Royal School of Library and Information Science

# Adding user context to IR test collections


Birger Larsen

Information Systems and Interaction Design

Royal School of Library and Information Science

Copenhagen, Denmark

blar @ iva.dk

# Outline


- RSLIS and ISID
- A cognitive view on information needs
- iSearch – a test collection for task-based and integrated search

# RSLIS

Royal School of Library and Information Science

- A small higher education institution under the Danish Ministry of Culture
  - Full university status 2000
- BSc, MSc and PhD programmes
  - 200 BSc students accepted a year
  - 75 MSc students accepted a year
    - 1 degree course taught in English
  - Continuing education courses
- Two locations
  - Copenhagen and Aalborg
- 40 faculty members

# Research @ RSLIS

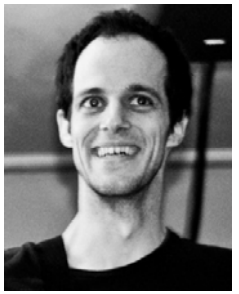
Royal School of Library and Information Science

- Seven research groups since April 2010 (including cultural mediation, library and knowledge management, research analysis and bibliometrics)
- Information Systems and Interaction Design
  - Interaction between users, information and systems in given contexts
  - Research in theories, models and tools for design, construction and evaluation of interactive and motivated solution for access to information
  - 9 faculty members
- RSLIS ITlab
  - OpenLab for students (10 desktops)
  - USElab (4 powerful desktops; Tobii EyeTracker)
  - ITlab (5 racked linux servers; 4 power desktops; 20 TB storage)

*Peter Ingwersen*

*Birger Larsen*

*Haakon Lund*



*Toine Bogers*

# ISID research interests

Royal School of Library and Information Science

- Task-based IR and domain specific IR
- Aggregated and integrated search
- Citation analysis and bibliometrics for IR
- User-based IR and personalisation, including interactive approaches e.g. eye-tracking
- Recommender systems
- Cultural heritage information access and multimedia metadata
- Collaboration on formal approaches IR, e.g. quantum IR and subjective logic for IR

