

Institutional Scientific Output and Citation Coverage in OpenAlex: The Case of the Czech Technical University in Prague

Jan Dvořák ^{1,2,3}, Tomáš Chudlarský ^{1,2}, and Josef Špaček ¹

¹ Czech Technical University in Prague, Computing and Information Centre, Czechia
jan.dvorak.2@cvut.cz, tomas.chudlarsky@cvut.cz, josef.spacek@cvut.cz

² Charles University, Institute of Information Studies and Librarianship, Czechia

³ euroCRIS, The Netherlands

Abstract

Institutional Current Research Information Systems (CRIS) increasingly need reliable, affordable data on the scientific output of their institution and on the citations that output receives. OpenAlex, a fully open and freely accessible scholarly index launched in 2022, is a compelling candidate to complement subscription databases such as Web of Science (WoS) and Scopus for these purposes. We report a case study at the Czech Technical University in Prague (CTU), whose in-house CRIS (V3S) is used as a curated baseline. We describe how OpenAlex was integrated with V3S through the institution's ROR identifier and a DOI-based matching procedure, and we measure publication coverage over the period 1996–2025 together with the citations recorded against the institution's outputs. We find that OpenAlex matches 93% of the CTU outputs that carry a DOI, that for recent DOI-bearing output its coverage is competitive with the subscription databases, and that it contributes a substantial number of citations not seen in WoS or Scopus (raising the count of distinct citations to CTU output by about 21%). At the same time, problems appear beneath the surface: aggregate figures that are internally inconsistent, and individual citation counts that are grossly inflated; and even among the outputs all three databases index, none holds the whole citation record. OpenAlex is a highly valuable and freely available research information infrastructure. Institutions that get their DOIs right can expect good publication coverage, which a curated CRIS can confirm against their own records. OpenAlex's citation counts, however, should be treated with caution, and a CRIS cannot validate them on its own.

1 Introduction

OpenAlex [9] is a promising source of research information with global coverage and free access. For operators of an institutional Current Research Information System (CRIS), such a source is doubly interesting: it can help pre-fill the metadata of research outputs for their authors to verify and amend, and it can supply the citations that those outputs receive. Both functions

have typically depended on the subscription databases Web of Science (WoS) [2] and Scopus [1], which OpenAlex is a compelling candidate to complement.

The Czech Technical University in Prague (CTU) is the largest technical university in Czechia and an important one in Central Europe. Its research information is managed in an in-house CRIS developed and operated by the CTU Computing and Information Centre. The system consists of two connected applications: V3S, which records research outputs together with researchers' service to the research community and the honours they receive, and EZOP, which records projects and contracts for contract research. The CRIS has a long history, with its technology stack migrating from Oracle Forms (1995) through a PHP application (2003) to the current pair of Java EE applications (2016); however, the metadata content (held in a common Oracle Exadata database) is what stays, grows, and gets enriched. A fuller description of the system is given in [6]. As of August 2026, V3S holds about 209 thousand research outputs, 32 thousand records of service and honours, and 1.04 million citation records, while EZOP holds about 15 thousand projects and 8 thousand contracts.

Operating such a CRIS, we expect three qualities from a source of research information: (1) comprehensive coverage of the publication output of the institution's researchers; (2) correct affiliation information, so that the institution's output can be identified reliably; and (3) dependable information on the citations that the output receives. In this paper we ask how well OpenAlex meets these expectations for a single institution, using the curated institutional CRIS as the baseline. We describe the integration of OpenAlex with V3S, we quantify publication and citation coverage, we compare OpenAlex with WoS and Scopus, and we examine the reliability of the data it provides.

2 Background

OpenAlex was established in 2022 by the non-profit OurResearch as a successor to the Microsoft Academic Graph, discontinued at the end of 2021 [9]. OpenAlex is fully open: its data are available under CC0 through a well-documented API and as periodic data dumps. It describes several interlinked entity types, including scholarly works (about 320 million records, spanning journal articles, conference papers, preprints and datasets, and carrying citation links between works), authors, sources (journals, repositories and other venues), institutions, publishers and funders. A substantial part of the metadata originates from Crossref [7], which OpenAlex sources and enriches.

The relative coverage of bibliographic data sources has been compared at large scale and across disciplines [13, 8], consistently finding substantial differences between databases and no single source that subsumes the others. While both of those studies predate OpenAlex and assess Microsoft Academic, the questions they raise about coverage, metadata completeness and citation accuracy carry over. So do their answers — in part — as Scheidsteger and Haunschild [12] document that practically all Microsoft Academic works were carried into OpenAlex with their publication year, volume, pages, DOI and reference counts preserved. For publication years before 2021 OpenAlex therefore inherits the corpus those studies assessed.

OpenAlex itself has since been assessed directly. Delgado-Quirós and Ortega [5] compare the completeness of publication metadata across the free-access scholarly databases and place OpenAlex among the more complete of them, well ahead of the academic search engines, while noting the characteristic weakness of such aggregators: information lost in integrating their sources. Culbert et al. [4] compare the reference coverage of OpenAlex, WoS and Scopus over the publications all three hold and find them on a par, while noting that OpenAlex's much larger corpus raises the question of what the additional content is. Others report gaps that

fall unevenly: in the reference lists OpenAlex holds [15], and in its coverage of particular countries [14].

Comparatively little of this work takes the perspective of a single institution that can validate the data against its own curated records. The nearest exception is Rodrigues et al. [11], who start from the DOIs held in their institution’s CRIS and cross-reference the three platforms to discover a substantial number of citation links missing from Web of Science and from Scopus.

