Automated affiliation identification for Converis using Web of Science core collection data

Marcus Walther and Bastian Melsheimer
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Schlossplatz 4, 91054 Erlangen, Germany

I. Introduction

This poster presents a technical development for an automated processing of affiliation data using Web of Science core collection data. This enables our Converis installation to reflect external associations without adding noticeable additional work load to scientists while adding publications. Our method uses standard tools that make the procedure portable and sustainable. So far we processed 1.600 publications with 10.000 authors.

II. Affiliation data retrieval

Publications listed in Web of Science (WOS) have a unique ID ("Accession Number") that is saved by Converis during the import process. Additionally Converis also fetches the publication's XML data-set including affiliation information. In order to use Thomson Reuter's "article match retrieval" (AMR) service, a Web of Science Core Collection subscription is necessary.

If a publication is entered manually, we need the DOI for a look-up inside Web of Science (described in [2]). Afterwards the affiliation data can be downloaded using the AMR API[3].

III. Affiliation mapping

Affiliation data from WOS contains several levels reflecting the organization structure. You'll find a notation for the whole organization as well as for subordinate parts of it.

From the CRIS's point of view there is a big difference between internal and external authors. For the latter we neglect sub-organizations so that only the main organization must be identified. In order to allow identification of internal units of our university and match person data from the HR system, all levels of affiliation data must be taken into account.

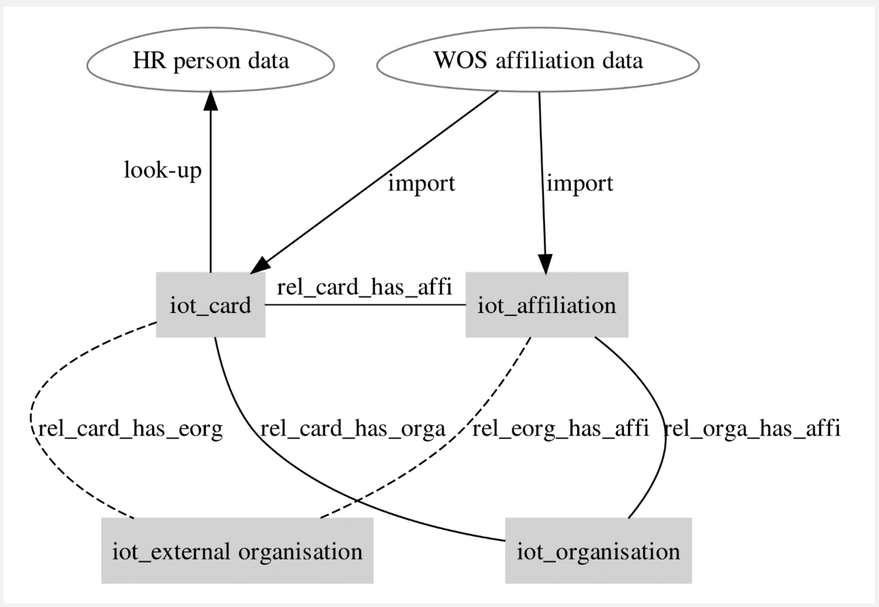
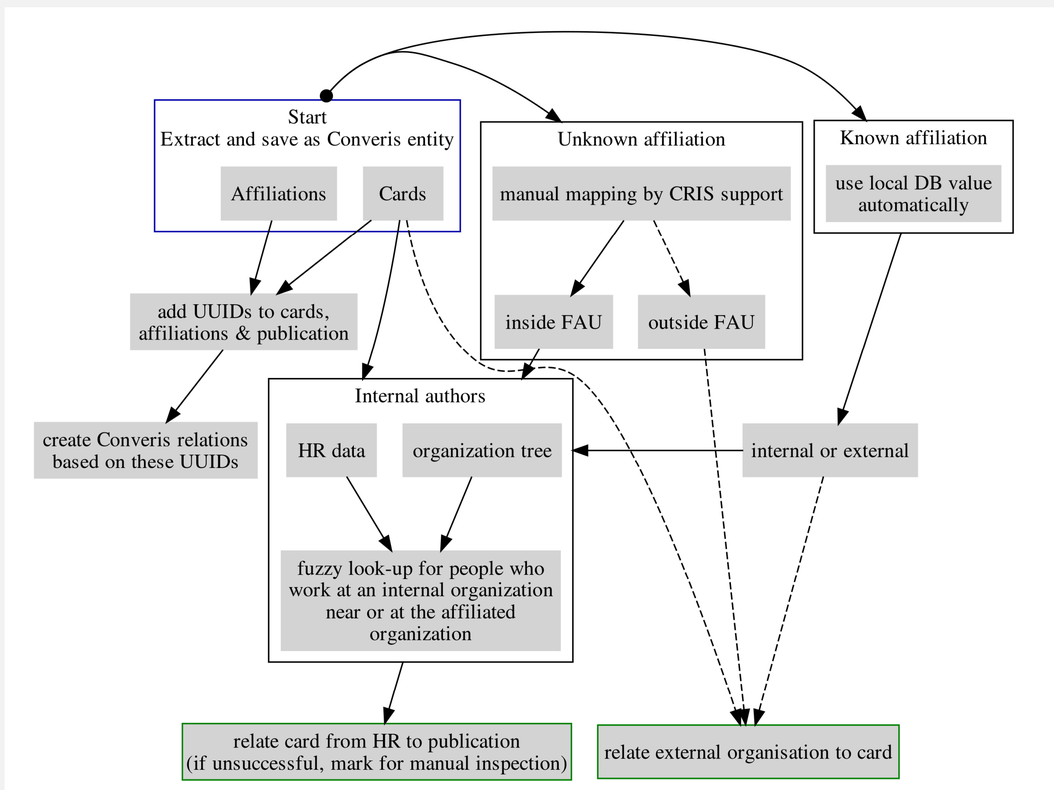


Figure 1: Converis work flow and data model overview: Once new affiliations are mapped manually, it's type is known (internal (solid) or external (dashed)) and the relation between card and related organization can be created automatically. Any subsequent occurrence of the same affiliation requires no manual action.

IV. Technical workflow

Converis includes Pentaho Kettle[1] for data processing and integration. Converis supports "Kettle transformations" only, so the process has to be split-up into several independent parts. Relations between publication, cards and affiliations are specified by UUIDs at first and transferred into Converis relations in a separate step.



V. Limitations and outlook

Although authors and affiliations are generally available in WOS, relations between these are only available in a good quality from 2008 onwards.

WOS XML data doesn't contain umlauts. In some cases they were transcribed ('ö' to 'oe'), in other cases the diacritic signs were dropped ('ö' to 'o'). This can be adjusted for internal authors automatically using HR data. External authors need to be corrected manually using publisher information or other resources.

A broader usage of CERIF organization data would result in a big simplification of affiliation identification. Today we maintain our own lists of external and internal organizations. If available we store the internet domain name (URL) and the International Standard Name Identifier (ISNI) of the organizations for future interoperability.

Additional resources

[1] http://www.pentaho.com/product/data-integration
[2] http://kitchingroup.cheme.cmu.edu/blog/2015/06/08/Getting-a-WOS-Accession-number-from-a-DOI/
[3] http://science.thomsonreuters.com/tutorials/wsp_docs/soap/Guide/

Author contact: marcus.walther@fau.de