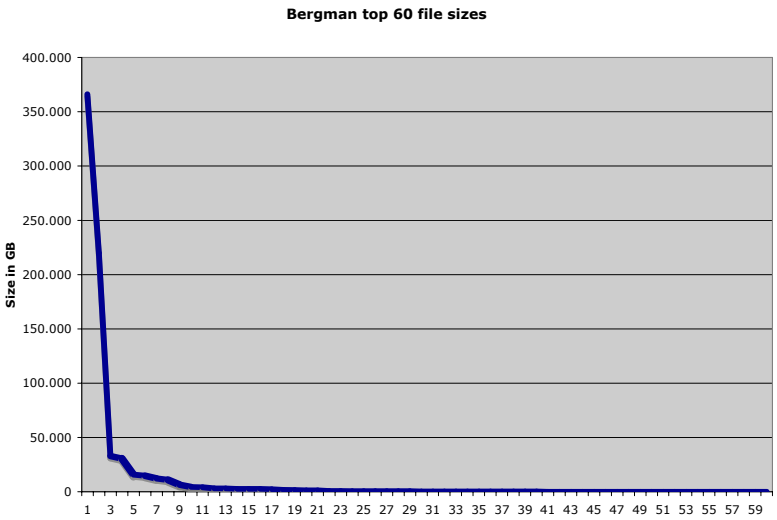
  100.000 databases

  * 5,43 Mio. documents

  = total of **543 billion documents**.
- Size of the surface Web: 1 billion documents (2001).

→ The Invisible/Deep Web is **550 times larger** than the surface Web.

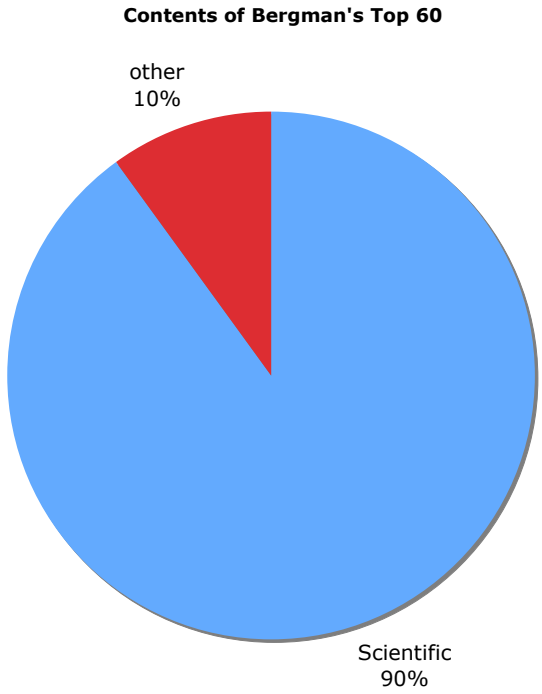
Bergman's calculation

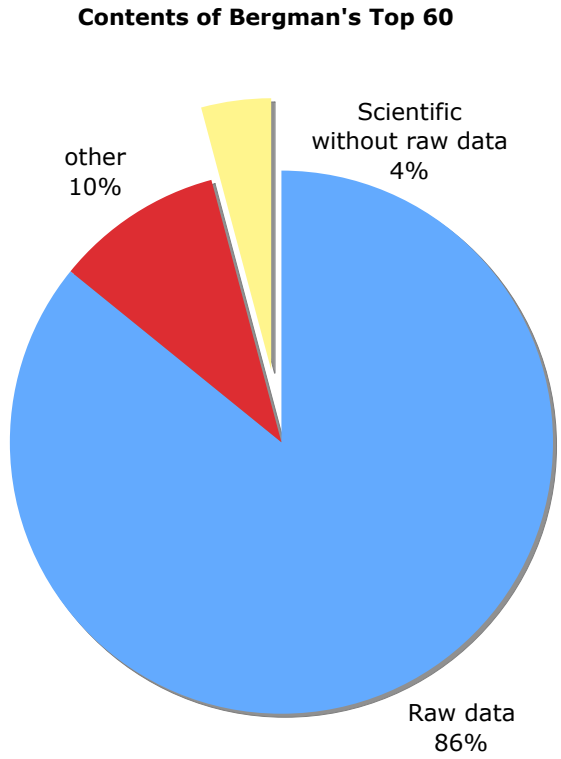


But:

- Use of the mean, although distribution of sizes is highly skewed.
  - 5,43 million documents (mean)
  - 4.950 documents (median)
- Top60 contain 85 billion documents, 748.504 GB.
- Top2 contain 585.400 GB (>75% of Top60).

