OPEN METADATA SOURCES

COMPARING OPENALEX (MAG FORMAT) TO CROSSREF

DATE: 27 MARCH 2022

[PRELIMINARY VERSION]

Executive Summary

In January 2022, OpenAlex was launched as a source of open bibliographic metadata. Intended both as a replacement of and improvement on Microsoft Academic, it provides structured data on publications, authors, institutions and publication venues.

In this project, we assess and compare the value added by OpenAlex to Crossref metadata, both in coverage of publications and other research output (with and without DOIs) as well as in coverage of metadata (including identifiers) for authors, institutions, publication venues and disciplines.

The report currently contains all the graphs comparing metadata coverage of OpenAlex (MAG format) compared to Crossref, and of DOIs vs non-DOIs in OpenAlex (MAG format), as well as some basic tables. More explanatory text and interpretation of findings will be added in a later version.

Complete data and code are available at:
https://github.com/Curtin-Open-Knowledge-Initiative/open-metadata-report
with all images and data belonging to this report located in /reports/run_20220327_2

Introduction and Background

In January 2022, OpenAlex was launched as a source of open bibliographic metadata. Intended both as a replacement of and improvement on Microsoft Academic, it provides structured data on publications, authors, institutions and publication venues.

Many tools, projects and services relied on Microsoft Academic as source of largely open metadata, and might consider switching to OpenAlex. More broadly, the launch of OpenAlex has increased interest in the potential of open metadata to enable discovery, linking and integration of data on research processes and outputs.

Unlike metadata from closed sources, open metadata can be combined and enriched to provide a rich open metadata landscape. Transparency and provenance allow identifying and addressing existing gaps and biases in coverage and quality.

In this project, we assess and compare the value added by OpenAlex to Crossref metadata, both in coverage of publications and other research output (with and without DOIs) as well as in coverage of metadata (including identifiers) for authors, institutions, publication venues and disciplines.

Data sources

This report was run using the following tables as source data:

- Crossref: academic-observatory.crossref.crossref_metadata20211007
- Crossref Member Data: utrecht-university.crossref.member_data with date 20220311
- OpenAlex (MAG format): {'Papers': 'utrecht-university.OpenAlex.Papers20211011', 'Affiliations': 'utrecht-university.OpenAlex.Affiliations20211011', 'Authors': 'utrecht-university.OpenAlex.Authors20211011', 'Journals': 'utrecht-university.OpenAlex.Journals20211011', 'ConferenceInstances': 'utrecht-university.OpenAlex.ConferenceInstances20211011', 'ConferenceSeries': 'utrecht-university.OpenAlex.ConferenceSeries20211011', 'PaperAuthorAffiliations': 'utrecht-university.OpenAlex.PaperAuthorAffiliations20211011', 'FieldsOfStudy': 'utrecht-university.OpenAlex.FieldsOfStudy20211011', 'FieldOfStudyExtendedAttributes': 'utrecht-university.OpenAlex.FieldOfStudyExtendedAttributes20211011', 'PaperAbstractsInvertedIndex': 'utrecht-university.OpenAlex.PaperAbstractsInvertedIndex20211011', 'PaperFieldsOfStudy': 'utrecht-university.OpenAlex.PaperFieldsOfStudy20211011', 'PaperExtendedAttributes': 'utrecht-university.OpenAlex.PaperExtendedAttributes20211011', 'PaperResources': 'utrecht-university.OpenAlex.PaperResources20211011', 'PaperUrls': 'utrecht-university.OpenAlex.PaperUrls20211011', 'PaperMeSH': 'utrecht-university.OpenAlex.PaperMeSH20211011', 'doi': 'utrecht-university.OpenAlex.doi20211011', 'Author': 'utrecht-university.OpenAlex.Author20211011', 'Concept': 'utrecht-university.OpenAlex.Concept20211011', 'Institution,Venue': 'utrecht-university.OpenAlex.Institution,Venue20211011', 'Work':

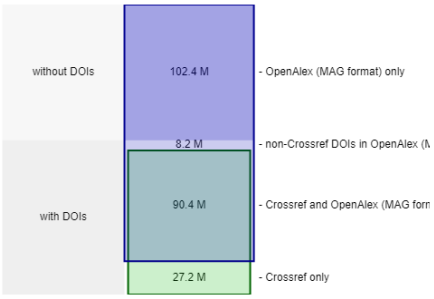
'utrecht-university.OpenAlex.Work20211011', 'additional_source_journal_fields': ', journal.Issns', 'additional_source_org_fields': ', affiliation.RorId, author.Orcid', 'additional_truthtable_fields': '\n , CASE\n WHEN (SELECT COUNT(1) from UNNEST(journal.Issns) as issn WHERE TRIM(issn) !="") > 0\n THEN TRUE\n ELSE FALSE\n END\n as has_venue_issn\n , (SELECT COUNT(1) from UNNEST(journal.Issns) as issn WHERE TRIM(issn) !="") as count_venue_issn\n , CASE\n WHEN TRIM(journal.Issn) != ""\n THEN TRUE\n ELSE FALSE\n END\n as has_venue_issnl\n , CASE\n WHEN TRIM(journal.Issn) !="" \n THEN 0\n ELSE 1\n END\n as count_venue_issnl\n , CASE\n WHEN (SELECT COUNT(1) FROM UNNEST(authors) AS authors WHERE authors.Orcid is not null) > 0 THEN TRUE\n ELSE FALSE\n END\n as has_authors_orcid\n , (SELECT COUNT(1) FROM UNNEST(authors) AS authors WHERE authors.Orcid is not null) as count_authors_orcid\n , CASE\n WHEN (SELECT COUNT(1) FROM UNNEST(authors) AS authors WHERE authors.RorId is not null) > 0 THEN TRUE\n ELSE FALSE\n END\n as has_affiliations_ror\n , (SELECT COUNT(1) FROM UNNEST(authors) AS authors WHERE authors.RorId is not null) as count_affiliations_ror\n'}

