The iSearch test collection
- an Information Retrieval benchmark with citations

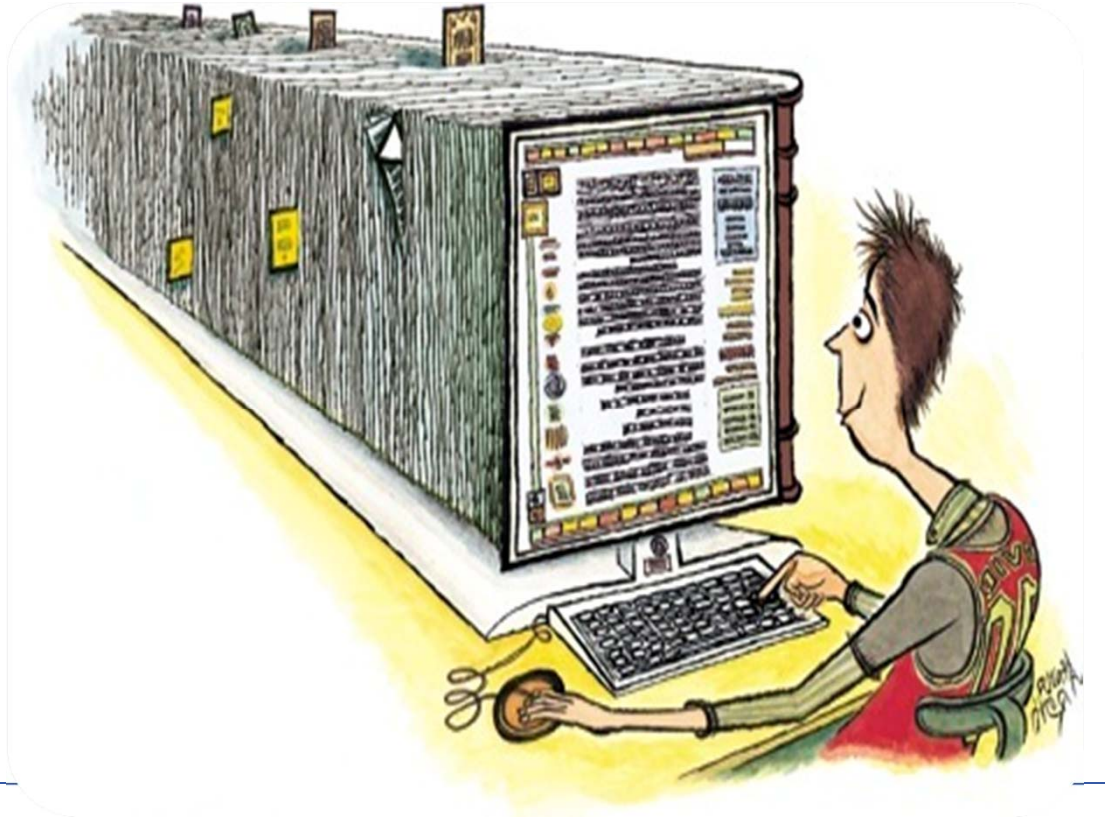
Birger Larsen and the iSearch team
Royal School of Library and Information Science
University of Copenhagen