3 Data and Methods

3.1 The institutional CRIS as the baseline

The V3S component of the CTU CRIS records research outputs entered by the authors themselves or by staff at their research groups and departments. The content is further validated by faculty and university staff, and reported to the Czech national research information system (RIV). Each output carries structured metadata, including an output type and, where available, a Digital Object Identifier (DOI). Because the set of outputs is curated locally and is authoritative for the institution, it can serve as the baseline for the question “which of our outputs does OpenAlex know about?” in a way that a global comparison between databases cannot. We study the 30-year period between 1996 and 2025. This window bounds the publication coverage analysis; the citation counts of Section 4.3 are deliberately not bounded in time.

The data used in this study reflect the state of the CRIS on 24 August 2026. The OpenAlex records were fetched from its API between 8 May and 24 August 2026.

3.2 Integrating OpenAlex with V3S

The integration follows a pipeline built on top of the OpenAlex API. First, the institution record is located in OpenAlex by searching for the CTU Research Organization Registry (ROR) [10] identifier [03kqpb082](#). Second, the OpenAlex identifiers of all works with an affiliation to the institution are fetched, and the corresponding work records are retrieved. These records are then matched against the research outputs in the CRIS. To catch outputs that OpenAlex does not associate with the institution’s affiliation, the DOIs of the outputs that remained unlinked are also queried directly against OpenAlex, and the returned records are matched as well. Finally, for the linked outputs, the identifiers of the works citing them are fetched, the unique citing records are retrieved, and the citing records are inserted into the CRIS as citations.

3.3 Matching procedure

The DOI is the key for matching. Candidate links are proposed by the affiliation-driven harvest and the direct DOI queries, both described in Section 3.2, and a candidate is considered only when the DOI recorded in V3S and the DOI recorded in OpenAlex agree.

However, a DOI entered in the CRIS sometimes belongs to a different publication. Each candidate is therefore corroborated on four further criteria before the link is written: (1) the publication years must not differ by more than two; (2) the OpenAlex work type and the V3S result kind must be compatible, according to a compatibility matrix maintained for the purpose; (3) the two titles must reach an edit-distance similarity of at least 90%; and (4) the DOI must identify a single OpenAlex record, since a duplicate on the OpenAlex side leaves the target ambiguous. The similarity threshold is set conservatively, since in a CRIS a missed match can be recovered later whereas a wrong one propagates into reported output

and citation counts and is laborious to trace and repair. The score is computed using Oracle's `utl_match.edit_distance_similarity` function over titles first normalised by stripping markup and inline mathematics, collapsing whitespace and folding case. As subtitles are generally absent from OpenAlex titles, a title that is between 40% and 55% of the length of its counterpart is compared not against the whole of it but against an equally long leading portion of it. The criteria are also diagnostic: a rejected candidate usually indicates a misstated DOI on the V3S side, and many were repaired on that evidence. Cleaning and matching therefore ran as a cycle, with each round of repairs and each newly harvested batch of OpenAlex records prompting a fresh matching pass. The criteria are applied when a link is written rather than continuously, and a link may also be recorded as manually checked, so a standing link need not satisfy them against later data.

Matching on the DOI alone means that not all records on each side can take part, and we define the two populations accordingly. On the OpenAlex side, a work is matchable if it carries a DOI. On the V3S side, an output is matchable if its result kind is one for which V3S defines a DOI field at all: this covers journal articles, conference papers, books, book chapters and proceedings volumes. These are groups of finer kinds rather than single types: the first also holds abstracts published in journals and editorials, and the third holds textbooks, lecture notes and dictionaries. As V3S is currently configured, the matchable set excludes research reports, qualifying works, patents, software, prototypes, methodologies, artistic outputs and lectures. The matchability predicate is a property of the output type rather than of the individual record, so it is decided independently of the matching outcome, and in our data it is also exact, in that no V3S output outside these types carries a DOI. Of the 193 thousand CTU outputs recorded in V3S for 1996–2025, approximately 145 thousand are matchable in this sense.¹

Coverage is reported over the matchable population as the match rate (the share of matchable V3S outputs matched to an OpenAlex work), the DOI coverage (the share of them that carry a DOI) and the DOI match yield (the share of the DOI-bearing outputs that are matched). Because a DOI is a precondition for matching, the first quantity is bounded by the second. The match rate measures how much of the output OpenAlex covers, the DOI coverage how much of it the CRIS has made matchable at all, and the yield how well OpenAlex does with what it is given. Alongside the match counts we record, for each matched output, whether its OpenAlex record also carries a CTU affiliation, and separately the number of CTU-affiliated OpenAlex works that matched no V3S output.

Earlier stages of this work admitted a handful of matches on title similarity where no DOI was available. Their negligible number (≤ 15) and the gain in clarity led us to remove those links from the dataset, so every match reported here rests on an agreeing DOI, corroborated when the link was written or confirmed by hand.

4 Results

4.1 Publication coverage

Figure 1 shows the coverage of matchable V3S outputs in OpenAlex by year of publication. The overall match rate rises steadily from under 3% for output published in the late 1990s to 63% for 2024 and 66% for 2025. The DOI coverage of V3S follows the same shape one step above it, reaching 66% and 70% for those two years. The DOI match yield is high and remarkably

¹About one matchable output in thirteen is of a kind — mainly conference abstracts, proceedings volumes and university lecture notes — for which V3S itself records a DOI in only about 2% of cases.