Complete data and code are available at:
https://github.com/Curtin-Open-Knowledge-Initiative/open-metadata-report
with all images and data belonging to this report located in /reports/run_20220327_2

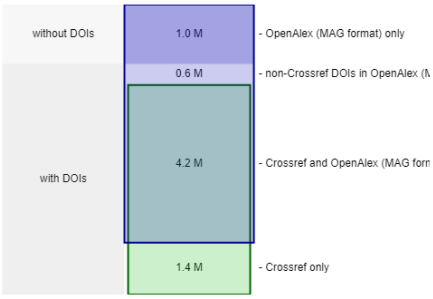
Coverage of OpenAlex (MAG format) vs Crossref

Comparing coverage

Overview

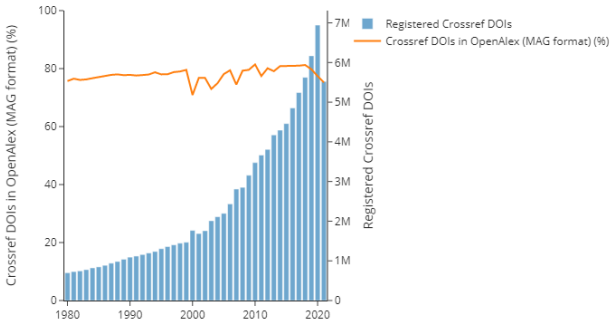


overall comparison - all time

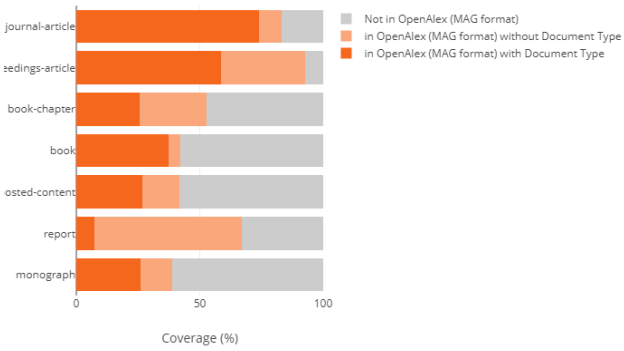


overall comparison - 2021

By year and publication type



coverage by publication date - all time

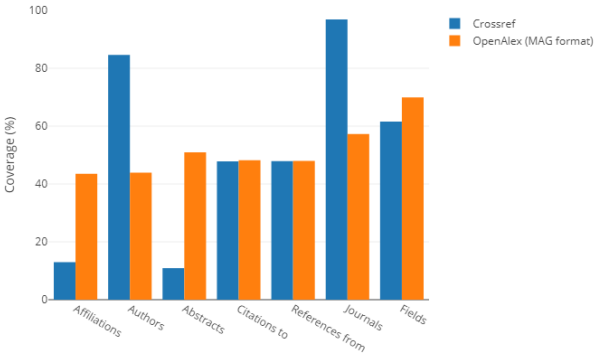


coverage by publication type - all time

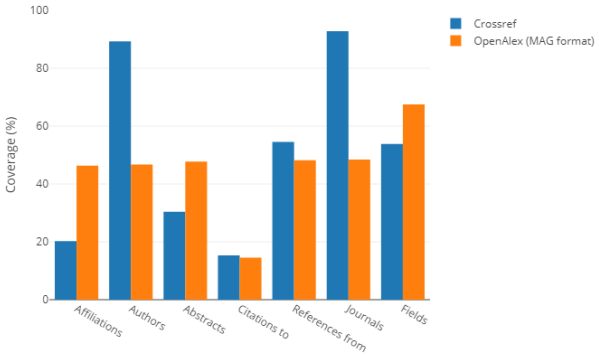
Value Add of OpenAlex (MAG format) to Crossref

Overview

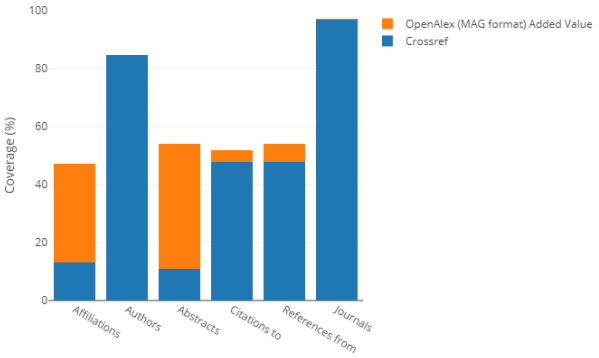
Comparing coverage of metadata types in Crossref and OpenAlex (MAG format)



