# Seven years of INEX interactive retrieval experiments – lessons and challenges

Ragnar Nordlie and Nils Pharo,
Oslo and Akershus University College

# INEX – the initiative for the evaluation of XML retrieval

- Started in 2002
- Benchmarking activity to evaluate the effectiveness of content-oriented XML retrieval systems
- Experimental infrastructure following the TREC model
- Several tracks over the years

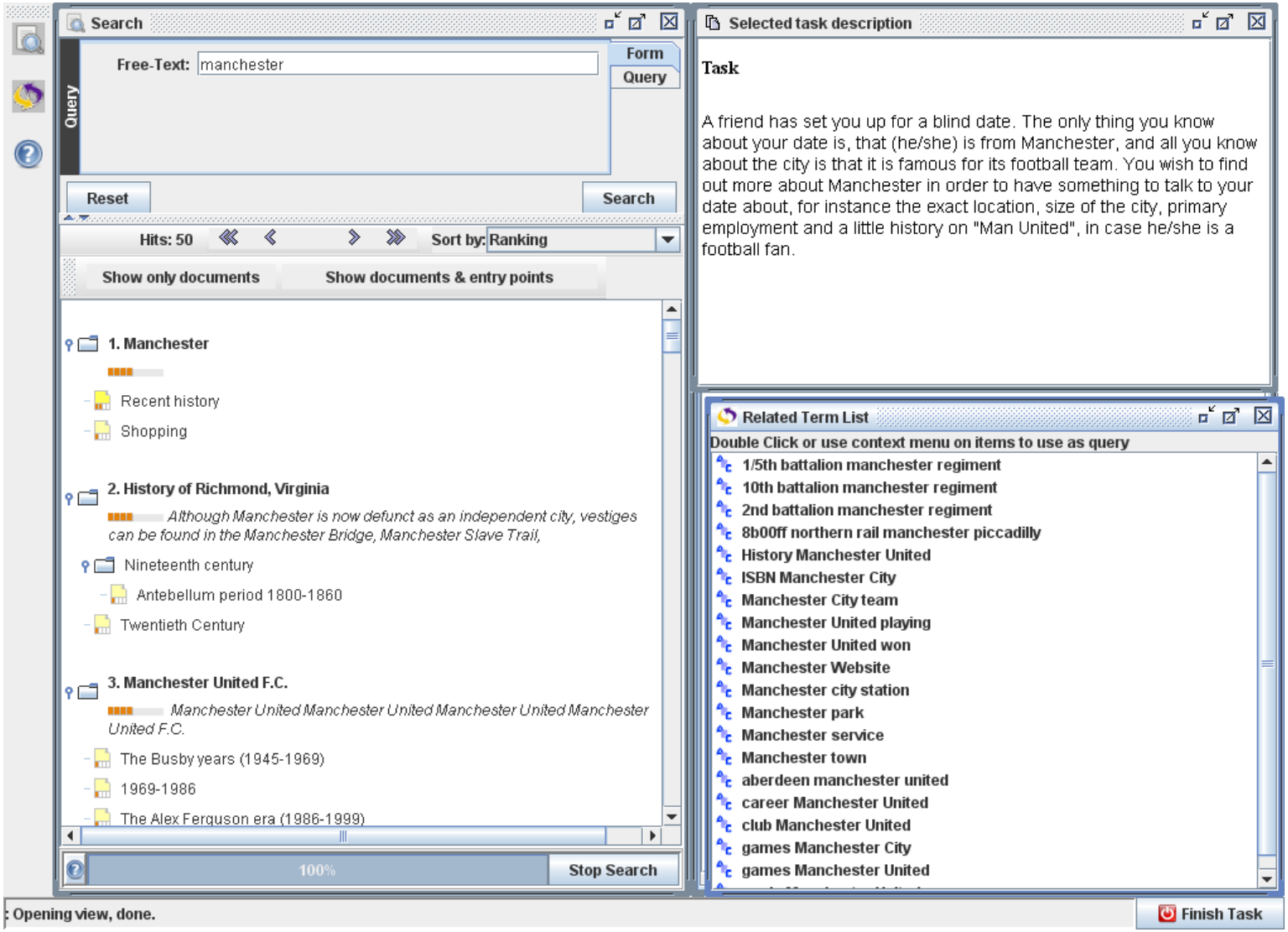
# The INEX interactive experiments

- Cooperative data collection effort
- 7 years/rounds of data collection
- 150-300 individual search sessions each year
- Changes in data corpora, task sets, search interfaces
- Common experimental framework
  - Supervised experiment
  - Common task: to find relevant documents (or parts of documents)
  - Pre and post experiment, pre and post task questionnaires
  - Detailed search logs