**iSearch**

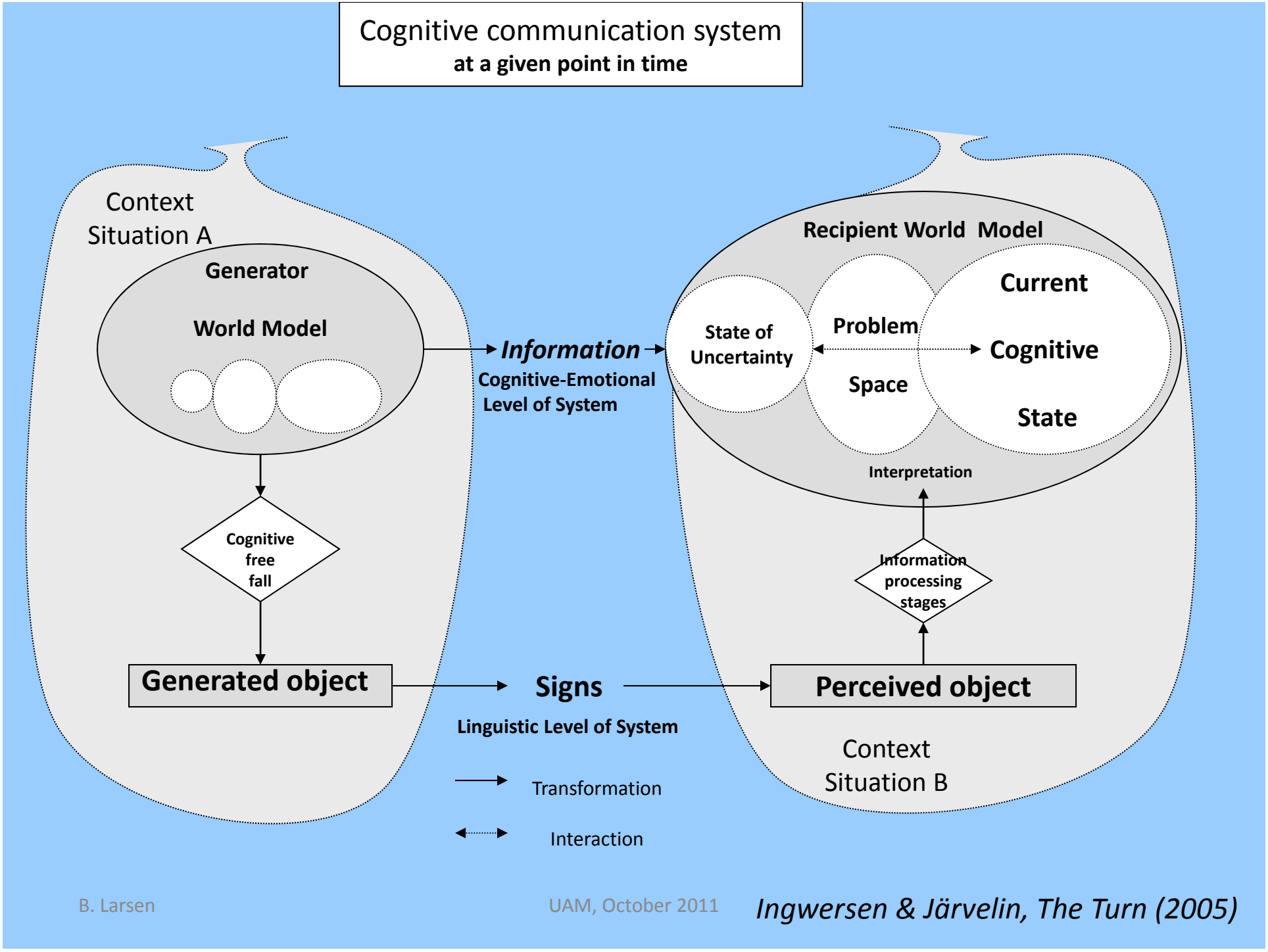
**MUMIA**

**LARM**

**CoSound**

Royal School
of Library and
Information
Science

# A COGNITIVE VIEW ON INFORMATION NEEDS

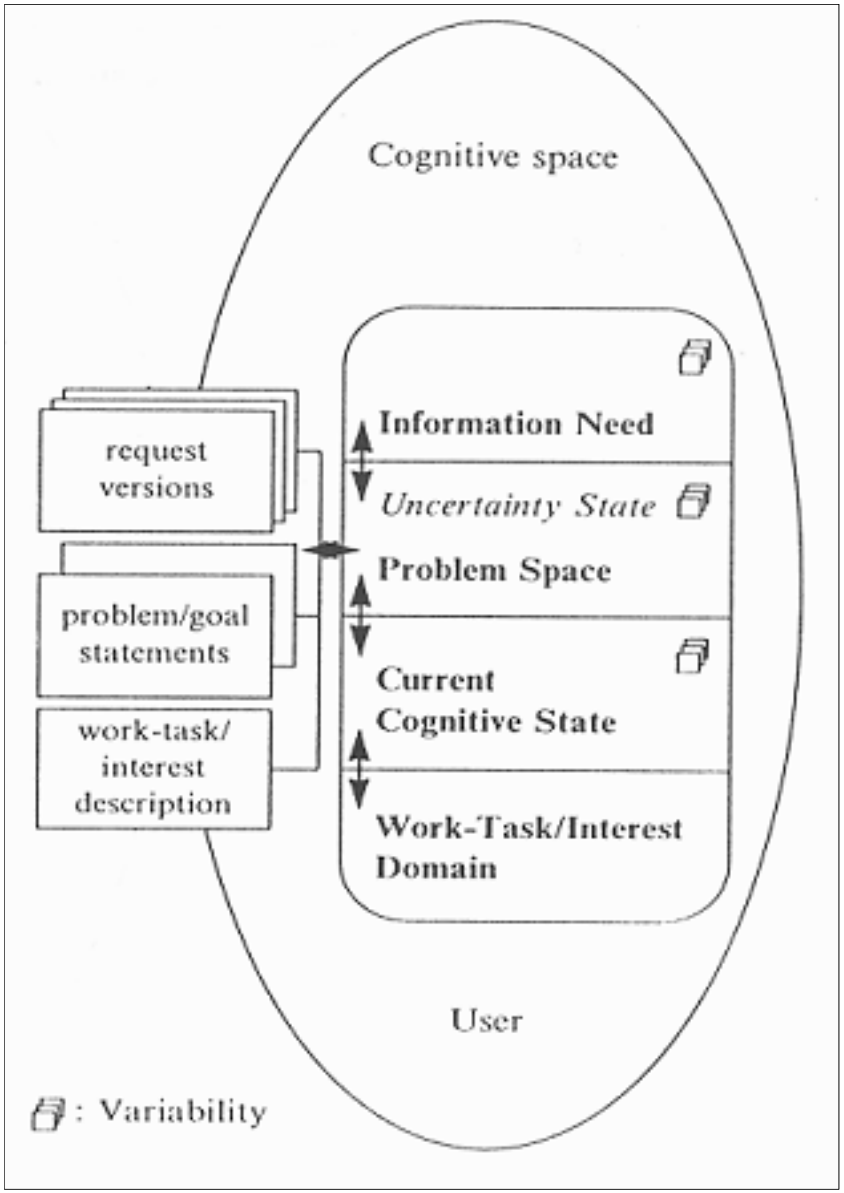




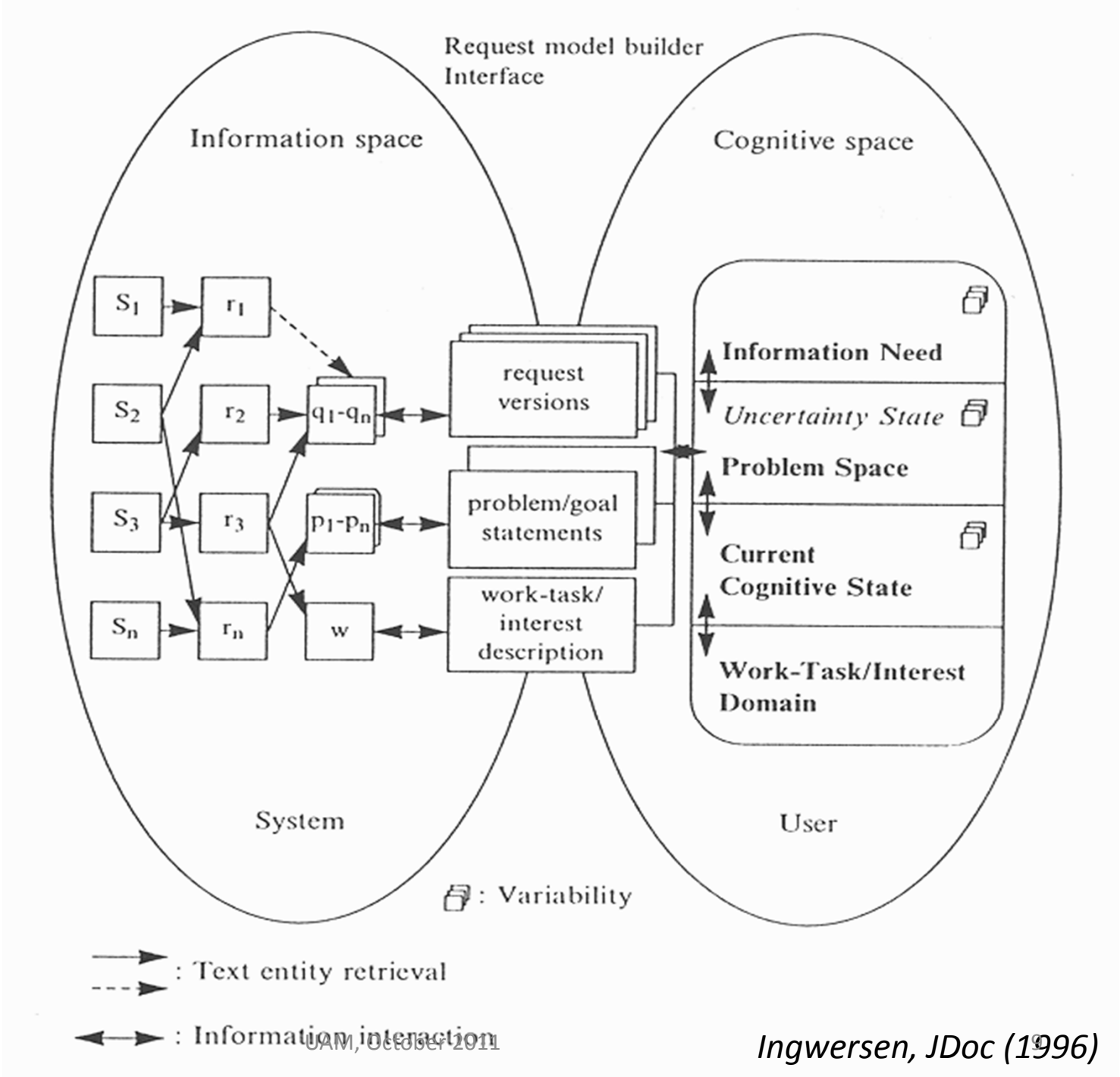

*Ingwersen & Järvelin, The Turn (2005)*

# Information needs

- Queries have a background and a context
- *Work tasks* as an important trigger

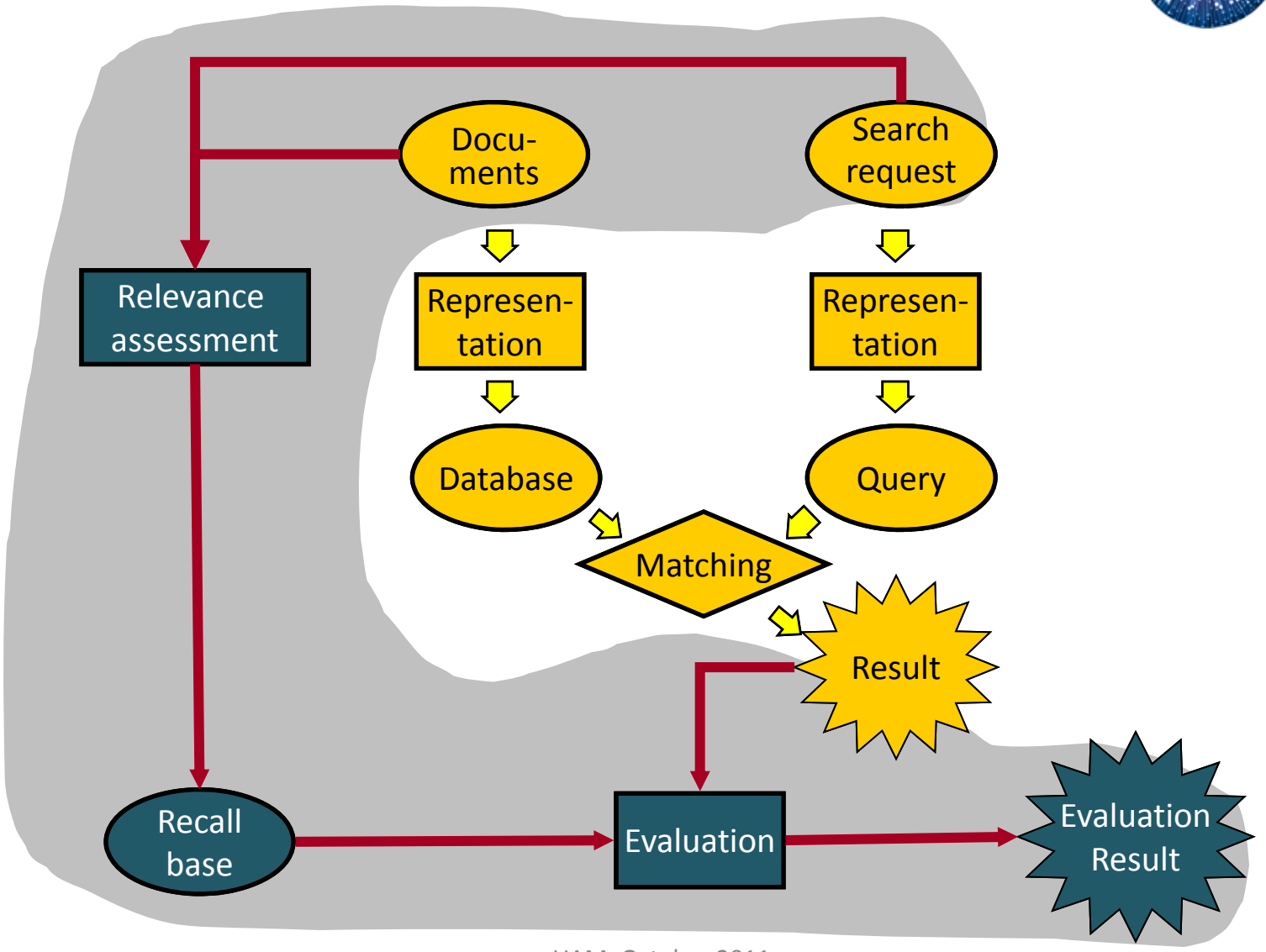



*Ingwersen, JDoc (1996)*