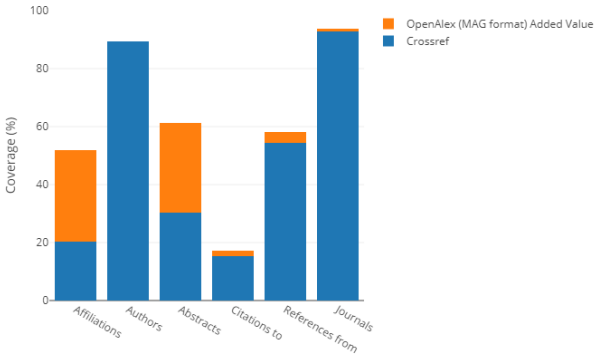
coverage comparison - all time



coverage comparison - 2021



coverage added value - all time

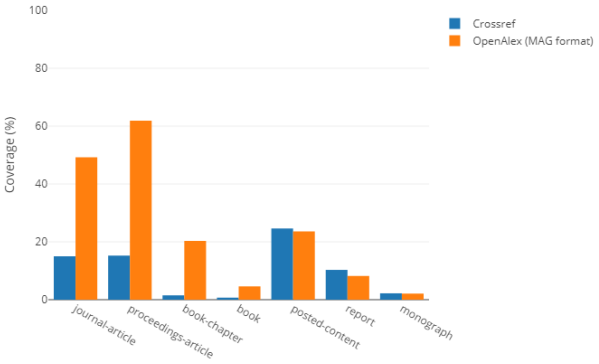


coverage added value - 2021

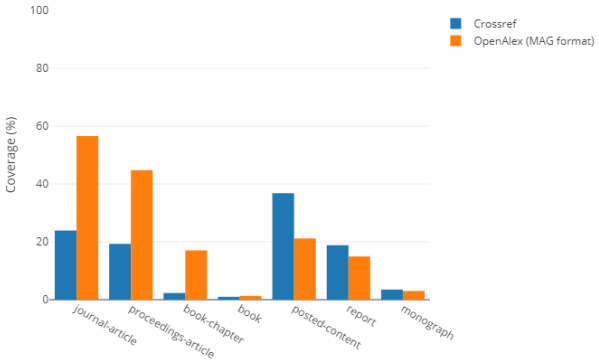
Details

Metadata coverage in OpenAlex (MAG format) and Crossref by publication type

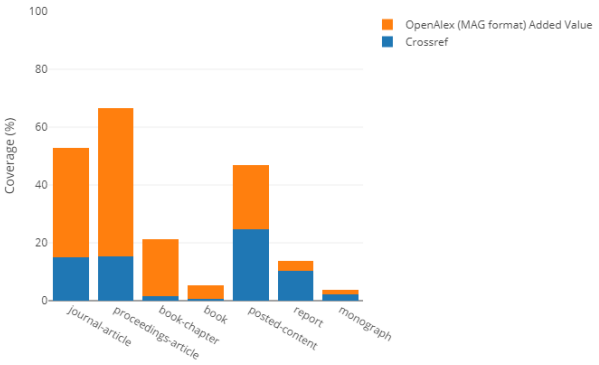
Affiliations



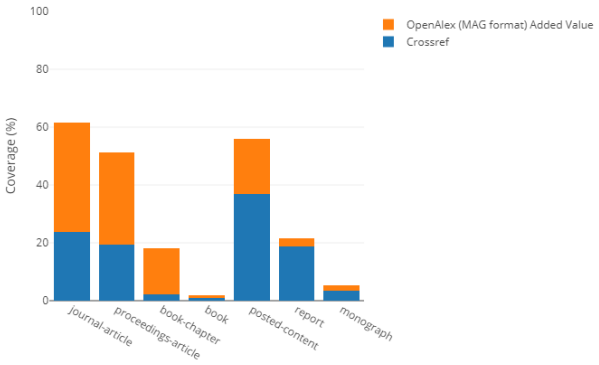
coverage comparison - all time



coverage comparison - 2021

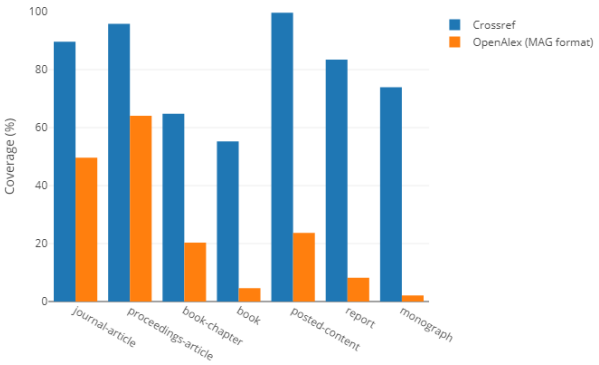


coverage added value - all time

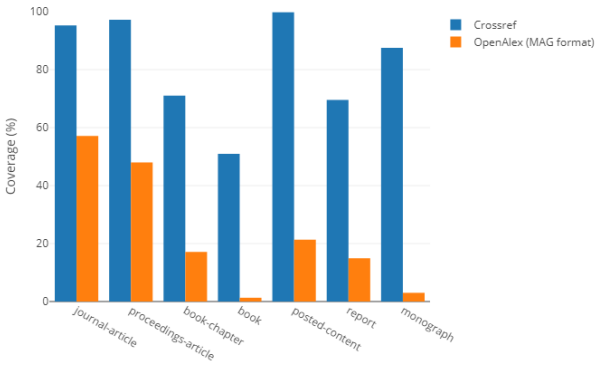


coverage added value - 2021

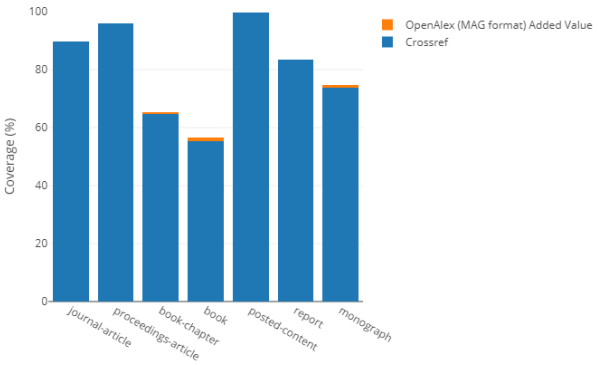
Authors



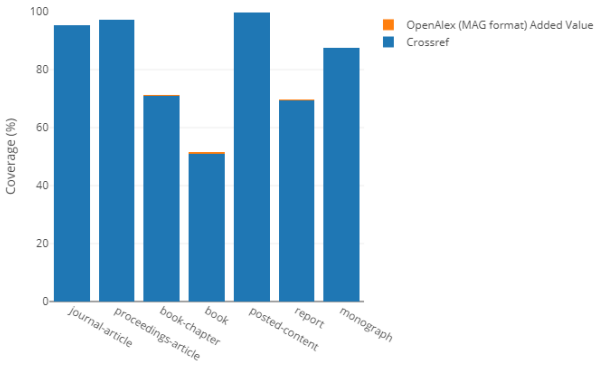
