# Interactive information retrieval in XML documents

Nils Pharo, Oslo University College

# Outline

- XML retrieval
- INEX
- INEX interactive track
- Recent study:
  - research questions
  - results
  - conclusion

# XML retrieval

```
<?xml version="1.0" encoding="UTF-8 "?>
<book class="H.3.3">
<author>John Smith</author>
<title>XML Retrieval</title>
<chapter>
<heading>Introduction</heading> This text explains all
    about XML and IR. </chapter>
<chapter>
<heading> XML Query Language XQL</heading>
<section> <heading>Examples</heading> </section>
<section> <heading>Syntax</heading> Now we describe
    the XQL syntax. </section>
</chapter>
</book>
```

From: http://woodstock.is.informatik.uni-duisburg.de/~najeeb/doc1.html

# XML-markuped documents

Facilitates

- possibilities to exploit weighting
- the retrieval of document parts

# INEX – the Initiative for the Evaluation of XML Retrieval

- International community whose main goal is ”to promote the evaluation of focused retrieval by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results”
- Since 2002
- Partly financed by DELOS → 2007
- http://www.inex.otago.ac.nz/

# INEX organization

- Test collection
  - Journal articles from IEEE (→2005)
  - Wikipedia articles (2006→)
  - Topics created by participants
  - Document assessed by topic creators
- 186 participating institutions (in 2008)

# INEX tracks

- Ad Hoc
- Book
- Efficiency
- Entity Ranking
- *Interactive (iTrack)*
- Question Answering
- Link-the-Wiki
- XML-Mining

# INEX Interactive track

- investigate end-users interaction with elements of XML documents
- develop approaches for element retrieval which are effective in user-based environments

# Organisation of INEX iTrack

Distributed setup:

- a common subject recruiting procedure
- a common set of user tasks and data collection instruments such as interview guides and questionnaires
- a common logging procedure for user/system interaction

# 2008: two task categories

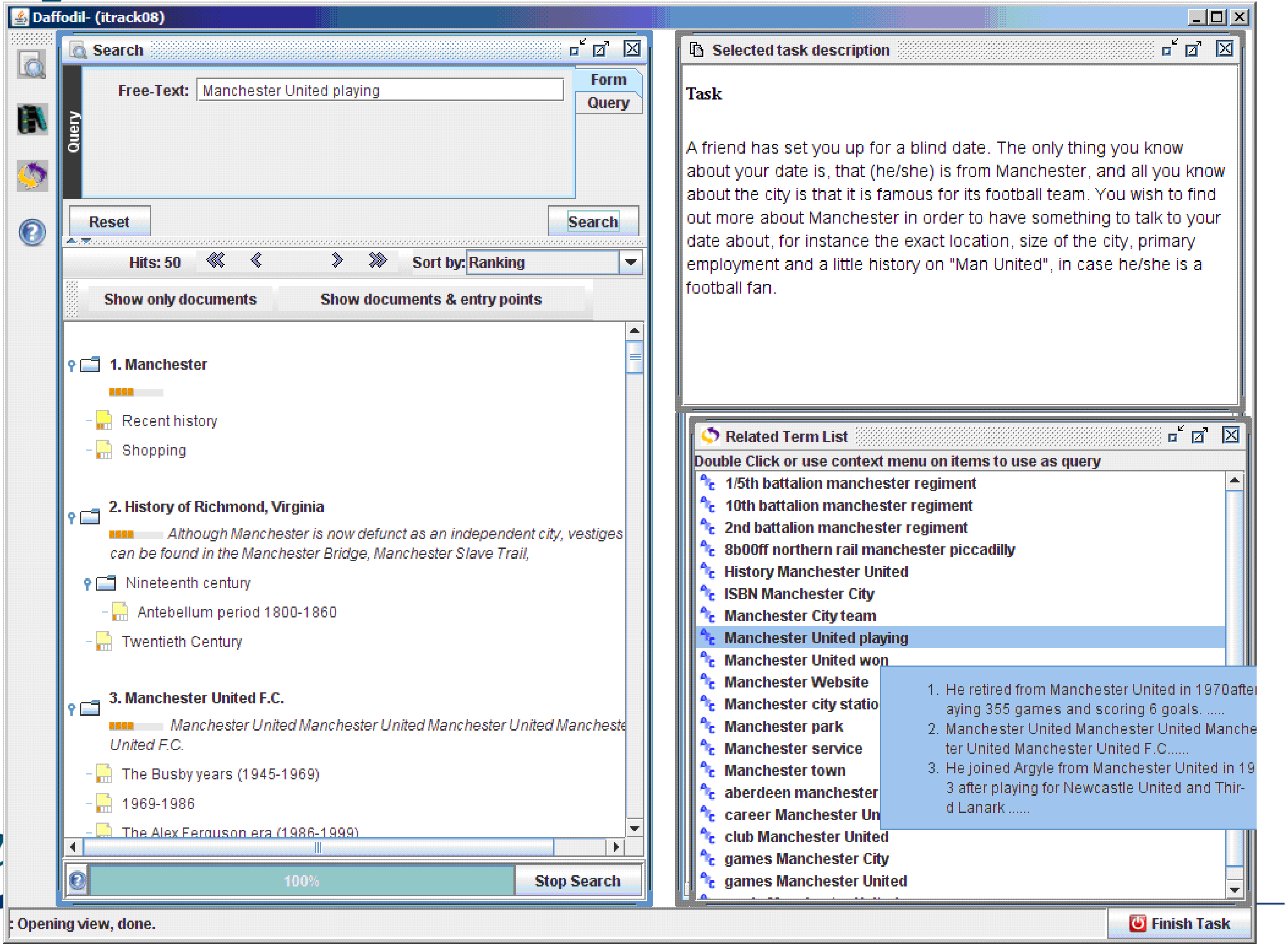
**Fact finding**: ”In the recent Olympics there was a controversy over the age of some of the female gymnasts. You want to know the minimum age for Olympic competitors in gymnastics.”

