“Where have all the papers gone?”: Investigating why some papers are not reported in a CRIS

Extended abstract submission for the CRIS 2022 conference

Jan Dvořák[1,2]
Tomáš Chudlarský[1,2]
Josef Špaček[1]

[1] Czech Technical University in Prague, Computing and Information Centre, Jugoslávských partyzánů 3, CZ-16000 Praha 6, Czechia
[2] Charles University, Institute of Information Science and Librarianship, Na Příkopě 29, CZ-11000 Praha 1, Czechia

Keywords: Current Research Information System, data completeness, affiliation recognition, record linkage

We investigate the extent, the variation by discipline and the reasons for some papers to have affiliation to an institution, but not be reported in the institutional CRIS. We compare the contents of the institutional CRIS of the Czech Technical University against the Web of Science database.

Institutional Current Research Information Systems (CRISs, also known as Research Information Management Systems, RIMS) help institutions systemize information about their on-going and past research and support many diverse reporting requirements that would otherwise have landed on researchers themselves. This effort typically involves keeping the publication record of the institution’s staff. An important question is how complete such a record is. We will investigate this question in the case of our institution, the Czech Technical University in Prague.

Some countries, Czechia among them, run national (or governmental) CRISs that serve as research information hubs for the local communities and often find their use in supporting nation-wide research assessment exercises. Where the body of scientific publications is considered, methods of bibliometrics are necessarily called upon. The Web of Science database has traditionally been and is still being considered a reliable source of information by the community of bibliometricians. Also internally, bibliometric criteria often play a role in supporting research funding allocation decisions and academic staff promotions. Not the least, CRISs serve as the environment for the researchers’ publication records, which allow them to generate CVs, a rather frequent requirement.

All these incentives, both internal and external, seemingly guarantee that the publication records in CRISs are complete. Our aim is to verify this hypothesis. V3S, our in-house built CRIS solution, supports fetching publication metadata from the Web of Science database

operated by Clarivate Analytics through a web service subscription. The CRIS regularly fetches the metadata of publications based on the affiliation of an author to our institution. These records are then offered to our staff as bases for full records in the CRIS itself: one typically needs to link the authors to real identities within our user account management system (ORCID iDs help here considerably), resolve affiliations and funding and supply the original abstract and a few other data items. This cannot be automated so it needs to be done by the authors themselves or by assigned personnel at departments. In rare cases an error in the Web of Science record is spotted and its correction is requested through the institution's Central Library.

However, in our institution we estimate that around 10% of publication records we fetch from the Web of Science (so they have authors affiliated to our institution) go unclaimed by our authors. This amounts to hundreds of articles every year.

We will investigate the extent, the variation by discipline and the reasons for this phenomenon. Our working hypotheses include:

- All the authors have left the institution.
- Authors with multiple affiliations are finding themselves under bureaucratic restrictions that prevent them from reporting their publications in another institution. Such requirements do appear.
- A lag between the publication of a paper and its indexing in the citation database make the papers available for claiming too late to be interesting for its authors.
- The unclaimed papers are outputs of large collaborations, where the practice of listing all researchers as authors of all papers has developed. In the case of our institution, this includes papers in High Energy Physics where large researcher collaborations have formed around the CERN. The researchers are even not able to follow all the papers their name appears on.
- Some researchers do not recognize the importance of research output reporting.

We will use the methods of both quantitative and qualitative research.
Our data sources are: the CRIS (V3S) and the Web of Science database which we access through the SOAP webservice interface ("Web of Science API Expanded Premium"). We use the DOIs and an approximate string-matching algorithm applied to the titles of the works to match records between the external database and the CRIS's own records. We may also support this research with data from IS VaVaI, the national CRIS in Czechia.

This research can help better understand the practice of research publication reporting. We believe it can consequently also lead to an improvement. A more complete picture of research at their institution will allow the management to base their decisions on more accurate data and also enhance the public image of the institution.

We trust this proposed paper fits the theme "Linking Research Information Across Data Spaces" quite well. Of the suggested conference topics, the following ones are addressed:

- Best practices in system interoperability and research information exchange at a regional or a national level
- Value, impact and outcomes of universities