coverage comparison - all time



coverage comparison - 2021

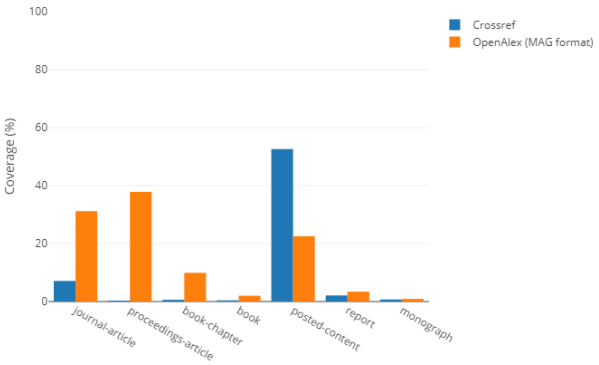


coverage added value - all time

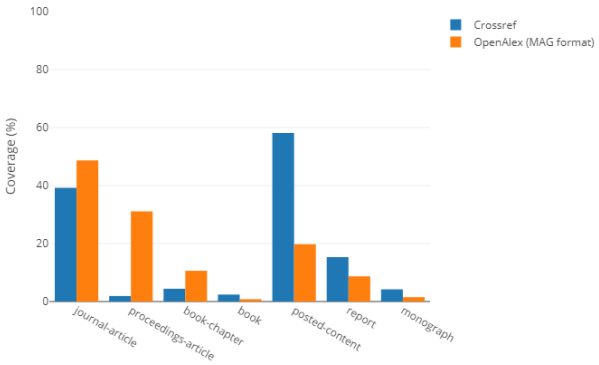


coverage added value - 2021

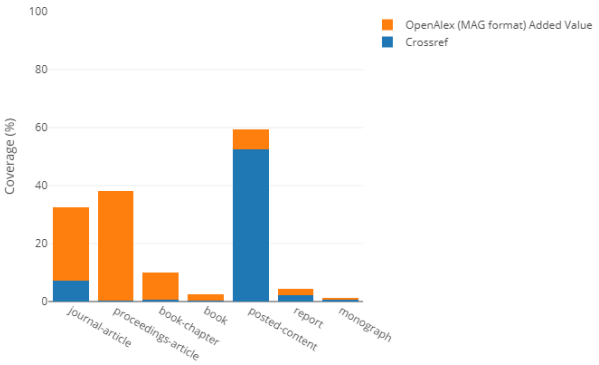
Authors ORCIDs



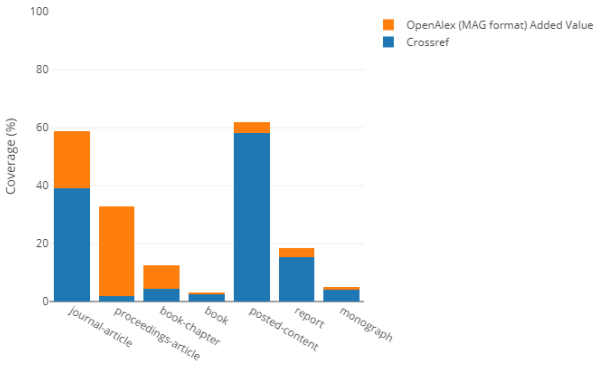
coverage comparison - all time



coverage comparison - 2021

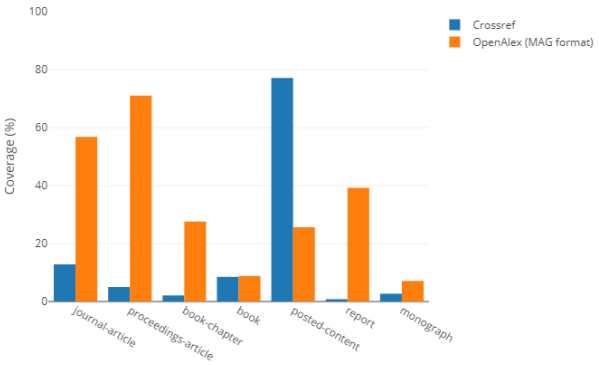


coverage added value - all time

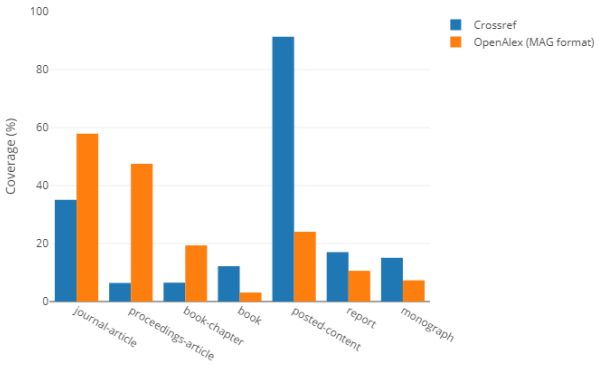


coverage added value - 2021

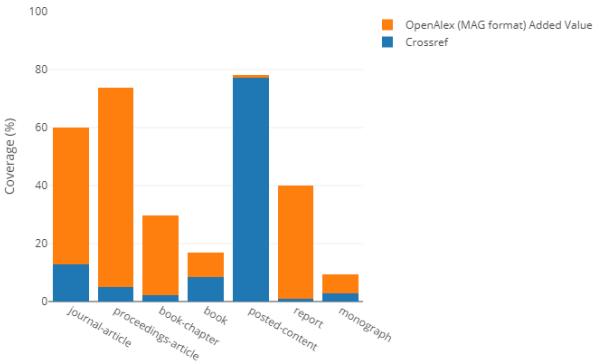
Abstracts



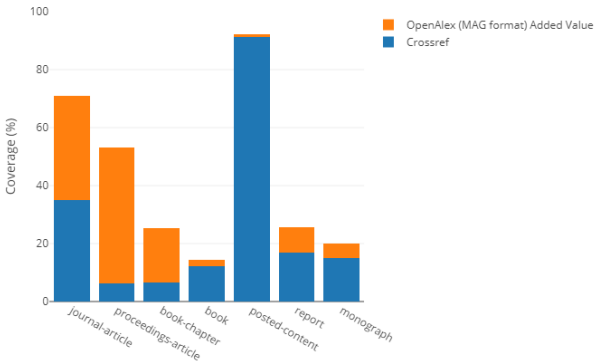
coverage comparison - all time



coverage comparison - 2021

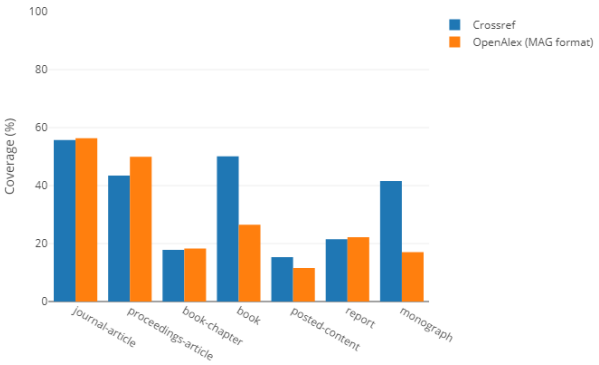


coverage added value - all time