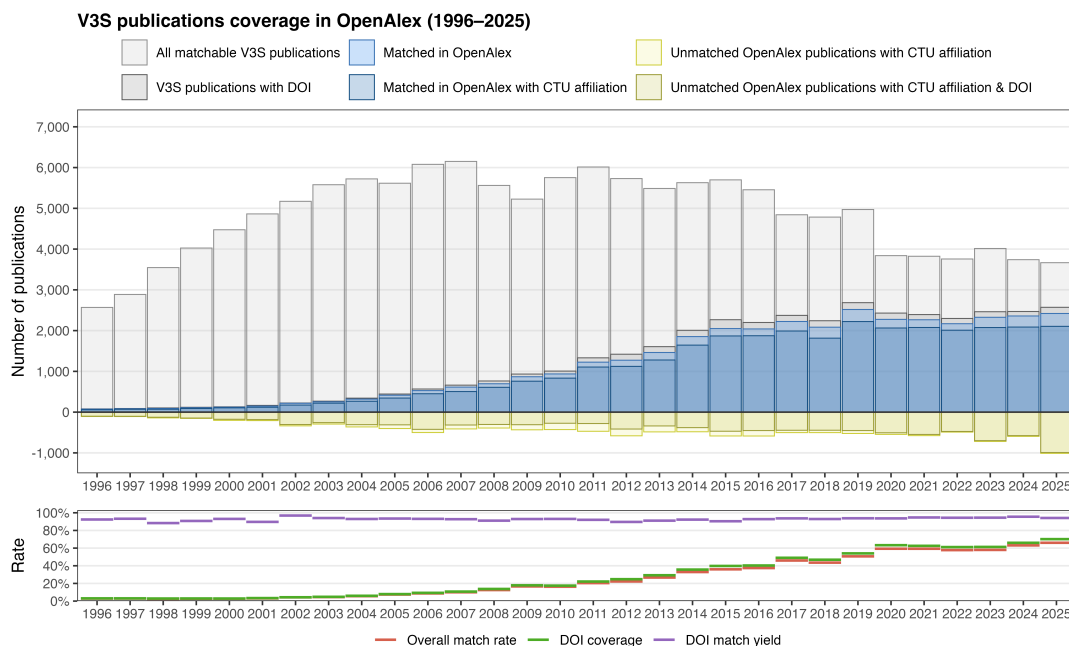


Figure 1: Coverage of matchable V3S publications in OpenAlex by year of publication, 1996–2025. Top: counts per year as nested rectangles, named in the legend, with V3S outputs above the axis and unmatched OpenAlex works below it; matched outputs are counted by their DOI, which is the matching key. Bottom: overall match rate (matched over all matchable outputs), DOI coverage (matchable outputs with a DOI over all matchable) and DOI match yield (matched over matchable outputs with a DOI).

stable across the whole period: 93% over the thirty years, and between 88% and 96% in every individual year.

The yield understates what OpenAlex holds. Of the 2,636 DOI-bearing outputs that remain unmatched over the period, only 122 carry a DOI that OpenAlex does not hold at all. For the other 2,514 the two DOIs agree, but the candidate failed one of the corroboration criteria of Section 3.3 and the link was withheld. Once an output carries a DOI, OpenAlex knows it in 99.7% of cases, and the yield measures not the reach of OpenAlex but how much of that reach we are prepared to trust. As candidate pairs rather than outputs, the two DOIs agree in 38,699 cases, of which 36,171 were linked and 2,528 withheld. Almost all of the withholding is done by the title criterion, which fails for 2,174 of them and is the only check to fail for 1,904; type compatibility fails for 488, the publication year for 107, and in 34 the DOI does not identify a single OpenAlex record. The distribution of those title scores is instructive: only 210 fall below 20, where the two records plainly describe different works and the V3S DOI is simply wrong, while 1,408 score 70 or above, 613 of them between 85 and 89, just under the threshold, which was set by judgement rather than measurement. Agreement on a DOI between a curated CRIS and OpenAlex is therefore right about the work in roughly 99.5% of cases — 212 of the 38,699 pairs score below 20 — and the price of catching the remaining half per cent is paid in recall: the withheld candidates are far more often the same work under a different rendering of its title

than a different work altogether. Publication coverage is therefore, to a first approximation, a function of DOI availability: the poor coverage of older output is driven mainly by missing DOIs rather than by gaps in OpenAlex itself. The denominator is deliberately inclusive (Section 3.3), so the match rate reported is a conservative floor.

Figure 1 also shows two mismatches in the affiliation link, which coverage rates alone do not reveal. First, not every matched work is attributed to CTU in OpenAlex: over the period 88.9% of the matched outputs carry a CTU affiliation in their OpenAlex record, and 86.9% for 2025 (2,105 of 2,421), so an institutional harvest driven by the ROR identifier alone would miss roughly one matched output in nine; the direct DOI queries described in Section 3.2 recover them.

Second, the rectangles below the axis show the converse case: works that OpenAlex attributes to CTU but that were not matched to any V3S output. These number several hundred to a thousand in every year of the period — 595 for 2024 and 1,012 for 2025 — and almost all of them carry a DOI, so they are not an artefact of missing identifiers on the OpenAlex side.

A breakdown by the expanded OpenAlex work types shows what some of these unmatched works are. Across all years, 1,482 of the 1,634 CTU-affiliated preprints are unmatched, as are all 724 datasets, all 147 software records and all 64 reports: kinds of output that V3S either does not record or does not record as publications — dataset records were introduced in V3S only in 2026 — and which an institution should not expect to reconcile.

