

One University Three Positions: A Reproducible Simulation Study of Database Dependent Institutional Rankings across OpenAlex Scopus and Web of Science

Mustafa Kayyali

Kayyali@heranking.com

Maaref University of Applied Sciences

Research Article

Keywords: OpenAlex, Scopus, Web of Science, university rankings, database coverage, Monte Carlo simulation, research evaluation, reproducibility

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11088900/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

One University Three Positions: A Reproducible Simulation Study of Database Dependent Institutional Rankings across OpenAlex Scopus and Web of Science

Mustafa Kayyali

Maaref University of Applied Sciences, Syria

Email: Kayyali@heranking.com

Abstract

Institutional research rankings are often reported as if a university occupies one stable position. In practice, the position is partly produced by the bibliographic database through coverage, affiliation resolution, citation scaling, document classification, and residual metadata error. This study develops a database-choice uncertainty framework and evaluates it through a fully reproducible Monte Carlo simulation. A fixed population of 700 synthetic universities was generated with heterogeneous size, latent research quality, disciplinary portfolio, language profile, and regional location. OpenAlex, Scopus, and Web of Science were represented by prespecified observation functions informed by documented differences rather than fitted to vendor data. Identical indicator definitions and weights were then applied in 1,000 replications for the reference model and for six sensitivity scenarios. In the reference model, OpenAlex and Web of Science rankings had a mean Spearman correlation of 0.954, yet their mean median absolute rank gap was 40.8 places and their mean top-100 overlap was 81.7%. The apparent contradiction reflects high system-level concordance alongside consequential institution-level movement. Directional effects were strongly patterned: the median synthetic Global South institution ranked 34 places better in OpenAlex than in Web of Science, whereas the median institution in other regions ranked 52 places worse. A Shapley decomposition assigned 65.4% of explained OpenAlex-Web of Science discordance to coverage, 24.6% to residual noise, 11.7% to citation scaling, and -1.7% to an affiliation suppressor interaction. Removing language and geography penalties reduced the mean median rank gap from 40.8 to 17.9 places. The study proposes rank envelopes, membership probabilities, layer-specific decomposition, and a minimum reporting standard. It demonstrates mechanisms rather than measuring current vendor performance; all code, parameters, synthetic data, and replication outputs accompany the manuscript.

Keywords: OpenAlex; Scopus; Web of Science; university rankings; database coverage; Monte Carlo simulation; research evaluation; reproducibility

1 Introduction

A university does not enter a bibliographic database as an intact analytical object. It enters as a changing collection of works, names, addresses, identifiers, references, versions, document types, and subject labels. Each element can be absent, duplicated, merged, split, or classified differently. When those records are aggregated into indicators and the indicators are sorted into a league table, the database becomes part of the ranking method even if it is mentioned only in a methodological footnote.

This dependence is not a theoretical curiosity. Database comparisons have repeatedly found differences in journal coverage, linguistic representation, affiliation assignment, reference linking, and citation counts. Early work showed that Scopus and Web of Science can produce different citation counts and rankings for the same scholars (Meho & Yang, 2007), while longitudinal and large-scale comparisons extended the observation across disciplines and data sources (Harzing & Alakangas, 2016; Martin-Martin et al., 2021; Visser et al., 2021). Mongeon and Paul-Hus (2016) demonstrated that the journal universes of Scopus and Web of Science share biases but differ materially by field, country, and language. Those differences do not disappear when the unit of analysis becomes a university; aggregation can amplify them.

OpenAlex has made the problem more visible because it offers an openly accessible alternative to commercial infrastructures (Priem et al., 2022). Comparative studies indicate broad coverage and substantial reference completeness, but also uneven institution linking, document classification, and metadata quality (Culbert et al., 2025; Haupka et al., 2024; Zhang et al., 2024). Its linguistic coverage is more balanced than that of Web of Science, although language metadata remain imperfect (Cespedes et al., 2025). Studies of African publications

and open-access journals likewise show that broader representation can coexist with incomplete metadata (Alonso-Alvarez & van Eck, 2025; Maddi et al., 2025). Openness therefore changes the conditions of verification; it does not abolish measurement error.

