# ESF Short Visit Grant 5667 – Final Report

**Petr Sojka**

## Contents

## 1 Purpose of the visit: Math Information Retrieval (MIR) Evaluation

Aim of the visit was to present and discuss evaluation of the MIaS, the Math Indexer and Searcher system, that has been developed at Masaryk University under my supervision since 2008, and get new directions at further developments. Our team at Masaryk UNiversity (MIRMU) registered for the new (first ever) Math task at the NTCIR-10 (Evaluation of Information Access Technologies) conference held at the National Institute of Informatics, Tokyo, Japan (`http://research.nii.ac.jp/ntcir/ntcir-10/index.html`). It was a unique opportunity to compare our MIR research, approaches, engine and evaluation results with leading experts in the field, which are working specifically in the Math Information Retrieval domain, and with participants of The 5th International Workshop on Evaluating Information Access (EVIA 2013), a collocated satellite workshop of the NTCIR-10 Conference.
Hosting teams of prof. Noriko Kando and prof. Akiko Aizawa also organized two workshop seminars before and after the main two conferences:

- NTCIR-10 MATH Events: Pre Workshop Seminar on MKM Challenges
  Date : June 17, 2013

Place: Grace Center 1 (20th floor) at NII
Web: http://ntcir-math.nii.ac.jp/?page_id=376

- NTCIR-10 MATH Events: Post Workshop Seminar on DML
  Date : June 24, 2013
  Place: 19F Presentation Room at NII
  Web: http://ntcir-math.nii.ac.jp/?page_id=395

I have attended and took advantage of participation and discussions at the all four events held at NII from Jun 17th to Jun 24th.

## 2 Description of the work carried out during the visit

I have attended all four events (NTCIR, EVIA, pre and post workshops) and actively participated in them.
We have discussed how to push forward the frontiers of evaluation of mathematics retrieval at breakout session with prof. Noriko Kando, NTCIR 10 Programme co-chair, prof. Akiko Aizawa, NTCIR-10 Math task organizer and experts in Math NLP techniques at NII, with Michael Kohlhase (Jacobs University Bremen) Iadh Ounis (University of Glasgow) Fredric C. Gey (University of California, Berkeley) Douglas W. Oard (University of Maryland, USA) and others.
I consulted with prof. Masakazu Suzuki, primary developer of Infty system capable of Math OCR about developments on how to automate workflow to get Math data (in MathML) from PDF.
With Martin Líška, we

- took part at Math task breakout session (NII, Jun 18th, Room 2004, 10:30AM) and helped to form next year's Math task setup;
- presented paper Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task Martin Líška, Petr Sojka and Michal Růžička (Masaryk University, Czech Republic) as part of Parallel Session B-2 (10:45-12:00): NTCIR-10 Math Task (MATH) on July 21st; cf. slides: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MATH/06-NTCIR-10-Math-LiskaM_slides.pdf
- presented poster MATH06 named Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task on July 21st, http://research.nii.ac.jp/ntcir/ntcir-10/program-poster.html#math, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MATH/06-NTCIR-10-Math-LiskaM_poster.pdf about Evaluation of Math Information Retrieval (MIR) of our MIaS.

At the NTCIR-10 MATH Events: Post Workshop Seminar on DML on June 24, 2013 I have presented an invited talk *The European Digital Mathematical Library: An Overview of Math Specific Technologies*

> Abstract: There were developed and deployed several math-aware technologies, methodologies and approaches during the project of the European Digital Mathematical Library `http://eudml.org`. More than 220,000 mathematical papers were collected, math OCRed, math indexed, and processed with leading edge methods to provide modern digital library with math formulae and similarity search, accessibility, browsing, visualization and annotation features.

Slides are available at `http://www.fi.muni.cz/usr/sojka/presentations/sojka-eudml-nii-pres2013.pdf`.

## 3 Description of the main results obtained

We have evaluated MIaS as a promising system with high recall. We have realized that further developments and directions have to deal with:

1. MathML canonicalization and semantic annotation additions (e.g. linking named entities to math formulae) to increase precision;
2. implementation of Presentation to Content MathML conversion with disambiguation of formulae semantic markup: using of NLP, machine translation, machine learning techniques seems necessary;
3. develop ground truth for testing our engine to implement evaluation driven development;
4. evaluate possibility of metric similarity search indexing to have even better recall.

## 4 Future collaboration with host institution (if applicable)

We promised further joint cooperation and participation at NTCIR-11 (Math task), and discussed granting possibilities in MIR evaluation within Masaryk University and Japanese institutions. We evaluate NII International exchange activities `http://www.nii.ac.jp/en/about/international/` especially NII International Internship Program `http://www.nii.ac.jp/en/about/international/mouresearch/` for Ph.D. students' (Martin Líška and Michal Růžička) exchange.

I have mentored NII student Minh-Quoc Nghiem: Semantic enrichment for mathematical expressions and its application to math search problem at Doctoral programme of DML/CICM 2013, Bath, UK, in July 2013, as a follow up activity.

## 5 Projected publications/articles resulting or to result from your grant

Martin Líška, Petr Sojka, Michal Růžička. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task. In: Noriko Kando, Kazuaki Kishida. Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan, 2013. pp. 686–691. ISBN 978-4-86049-062-1.

> This paper describes and summarizes experiences of Masaryk University team MIRMU with the mathematical search performed for the NTCIR pilot Math Task. Our approach is the similarity search based on enhanced full text search utilizing attested state-of-the-art techniques and implementations. The variability of used Math Indexer and Searcher (MIaS) system in terms of the math query notation was tested by submitting multiple runs with four query notations provided. The analysis of the evaluation results shows that the system performs best using TEX queries that are translated to combined Presentation-Content MathML.

For further details, see `https://is.muni.cz/auth/publication/1112631/en`

We expect to participate at NTCIR-11 and publish our MIR results there, and at CICM 2014 conference in Coimbra next year.

## 6 Other comments (if any).

We thank for the perfect organization and hospitality of NII, especially prof. Akiko Aizawa, and are grateful for the ELIAS support.