Automated author affiliation processing using Scopus data

Abstract

In this paper, we present a technical development for an automated processing of Scopus' affiliation data. The enables our users to record all external associations of research output without adding noticeable additional work load. The method uses standard tools that make the procedure portable and sustainable.

Motivation

The Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) uses the Converis software (Maier et al. 2012) as research information system (CRIS, located at https://cris.fau.de/). Currently the operational data model includes publications, research projects, patents, awards, researcher activities and memberships.

From the administration's point of view, CRIS is vital for internationalization strategies, intra-university collaborations, bibliography, easy data collection for rankings and more. From the researcher's point of view, CRIS generates additional workload. Therefore, besides targeting administrative processes we need to generate extra value for the researchers. While major awards and funded projects are registered directly by the appropriate division of the university, easing publication data input is a bit more complex.

At FAU, it is required to identify each author of a publication including internal or external affiliation. This information is usually not included in regular bibliography formats like BibTeX (Patashnik 1988) nor available inside ORCiD (Haak et al. 2012) or Researchgate (Madisch 2008) profiles.

Hence, we offer an automated process for author relation of research papers and conference contributions. This has major advantages: We are able to maximize interpretability of CRIS data and we can provide several citation styles including APA, which demands full first names. Furthermore this automation allows coverage of publications with a huge number of authors, which is common in physics (e.g. Acharya et al. 2013: 976 authors located at 130 research organizations), large medical panel studies and other fields of research.

Since 2015, we offer this service using data from Web of Science (a similar database offered by Clarivate Analystics). Because Scopus offers a larger coverage of scientific journals (Mongeon, P. & Paul-Hus, A. 2016) an implementation of an interface to our research information system was desirable.

Data recorded in CRIS is easy reusable by scientists for any purpose (e.g. reports, raw data export). One popular feature is the automated update of the scientists' website.

Approach

Affiliation information is usually available online at the publisher's website for every single publication. Unfortunately, there is no common format in broad use for transferring this data into other databases. The obvious format CERIF (Common European Research Information) is also not available from large information providers.

There is an existing agreement between FAU and Elsevier, so we have access to the REST based web service interface (Elsevier 2018). It delivers affiliation information of each author for up to two levels: Main organization (e.g. university name) and sub-division (example 1).

```
<affiliation afid="60000765" country="deu">
      <organization>FAU Erlangen-N&#xFC;rnberg</organization>
      <organization>i-MEET</organization>
      <address-part>Martensstra&#xDF;e 7</address-part>
      <city>Erlangen</city>
      <postal-code>91058</postal-code>
      <affiliation-id afid="60000765"/>
      <country>Germany</country>
</affiliation>
```

Example 1: Part of Scopus XML data for affiliation data

Although this data does not include common identifiers for organizations like Ringgold ID (Ringgold 2004) or ISNI (MacEwan et al. 2013) the affiliation information can be used for our automated process. We manually need to define a mapping between organization name from Scopus and our list of organizations only once. Furthermore Scopus delivers their internal affiliation id (`afid="60000765"` in example 1).

Technical solution

1. Publication identification

A unique identifier must distinguish each publication. Usually this is the Digital Objects Identifier (DOI). Either the user enters it directly or uses the Converis' built-in DOI look-up feature. Other sources might be imported publication data from Web of Science or PubMed.

Once a DOI is present, we use the Scopus API in order to fetch the Scopus publication ID (example 2). Using the Scopus API is not free; one will need a license or institutional access. All requests need to be authenticated by an APIKEY that must be included as HTTP header value of "X-ELS-APIKey". Scopus' API responds JSON formatted data by default, adding the HTTP-header "Accept: application/xml" requests XML data.

```
GET
http://api.elsevier.com/content/search/scopus?query=doi('10.1002/zaac.201500570')
```

Example 2: DOI to SID look-up request for “10.1002/zaac.201500570”

The service delivers a comprehensive XML dataset for the publication, which includes the Scopus publication identifier (example 3).

```
…
  <opensearch:Query role="request"
searchTerms="doi('10.1002/zaac.201500570')" startPage="0"/>
…
  <entry>
    …
    <dc:identifier>SCOPUS_ID:84943574153</dc:identifier>
    <eid>2-s2.0-84943574153</eid>
    <dc:title>Strontium Chemistry with Silicone Grease</dc:title>
    …
  </entry>
</search-results>
```

Example 3: Part of API XML response including identifier “SCOPUS_ID:84943574153”

This look-up is not necessary if the publication was imported using Converis’ Scopus interface: the Scopus publication ID is already present.

2. Fetching affiliation data

After the Scopus identifier is known we use another API request for retrieval of the XML meta-data for the publication (example 4).

```
GET http://api.elsevier.com/content/ abstract/SCOPUS_ID:84943574153
```

Example 4: download request for publication “84943574153”

The XML data set includes author and affiliation data (see example 1) that can be accessed using XPath query language (Clark et al. 1999). Finally, we attach the extracted information to the publication’s attributes.

CRIS holds a controlled list of external organizations. The support team needs to match the organization name to an entry of that list only once for new unknown affiliations.

We include two additional identifiers for interoperability reasons for organizations (if available): domain name (e.g. “uni-bamberg.de”) and ISNI code in ISO 27729 standard (ISO 27729:2012). The latter will guarantee a future look-up inside ORCID (ISNI & ORCID, 2013).

There is a big difference between affiliations of FAU staff and external scientists: For our use-cases, it is sufficient to identify e.g. other universities as a whole without reflecting internal

divisions (e.g. Department for Mathematics). In this case, we evaluate only the first level of affiliation. Using Scopus' organization affiliation ID as reference (see example 1) simplifies the assignment effort dramatically, because different notation of the organization's name has no impact.