*Ingwersen, JDoc (1996)*

# The Cranfield/TREC IR model




*Ingwersen & Järvelin, The Turn (2005)*

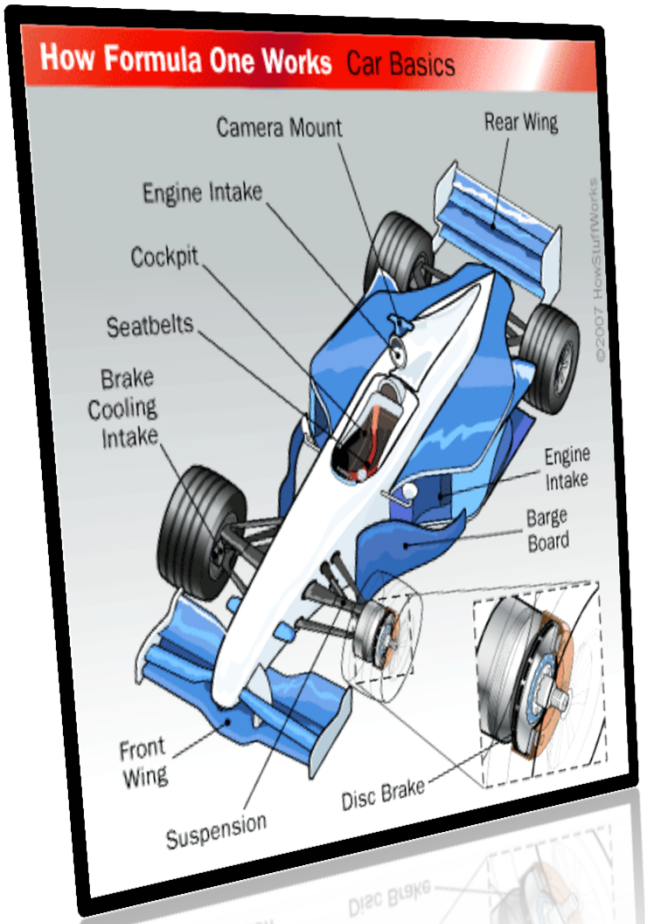
# IR test collections


- Three parts
  1. **Document corpus**
     a collection of documents (text, audio, video, web pages)
  2. **Topics**
     a set of information needs, usually described at several levels (TREC: title; description; narrative)
  3. **Relevance assessments**
     judgements of the relevance of documents for the topics

- IR test collections facilitate
  - Comparison of the performance of different IR models and variants
  - Using a set of standard performance measures, e.g.,
    - Mean Average Precision (MAP): Precision calculated after each relevant retrieved document, then averaged over each topic and mean over all topics
    - Precision @ 10 (P@10): Precision after 10 retrieved documents
    - Cumulated Gain (NDCG): Gain cumulated after each retrieved document; allows comparison to 'ideal' model + use of non-binary relevance assessments

# Information Retrieval (IR) as Formula 1?

Royal School of Library and Information Science




# A cognitive view on information needs

Royal School of Library and Information Science

- Human interaction with information systems including IR systems is a cognitive activity
- Queries and information needs have a background and context that might be important for retrieval
  - Several distinct aspects may be identified
  - Most IR test collections contain only a few of these

