

# Bridging data gaps in a Research Information Management System with OpenAlex

Author

Søren Vidmar  
sv@aub.aau.dk  
0000-0003-3055-6053



Affiliation

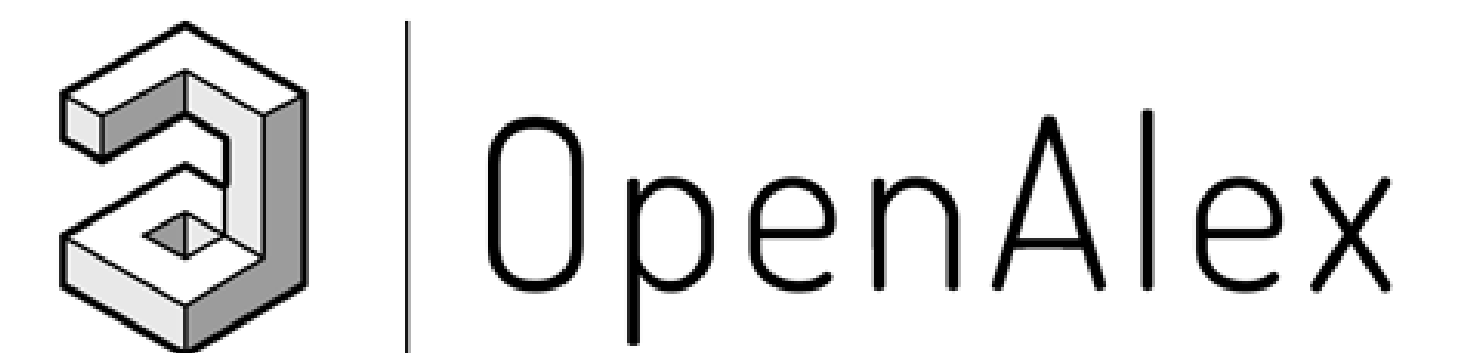
Aalborg University Library



How can you discern the absence of something when you lack a comprehensive source of all necessary information?

Identifying data gaps within a research information management system presents a considerable challenge. It hinges on the available resources, licenses, and organizational procedures, not to forget the ever-precious commodity of time—a resource universally coveted for its scarcity. Promising endeavors such as OpenAlex could offer us significant assistance.

Consequently, we have begun experimenting with new tools that could aid us in both detecting and bridging these gaps eventually. CRISearch and OpenAlex2RIS are experimental tools that can help you identify and close the data gap in your Current Research Information System (CRIS)



## 1 Introduction

Data within a CRIS must be as complete as possible to provide a solid base for distributing, showcasing and evaluating an institution's research contributions. Despite solid institutional processes and access to multiple systems to help us out, some records inevitably slip through the cracks each year. Thankfully, initiatives like OpenAlex are making significant progress in this domain. By utilizing persistent identifiers such as ROR, ORCID, and DOIs, OpenAlex provides a valuable resource with well-structured data on open research infrastructure platforms. This aids us in identifying missing records and potentially enhancing our local metadata quality in our CRIS.

## 2 Objective

To initiate the creation of an easy-to-use tool that facilitates the identification and closure of data gaps within our CRIS system,

## 3 Identifying missing content with CRISearch - A PowerBI tool for the OpenAlex API

First step in closing the gap is identifying the publications missing in the CRIS.

The primary function of CRISearch is to compare the publications from our institution in OpenAlex against those in our CRIS, providing us with a comprehensive overview of any research outputs our institution may be lacking. Its design requires further refinement, and while it initially was built within PowerBI for its ease of implementation, it should evolve to have a presence beyond PowerBI.

We use Elsevier's Pure as our CRIS, but the tool is system-agnostic as such, relying on DOI matching and a supplementary Lucene search for title matching to account for missing or faulty DOIs. By utilizing the APIs of OpenAlex and our CRIS, we can systematically identify gaps in our database, by narrowing down a list of publications which belong to our institution, but which we don't have in our CRIS.