Recent research has moved beyond record counts. Scheidsteger et al. (2025) found strong paper-level relationships but weak agreement in university-level mean normalized citation scores across OpenAlex and commercial databases. Abramo et al. (2026) directly compared researcher and institutional evaluations across OpenAlex, Scopus, and Web of Science and reported substantial rank volatility for impact. Rodrigues et al. (2025) documented incoherent citation data among the same three platforms. The accumulated evidence rejects the idea that a database is a passive container. Yet ranking practice still tends to report a single position without expressing how far that position depends on the source beneath it.

This study addresses that gap through a simulation whose assumptions, code, and outputs are fully disclosed. Its purposes are to: (1) formalize database-choice uncertainty as a component of institutional rank uncertainty; (2) specify a harmonized, reproducible protocol for comparing OpenAlex, Scopus, and Web of Science; and (3) quantify how coverage, affiliation resolution, citation scaling, and residual metadata noise can generate system-level agreement alongside institution-level discordance. The study does not estimate the current quality of any vendor database. Instead, it asks a methodological question: under plausible and explicitly declared observation regimes, what kinds of rank instability should analysts expect, how should they diagnose it, and how should they communicate it?

2 Conceptual framework

2.1 One institution as three statistical objects

Bibliographic databases differ before an indicator is calculated. Coverage policies determine which journals, conference proceedings, books, preprints, regional sources, and years enter the analytical universe. Entity-resolution systems decide whether a department, hospital, branch campus, translated name, or affiliated laboratory belongs to a parent university. Citation graphs differ because references may be absent, malformed, duplicated, or connected to different versions. Field systems create distinct expected-citation baselines. Document taxonomies decide whether an item is an article, review, editorial, proceedings paper, letter, or other output.

These layers interact. A coverage difference can alter a university's disciplinary portfolio, which changes the denominator used for field normalization. An affiliation error can remove a highly cited laboratory and thereby affect output, impact, and collaboration simultaneously. Broader coverage may improve representation while introducing records with incomplete metadata; selective coverage may be internally consistent yet omit legitimate regional or multilingual scholarship. Accordingly, coverage is not synonymous with quality, and curation is not synonymous with neutrality.

Table 1 Database layers that can alter an institutional position

Layer	Typical discrepancy	Ranking consequence	Required audit
Coverage	Sources or output families appear in one database only	Changes visible output and disciplinary portfolio	Shared-source and exclusive-output profiles
Document type	Articles, reviews, proceedings, and editorial content are labeled differently	Changes eligibility and reference sets	Publisher-page validation sample
Affiliation	Names, hospitals, campuses, or historical units are split or merged	Moves works between institutions	ROR-linked crosswalk and high-impact review
Citation graph	References or versions are linked differently	Changes raw and normalized impact	Shared-DOI citation comparison
Field taxonomy	Subject granularity and assignment rules differ	Changes expected citation rates	Common external map and native-map sensitivity

Layer	Typical discrepancy	Ranking consequence	Required audit
Versioning	Preprint, early-access, and final records are consolidated differently	Duplicates or reallocates outputs	Version-of-record rules and date freeze

2.2 Database-choice uncertainty

Database-choice uncertainty is defined here as the variation in institutional results attributable to defensible choices among bibliographic sources after the analyst has fixed the population, time window, document eligibility, affiliation perimeter, field normalization, counting method, and indicator formula. The concept treats no source as an automatic ground truth. Agreement raises confidence in a signal; disagreement identifies a result whose provenance requires investigation.

Rankings add two nonlinear operations to ordinary database comparison: aggregation and sorting. A small number of missing highly cited papers may matter more than hundreds of missing low-impact records. The positional consequence also depends on score density. A modest score change in a crowded middle band can produce a large rank movement, whereas a larger score change at the sparsely populated top may not alter position. Spearman correlation therefore answers only whether two full orderings move together. It does not show whether the scores are interchangeable, whether a particular university is stable, or whether membership of a policy-relevant group changes.