Contents of Bergman’s Top 60





Basis: Database sizes in GB

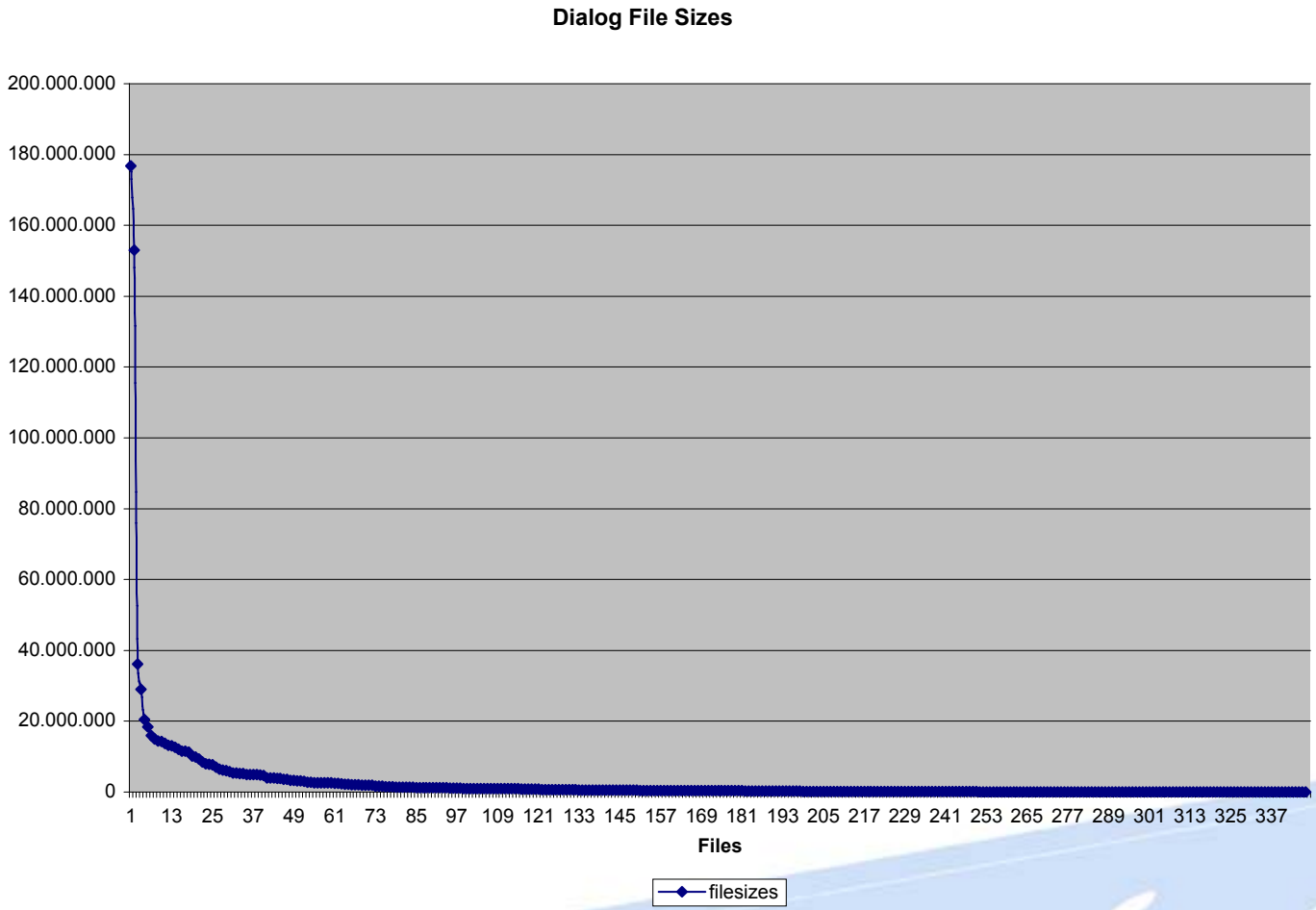
Summary Bergman criticism

- Database selection
  - Database types
  - Database content
- Calculation

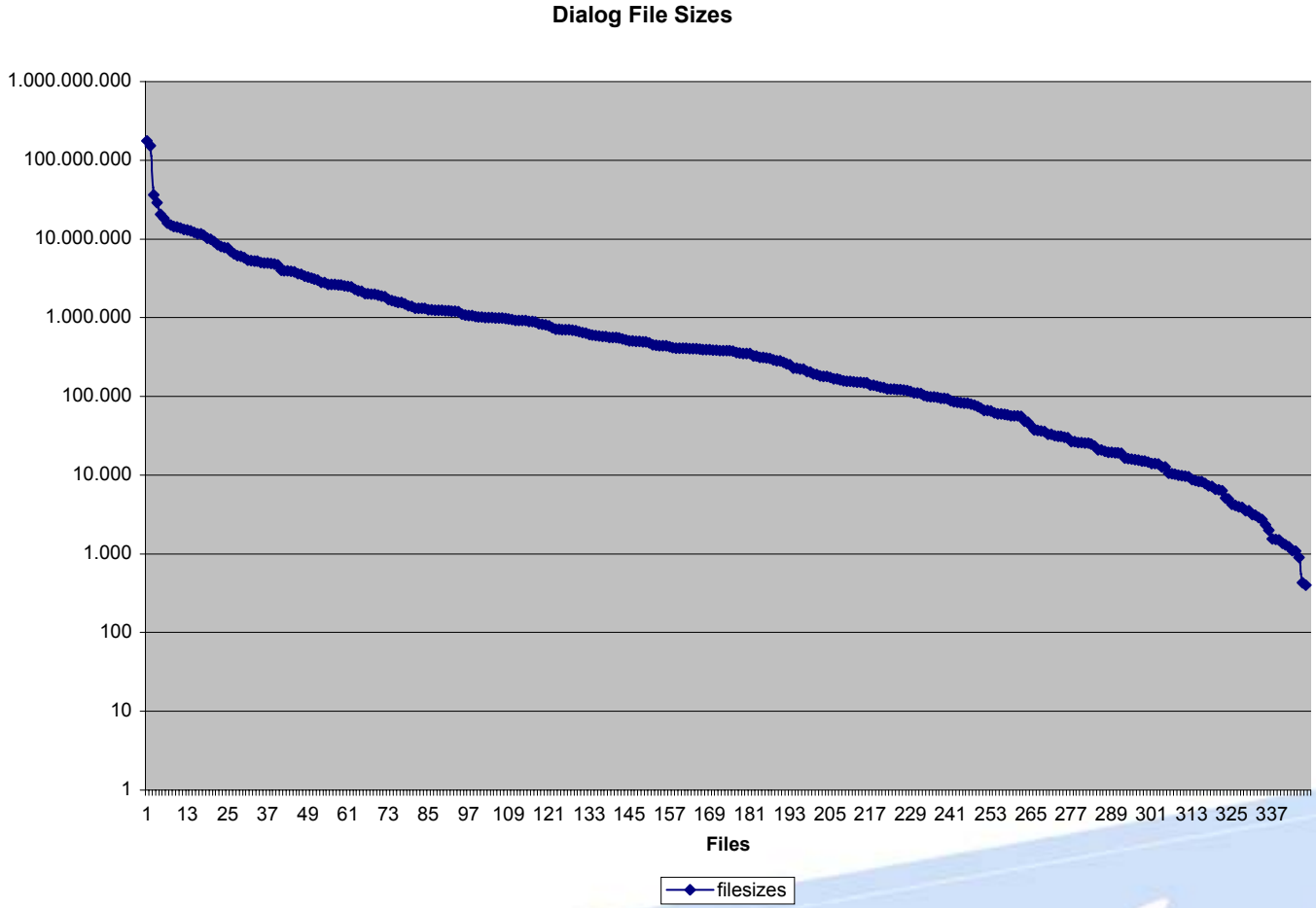
Size comparison:
Gale Directory of Databases

- Contains approx. 16.000 databases (2003); covers all major academic databases.
- Total size estimate for all databases: **18,55 billion documents** (includes CD-ROM databases).
- Estimate is based on less than 10 percent of all databases.
- 5 percent of all databases contain >1 million documents, some more than 100 million.
- Some of the databases included in Bergman's top 60 are missing in Gale.

Will AIW show also an exponential distribution?



Will AIW show also an exponential distribution?



Conclusion: Size of the Invisible Web

- Bergman's size of 550 billion documents is highly overestimated.
- An exact calculation from the distribution of Bergman's top 60 is not possible.
- The size estimate from Gale directory includes databases beyond the web, but does not include all web databases.
- The estimate from Gale is probably too low.

4 AIW relevant for scholars, searchers, librarians, information professionals

4 AIW relevant for scholars, searchers, librarians, information professionals

- Everything relevant for the scientific process
  - Literature (articles, dissertations, reports, books, …)
  - Data
  - Pure Online content (e.g. OA)
- Providers of AIW content
