Mathematical Indexing and Querying

Martin Líška et al.

Masaryk University, Faculty of Informatics
Brno, Czech Republic
<255768@mail.muni.cz>

4th May, 2011

Motivation

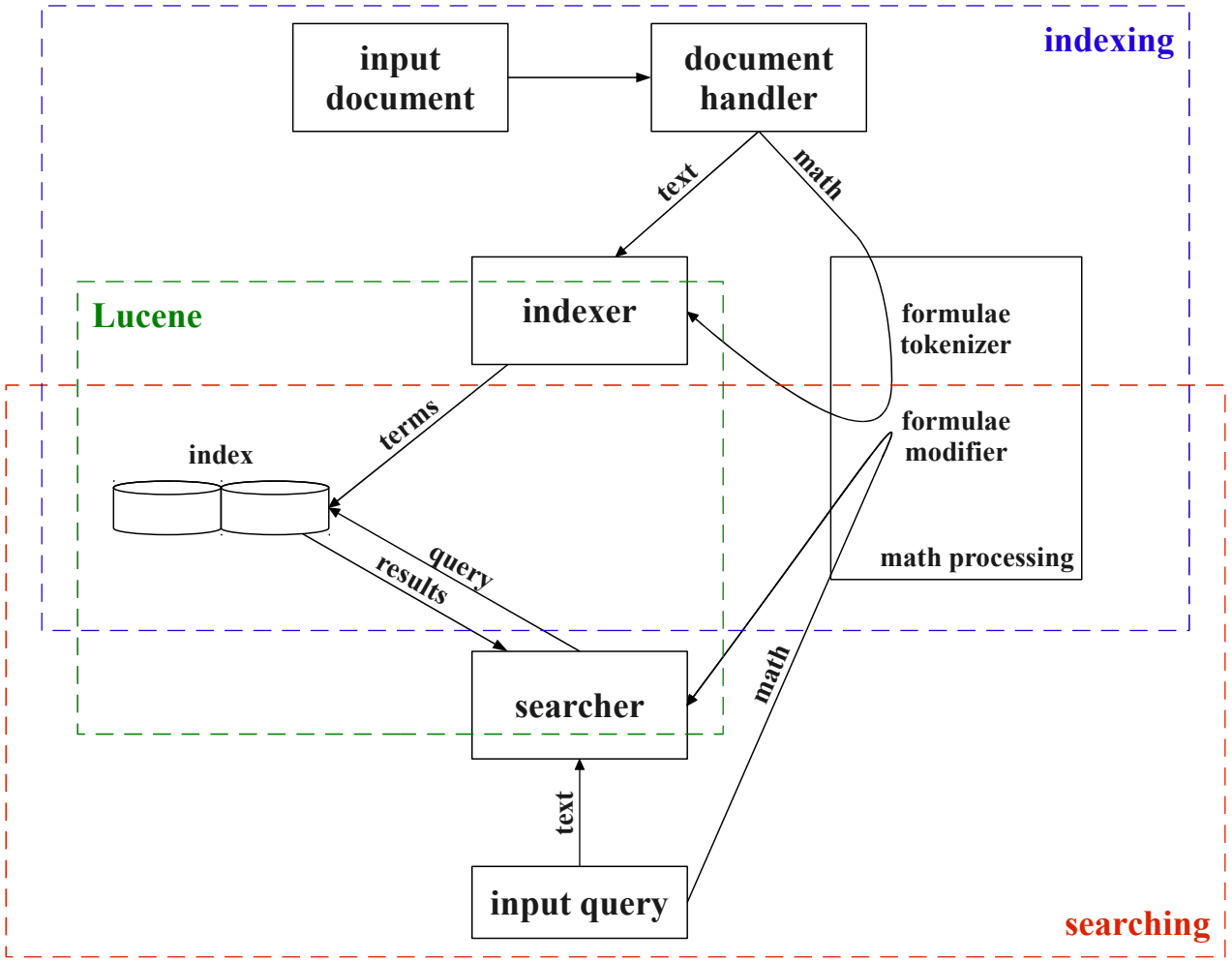
- European Digital Mathematics Library – mathematical searching *should* be provided
- Conventional searching approaches are not applicable
- WP 5 Metadata repository and search engine implementation
- Deliverables D5.2 – The EuDML Search and Browsing Service (Demo due in M12, 18 PMs) and D5.3 (final, in M30, 26 PMs)

Task	IST	FIZ	MU	ICM	Total	Start	End
⋮							
T5.3 Searching and browsing – simple version	2			9	11	M04	M12
T5.4 Searching and browsing – advanced version	2		1	10	13	M09	M24
Total PMs per WP5	24	10	10	24	68		
⋮							