The remainder of these unmatched works, approximately 13 thousand typed as articles, conference papers or book chapters, is the genuinely problematic residue. These are either outputs missing from the CRIS, records whose registration in V3S lags behind publication, or affiliations wrongly assigned by OpenAlex. Distinguishing between these cases requires manual inspection, which is only viable if distributed among the relevant faculty and research group staff. For now this limits how closely an institution’s OpenAlex profile can be reconciled with its CRIS.

4.2 Overlap with Web of Science and Scopus

Considering the outputs published in the studied period that can be linked to at least one of the three databases, the three sources overlap heavily (Figure 2). Over the whole period, Scopus is the broadest single source, covering about 41 thousand CTU outputs, while WoS and OpenAlex each cover about 36 thousand; roughly 47 thousand outputs are covered by at least one of the three. A large common core of about 27 thousand outputs is present in all three databases simultaneously. The number of outputs found only in OpenAlex is small (about 1.8 thousand), which indicates that, at the level of publications, OpenAlex adds little unique coverage beyond WoS and Scopus.

The aggregate is dominated by the older output, where the DOI coverage of V3S is thin and Scopus holds the lead. Restricted to the outputs that carry a DOI, the recent years look different (Figure 3): OpenAlex has been the broadest single source in every year since 2015, reaching 2,421 of the 2,572 DOI-bearing outputs published in 2025, against 2,314 for Scopus and 1,891 for WoS, and its margin over WoS has widened since 2022. The three sources nevertheless stay close. The same view shows that only 1.9% of the DOI-bearing outputs of 2025 (49 of 2,572) are absent from all three databases, which confirms from the other side what Section 4.1 found: what limits coverage is the availability of the DOI, not the reach of the databases.

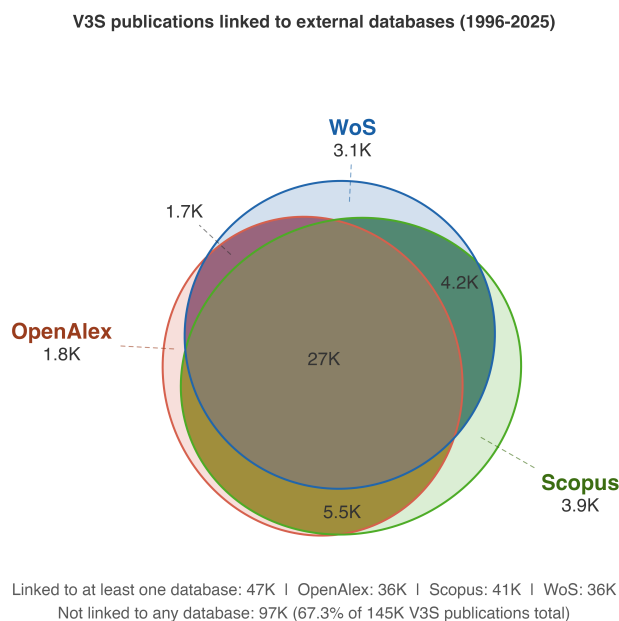


Figure 2: Overlap of CTU publication coverage across WoS, Scopus and OpenAlex, over the outputs published in 1996–2025 that at least one of the three indexes.

4.3 Citation coverage

The picture is different for citations. The counts that follow are not restricted to the matched outputs: they count every citation recorded against any of the 40,769 CTU outputs that carry one, from whichever database. They are deliberately not bounded in time either: a citation counts whatever the publication year of the cited output or of the citing work. Citations entered by hand are set aside throughout: a citation with no source database has no place in a comparison between them. While manual entry is now very rare, the records that accumulated under it remain; they account for the gap between the totals here and the count of citation records in Section 1.

Only an output matched to OpenAlex can carry a citation harvested from OpenAlex, so that database’s contribution is bounded by the matching. From WoS and Scopus alone, CTU outputs carry about 855 thousand distinct citations (about 669 thousand from WoS and 698 thousand from Scopus, with about 511 thousand shared). Adding OpenAlex, which itself contributes about 679 thousand citations, raises the total number of distinct citations to about 1.03 million (Figure 4, upper panels). OpenAlex thus adds roughly 176 thousand citations that appear in neither WoS nor Scopus, an increase of about 21% over the WoS-plus-Scopus total. Section 4.4 looks into these additional citations in more detail.

Of the OpenAlex citations, about 368 thousand² are present in all three databases. Seen

²The attentive reader may have spotted a gap of approximately 5 thousand citations between the intersections in the upper and lower right panels of Figure 4. It is caused by citations where the cited publication was co-authored by CTU researchers but was represented in WoS only as a bibliographic reference string, so the citation is counted while the cited output does not qualify for the lower row. Such citations were common in the past