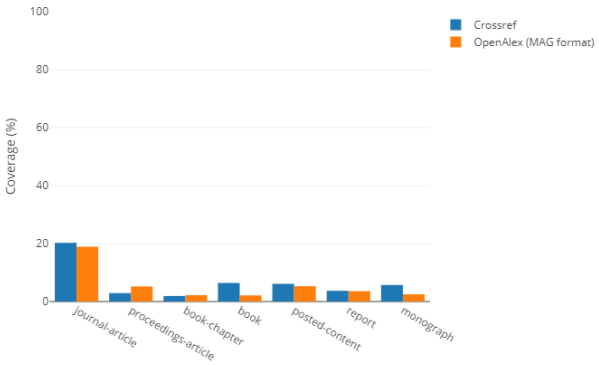


coverage added value - 2021

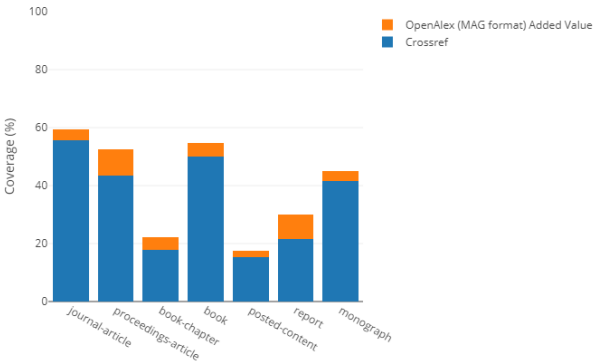
Citations to



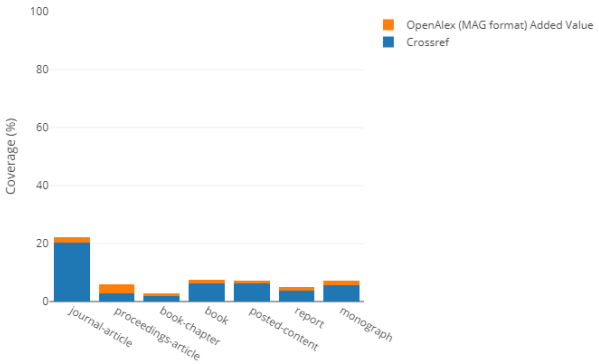
coverage comparison - all time



coverage comparison - 2021

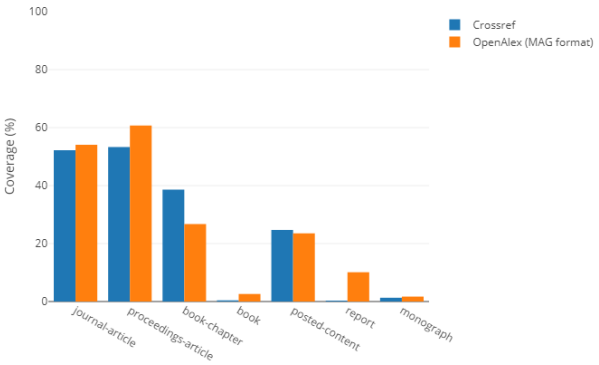


coverage added value - all time

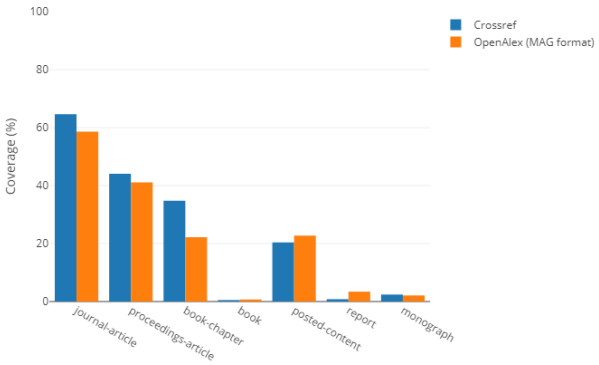


coverage added value - 2021

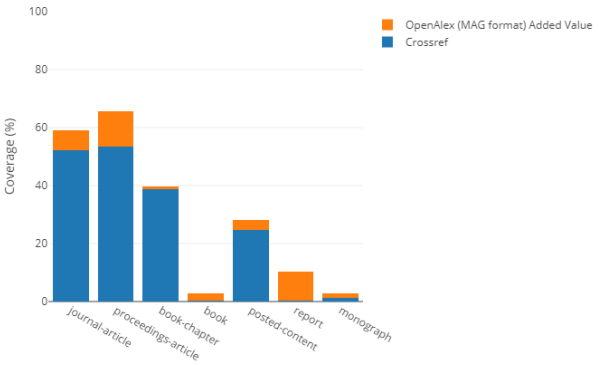
References from



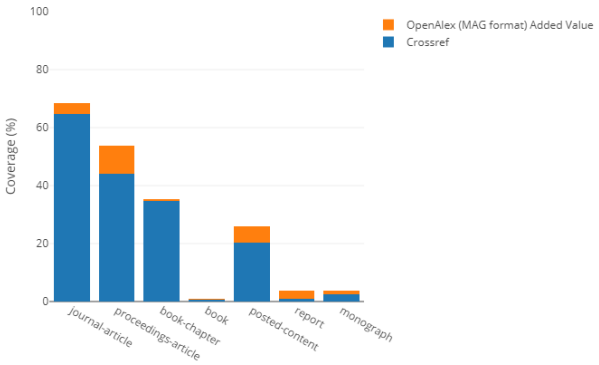
coverage comparison - all time



coverage comparison - 2021

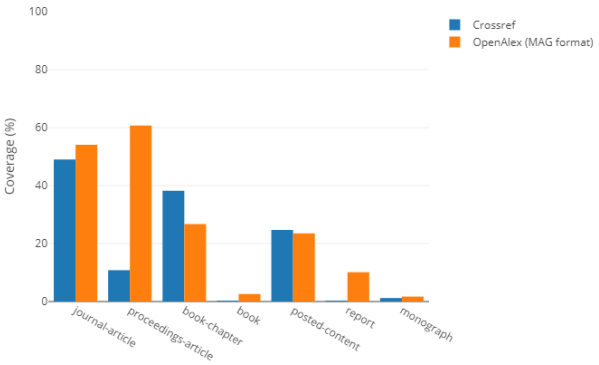


coverage added value - all time

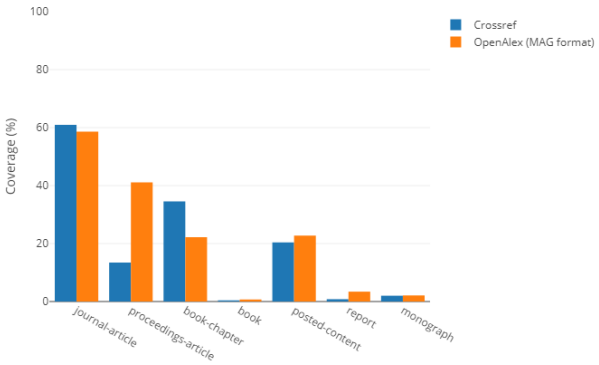


coverage added value - 2021

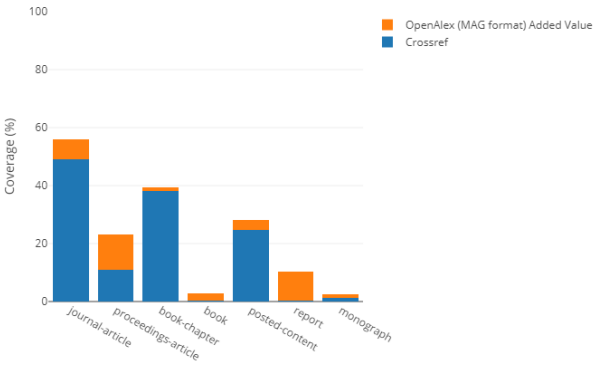