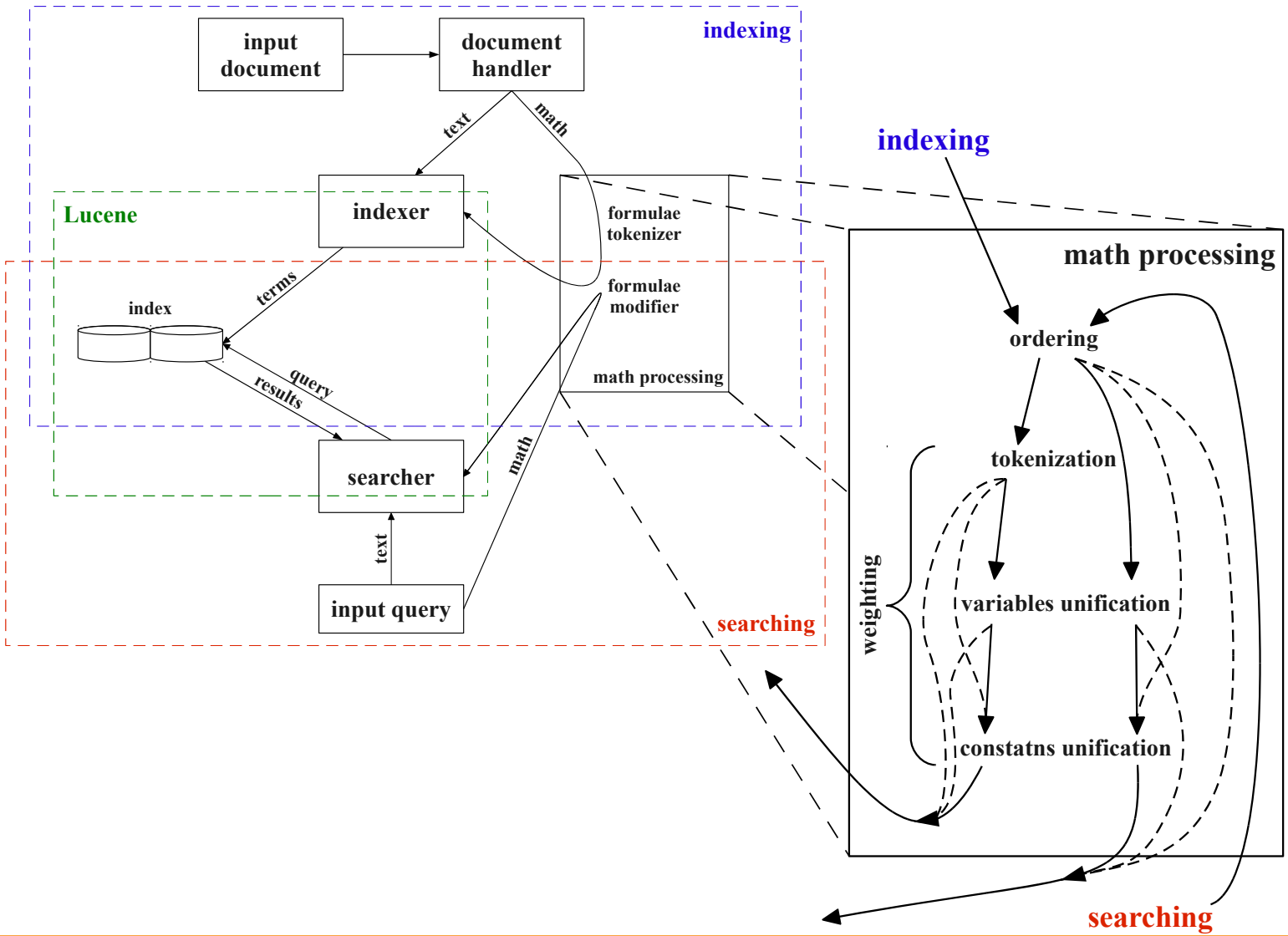
Math Indexer and Searcher – Features

- Based on full-text core Apache Lucene
- Presentation MathML
- Allows similarity (not only exact match) between query and matched term
  - Commutativity
  - Unification of variables and number constants
  - Subformulae matching
- Level of similarity calculation for expressions
- Mixed mathematical-textual queries
- Match snippet generation