**Research**: “Every year there are several ranking lists over the best universities in the world. These lists are seldom similar. You are writing an article discussing and comparing the different ranking systems and need information about the different lists and what criteria and factors they use in their ranking.”

# Method

- The searchers perform simulated tasks
- Documents and elements can be relevance judged
- Questionnaires to collect background data such as; demographic data, search experience, topic knowledge

# System



Free-Text: Manchester United playing

Hits: 50 Sort by: Ranking

1. Manchester
- Recent history
- Shopping

2. History of Richmond, Virginia
*Although Manchester is now defunct as an independent city, vestiges can be found in the Manchester Bridge, Manchester Slave Trail,*
- Nineteenth century
  - Antebellum period 1800-1860
- Twentieth Century

3. Manchester United F.C.
*Manchester United Manchester United Manchester United Manchester United F.C.*
- The Busby years (1945-1969)
- 1969-1986
- The Alex Ferguson era (1986-1999)

**Task**

A friend has set you up for a blind date. The only thing you know about your date is, that (he/she) is from Manchester, and all you know about the city is that it is famous for its football team. You wish to find out more about Manchester in order to have something to talk to your date about, for instance the exact location, size of the city, primary employment and a little history on "Man United", in case he/she is a football fan.

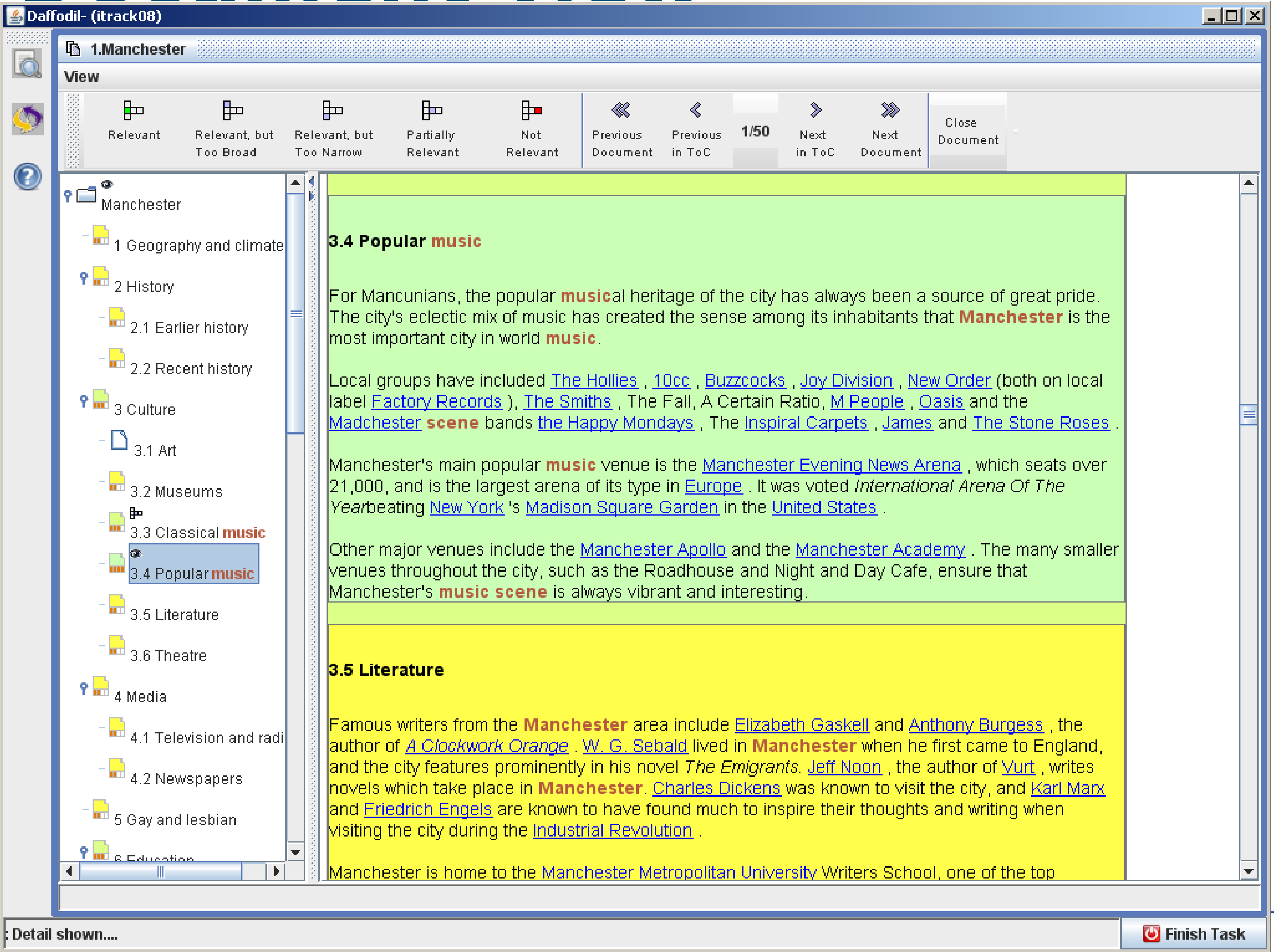
Double Click or use context menu on items to use as query