The framework distinguishes a database-native estimand from a common-corpus estimand. The native estimand asks what performance looks like inside each database's own scholarly universe. The common-corpus estimand restricts the analysis to matched works and a common field map, thereby isolating differences in metadata, citation linking, and classification. An empirical audit should report both. The present simulation concentrates on the native estimand because its purpose is to expose the combined positional consequence of source-specific observation.

3 Methods

3.1 Simulation aims and estimands

The design follows the aims-data-generating mechanisms-estimands-methods-performance measures structure recommended for simulation studies (Morris et al., 2019). The primary estimand is the distribution of disagreement between institutional rankings produced by three source-specific observation systems applied to the same latent population. Secondary estimands are institution-specific rank intervals, top-100 membership probabilities, subgroup-specific directional shifts, sensitivity of disagreement to prespecified assumptions, and the contribution of four observation layers to the OpenAlex-Web of Science rank gap.

The simulation uses one fixed population of 700 institutions. Repeated draws represent alternative realizations of database observation error for that same population, not new samples of universities. This separation is important: population heterogeneity is held constant while database-conditioned measurement changes. Every stochastic component is controlled by master seed 20260918, and every scenario uses common replicate numbers so that differences between scenarios are not driven by unrelated random streams.

Table 2 ADEMP specification of the simulation

Element	Prespecified choice
Aims	Measure source-dependent rank disagreement and diagnose its mechanisms
Data-generating mechanism	Fixed heterogeneous population of 700 synthetic universities
Estimands	Rank agreement, absolute gaps, top-100 overlap, rank intervals, directional shifts, Shapley contributions

Element	Prespecified choice
Methods	Identical four-component ranking model applied after three observation functions
Performance measures	Spearman rho, Kendall tau, median and 90th-percentile gaps, top-100 overlap, 95% Monte Carlo intervals
Replications	1,000 per reference and sensitivity scenario; 1,000 for each of 16 Shapley coalitions

3.2 Synthetic institutional population

Each institution i has a cumulative five-year publication scale, latent research quality, regional indicator, non-English output share, and a five-part disciplinary portfolio. Global South status is generated as Bernoulli(0.55). Conditional on that indicator, non-English output share follows Beta(5,3) for Global South institutions and Beta(2,6) for other institutions. Field shares follow a Dirichlet distribution over natural sciences, engineering, medicine, social sciences, and arts and humanities. Global South profiles receive modest additional concentration in the latter two fields. These are design devices, not estimates of real regional publishing systems.

The publication scale and latent quality are generated as follows, where z denotes standardization within the synthetic population."

$$\log S_i \sim \mathcal{N}(7.50, 0.72^2), \quad q_i \sim \mathcal{N}(0, 0.62^2) + 0.15z(\log S_i) \quad (1)$$

True output is the rounded exponential of log scale plus a quality coefficient of 0.18 and a small field-composition term. It is truncated to 250-20,000 eligible works to prevent implausible extremes. True mean normalized citation score is log-linear in quality and field composition. True top-decile share and international collaboration share are generated through logistic functions. The average top-decile probability is anchored near 0.10 before quality and field heterogeneity. The complete coefficients appear in the executable script and parameter file.

Table 3 Fixed population-generating settings

Quantity	Distribution or rule	Purpose
Institutions	700	Large enough to study top-100 boundaries and dense middle ranks
Global South	Bernoulli probability 0.55	Regional heterogeneity
Non-English share	Beta(5,3) if Global South; Beta(2,6) otherwise	Continuous language profile
Field mix	Dirichlet(3.0, 2.5, 2.5, 2.0, 1.0) plus regional adjustment	Portfolio heterogeneity
Log scale	Normal(7.50, 0.72 squared)	Right-skewed institutional size
Quality	Normal(0, 0.62 squared) plus 0.15 standardized log scale	Correlated size and performance
Output range	250-20,000 eligible works	Controls unrealistic extremes

3.3 Database observation functions

Database d observes institution i through coverage probability c and affiliation completeness a . L is the non-English output share, G is the Global South indicator, and f is the field-share vector. The coverage function is:"

$$c_{id} = \text{logit}^{-1}(\alpha_d + \lambda_d L_i + g_d G_i + \mathbf{f}_i^T \mathbf{y}_d) \quad (2)$$