blar@iva.dk - www.iva.dk/blar

http://itlab.dbit.dk/~isearch

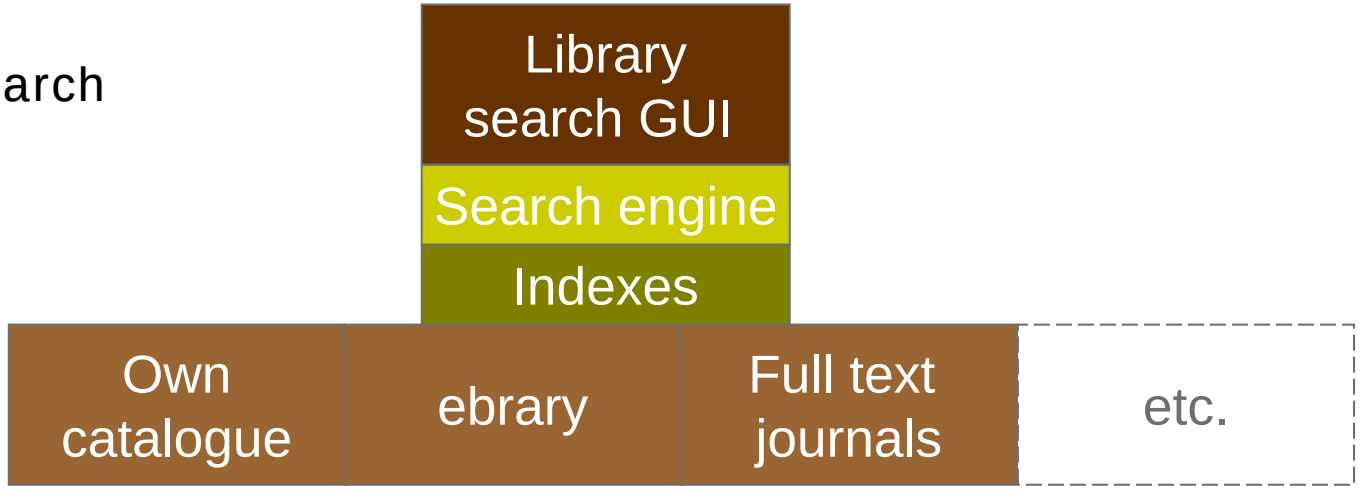
Why is Google so easy and the library so hard?
(Claire Duddy - student)

United Kingdom Serials Group 2009 Annual conference



What is Integrated Search?

Integrated search scenario



- BUT how to test existing and develop new integrated search solutions (= IR research)?
  → build a **test collection**
- Test collection = Information Retrieval benchmark
  1. **Documents**
  2. **Topics**
  3. **Relevance assessments**

