



**Semantic Analysis with Automatic Tools:
Progress Report
1st EERQI Workshop Part 3
17th-18th March 2009**

Ágnes Sándor
Xerox Research Centre Europe
Agnes.Sandor@xrce.xerox.com

Deliverable 2 proposes the initial prototype of the EERQI research quality indicators. Within this framework automatic semantic analysis is used to provide certain kinds of contextual information, content-based information and citation information.

As for contextual information, we can automatically detect the names of the authors and their affiliations from the first pages of the papers. This is based on the Xerox named entity extraction system that automatically recognizes persons' names and organization names and on the fact that these names in the first page of a publication have a great probability of being the author(s)' names and the affiliations respectively. The system does not give a 100% detection, so its usability should be decided by the partners. This functionality of the Xerox parser is not included in any work package, but if it is of help, we would be pleased to provide it.

Another contextual information is the language of the publication. In this respect we are going to work out in collaboration with RRZN and DIPF query expansion so that the search engine returns articles in languages different from the language of the query as well as articles that do not contain the query expressions but their synonyms.

Providing content-based information by automatic semantic analysis consists in the detection of key sentences and key concepts as evidence for the paper's meeting quality criteria. Automatic analysis does not claim to be a tool that assesses quality automatically but an aid to evaluators for accelerating processing time by highlighting the outline of the article.

We consider that key sentences are the sentences which contain the most important information a paper contains, and these are the sentences that describe the research problems the article handles and the conclusions. The detection of the key sentences is carried out by the Xerox methodology: concept matching.

We hypothesize that by highlighting these sentences the peer-reviewer's attention can be focused on the main message of the paper, which allows him/her to rapidly judge if the paper meets the quality criteria of significance, originality and style. The sentences detected, however do not account for rigour and integrity. We propose helping the peer-reviewer to be able to judge rigour by highlighting the main concepts and the argumentation, and we cannot propose automatically detect evidence for integrity.

Key concepts are extended noun phrases that give a wider context than keywords.

We propose testing the automatically detected contextual information in collaboration with the partners in two ways: by evaluating if the highlighted articles do really help in processing articles more effectively and by applying the tool on peer-reviews.

Citation information is provided through typing citations. The citation types that have been established prove the scientific communication model set up by Fredrik Åström.¹ According to this model scientific communication may be cumulative, - which is mainly the case in natural sciences -, negotiating - which characterizes social sciences -, or distinctive - which can often be said of humanities.

We have distinguished 5 citation types according to the relationship between citing and cited articles. These are the following:

Evidence: the cited work provides evidence for the cited work. In this case the communication is cumulative.

Argumentation: argumentation between the citing and the cited work. In this case the communication is negotiating.

In the following 3 types no communication type can be associated with the citation type:

Importance: the author of the citing work finds the cited work important

Qualification: the cited work is qualified by the citing work

Surprise: the author of the citing work is surprised by the cited work

At this stage that the basic principles and goals of semantic analysis in EERQI have been established. The initial language analysis systems have been developed and have been tested on the content base with the collaboration of RRZN who transform the pdf documents into xml that is processed automatically. These systems need testing and incremental improvement through interactions with the evaluators. We have set up evaluation plans for the coming months.

¹ Article accepted at ISSI conference: Fredrik Åström and Ágnes Sándor: Models of Scholarly Communication and Citation Analysis