# Document corpora

| Year | Collection | Size (items) | Use |
|---|---|---|---|
| 2004-2005 | IEEE journal articles | 12 000 - 16000 | Ad hoc / i-track |
| 2006-2008 | Wikipedia articles | 660 000 | Ad hoc / i-track |
| 2009-2011 | Amazon / LibraryThing | 2. 8 mill. | I-track / book track |

# Search systems and tasks



# Task types used

| Year | Task categories | Tasks per category | Tasks per searcher |
|---|---|---|---|
| 2004 | Background; comparison | 2 | 2 |
| 2005 | General; challenging; own | 3 + own | 3 |
| 2006 | Decision making; fact finding; information gathering (further divided by structure: “hierarchical” and “parallel”) | 4 (2 of each structure) | 4 |
| 2008 | Fact finding; research | 3 | 2 |
| 2009 | Broad; narrow; own | 3 + own | 3 |
| 2010/ 2011 | Explorative; data gathering; own | 3 + own | 3 |

# Searchers' relevance judgements

- Variation of relevance scales have been used
  - 10-point (very / fairly / marginally *useful* in combination with very / fairly / marginally *specific*)
  - Relevant / partly relevant / not relevant
  - Same in combination with "too broad" / "too narrow"
  - Same in combination with "shopping basket"

# Data analysis

- For a large part quantitative analysis
- Interpretation on three levels
  - *Types* of transactions over all searches
    - "how many times is an element of a certain granularity viewed / judged relevant?"
  - *Pattern* of transactions over all searches
    - " where in a session does a certain reading pattern occur?"
  - Behaviour in individual *sessions*
    - "how does search purpose influence transaction patterns?"

# Examples of i-track research questions

- What element types / level of granularity do searchers chose to see? In what sequence?
- How do users make use of document structure
  - in making relevance judgements?
  - in choosing level of granularity to view?
- What level of element granularity constitutes the basis of a relevance decision? With what degree of certainty?
- How do factors such as topic knowledge influence
  - choice of element granularity
  - number of elements viewed / amount read
  - relevance judgements

# Quasi-experiment or experiment?

- INEX i-track effort lacks
  - Controlled selection of subjects
  - Firmly stated research goals, with necessary control of variables (interface features, tasks ….)
- I-track offers
  - An environment to collect a relative large number of recorded sessions without too much burden on each researcher
  - A framework which allows comparison across years and participating institutions to a certain extent
  - Relatively rich information on searcher background etc

# Quasi-experiment or field study?

- I-track lacks
  - Authentic user behaviour
    - Personal interest in result
    - Understanding of task and relevance criteria
    - Individually determined session length
    - Multiple sessions, *search* as part of *seeking*
- I-track has
  - Comparable sessions
  - Access to user background and user experience

# Challenges

- How do we isolate features (of users, interfaces…) which may influence/explain behaviour?
  - Is it task variations / different understandings of the interface / different level of training / different level of interest in the experiment which prompts certain actions to be taken / features to be used?
- How do we create tasks which
  - match actual search tasks
  - are uniformly understandable and not prone to individual interpretation
  - Are sufficiently engaging / challenging
  - Reflect some theoretical division of task types?
- How do we make user-defined tasks comparable?

# Challenges ctd.

- How do we define units of analysis?
  - Actions
  - Sequences of actions (which?)
  - Time
  - Outcomes
- How do we handle factors like reading speed, ability to handle disruptions etc?
- How do we measure relevance / success?

# Challenges ctd.

- How do we collect data which help us understand user actions?
  - Video, screen capture, think aloud, review of session together with user----
- How do we create a database for searching which is both realistic and controllable/ measurable?
- How do we cater for individual researchers' research interest and ensure shareability of data?

# Final challenge

- Is it *possible* and *desirable* to develop research framework(s) for user search experiments to ensure comparability / shareability of results?
- Can the INEX i-track experience assist in this?