Math Indexer and Searcher – Design



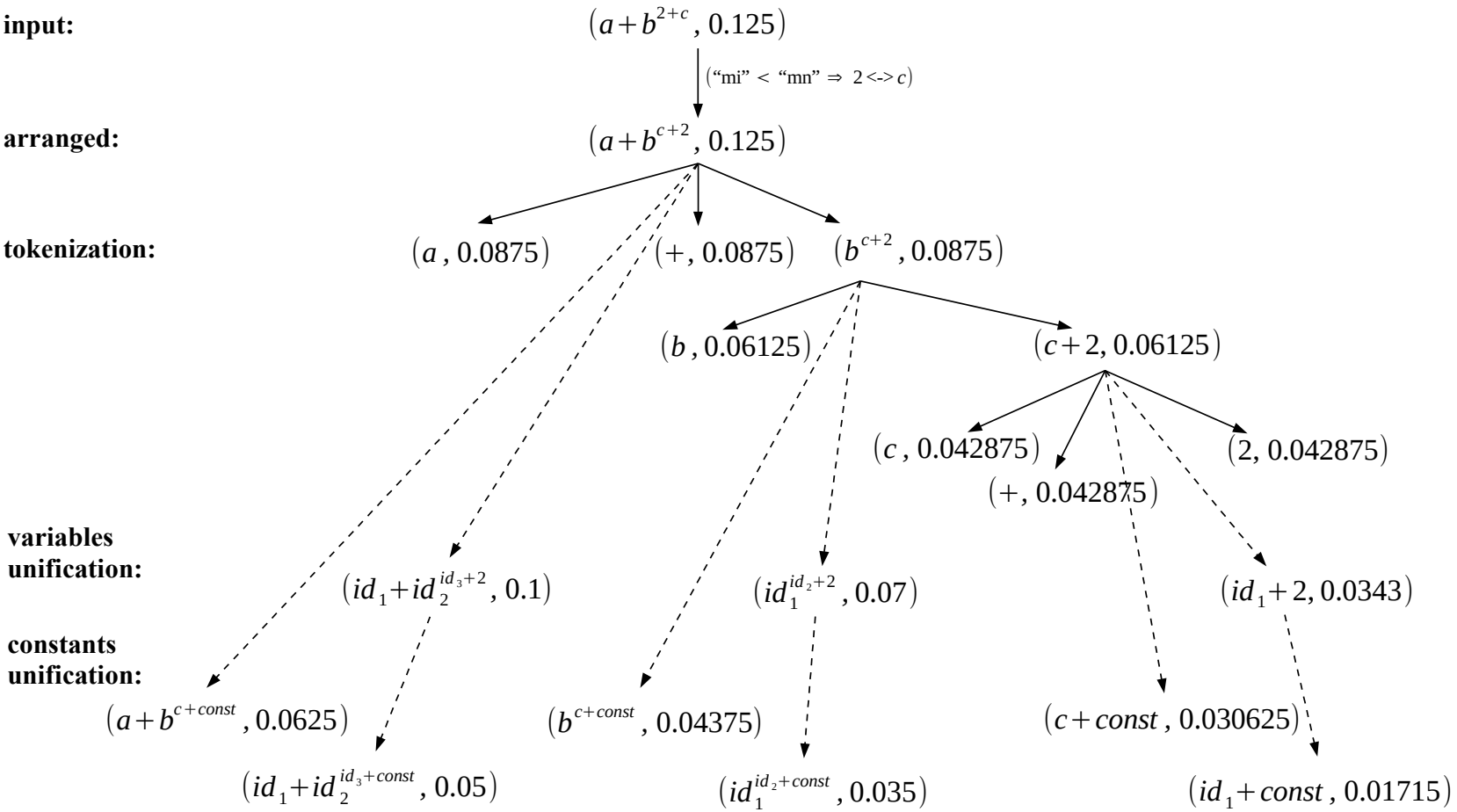
Math Indexer and Searcher – Design II



Math Indexer and Searcher – Weighting

- We used the weighting utility
- Indexing
  - initial weight $= \frac{1}{number\_of\_nodes}$
  - level coefficient $l = 0.7$
  - variables coefficient $v = 0.8$
  - constants coefficient $c = 0.5$
- Searching
  - $result * number\_of\_query\_nodes$

Formula Processing Example



Implementation

- Java
- Lucene 3.1.0
- Mathematical part implements Lucene's interface Tokenizer – able to integrate to any Lucene based system
  - MIaS4Solr plugin was created for the use in SOLR in EuDML

Evaluation

- MREC 2011.3.324
  - 324 060 documents
  - Uncompressed size 53 GB, compressed 6.3 GB
  - 112 million input formulae, over 2 bilion expressions indexed
  - Index size 45 GB
  - Download here
- MREC 2011.4.439
  - 439 423 documents
  - Uncompressed size 124 GB, compressed 15 GB
  - 158 million input formulae, 2.9 bilion expressions indexed
  - Index size 63 GB
  - Download here

WebMIaS

- Demo web interface: WebMIaS
  - MathML/TEX input
  - Canonization of the query – UMCL
  - Matched document snippet generation

Conclusion

MIaS is

- *text+math IR compatible* (fits mathematician's needs)
- *scalable* (index with almost 3 billions formulae tested)
- *Lucene/SOLR compatible* system

Conclusion

- Project pages – MIaS/WebMIaS
- Future work
  - Reindex new corpus with canonized mathematics
  - Optimization
  - Mathematical eqivalence computation via symbolic algebra system?
  - Suggestions welcomed

Questions?