Affiliation completeness is generated separately:

$$a_{id} = \text{logit}^{-1}(\eta_d + \lambda_d^{(a)} L_i + g_d^{(a)} G_i) \quad (3)$$

The coefficients impose relative differences, not empirical estimates of current database performance. OpenAlex receives the most inclusive coverage surface but lower affiliation completeness for complex multilingual profiles. Scopus occupies an intermediate position. Web of Science is more selective and carries larger simulated penalties for non-English, Global South, social-science, and arts-and-humanities portfolios. These directions reflect recurring patterns in comparative research (Vera-Baceta et al., 2019; Alonso-Alvarez & van Eck, 2025; Maddi et al., 2025), while the exact magnitudes remain deliberately testable assumptions.

Observed publication output is the latent output multiplied by coverage, affiliation completeness, and lognormal residual error:"

$$Y_{idr}^{(O)} = N_i c_{id} a_{id} \exp(\varepsilon_{idr}^{(O)}) \quad (4)$$

Observed normalized citation impact combines the latent impact M , a source-specific citation-scaling term κ , a coverage-composition term, and residual error:

$$Y_{idr}^{(C)} = M_i \exp(\kappa_{id} + 0.30[\log c_{id} - \overline{\log c_d}] + \varepsilon_{idr}^{(C)}) \quad (5)$$

The top-decile and international-collaboration components use corresponding logistic transformations. Coverage affects them modestly because a source-specific portfolio changes both the set of eligible works and the composition of collaboration records. Residual standard deviations decrease with institutional size through a bounded square-root factor, representing the dilution of isolated record errors in larger portfolios. Random errors are independent across the four components and across databases within a replication; the same deterministic institutional profile persists throughout.

Table 4 Prespecified parameters of the three observation systems

Parameter	OpenAlex	Scopus	Web of Science
Coverage intercept	3.000	2.400	2.000
Coverage language	-0.150	-0.650	-1.200
Coverage Global South	-0.100	-0.300	-0.550
Coverage field vector	0.050, 0.030, 0.050, 0.100, 0.150	0.040, 0.020, 0.050, -0.150, -0.400	0.080, 0.000, 0.080, -0.450, -0.900
Affiliation intercept	2.350	3.150	3.000
Affiliation language	-0.600	-0.150	-0.250
Affiliation Global South	-0.350	-0.100	-0.150

Parameter	OpenAlex	Scopus	Web of Science
Citation intercept	0.015	0.000	-0.010
Citation language	-0.030	-0.040	-0.060
Citation Global South	-0.020	-0.020	-0.040
Citation field vector	0.010, 0.000, 0.015, -0.005, -0.020	0.005, 0.010, 0.010, -0.015, -0.035	0.015, 0.000, 0.020, -0.025, -0.050
Collaboration bias	0.020	0.000	-0.020
Residual noise SD	0.050	0.040	0.035

3.4 Indicators composite score and ranks

Each observed database produces four components: log publication output, log mean normalized citation score, top-decile publication share, and international collaboration share. Components are standardized within database and replication. The composite score uses illustrative weights of 0.32, 0.33, 0.22, and 0.13. The weights sum to one and remain identical across databases."

$$S_{idr} = 0.32z(\log[1 + Y^{(O)}]) + 0.33z(\log Y^{(C)}) + 0.22z(Y^{(T)}) + 0.13z(Y^{(I)}) \quad (6)$$

Scores are ranked in descending order. Because the weights are not asserted as universally optimal, equal weighting is examined as a prespecified sensitivity scenario. Primary reporting retains the component scores; the composite is used to study positional consequences.

3.5 Repetitions uncertainty intervals and robustness envelope

The reference model is repeated 1,000 times. For each database pair and replication, the analysis calculates Spearman rank correlation, Kendall tau, the median absolute rank gap, the 90th percentile of absolute rank gaps, and the proportion of the top 100 shared by both databases. The reported 95% Monte Carlo interval is the 2.5th to 97.5th percentile of the 1,000 replication-level statistics. It expresses variability under the specified observation process; it is not a confidence interval for a vendor population parameter.