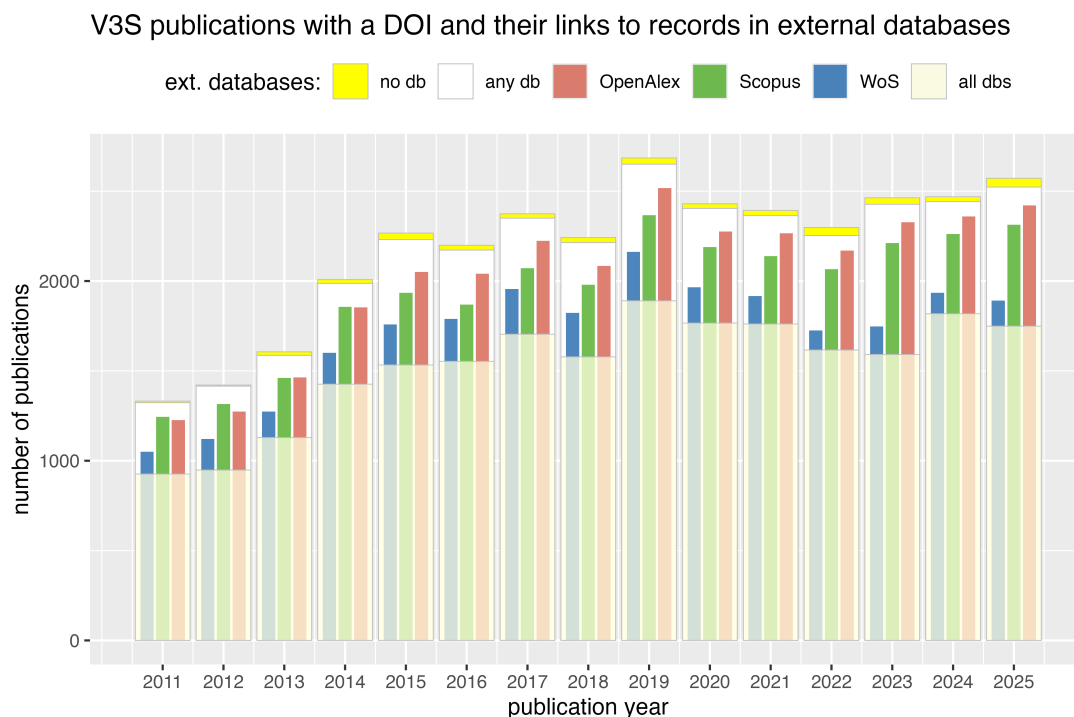


Figure 3: V3S publications carrying a DOI and their links to records in external databases, 2011–2025. For each year, the outlined bar gives all DOI-bearing publications, the yellow cap those linked to no database; the three coloured bars give the publications linked to WoS, Scopus and OpenAlex respectively; and the pale overlay gives those linked to all three at once.

from the other side, no single database holds the whole record: of the 1.03 million, about 105 thousand are recorded only by WoS, 105 thousand only by Scopus and 176 thousand only by OpenAlex.

Those counts still reflect which outputs each database indexes at all, so a comparison needs a common basis. Confining it to the 24,651 of those cited outputs that all three index, which carry 767 thousand citations between them, leaves the shortfall large and no longer even: WoS lacks 35% of those citations, Scopus 26% and OpenAlex 22%, and fewer than half of them (47%) are recorded by all three (Figure 4, lower right). The asymmetry is sharper in what each source contributes alone: 145 thousand of these citations are recorded by OpenAlex and by neither of the others, against 64 thousand for Scopus and 29 thousand for WoS. The same restriction separates the two subscription databases, which the unrestricted view does not: over all cited outputs each lacks about a fifth of the citations the two record between them, 22% for WoS and 18% for Scopus, while over the restricted set WoS lacks 20% and Scopus 9% of the pair's total (Figure 4, left panels). The restriction controls the cited side only — a database records a citation only if it holds the citing work as well — so the differences between the three are in part a difference in how much of the citing literature each indexes. Nor is this population the common core of Section 4.2, which is bounded to the studied period and counts outputs

and have become rare in recent years.