- → our work on the iSearch test collection

Royal School of Library and Information Science

# THE ISEARCH TEST COLLECTION

Royal School of Library and Information Science

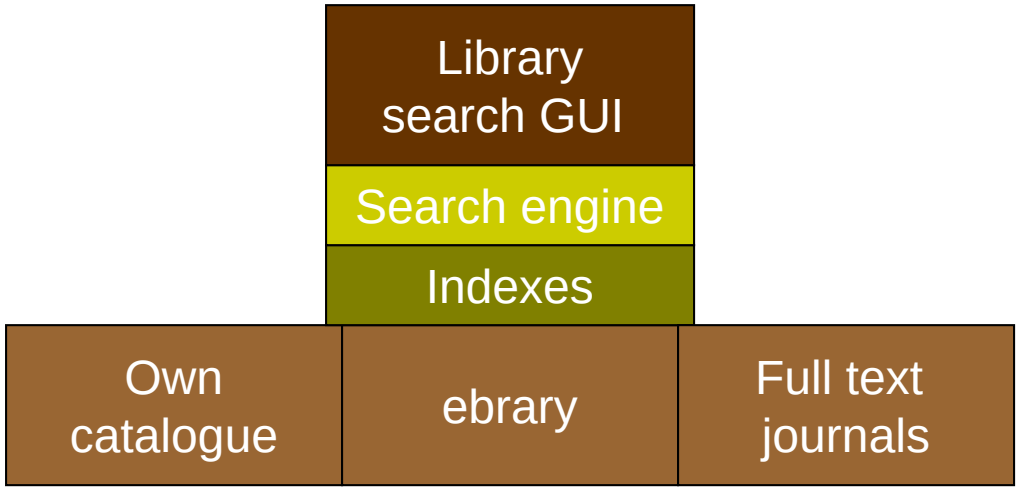
*'Why is Google so easy and the library so hard?'*
(Claire Duddy - student)

*United Kingdom Serials Group*
*2009 Annual conference*
*(Stone, 2010)*

# What is Integrated Search?

Royal School of Library and Information Science

- Integrated search scenario

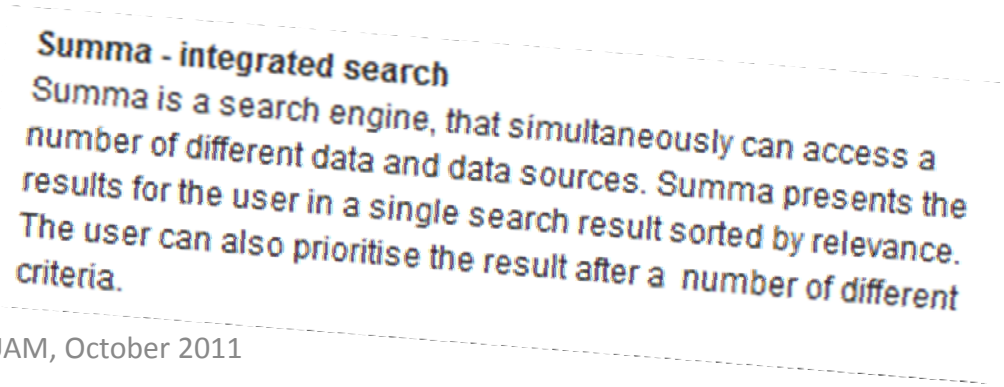



- BUT how to develop and test existing and new integrated search solutions? → build a test collection

# Integrated Search

Royal School of Library and Information Science

- Move from *federated search*
  - search conducted in several disjoint sources and top results presented for each
- ... to *integrated search*
  - Search conducted in one system integrating several sources and document types
  - Same query across all document types, presenting results in one ranked list
- Challenge: different documents types, genres and levels of descriptions
  - For instance, articles, conference papers, books, audio, images, video, web pages
  - Full text, title only, metadata with/without abstracts, citations, links, anchor text, user tagging, intellectual indexing and controlled vocabulary
- Real world digital library need, e.g.,
  - Danish State Library (http://www.statsbiblioteket.dk/summa/features-text-in-english)
  - Similar to universal search in web retrieval



# Integrated Search


- Comparison to standard web retrieval
  - Similarities
    - Heterogeneous document corpus
  - Differences
    - Sources provide different but distinct levels of description
      - For instance, very sparse data for a full book (title, authors, classification codes, controlled and uncontrolled index terms and notes)
      - versus the full text of an article, with self-assigned classification and uncontrolled keywords
    - Conscious effort to include each type in the results regardless of level of description
    - Different scale – cannot rely on large quantity data
- Challenges
  - How to rank and integrate different types of documents?
  - How to exploit different levels and types of descriptions?
  - How to present results?