For each institution and database, the analysis retains the median rank and the 2.5th and 97.5th percentiles of its 1,000 ranks. It also calculates the probability of top-100 membership. Database-choice uncertainty is summarized with a robustness envelope formed from the best and worst database-specific median ranks:"

$$E_i = \left[\min_d \{ \tilde{R}_{id} \}, \max_d \{ \tilde{R}_{id} \} \right], \quad \tilde{R}_{id} = \text{median}_r(R_{idr}) \quad (7)$$

3.6 Sensitivity analysis

Six scenarios examine whether the conclusions depend on a single parameterization. They halve cross-database coverage differences; remove language and geography penalties; replace affiliation functions with their pooled parameter values; replace citation functions with pooled values; halve residual noise; and use equal component weights. Each scenario receives 1,000 replications with the same population and master-seed structure as the reference model. The sensitivity analysis therefore changes one assumption family at a time while preserving all other design choices.

Table 5 Prespecified sensitivity scenarios

Scenario	Change from reference model	Diagnostic purpose
Half coverage differences	Move every coverage coefficient halfway toward the across-database mean	Strength of coverage contrast
No language or geography penalties	Set all language and regional coefficients to zero	Distributional source effects
Common affiliation model	Use pooled affiliation parameters	Entity-resolution contribution
Common citation model	Use pooled citation and collaboration parameters	Citation-scaling contribution
Half residual noise	Multiply all residual SDs by 0.50	Random metadata variation
Equal weights	Set all four component weights to 0.25	Composite-model dependence

3.7 Shapley decomposition

The OpenAlex-Web of Science mean absolute rank gap is decomposed into coverage, affiliation, citation scaling, and residual noise. For every one of the 16 possible coalitions of the four layers, included layers are harmonized: deterministic parameter families are replaced by across-database pooled values, while harmonized residual noise is set to zero. The value function $v(A)$ is the mean absolute rank gap after harmonizing coalition A . The marginal reduction attributable to layer j is averaged over every order in which the layers can be harmonized:"

$$\phi_j = \sum_{A \subseteq K \setminus \{j\}} \frac{|A|! (|K| - |A| - 1)!}{|K|!} [v(A) - v(A \cup \{j\})] \quad (8)$$

The decomposition is calculated separately within each of 1,000 replications. Reported shares are the mean and percentile interval of each layer's contribution divided by the total reduction. Negative values are allowed; they indicate a suppressor interaction in which harmonizing one layer alone increases discordance even though the fully harmonized system converges.

3.8 Reproducibility and validation

The complete Python script writes the fixed population, parameter file, all reference-scenario replication metrics, scenario summaries, institution-level rank intervals, subgroup results, Shapley results, and publication figures. Independent reruns require only NumPy, pandas, SciPy, and Matplotlib. The accompanying README provides the command and package versions. Internal validation checks require 700 unique institutions, ranks that form complete permutations within every database and replication, component weights that sum to one, top-100 sets of exactly 100 institutions, and Shapley shares that sum to one within numerical tolerance. No proprietary records are included or implied.

4 Results

4.1 Agreement at the system level

All database pairs preserved much of the broad institutional hierarchy, but none was positionally interchangeable. OpenAlex and Scopus had the strongest apparent proximity to OpenAlex in the correlation matrix: their mean Spearman correlation was 0.986. The OpenAlex-Web of Science correlation was 0.954, with a 95% Monte Carlo interval from 0.951 to 0.958. Even at that high correlation, the mean of the replication-specific median absolute rank gap was 40.8 places, and the mean 90th-percentile gap exceeded 101.3 places.

Top-group agreement was less reassuring than the correlations. OpenAlex and Web of Science shared an average of 81.7% of their top-100 institutions, with a 95% interval from 78.0% to 85.0%. Thus, roughly 18 institutions on average crossed the top-100 boundary solely because the observation system changed. Scopus and Web of Science were closer, but their mean median gap remained 32.9 places. High correlation and consequential mobility were not competing descriptions; they summarized different levels of the ranking.