Integrated Search Test Collection

DEff Denmark's Electronic Research Library

Co-funded project with *Denmark's Electronic Research Library*

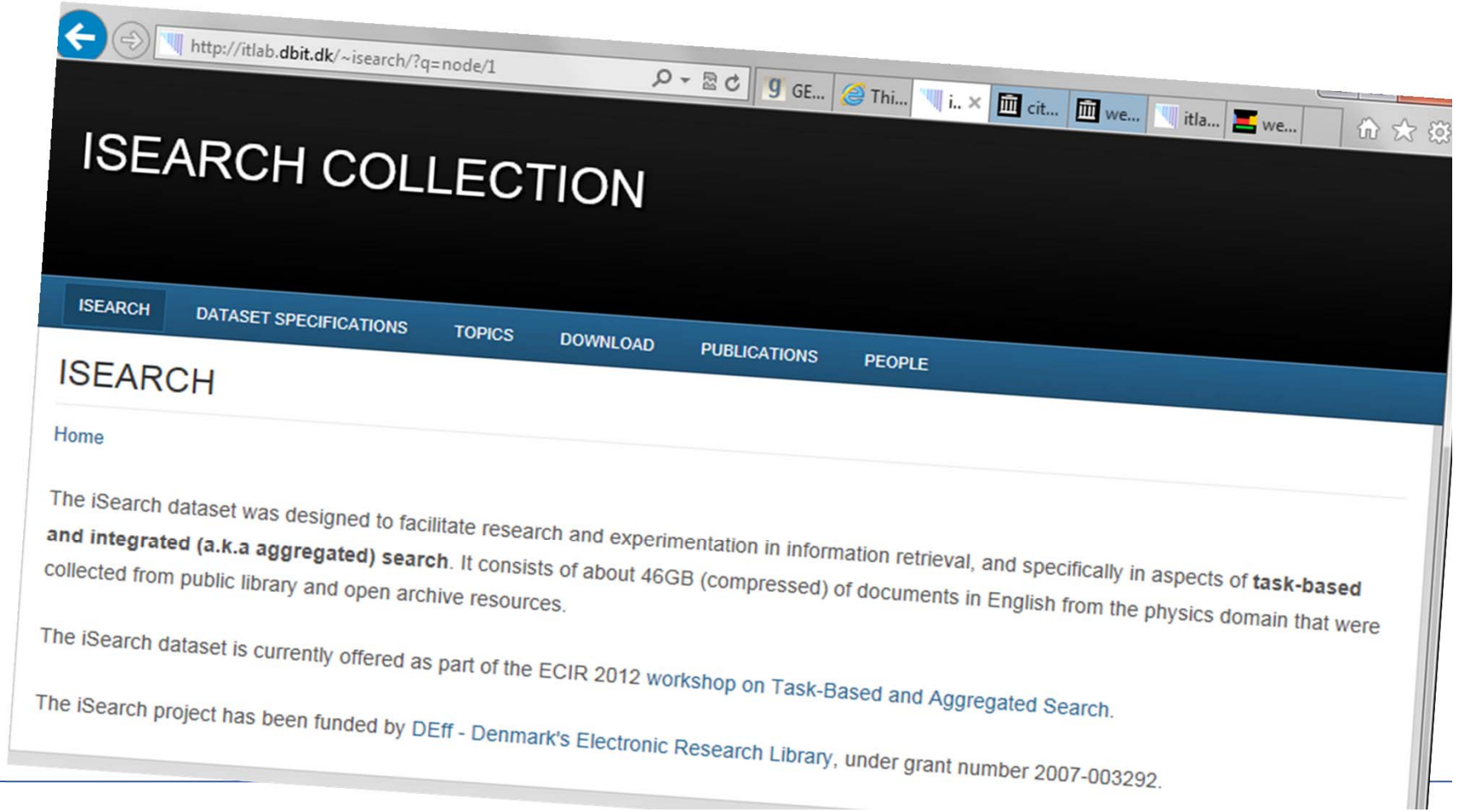
- Research team: Marianne Lykke, Birger Larsen, Haakon Lund, Toine Bogers, Peter Ingwersen & Christina Lioma

Goals

- to **create a test collection** of **scholarly documents** that facilitates the design and test of integrated search IR models, e.g.,
  - Which standard IR models perform best for this task?
  - Is parameter tuning sufficient to avoid over-emphasis of some document types (e.g., full text), or do special measures need to be taken?
- to base topics on **realistic information needs** and to obtain **relevance assessments** from actual users
- *(secretly: **include citation network**)*

iSearch is freely available

Collection released November 2011 – see
http://itlab.dbit.dk/~isearch for how to obtain it



1. Domain and document subsets

Physics chosen as domain

- Availability of documents because of self archiving in open access repositories and e-print archives
  - arXiv.org - 850,000+ documents (metadata + full text)
- Complex and specific information needs
- Sufficiently large research field from which to recruit topic authors

Document subset extracted (453,254 in total)

- **143,569** (32%) arXiv.org **full text PDF** ePrints + metadata
- **291,244** (64%) arXiv.org e-print **metadata** (title, authors, subject, source, abstract)
- **18,441** (4%) **book records** (title, authors, subject, source)

Citations from citebase.org

Approx. 47 GB (6.2 GB if text extracted from PDFs)

1. Domain and document subsets

Citation data from www.citebase.org (now defunct) developed by Tim Brody, University of Southampton, UK

- "semi-autonomous citation index, harvests pre/post-prints from free OAI-PMH compliant archives, parses and links their references"
- For 378,147 of the arxiv.org documents in iSearch references could be extracted
  - 12,727,716 references (33,6 per paper)
  - 3,768,410 are linked to iSearch documents (= internal citations; 10 per paper)
- Rich citation network for experimentation

