Introduction
Math information retrieval (MIR) starts to be recognized as an important very domain-specific sort of information retrieval. research field. Monash University (MUM) has entered the area of MIR during the development of the Czech Digital Mathematics Library DML-CZ in mid nineties. It became obvious that Digital Mathematical Libraries (DMLs) are specific in many aspects.

Some papers in DMLs consist of more formulas than texts, and we started to think about representation and indexing of mathematical formulas in addition to texts. There was no widely accepted user interface and representation for mark formulas in information retrieval (IR). We have designed and developed first such formula indexing and retrieval prototypes in the series of Bachelor Theses. Mark formulas in structure appearing within texts is a natural part of information retrieval. The relationships between objects mentioned in the text and objects appearing in the formulas can define formula similarity as tree structure similarity. MUM has participated in the development of the European Digital Mathematics Libraries EMLs, where it has been decided to support math formula search.

Canonicalization
Full text papers have to be *canonicalized*, converted to some uniform representation, in order for mathematical formulas to be found by searching. A canonical form is a canonical representation for a DML, which is for mathematical domain. To reliably represent a paper for DML processing, including handling the mathematics, it is necessary to

1. select a canonical representation of the non-structural entities appearing in full-text documents (mathematical symbols, formulas, and equations);
2. define an equivalence classes for these entities (e.g., for which formulas should be considered equal or isomorphic); then, the full-text processing, similarity computations, formula editing, and conversion to a canonical form are performed.

Using our public working demos of the WebMiaS system, we invented the concept of MIR of MathML generated by the real-time TeX to MathML converter, which we tested on several discrete math and logic examples.

Indexing and Searching
Our approach to searching mathematical content in documents is based on similarities of math structures through conventional full-text searching. As mathematical notation, e.g. expressions and formulas, is highly structured, we propose to use indexing and searching methods for full-text searching methods. The preprocessing of the formulas has an important role in order to find matching formulas in text documents.

We developed a search engine according to these principles: Mathset (Math Indexer and Searcher) is a math-aware full-text based search engine. It is based on the state-of-the-art searching library Lucene. It supports combined text and math searching. Refinement of many text query results by adding a math query is believed to be a very powerful tool. Mathematical preprocessing is a plug-in that can be used with any Lucene or Solr-based search engine. The documents with mathematical formulas encoded in Presentation or Content MathML. At the end of the preprocessing, expression trees are incompactated string form to reduce index space requirements.

The very straightforward query interface of Mathset consists of only one input field. Users can type in textual queries together with math queries encoded using LaTeX notation as well as MathML notation. Query is on-the-fly simulated as ‘typo-sets’ formulas in user’s web browser to allow users to verify the correctness of the mathematical part of the query. Along the basic information about matching(s) is generated in the subsequent highlighted text and math tokens that match the query in the document are marked with an overlay.

This allows for quick primary evaluation of the documents relevance to user’s query. Alongside interactive with querying interface Mathset offers searching query services. This is an indispensable feature for automated querying that was used to derive relevance results for the NCBO Math Task.

Evaluation
WEBMiaS team participated in the Math Retrieval Subtask with contribution to all three types of subtasks: Formula, Search, Full Text Search and Open Information Retrieval for the NCBO-11. Full Text Search simulated the standard use of a search system – queries composed of math expressions as well as text. For each query, the system returned a list of documents as was provided to them in the test collection.

New modifications were introduced for the Formula Search task. The NCBO-11 Formula Search task aimed at retrieving independent formulas located in the provided documents. For example, a document contained 100 formulas, such of them could be retrieved as a hit on its own. This is a difference to the normal workflow: However, feasible indexing of MathML used to index every formula was an independent index document containing only that formula by adding a separate index document.

For the needs of Math Retrieval Subtask, we created two indexes from the provided test collections that contained 36,679,971 math expressions and had 7.5 GB in size.

After preprocessing, both indices stored more than 1.5 billions index entries. The first index, NCBO Fragments, was created from single formulas to complete Formula Search specific type. Entry index documents represented only one formula from the input file, therefore, the resulting index contained more than 72,881,509 entries.

The second index, NCBO Indexes, was created for regular document indexing. It contained both text and formulae where one index document represented exactly one physical document from the collection. It took 5 hours to complete the index sized around 39.5 GB. This comparison shows an interesting overhead of the Formula Search index. It contains less data but has a better recall of the logical units which results in the longer indexing time and a larger index.

Alongside test, Mathset accessed LaTeX and both Content and Presentation MathML as a query notation for mathematics. LaTeX queries are combined to be Presentation-Content MathML by LaTeX2e convertor.

We decided to utilize the possibility of submission of four runs to analyze the difference in the performance of the system with regard to the query language. This was supported by the test query collection that provided all of the mentioned formulas for each query.

Overall scores of Mathset were above average of the Math task results. Precision at rank five (P@5) of Mathset in Run 4 was the highest from the all competing submissions. Table 3 shows all four reported metrics for relevance level relevant. Table 4 shows the same metrics for relevance level partially relevant.

We discovered, that ability to evaluate is very valuable in information retrieval. It is a driving force in the evolution process of IR systems, more so if it is imparted as for example at the NCBO conference track. But, in mathematics, we have noticed at NCBO-11 a real progress in our own collection with gold standards against which we could evaluate our development stage. Our future goal is to create our own, gold standard evaluation collection that is a provo of the further development of retrieval techniques.