Table 6 Reference-scenario agreement across 1,000 replications

Database pair	Spearman rho	Kendall tau	Median gap	90th-percentile gap	Top-100 overlap
OpenAlex vs Scopus	0.986 [0.985, 0.988]	0.900 [0.894, 0.906]	19.0 [17.0, 21.0]	55.8 [51.1, 60.1]	90.0% [86.0%, 93.0%]
OpenAlex vs Web of Science	0.954 [0.951, 0.958]	0.814 [0.807, 0.821]	40.8 [38.0, 44.0]	101.3 [96.0, 107.0]	81.7% [78.0%, 85.0%]
Scopus vs Web of Science	0.969 [0.967, 0.972]	0.847 [0.842, 0.853]	32.9 [30.5, 36.0]	83.6 [79.1, 88.0]	86.4% [83.0%, 90.0%]

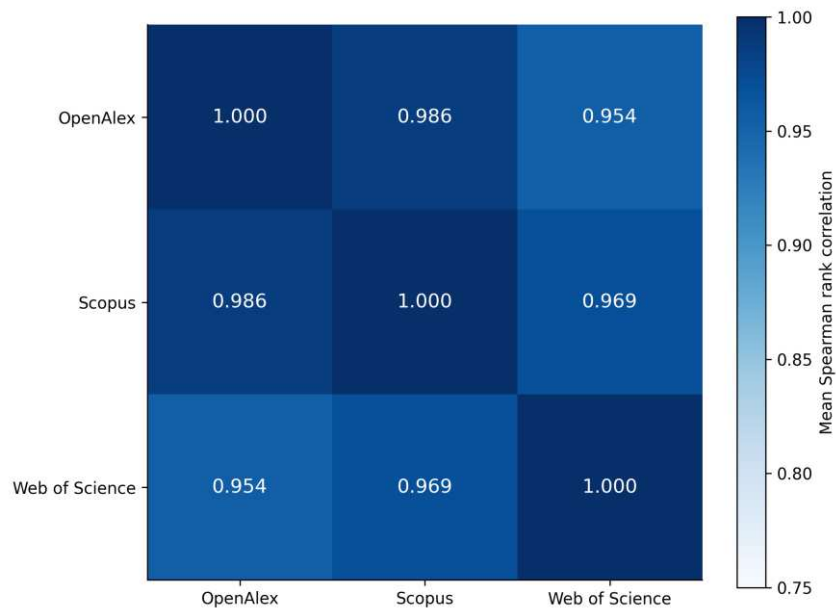


Fig. 1 Mean Spearman correlations across 1,000 reference-model replications. The diagonal is unity by definition.

4.2 Institution-specific rank envelopes

Institution-level envelopes revealed the variation concealed by pairwise averages. The median cross-database envelope width was 41.0 places, the 90th percentile was 84.0, and the widest synthetic institution spanned 141 places across database-specific median ranks. The trajectory plot focuses on the 18 widest envelopes. The crossings show that no source simply shifted every institution in the same direction. Movement depended on the joint configuration of language, geography, field composition, affiliation completeness, and score density.

Monte Carlo rank intervals added a second source of uncertainty. Some institutions had similar median positions across databases but wide within-database intervals because residual observation error interacted with crowded score regions. Others had narrow within-database intervals yet widely separated database medians, indicating systematic source dependence. The robustness envelope and the database-specific interval therefore answer different questions and should be reported together.