2. Topics (= information needs)

23 physics master's and PhD students + lecturers recruited

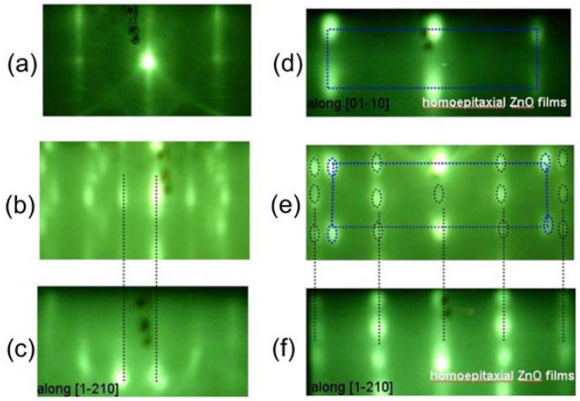
Created **65 topics** based on own tasks

Thorough descriptions in five fields based on user studies:

- Which central search terms would you use to express your situation and information need? ← **Keywords**
- What are you looking for? ← **current Information Need**
- Why are you looking for this? ← **underlying Work Task**
- What is your background knowledge of this topic? ← **current knowledge state**
- What should an **Ideal Answer** contain to solve your problem or task?

Formed the basis for relevance assessments

Example: iSearch topic no. 49



1. **Keywords**: ZnO, rf magnetron sputtering, photo luminescence, al doped, green luminescence
2. **Information Need**: Information on characterization by photo luminescence of highly doped ZnO films
3. **Work Task**: For my master thesis I work with characterization of ZnO films by photo luminescence. The films are manufactured by RF magnetron sputtering and have thicknesses of approximately 100 nm. The films are either intrinsic or doped with Al. Green luminescence are of particular interest, but other defect modes are also of interest. The aim is to document a simple way of characterizing films in a non intrusive manor, and maybe to implement the technique in the production to monitor film growth. In particular information on sub band gab excitation is interesting as only a 405 nm laser is readily available at the institute
4. **Background**: I have worked with the topic for a year and a half. We have made experiments with photo luminescence and have observed green luminescence. I have read quite a lot of review articles on the subject and have been seeking articles with comparable parameters
5. **Ideal Answer**: An article containing examples of luminescence from samples made by rf magnetron sputtering. Graphs with photoluminescence data from ZnO films are essential. Ideally Al doped ZnO films would be featured in the article

Relevance Assessment

Registration of relevance assessments for documents retrieved for physics search scenarios

Topic authors agreed to assess up to **200 documents** per topic

These were identified through iterative subject searches by the research team

- Similar to searches by information specialists (using document fields, Boolean combinations, classification codes etc. in *Lucene*)
- Separate searches adapted to each document subset
- Proportional to the corpus distribution where possible

Assessments collected through online interface

- Graded relevance:
  Highly, **Fairly**, **Marginally** + **Non-relevant**
- Additional background and satisfaction data collected through questionnaires

Summary

- iSearch is a **freely available** IR benchmark test collection
- It contains a variety of document types and representations
  - Including a rich citation network
- It includes topics (information needs) and relevance assessments

➢ iSearch is thus one of the few test collections that support information retrieval experiments both with citation networks as well as full text

➢ Not a very large collection, can we build bigger ones from *CiteSeerX*, *NASA ADS*, *PMC Open Access Subset*?


Thank you!
Questions?

References and links

iSearch test collection website: http://itlab.dbit.dk/~isearch

Brody, T. (2006): **Evaluating Research Impact through Open Access to Scholarly Communication**. *University of Southampton, Electronics and Computer Science, Doctoral Thesis. Available:* http://eprints.soton.ac.uk/263313/1.hasCoversheetVersion/brody.pdf

Lykke, M., Larsen, B., Lund, H., Ingwersen, P. (2010): **Developing a Test Collection for the Evaluation of Integrated Search**. *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010, Proceedings*. Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S. & van Rijsbergen, K. (ed.). 5993. Springer s. 627-630. (Lecture Notes in Computer Science; 5993).

Task Based and Aggregated Search Workshop at ECIR 2012 – with several participants using iSearch: http://itlab.dbit.dk/~tbas2012/

Acknowledgments

DEff - Denmark's Electronic Research Library (grant number 2007-003292)

Tim Brody, University of Southampton, UK