WebCLEF… the future

Maarten de Rijke

Three parts

- Between ad hoc and known-item
- Exploring the news
- Exploring cultural heritage

WebCLEF 2005/2006

- **Two tasks using known-item topics**
  - **Mixed Monolingual task**
    - stream of monolingual KI topics covering range of languages
    - language of topic statement is language of desired page
  - **Multilingual task**
    - stream of the English topics (translated from the monolingual)
    - the desired pages may cover any language/domain

WebCLEF 2005/2006

- **Great success!**
  - many known-item topics
  - CLIR effective for mixed-monolingual retrieval
  - multilingual retrieval is the hard nut to crack

- **But …**
  - no informational topics yet, only navigational known-item topics
  - naturalness of the tasks?

WebCLEF 2007

- No more known-item retrieval against EuroGOV
- But then what?

Topics first, task later

Jaap Kamps

Topics First, Task Later

- Reverse standard practice of task definition and topic creation:

  - Topic creator gets access to a search box, and is left to explore the collection and may create any search topic that she thinks is natural

- Questionnaire data is collected

  - The sort of topic, the language preferences
  - Topic set will be close reflection of actual CLWR
  - Mix of topics + metadata

    - Breakdown in familiar categories (ad hoc, known-item, ...)

Limited Assessments

- **Judgments as relevance assessments for informational topics, as duplicate detection for navigational topics**
  - **Very shallow pools: 10–50 results, or ranked pools**
    - **leave it to the assessor to decide when to stop judging**
  - **Page relevant only if judge understands language**
  - **Graded judgments: page may be non-relevant because it contains no useful information, or is not exact same page as visited before, or the language is not understood, …**

Proposal 1

- **Reversal of standard IR practice:**
  - Topics first, task later
- **Retrieval task basically the same with**
  - Natural CLWR topic stream
  - Broad range of topic types (including ad hoc)
- **Questionnaires → contextual metadata**
  - User context (e.g., languages) crucial for CLWR
  - Beyond generalizing one-size-fits-all approach
  - See K. Spärck Jones, *SIGIR Forum*, 40:10–20, 2006

In a Nutshell

- **So far, WebCLEF restricted to known-item search**

  - No informational ad hoc search topics
  - Current topic set may not reflect actual CLWR

- **Proposal: invert traditional topic creation**

  - Collect topics first, then classify according to type

Advantage

- **Allows for mixed-monolingual and multilingual tasks**
    - Evaluated over a (more) realistic stream of topics
    - Or over individual topic types (known-item, ad hoc, etc)