Plus, it finds (legal!) Open Access versions of articles through Unpaywall, hinting at the potential for further automation in writing Unpaywall URLs directly to our CRIS

DOI	Publication Year	Possible match on title in CRIS?	Title	Type	Journal/Venue	Volume	Has volume?
<a href="https://doi.org/10.1001/jama.networkopen.2023.55716">https://doi.org/10.1001/jama.networkopen.2023.55716</a>	2024	No	Safety and Efficacy of Midline vs Peripherally Inserted Central Catheters Among Adults	journal-article	JAMA network open	7	Yes
<a href="https://doi.org/10.1002/9781394188789.ch1">https://doi.org/10.1002/9781394188789.ch1</a>	2024	Yes	The Necessity for Modernizing the Coupled Structure of Intelligent Transportation Systems and Multi-Energy Networks	other			No
<a href="https://doi.org/10.1002/9781394188789.ch6">https://doi.org/10.1002/9781394188789.ch6</a>	2024	Yes	Flexible Operation of Power-To-X Energy Systems in Transportation Networks	other			No
<a href="https://doi.org/10.1002/adfm.202313850">https://doi.org/10.1002/adfm.202313850</a>	2024	No	Cyanocentrone Based Low-Cost Polymer Donors for High Efficiency Organic Solar Cells	journal-article	Advanced Functional Materials		No
<a href="https://doi.org/10.1002/advs.202304834">https://doi.org/10.1002/advs.202304834</a>	2024	No	De Novo Atomistic Discovery of Disordered Mechanical Metamaterials by Machine Learning	journal-article	Advanced Science		No
<a href="https://doi.org/10.1002/alz.13681">https://doi.org/10.1002/alz.13681</a>	2024	Yes	Mapping morbidity 10 years prior to a diagnosis of young onset Alzheimer's disease	journal-article	Alzheimer's & Dementia		No
<a href="https://doi.org/10.1002/ctm2.1565">https://doi.org/10.1002/ctm2.1565</a>	2024	No	Pericardial delta like non-canonical NOTCH ligand 1 (Dlk1) augments fibrosis in the heart through epithelial to mesenchymal transition	journal-article	Clinical and translational medicine	14	Yes
<a href="https://doi.org/10.1002/dmrr.3775">https://doi.org/10.1002/dmrr.3775</a>	2024	No	The impact of sodium-glucose co-transporter-2 inhibitors on dementia and cardiovascular events in diabetic patients with atrial fibrillation	journal-article	Diabetes/Metabolism Research and Reviews	40	Yes
<a href="https://doi.org/10.1002/ehf2.14688">https://doi.org/10.1002/ehf2.14688</a>	2024	Yes	Computed tomography or chest X-ray to assess pulmonary congestion in dyspnoeic patients with acute heart failure	journal-article	Esc Heart Failure		No
<a href="https://doi.org/10.1002/ijc.34851">https://doi.org/10.1002/ijc.34851</a>	2024	Yes	Early mortality in children with cancer in Denmark and Sweden: The role of social	journal-article	International Journal of Cancer		No

## 4 Filling the gap with Python - OpenAlex2RIS

So identifying what's missing is one thing - filling the gap is another. To facilitate the integration of missing records, I've created two Python scripts. These scripts are capable of submitting either an individual DOI or a collection of DOIs in a CSV file to the OpenAlex API, producing RIS files in response. These RIS files can then be imported into our CRIS system. Although a collection of RIS files may not be the perfect solution for a large gap, it's a step in the right direction and certainly better than nothing.

```
Enter the DOI: 10.1016/S0821-9258(19)52451-6
Saved RIS file: output.ris
TY - JOUR
TI - PROTEIN MEASUREMENT WITH THE FOLIN PHENOL REAGENT
J2 - Journal of Biological Chemistry
SN - 0821-9258
A1 - Oliver H. Lowry
A1 - N. J. Rosebrough
A1 - A. Farr
A1 - Rose J. Randall
PY - 1951
VL - 193
IS - 1
SP - 265
EP - 275
SO - https://doi.org/10.1016/S0821-9258(19)52451-6
LA - en
KW - folin phenol reagent
KW - protein
SO - https://doi.org/10.1016/S0821-9258(19)52451-6
AB - Since 1922 when Wu proposed the use of the Folin phenol reagent for the measurement of proteins (1), a number of modified analytical procedures utilizing this reagent have been reported for the determination of proteins in serum (2-6), in antigen-antibody precipitates (7-9), and in insulin (10).
```



## 5 Conclusion

Based on initial tests, it's advisable to approach publication import automation cautiously and ensure manual quality checks are in place – not all metadata are equal, and besides complete, we also want our data to be correct. No database can encompass all sources. While an open, comprehensive database by itself doesn't close the gap, it represents an important step in the right direction. Therefore, it's crucial that we explore OpenAlex more extensively in our daily work with information management. OpenAlex offers many opportunities we have yet to fully explore. Let's continue to delve into this resource and see what we can develop from it.