# Integrated Search Test Collection


- 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 realistic relevance assessments
  - to obtain a realistic and diverse document corpus
    - Focus on textual documents, but other media could be used also

# 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 - 500,000+ 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 e-prints + metadata
  - 291,244 (64%) arXiv.org e-print metadata (title, authors, subject, source, abstract)
  - 18,441 (4%) book records (title, authors, subject, source)
- Approx. 6,2 GB

# Topics

Royal School of Library and Information Science

- 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

# Topics – 5 perspectives

Royal School of Library and Information Science

| Perspective | Question |
|---|---|
| a) Current information need | What are you looking for? |
| b) Work task situation | Why are you looking for this? |
| c) Current knowledge state | What is you background knowledge of this topic? |
| d) Ideal answer | What should am ideal answer contain to solve problem or task? |
| e) Adequate search terms | Which central search terms would you use to express situation and information need? |

*Based on the cognitive view on IR , e.g. Ingwersen & Järvelin (2005)*

# Example: iSearch topic No. 49

Royal School of Library and Information Science

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

iSearch Test Collection

# Collection Statistics

Royal School of Library and Information Science

| Features | Number | Mean no. of words per item |
|---|---|---|
| PDF items, arXiv.org | 160,168 | 4422* |
| Abstracts, arXiv.org | 274,749 | 272* |
| Library Records | 18,222 | 189* |
| References (Citations) | 3,748,555 | (23.4 on avg. for 160,168 PDF items only) |
| Work Task situations | 65 | 104.4* |
| a) Information need | 65 | 17.7* (13 tasks with 5-10 terms) |
| b) Work task context | 65 | 35.7* |
| c) Knowledge state | 65 | 22.2* |
| d) Ideal information | 65 | 19.3* |
| e) Search terms | 65 | 9.4* (20 tasks with 3-6 terms) |

* minus stop words (using the 318 word list at ir.dcs.gla.ac.uk/resources/linguistic_utils/)

*(Lykke et al., 2010)*

# Usage

Royal School of Library and Information Science

- Testing aspects of integrated search
  - How do standard IR models perform for this task?
    - How are relevant documents in different subsets ranked?
    - How does parameter tuning affect each subset?
  - Can better performing integrated models be developed?
    - How to weight or fuse subsets?
    - How to exploit different types and level of description?
- Studying other aspects
  - Can added user context improve retrieval?
  - IR in a restricted highly specialised scientific domain
  - High precision IR using graded relevance
  - Use of citations in IR
  - Theoretical aspects, e.g., effect of topic fields in Polyrepresentation

# Usage

Royal School of Library and Information Science

- Plan to release the collection in fall 2011
- Have proposed an ECIR 2012 workshop on ‘Task-based and aggregated search’
  - iSearch will be provided for participants
  - Facilitate exploratory experiments in both or either aspect

# Wrap up

Royal School of Library and Information Science

- ISID group and research interests
- iSearch
  - with extended user need descriptions
  - Integrated/aggregated search

# References



- Peter Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52:3-50, 1996.
- Peter Ingwersen and Kalvero Järvelin. *The turn: integration of information seeking and retrieval in context*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
- Marianne Lykke, Birger Larsen, Haakon Lund & Peter Ingwersen (2010): Developing a Test Collection for the Evaluation of Integrated Search. In: Proceedings of ECIR 2010*, 32nd European Conference on IR Research*. Berlin: Springer, p. 627-630.
- Graham Stone. Searching life, the universe and everything? The implementation of Summon at the University of Huddersfield. *Library Quarterly*, 20(1):24-52, 2010.

Royal School of Library and Information Science

# Thank you!