  - Database vendors (meta data) + human indexing
  - Library content (OPACs, collections) + human indexing
  - Publishers content (full text) + mixed indexing
  - Other repositories
- A lot of these materials are not necessarily AIW, but in fact uncovered by the main search engines and tools.

5 Opening the AIW – different models

- Commercial search engines
  - Google Scholar
  - Scirus
- Libraries & database vendors
  - BASE (Bielefeld Academic Search Engine)
  - Vascoda (Integration of library and database collections)
- Open Access repositories
  - Citebase
  - OpenROAR

Conclusion

Summary

- Existing search tools and approaches show potential to make AIW visible
- All protagonists should work together
  - Commercial search engine providers with their machine and financing power
  - Librarians with their experience in collection building and subject access (e.g. thesauri, classification, taxonomies)
  - Publishers and database vendors via opening their collections

Future research

- Building an AIW sample for further tests.
- Better size estimates from this sample.
- Classification of AIW content.
- Distinction between Academic Surface Web and AIW.

HEINRICH HEINE
UNIVERSITÄT
DÜSSELDORF

Vielen Dank.

dirk.lewandowski@uni-duesseldorf.de
www.durchdenken.de/lewandowski

References

- Bergman, M.K. (2001). The Deep Web: Surfacing Hidden Value. Journal of Electronic Pub-lishing, 7(1).
- Sherman, C., & Price, G. (2001). The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today.