Tracking European News

Maarten de Rijke

NewsExplorer

- **Publicly accessible news aggregation and analysis system**
  - Navigate across languages and over time
  - 13 languages, 9 countries
  - 1,000 newsfeeds, 15,000 articles/day
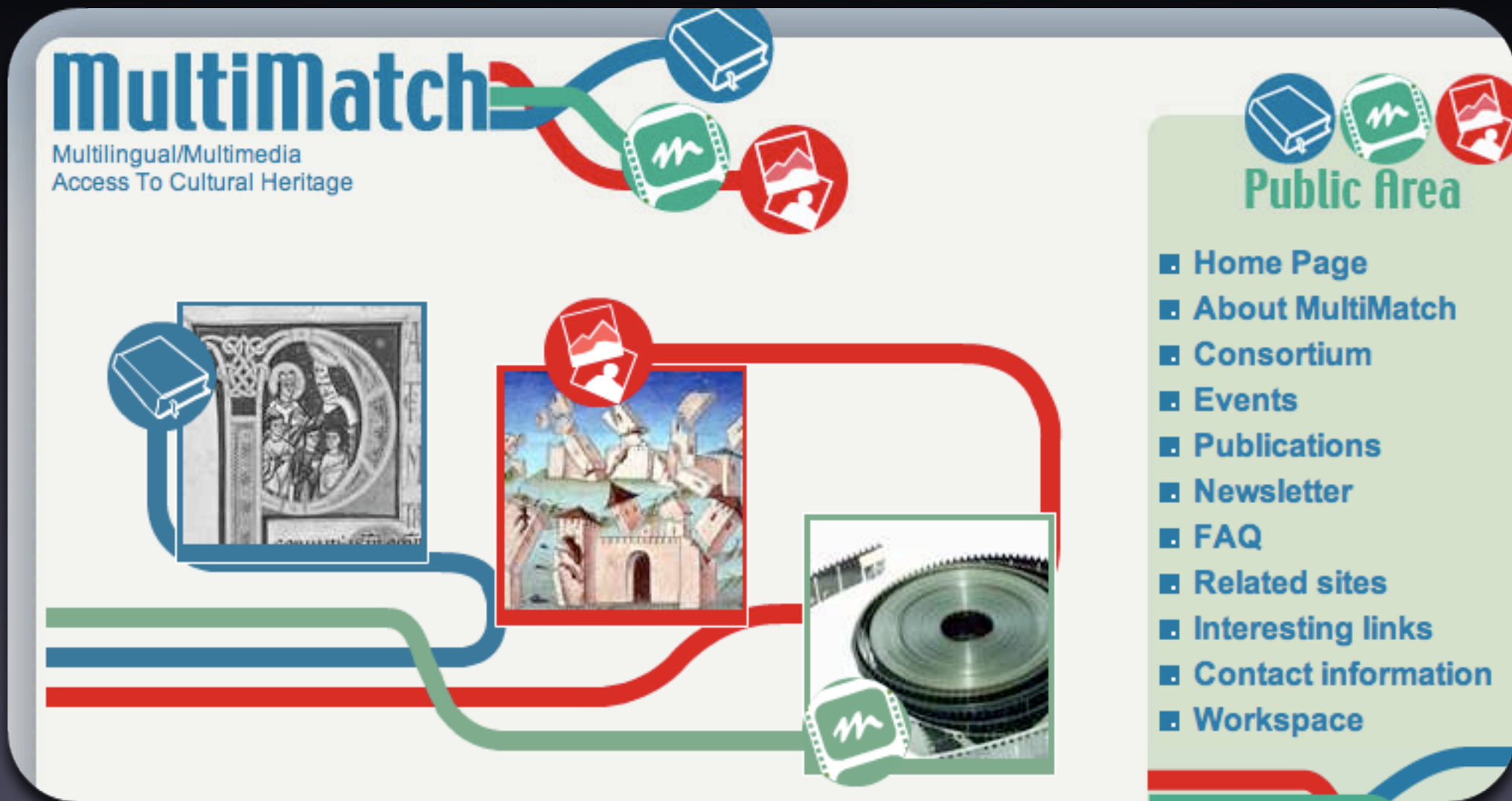- **http://press.jrc.it/NewsExplorer/**

NewsExplorer

NewsExplorer

- **News-related search tasks**

  - **E.g., track and detect topics or events across Europe**
  - **+ Can use logs and other resources from NewsExplorer**
  - **– CLEF would have to create a news corpus, but could use information on feeds from NewsExplorer**

MultiMATCH

Maarten de Rijke

MultiMATCH

MultiMATCH

- **FP6 project**
- **Multilingual search engine for access, organisation and personalised presentation of cultural heritage information**
  - **"Present the complete picture"**
  - **Explore and interact with online accessible cultural heritage content, across media types and language boundaries**

MultiMATCH

+ **"White-list crawl" of CH institutions to be available Dec 06/Jan 07**

+ **Log files and input on search behavior from CH institutions**

• **Search/access tasks**

  • **Faceted retrieval, CH event mining, link generation (e.g., Wikipedia), …**
  • **Later: link up to user generated content, reviews, …**

Now what?

- **Any thoughts at this point?**
- **Join discussions on the WebCLEF wiki**
  - **http://ilps.science.uva.nl/WebCLEF/**