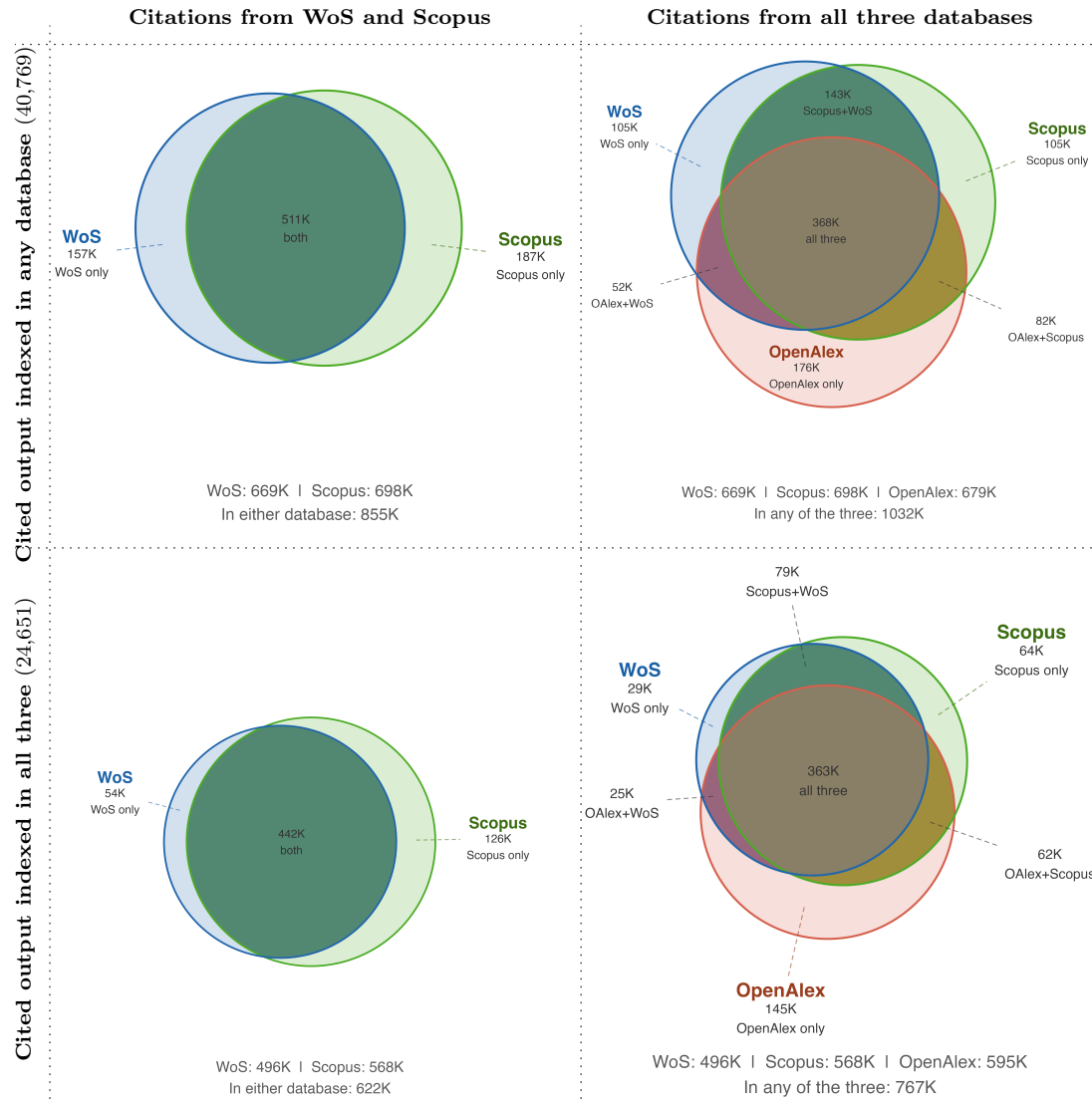


Figure 4: Distinct citations to CTU outputs. In the left column the citations recorded by OpenAlex alone are set aside, so the two columns differ by OpenAlex only. The four panels are drawn to one area per citation, so their sizes may be compared directly.

whether or not they have been cited; the two are close in size but neither contains the other.

4.4 Provenance of the OpenAlex citations

To assess the reliability of the OpenAlex citations, we returned to the full set of cited outputs and examined the source type of the citing works, distinguishing the citations that OpenAlex shares with another database from those that OpenAlex contributes alone (Figure 5). About 74% of OpenAlex citations (some 502 thousand) also appear in WoS or Scopus, and these overlapping

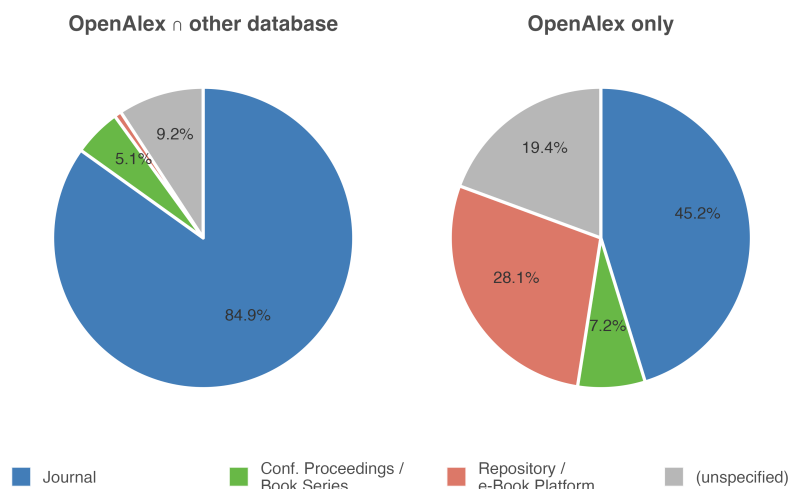


Figure 5: Source type of the citing works, for OpenAlex citations shared with another database (left) and for OpenAlex-only citations (right).

citations come overwhelmingly from journals (about 85%). The 26% of OpenAlex citations that are unique to OpenAlex (some 176 thousand) have a markedly different composition: only about 45% come from journals, while about 28% come from repositories or e-book platforms and about 19% come from sources of unspecified type, with the remainder from conference proceedings and book series. The citations that OpenAlex adds beyond the subscription databases therefore originate disproportionately from grey or less-curated sources, so their individual reliability is more variable than that of the citations shared with WoS and Scopus.

4.5 Data-quality anomalies

While the aggregate coverage rates look encouraging, some high-level numbers raise suspicion. The first anomaly is an internal inconsistency in OpenAlex itself. In May 2026 the institution profile page for CTU reported about 49,650 works and about 4,072,000 citations, while a works search filtered to the same institution returned about 47,450 works and only about 793,700 citations — a difference of more than fivefold in the citation count between two views of the same institution within the same system. The gap in the number of works has since widened rather than closed: by August 2026 the profile reported 52,087 against 49,510 from the search, and the search returned that same count whether it was filtered on the institution or on its lineage. The profile still reported a citation total, which had risen to about 4,355,000, but the comparison itself could no longer be made, because the works-search view had stopped displaying one.

The second anomaly concerns individual records. In May 2026 the most-cited CTU work according to OpenAlex was the 2023 conference paper “MizAR 60 for Mizar 50”, published in the Dagstuhl LIPIcs series for the 14th International Conference on Interactive Theorem Proving and credited with 75,682 citations, against 4 in Web of Science. That record has since been corrected³ and by August 2026 stood at zero, itself an undercount against the four in

³We had filed a curation request against this record; whether the correction followed from it we cannot say,

Web of Science. Its place at the top of the institution’s ranking was taken by another 2023 record from the same publisher, credited with 79,074 citations — 86 times the 917 citations of the next most-cited CTU work of that year. A single erroneous record of this kind is large enough to distort the institutional citation aggregate on its own; that correcting one record was promptly followed by another of the same shape suggests the fault lies less in the individual records than in how this class of source is processed, and it confirms that OpenAlex citation counts for individual works cannot be taken at face value.

