Contribution ID: **36** Type: **not specified**

Matching and merging

Wednesday, 22 March 2017 11:15 (15 minutes)

When harvesting information from different sources it is necessary to identify identical objects. If both have the same unique identifier like a DOI or a report-number this is trivial but unfortunately a rare case.

Most of the time matching is mainly based on author and title information. However, titles may change significantly from preprint to publication and depending on the type of the publication (journal paper, conference contribution, thesis) even identical basic metadata would lead to separate records.

In general a two-step process is needed:
a) search for potential candidates. Here it is necessary to define a search query with a high efficiency. However, if the search is too fuzzy, the number of records as search result is too large and matching becomes not feasible. Restriction to a limited scope of records is helpful.
b) confirmation of the match. Depending on the strategy clear results can be treated automatically, whereas doubtful cases might be presented to a human for final decision. In both cases it is essential to have enough information.

For a reliable match good quality of uniform metadata is essential and in many cases processing of content information like abstract, references or fulltext is needed.

Once two records have been identified as equal or existing information receives an update, the information needs to be merged. There are obvious cases where one source always supersedes another, maybe some information comes only from one source. But to add e.g. an ORCID from one source to the author and affiliation from another source requires the identification of corresponding information.

Experience from INSPIRE shows what is currently done (fields with controlled vocabulary), what is doable (fields where the content can be identified) and where merging is not feasible but one version simply overwrites another.

What can be done automatically, which tools are needed, when is human intervention necessary? When is it worthwhile to overwrite (i.e. delete) manually curated, high quality information?

Primary author: SACHS, Kirsten (DESY)

Presenter: SACHS, Kirsten (DESY)

Session Classification: Legacy

Track Classification: Main track