- 1/5th battalion manchester regiment
- 10th battalion manchester regiment
- 2nd battalion manchester regiment
- 8b00ff northern rail manchester piccadilly
- History Manchester United
- ISBN Manchester City
- Manchester City team
- Manchester United playing
- Manchester United won
- Manchester Website
- Manchester city statio
- Manchester park
- Manchester service
- Manchester town
- aberdeen manchester
- career Manchester Un
- club Manchester United
- games Manchester City
- games Manchester United

1. He retired from Manchester United in 1970after aying 355 games and scoring 6 goals. .....
2. Manchester United Manchester United Manchester United Manchester United F.C......
3. He joined Argyle from Manchester United in 19 3 after playing for Newcastle United and Thir-d Lanark ......

: Opening view, done.

# Document view



**3.4 Popular music**

For Mancunians, the popular musical heritage of the city has always been a source of great pride. The city's eclectic mix of music has created the sense among its inhabitants that Manchester is the most important city in world music.

Local groups have included The Hollies , 10cc , Buzzcocks , Joy Division , New Order (both on local label Factory Records ), The Smiths , The Fall, A Certain Ratio, M People , Oasis and the Madchester scene bands the Happy Mondays , The Inspiral Carpets , James and The Stone Roses .

Manchester's main popular music venue is the Manchester Evening News Arena , which seats over 21,000, and is the largest arena of its type in Europe . It was voted *International Arena Of The Year*beating New York 's Madison Square Garden in the United States .

Other major venues include the Manchester Apollo and the Manchester Academy . The many smaller venues throughout the city, such as the Roadhouse and Night and Day Cafe, ensure that Manchester's music scene is always vibrant and interesting.

**3.5 Literature**

Famous writers from the Manchester area include Elizabeth Gaskell and Anthony Burgess , the author of *A Clockwork Orange* . W. G. Sebald lived in Manchester when he first came to England, and the city features prominently in his novel *The Emigrants*. Jeff Noon , the author of Vurt , writes novels which take place in Manchester. Charles Dickens was known to visit the city, and Karl Marx and Friedrich Engels are known to have found much to inspire their thoughts and writing when visiting the city during the Industrial Revolution .

Manchester is home to the Manchester Metropolitan University Writers School, one of the top

# Example study

- Goal: identify searchers preferred level of granularity
- Compared two genre: journal articles (2005 data) and encyclopedic articles (2006 data)
- Log analysis using Excel

# Result list interaction

| *Granularity level* | *IEEE* | *Wikipedia* |
|---|---|---|
| Level 1/2 (article or metadata) | 71 % | 73.5 % |
| Level 3 (section) | 14.5 % | 15 % |
| Level 4/5 (subsections) | 14.5 % | 11.5 % |
| Total | 100 % | 100 |

# Within documents retrieval

| *Granularity level* | *IEEE* | *Wikipedia* |
|---|---|---|
| Level 1 (article) | 8 % | 9 % |
| Level 2 (metadata) | 6.5 % | |
| Level 3 (section) | 55.5 % | 49 % |
| Level 4/5 (subsections) | 30 % | 42 % |
| Total | 100 % | 100 % |

# Relevance assessments

| *Granularity level* | *IEEE* | *Wikipedia* |
|---|---|---|
| Level 1 (article) | 7 % | 40 % |
| Level 2 | 14 % | |
| Level 3 | 50 % | 33.5 % |
| Level 4/5 | 29 % | 26.5 % |
| Total | 100 % | |

# Two possible reasons for difference in relevance assessments

- The Size factor: Wikipedia articles are smaller than scientific articles
- The Genre factor: the structure of scientific articles and encyclopaedic articles differ

# Conclusions

- searchers prefer to use article as initial entry point
- within documents searchers look at elements relatively proportional to their distribution
- genre has a strong influence on what items are assessed for relevance