5 Discussion

Our experience with OpenAlex is, at the level of the system, very positive. The API is clearly designed, reliable and fast, and user support for the API is responsive; support for data-curation requests is, by contrast, much slower. Of the two CRIS uses set out in Section 1, OpenAlex serves the first well. For pre-filling the metadata of research outputs and for confirming publication coverage, it is competitive with the subscription databases for recent, DOI-bearing output and is freely available. That does not make the subscription databases dispensable: their coverage of our recent DOI-bearing output is high, and indexing in WoS and Scopus is an input to the national research evaluation procedure our institution is subject to, so they will remain necessary for the foreseeable future.

The single most effective lever for improving coverage is on the institution’s own side: getting the DOIs right in the CRIS. Matching on the DOI is then largely a question of how the identifier is corroborated, and the checks we apply are very unequal: the title comparison rejects four times as many candidates as the type check and twenty times as many as the publication year. That comparison is where the engineering effort belongs — normalisation, and the handling of missing subtitles — and equally where the risk of discarding good matches is concentrated.

OpenAlex’s publication coverage has been found to vary by country: Zheng et al. [14] report it to be incomplete and discontinuous for China, with comparable effects in other non-English-speaking countries and variation across regions. Our institution is in one such country, so the question is a live one for the coverage rates of Section 4.1. What our data can answer is bounded by the matching design. The 99.7% is one answer: for this institution the limit on coverage is the availability of the DOI rather than the reach of OpenAlex. It can say nothing about the outputs that carry no DOI — some three quarters of the matchable population over the whole period — because matching on the DOI alone leaves them unobservable, and a gap of the kind Zheng et al. describe, mediated by a national database, would fall exactly there. These are also the figures of a single large technical university, and an institution with a different output profile should expect different rates.

For augmenting citation data, OpenAlex adds a meaningful volume of citations that WoS and Scopus do not see. The sources of these citations are, however, markedly less journal-based: journals account for about 45% of them, against 85% of the citations OpenAlex shares with another database. This seems to give a partial answer — on the small scale of one institution, and only for the works that cite it — to the question Culbert et al. [4] raise about what the additional content is that OpenAlex covers but neither Web of Science nor Scopus does.

We also second Culbert et al.’s observation that OpenAlex data are a moving target: OpenAlex expanded its work-type vocabulary in mid-2026, with the new types appearing some weeks before the announcement, which obliged us to re-harvest records that our pipeline had previously been unable to ingest. The change was an improvement — conference papers, in

as we received no reply.

particular, are now a type in their own right rather than being subsumed under “article” — but it is a reminder that an integration with OpenAlex has to be maintained, not merely built. The same drift shows in the matching itself: of the links standing over the studied period, 245 would not pass our checks against the data as they now stand, most of them because a second OpenAlex record has since appeared for the DOI.

The caveats concern data quality. The two anomalies of Section 4.5 are errors of excess; the opposite error is documented as well. Zhou et al. [15] find pervasive reference missingness in OpenAlex — works whose reference lists are absent or incomplete — and every such gap withholds a citation from each work the missing list would have pointed to. The missingness is systematic rather than random: it is concentrated in articles from less prestigious institutions, lower-impact journals and countries of lower economic standing, and 57% of the articles below the median on both institutional prestige and journal impact that come from the Global South carry no references at all. We have not measured this effect on our own data, and the CRIS could not reveal it. An institutional citation count drawn from OpenAlex may therefore err in either direction, and the two errors differ in how they can be found — an inflated count announces itself as an outlier, whereas a missing one leaves no trace at all. Both are also systematic, so their effect on an institutional aggregate does not average out as it grows.

However, missing citations are not peculiar to OpenAlex. Rodrigues et al. [11] discover over a thousand citation links absent from each of Web of Science and Scopus, across some two thousand papers, by cross-referencing the three platforms. Our own figures point the same way, and more sharply once the cited outputs are held constant: among the outputs all three databases index, each of the three still lacks a fifth or more of the citations to them, and Web of Science over a third (Section 4.3). None of them is close to complete, and a citation count is therefore a property of the source it was drawn from at least as much as of the work it describes.

In 2020 two of us asked whether open citation data could replace Web of Science for research evaluation and found the coverage of the day insufficient [3], at some 54% of WoS citation links present in the Crossref Open Citation Index. OpenAlex now holds about 63% of the WoS citations to our outputs — a different basis, since the earlier study counted only links whose two ends both carried a WoS accession number and a DOI — so the share has risen without coming close to sufficiency. The findings reported here do not overturn that conclusion; they shift its grounds, from how much of the citation record an open source holds to how unevenly the missing part is distributed.

OpenAlex is therefore well suited to discovery and to maximising breadth of coverage, but its citation counts should not be taken directly into research evaluation or institutional metrics. The citations unique to it skew towards grey and less-curated sources, and independent validation is needed before the counts are relied on. Outlier checks alone will not supply it: by construction they catch only the error of excess.

This leads to an observation about the role of the CRIS. A curated institutional CRIS can validate the publication side of OpenAlex: because the CRIS holds an authoritative list of the institution’s outputs, we can confirm which of them OpenAlex covers and identify what is missing. Rodrigues et al. [11] put a CRIS to a different use: there the CRIS supplies the sample and the databases are checked against one another; here they are checked against the CRIS. The CRIS cannot, however, validate the citation links. It does not hold an independent, complete citation graph, so inflated or spurious citations such as the record of Section 4.5 cannot be detected from within the CRIS alone. Anyone tempted to use OpenAlex citation counts for evaluation should be aware of this asymmetry: publication coverage is verifiable locally, the citation links are not.

