DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Multilingual Information retrieval

for MICHAEL

Carol Peters

ISTI - CNR

Overview

- Multilingual Information Retrieval?
- Different Degrees of MLIR
- Architecture
- Why not MT?
- Requirements for Library Search
- Possible Directions
- Commercial Systems

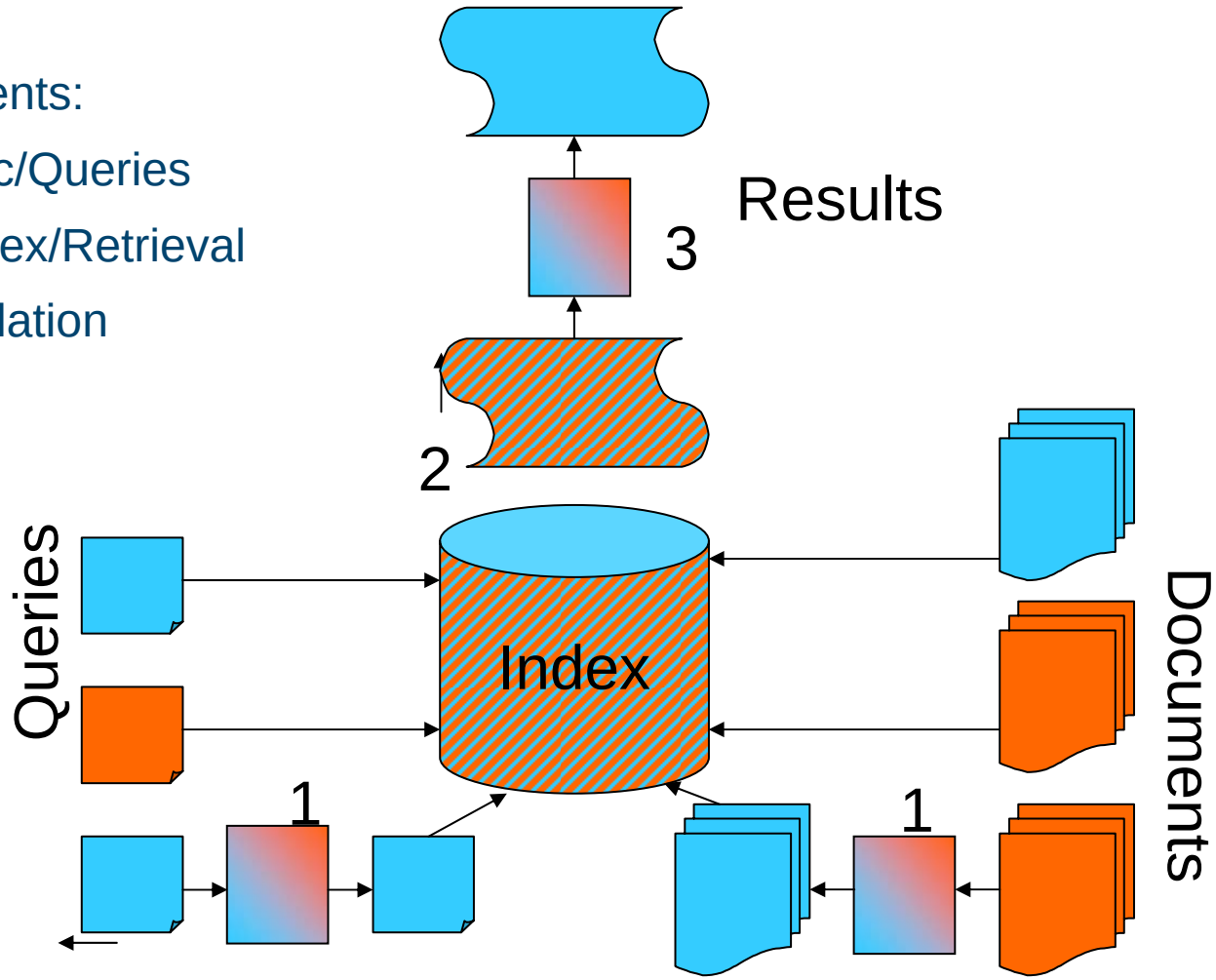
Different Degrees of MLIR

1. Localization of user interfaces, monolingual in different languages
2. Query in L1, all documents in L2
3. Query in L1, documents in L1+L2...Ln
4. Query in L1, documents multilingual

Architecture

- Key Requirements:

1. Translation Doc/Queries
2. Multilingual Index/Retrieval
3. Machine Translation



Different MLIR Architectures

MLIR systems can implement only parts of the overall architecture:

- Only 2: multilingual index/retrieval, monolingual querying
- Only 1+2: Full multilingual system, but results in original languages of the documents only
- 1+2+3: Full mulitilingual system, results in language preferred by the user

Requirements for Library Search

- Many languages (number of official EU languages: 20; 21 from 1-1-2007): System must handle large number of languages
- Coping with OCR errors
- Integration of existing metadata
- Search on metadata and/or full-text
- Presentation of search results: in original language, or translated

Many languages

- some query translation approaches do not scale well (query must be translated into n languages)
- document translation approaches do not scale well if more than one target language
- some approaches need interlingua (propagation of translation errors)

Search on metadata/full text

- issues of merging of result lists: how is the information from the metadata weighted against information from full text?
- potentially different translation strategy on metadata: thesaurus
- metadata can also potentially help to cluster documents → helpful for similarity search

Presentation of search results

- Translation of search results is a typical machine translation problem
- May be hard to offer for all desired language pairs
- There may be non-technical considerations on whether this feature should be offered (issues of quality of the information offered)

Possible directions

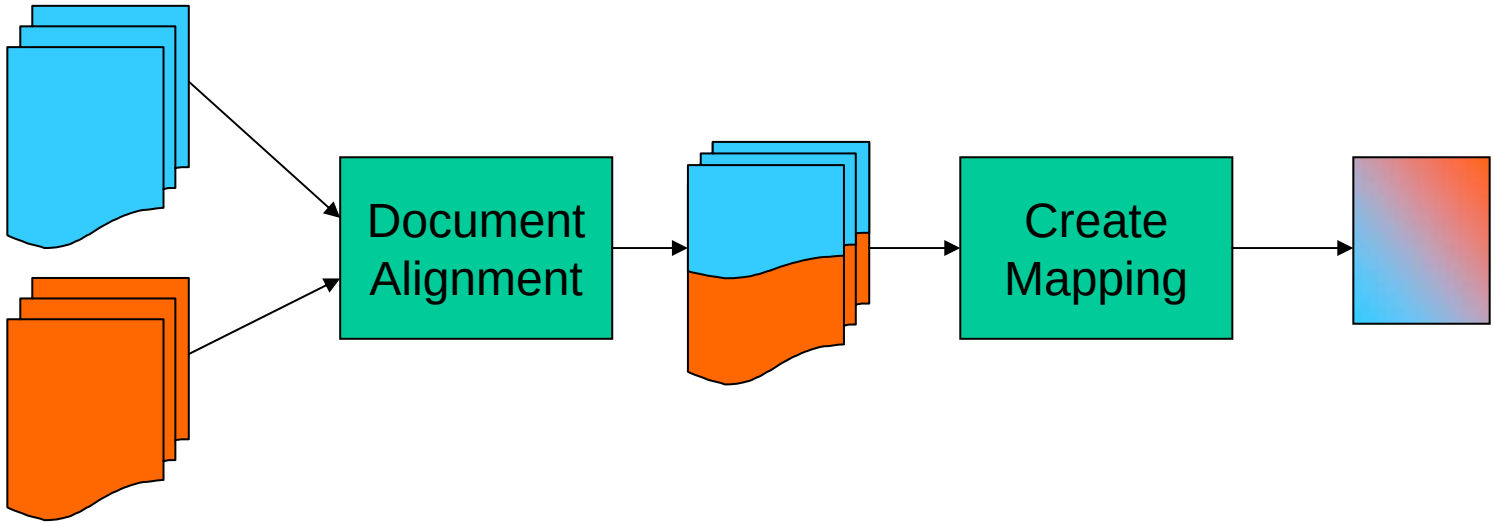
- Cross-Language on Metadata:
- Mapping between classifications of individual libraries
- The mapping can be generated from training data:
  - Aligned document set with classifications
- Meets user needs well

Mapping Metadata

- Steps:

1. Align training data
2. Create mapping

Results



Possible directions

- Similar documents search:

- User searches in her/his preferred language, identifies some relevant results
- If users is not satisfied with the result, the search is expanded to some additional languages
- Similar document search is relatively simple & robust
- May work well even despite OCR problems,

- This process may mirror actual user preferences fairly well

Similar Document Search

- Steps:

1. Monolingual Retrieval
2. Document selection
3. Similar document search