# Keystone IC1302

## STSM Report

### *Title: Keywords across domains*

*Host: Gabriella Pasi, Univ. of Milano Bicocca*
*Beneficiary: Mihai Lupu, TU Vienna*
*Duration: June 2nd, 2015 – June 9th, 2015*

**Purpose of the STSM**
Establish concrete collaboration plan for future join project proposal

**Description of the work**
The Short Term Scientific Mission established interaction paths between the information retrieval group at the Vienna University of Technology and at the Information Retrieval Lab at the University of Milano Bicocca covering three of the four work areas of Keystone:

- Keyword Search,
- User Interaction and Keyword Query Interpretation, and
- Research integrations, Showcases, Benchmarks and Evaluations.

Together with our colleagues in Milan we set out to explore two keyword-related domains: image tags and nano-publications for scientific articles. Image tags are special keywords generated by the concatenation of terms (e.g. *beautifulsunset* for *beautiful sunset*). As such, they are a special type of keywords and the group at TUW has obtained interesting results in interpreting them using statistical semantics. Another topic covered in the Mission has been nano-publications for scientific works. Nano-publications are structured data files, based on RDF, whose existence opens up new research directions: For instance, automatic scientific paper analysis would become possible, fundamentally changing research and innovation processes. Additionally, it would allow more complex keyword queries from the domain-expert user. Currently, the use of nano-publication is limited to bio-technology (and even there they are not frequently used) because of the complexity of creating them, searching for them, and ultimately reasoning on them. The work of the STSM has been in identifying, based on TUW experience in creating nano-publications and Prof. Pasi's experience at the intersection of information retrieval and fuzzy logic, new opportunities for structure data creation and search in this domain. In this sense we have found that the overlap is substantial enough to warrant work towards a joint proposal for a European Training Network (ETN) under the Marie-Curie Sklodowska actions, a new call for which will be issued over the summer. We have together written up the individual PhD research directions to be undertaken at TUW and at Milano Bicocca. We have also consulted with additional partners and obtained encouraging and concrete feedback. For instance, the work mentioned above in terms of nano-publications, complements perfectly the work on decision theory conducted at Milano Bicocca.

During the stay there, I interacted with the entire research group, consisting of Prof. Pasi, two post-docs (research on patent retrieval based on a flexible graph query language, and research on trust in information retrieval), and four PhD students. Two of the PhD students' work are extremely relevant for Keystone and are likely to find their place in our future collaboration. One on information exploration on graph databases, since our own research [ECIR:2015] has explored this idea and found that precision is extremely hard to automatically achieve and therefore an exploratory interface is probably preferable. The second is the evaluation of domain specific information retrieval for the scientific domain. This is not a new issue [RIAO:2007, WPI:2011] but it is one that has not yet been satisfactorily solved at scale.

In fact, all of these research avenues have to be placed in the context of benchmark-based evaluation, through the creation of domain-specific test collections. Some already exist (e.g. for the image tags the TUW has participated in the creation of MediaEval Retrieving Diverse Social Images task 2014/2015), some have to be created. For the scientific papers, we already have the necessary collection, but are missing the queries and relevance judgments.

### Main results obtained

A first draft towards a join project proposal.

### References

[ECIR:2015] S. Sabetghadam, M. Lupu, R. Bierig, A. Rauber, *Reachability Analysis of Graph Modelled Collections,* Advances in Information Retrieval, 2015, pages 370-381

[RIAO:2007] A. Ritchie and S. Robertson and S. Teufel. *Creating a Test Collection: Relevance Judgments of Cited and Non-cited Papers* , Proc. of RIAO, 2007

[WPI:2011] M. Lupu and J. Huang and J. Zhu and J. Tait. *TREC Chemical Information Retrieval - An Evaluation Effort for Chemical IR Systems*, WPI Journal, , 2011