Open References from



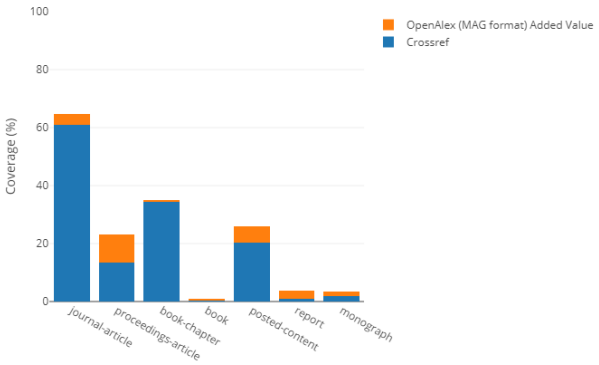
coverage comparison - all time



coverage comparison - 2021

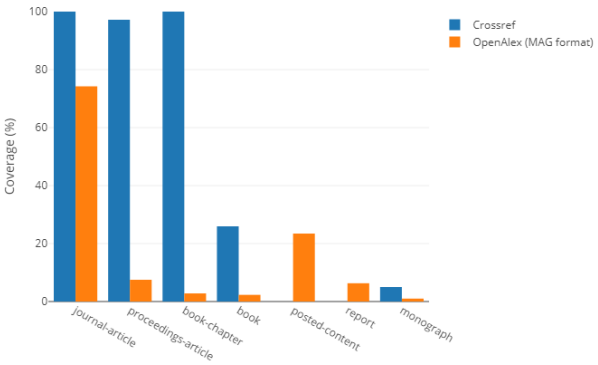


coverage added value - all time

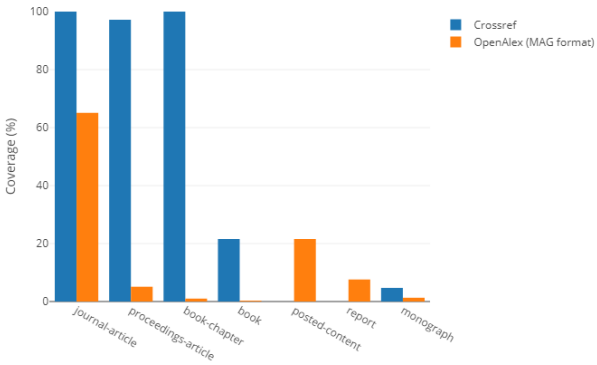


coverage added value - 2021

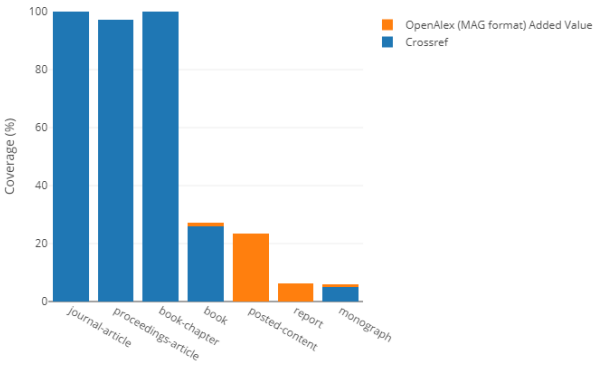
Journals



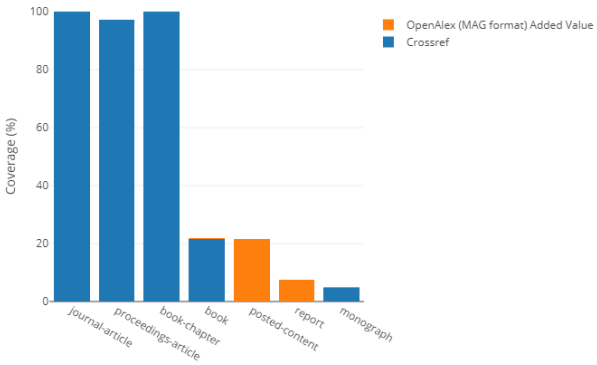
coverage comparison - all time



coverage comparison - 2021

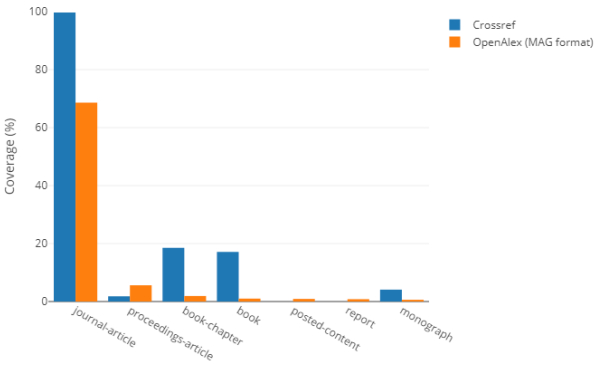


coverage added value - all time

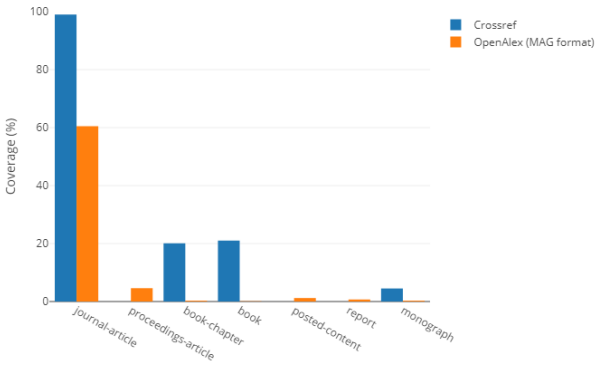


coverage added value - 2021

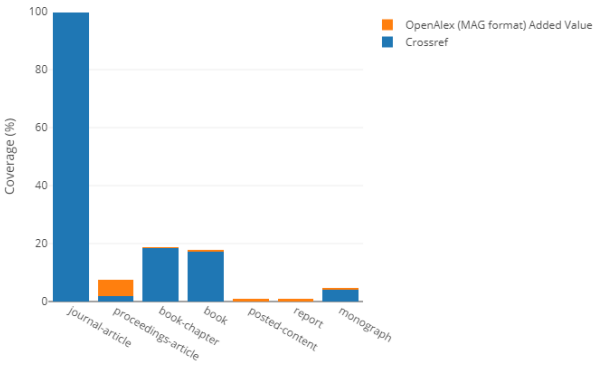
Journals ISSN



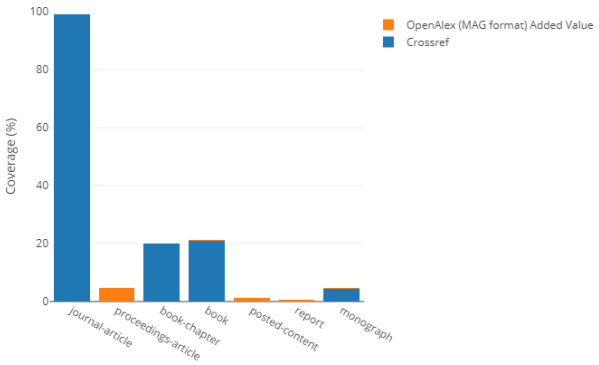
coverage comparison - all time