6 Conclusions

At the system level, OpenAlex is a highly valuable and freely available scholarly infrastructure; institutions that get their DOIs right will obtain publication coverage competitive with the subscription databases, and they will gain access to a substantial number of citations not available from WoS or Scopus. Beneath the surface, however, data-quality issues remain: aggregate figures that do not reconcile, and individual citation counts that are wildly overstated. Nor does any one source hold the whole citation record: even among the outputs all three databases index, each lacks from a fifth to over a third of the citations to them, so a citation count says as much about where it was drawn from as about the work it counts. For an institutional CRIS, the practical conclusion is twofold: while the institution's publication output can be verified against OpenAlex using the CRIS, the citation links cannot.

Future work includes developing an alternative matching procedure that does not rely on DOIs; calibrating the title-similarity threshold against the candidates that fall just below it, to balance precision against recall; auditing citation outliers, together with the harder problem of detecting the citations that are missing rather than inflated — where repeating the cross-referencing of Rodrigues et al. [11] on our own data would separate a link a database has missed from a work it does not index; establishing an effective feedback loop for curation requests to OpenAlex; and extending the analysis to author disambiguation.

AI usage statement

Anthropic Claude Sonnet and Opus were used in the preparation of this manuscript. They assisted in preparing the visualisations of the underlying data in R, in checking the data and the text for consistency, in retrieving the DOIs of the cited works, in typesetting the article using L^AT_EX, and in improving the style. Microsoft Copilot was used as an independent proof-reading assistant while finalising this paper.

Acknowledgements

This research was partly funded by the project *Strategic Setting of the Research Environment at CTU in Prague* (CZ.02.01.01/00/23_026/0011378) by the Ministry of Education, Youth and Sports of the Czech Republic with co-funding by the European Union.

References

- [1] Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1):377–386, 2020. doi:[10.1162/qss_a_00019](https://doi.org/10.1162/qss_a_00019).
- [2] Caroline Birkle, David A. Pendlebury, Joshua Schnell, and Jonathan Adams. Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1):363–376, 2020. doi:[10.1162/qss_a_00018](https://doi.org/10.1162/qss_a_00018).
- [3] Tomáš Chudlarský and Jan Dvořák. Can Crossref citations replace Web of Science for research evaluation? The share of open citations. *Journal of Data and Information Science*, 5(4):35–42, 2020. doi:[10.2478/jdis-2020-0037](https://doi.org/10.2478/jdis-2020-0037).
- [4] Jack H. Culbert, Anne Hobert, Najko Jahn, Nick Haupka, Marion Schmidt, Paul Donner, and Philipp Mayr. Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*, 130(4):2475–2492, 2025. doi:[10.1007/s11192-025-05293-3](https://doi.org/10.1007/s11192-025-05293-3).

- [5] Lorena Delgado-Quirós and José Luis Ortega. Completeness degree of publication meta-data in eight free-access scholarly databases. *Quantitative Science Studies*, 5(1):31–49, 2024. doi:[10.1162/qss_a_00286](https://doi.org/10.1162/qss_a_00286).
- [6] Jan Dvořák, Tomáš Chudlarský, and Josef Špaček. Practical CRIS interoperability. *Procedia Computer Science*, 146:256–264, 2019. doi:[10.1016/j.procs.2019.01.077](https://doi.org/10.1016/j.procs.2019.01.077).
- [7] Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427, 2020. doi:[10.1162/qss_a_00022](https://doi.org/10.1162/qss_a_00022).
- [8] Alberto Martín-Martín, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations’ COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1):871–906, 2021. doi:[10.1007/s11192-020-03690-4](https://doi.org/10.1007/s11192-020-03690-4).
- [9] Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. doi:[10.48550/arXiv.2205.01833](https://doi.org/10.48550/arXiv.2205.01833).
- [10] Research Organization Registry. ROR. <https://ror.org>, since 2019.
- [11] David Rodrigues, António Lopes, and Fernando Batista. Detecting incoherent citation data among three bibliometric platforms: OpenAlex, Scopus and Web of Science. *Journal of Information Science*, 2025. doi:[10.1177/01655515251330579](https://doi.org/10.1177/01655515251330579).
- [12] Thomas Scheidsteger and Robin Haunschild. Which of the metadata with relevance for bibliometrics are the same and which are different when switching from Microsoft Academic Graph to OpenAlex? *El Profesional de la información*, 32(2):e320209, 2023. doi:[10.3145/epi.2023.mar.09](https://doi.org/10.3145/epi.2023.mar.09).
- [13] Martijn Visser, Nees Jan van Eck, and Ludo Waltman. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1):20–41, 2021. doi:[10.1162/qss_a_00112](https://doi.org/10.1162/qss_a_00112).
- [14] Mengxue Zheng, Lili Miao, Yi Bu, and Vincent Larivière. Understanding discrepancies in the coverage of OpenAlex: The case of China. *Journal of the Association for Information Science and Technology*, 76(11):1591–1601, 2025. doi:[10.1002/asi.70013](https://doi.org/10.1002/asi.70013).
- [15] Huaxia Zhou, Lizhen Liang, and Daniel E. Acuna. Widespread reference missingness disparities in open scholarly metadata. *Quantitative Science Studies*, 7:53–65, 2026. doi:[10.1162/qss.a.400](https://doi.org/10.1162/qss.a.400).