In contrast, it is essential to match FAU's internal structure in order to identify all internal authors correctly. Therefore, we take all available levels of affiliation information into account.

3. Converis data integration

Converis includes Pentaho Business Intelligence tools (Bouman, van Dongen 2009) for data analysis as well as for data integration. In Converis context, it is called "DIS" (data integration server). It is recommended to use Kettle transformations (Casters et al. 2010) for adding data to Converis, without direct access to the database system. This makes our development sustainable: if the underlying database structure changes in new releases of Converis, there is no need to adjust the implementation.

Clarivate Analytics provides a plug-in for data exchange between Kettle and Converis. The plug-in saves the data into the database asynchronously. This method speeds up the process but causes one major limitation: The primary database key of new datasets is not returned to the transformation for further processing.

➔ Converis distinguishes between data entities, like publications or people, and relations between them (e.g. "PUBL_has_CARD"). For creation of a relation between data entities one needs the primary keys of the database representation.

Therefore, one needs to split the whole process into several sub-transformations. Between each step all data is made permanent inside the database.

1. Extract affiliation data, create new affiliations and author cards. Any relation is mapped using unique IDs as attribute of the involved entities.
2. Transfer relations from UUID mapping to Converis relation entries and clear helper UUID attributes.
3. Relate external organizations to authors whose affiliation is identified as external.
4. Exchange cards for internal affiliations with card supplied by identity management.

In presence of a new unknown affiliation, the process cannot execute steps 3 or 4. After successful manual mapping it will continue. Next time this mapping is available for the automated step and so manual effort drops significantly.

Step 1 will reuse cards already present in the system if author's name and affiliation are identical. This avoids unnecessary overhead inside the database.

Often authors have more than one affiliation while our data model supports only one. If possible, we use the first affiliation; in case of internal authors, we prefer the internal one.

Limitations

Reliable affiliation data is only present from 2008 on inside Scopus. Older publications often show only one address or affiliations are not related to the single authors.

Scopus XML data include numeric HTML entities in some cases (for example “Ä” for “Ä”). The Converis plug-in for Pentaho interprets these values while transferring data using JSON. So a value like “TU München” ends-up as “TU München” inside the database breaking comparison to the original value. As solution, we convert all HTML entities into UTF-8 right after XML parsing.

Sometimes Scopus data does not include the author’s full given name. This does not influence the automated affiliation identification, but leads to APA citation failure.

In rare cases, the organization affiliation ID is used for more than one organization. This leads to wrong relations to our list of external organizations that must be fixed manually.

Conclusion

We established a robust automated process for author identification including proper affiliations based on Scopus data. Up to now, more than 960 publications were already processed. By using this method, scientists at FAU save a lot of time for publication list maintenance. Furthermore, our method is less error-prone than manual work.

Outlook

In contrast to Web of Science, Scopus offers detailed author information for books and book chapters. We plan to enhance our procedure to for these publication types.

Remarks

The final paper will include more code examples and figures pointing out details of technical workflow.

References

Acharya, B. S.; Actis, M.; Aghajani, T.; Agnetta, G.; Aguilar, J.; Aharonian, F. et al. (2013): Introducing the CTA concept. In *Astroparticle Physics* 43, pp. 3–18. DOI: 10.1016/j.astropartphys.2013.01.007.

Bouman, Roland; van Dongen, Jos (2009): Pentaho Solutions. In *Business Intelligence and Data Warehousing with Pentaho and MYSQL*.

Casters, Matt; Bouman, Roland; van Dongen, Jos (2010): Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration: John Wiley & Sons.

Clark, James; DeRose, Steve; others (1999): XML path language (XPath) version 1.0.

Falagas, Matthew E.; Pitsouni, Eleni I.; Malietzis, George A.; Pappas, Georgios (2008): Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. In *The FASEB journal* 22 (2), pp. 338–342.

Haak, Laurel L.; Fenner, Martin; Paglione, Laura; Pentz, Ed; Ratner, Howard (2012): ORCID: a system to uniquely identify researchers. In *Learned Publishing* 25 (4), pp. 259–264.

ISO 27729:2012: Information and documentation - International standard name identifier (ISNI).

MacEwan, Andrew; Angjeli, Anila; Gatenby, Janifer (2013): The International Standard Name Identifier (ISNI): The evolving future of name authority control. In *Cataloging & Classification Quarterly* 51 (1-3), pp. 55–71.

Madisch, I. (2008): ResearchGate scientific network: A first step towards science 2.0. In : Clinical and Experimental Immunology, vol. 154, p. 214.

Maier, Jan C.; Höllrigl, Thorsten; Weiss, Rudolf (2012): CONVERIS 5: the Next Generation Current Research Information System.

ISNI & ORCID (2013): ISNI and ORCID Issue Joint Statement on Interoperation, April 2013. Available online at http://www.isni.org/content/isni-and-orcid-issue-joint-statement-interoperation-april-2013, checked on 2/27/2018.

Mongeon, P. & Paul-Hus, A. Scientometrics (2016) 106: 213. https://doi.org/10.1007/s11192-015-1765-5

Patashnik, Oren (1988): BibTeXing. documentation for general BibTeX users. In *Electronic document accompanying BibTeX distribution*.

Ringgold (2004): Identify Database. Available online at http://www.ringgold.com/identify, checked on 2/29/2016.

Elsevier B.V. (2018): Elsevier Scopus APIs. Available online at https://dev.elsevier.com/sc_apis.html, checked on 2/27/2018.