coverage comparison - 2021



coverage added value - all time

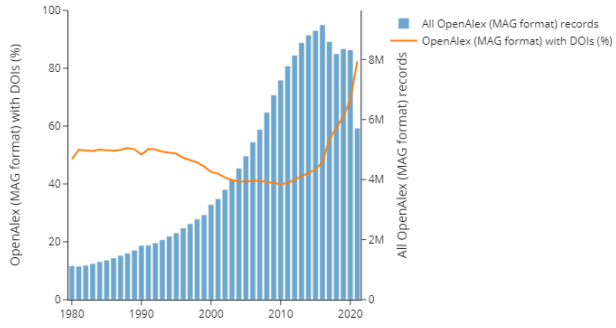


coverage added value - 2021

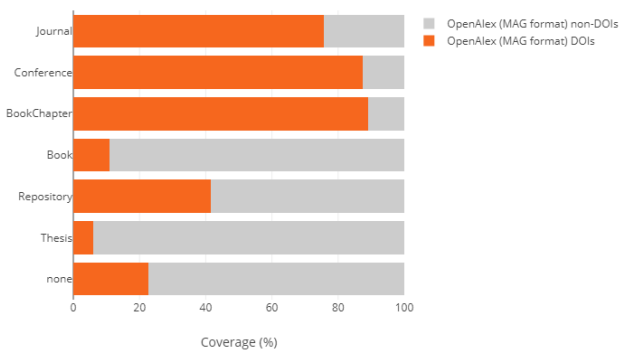
OpenAlex (MAG format) Coverage Beyond Crossref

DOIs vs non-DOIs

By year and publication type



coverage by publication date - all time

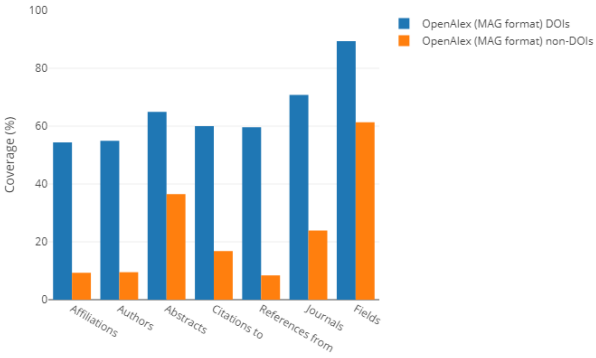


coverage by publication type - all time

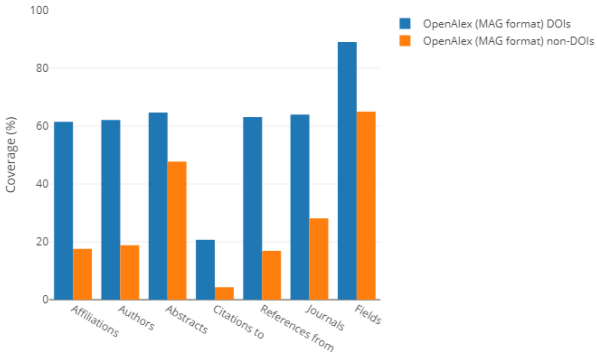
Metadata Coverage

Overview

Comparing coverage of metadata types for DOIs and non-DOIs in OpenAlex



coverage comparison - all time



coverage comparison - 2021

Appendix A - Tables

This section contains tables with summary counts. More tables will be added in a later version.

Crossref Current = 2019-2021
Focus Year = 2021

OpenAlex (MAG format) Coverage

Table 1. OpenAlex (MAG format) Metadata Coverage of Crossref DOIs

Time Frame	Crossref DOIs	OpenAlex Coverage of DOIs
All Time	117531122	90364816
Crossref Current	18635169	14448302
Focus Year	5522198	4153578

Crossref Coverage

Table 2. Crossref Metadata Coverage of Crossref DOIs

Time Frame	Crossref DOIs	Author Strings	Author ORCIDs	Affiliation Strings	Abstracts	Open Abstracts	Field Classification	Venue Names	ISSNs
All Time	117531122	99409110	7077220	15322758	12826809	50680820	72386377	113910712	93033865
Crossref Current	18635169	16243253	4820308	3665413	4723238	8729021	9099790	17469644	13507289
Focus Year	5522198	4929884	1740246	1117559	1679994	2776749	2970744	5122257	4170845