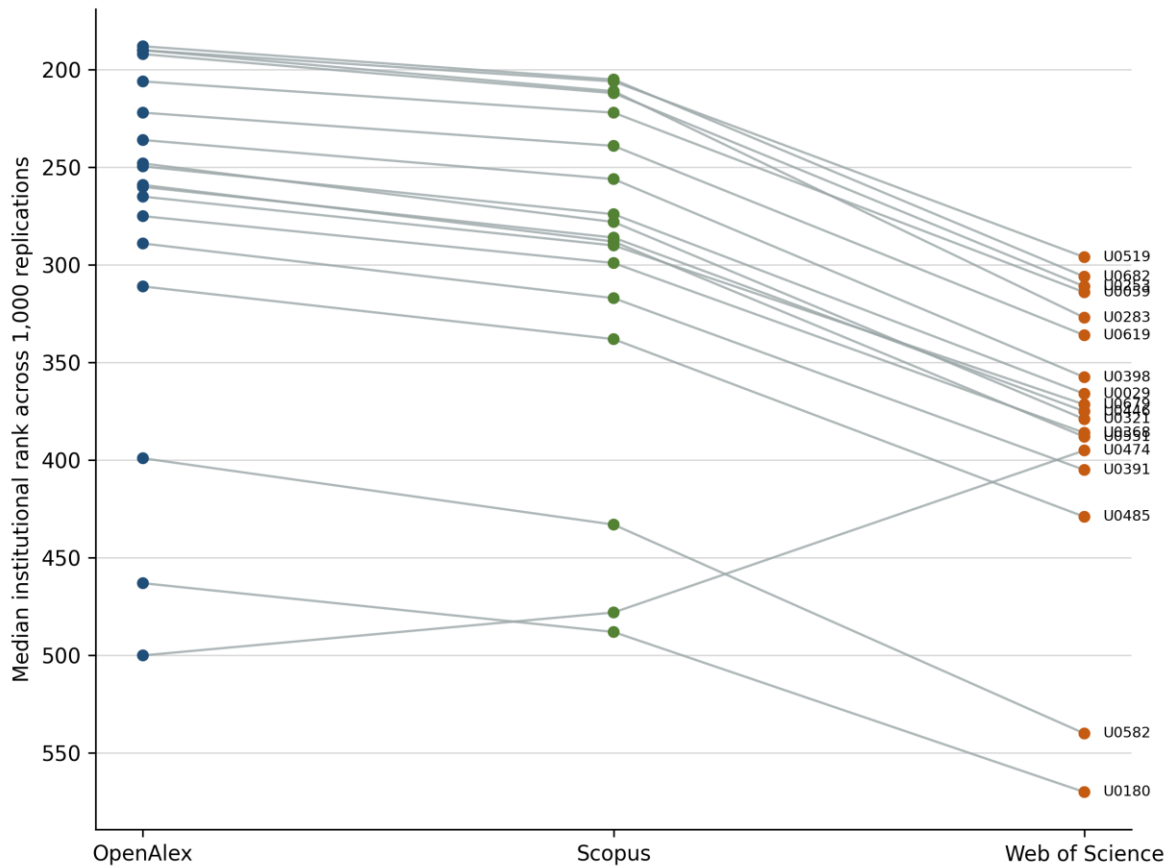


Fig. 2 Median rank trajectories for the 18 synthetic institutions with the widest cross-database envelopes. Lower numerical ranks appear higher; labels identify synthetic cases only.

4.3 Directional asymmetry by language and region

The subgroup pattern was directional rather than a simple increase in envelope width. A positive OpenAlex advantage means that an institution's Web of Science rank number exceeded its OpenAlex rank number. The median Global South institution had an OpenAlex advantage of 34 places [30, 37], whereas the median institution in other regions had an advantage of -52 places [-57, -48]. Likewise, institutions with at least half of their output represented as non-English had a median advantage of 39.5 places, compared with -39.0 among lower non-English-share institutions.

The magnitude of the robustness envelope was not uniformly larger for the groups that benefited from OpenAlex. This matters. Because rank is zero-sum, systematic visibility gains for one group create losses elsewhere; both groups can move substantially even when one experiences stronger simulated coverage penalties. The fair interpretation is therefore not that a database universally improves or harms a region. It is that the direction of database-conditioned movement follows the visibility assumptions built into the observation system.

Table 7 Subgroup rank envelopes and directional shifts

Group	n	Median envelope width	Median OpenAlex advantage vs Web of Science
Global South	397	34.0 [30.0, 37.0]	34.0 [30.0, 37.0]
Other regions	303	52.0 [48.0, 57.0]	-52.0 [-57.0, -48.0]
High non-English share	322	41.0 [37.0, 46.0]	39.5 [36.5, 46.0]

Group	n	Median envelope width	Median OpenAlex advantage vs Web of Science
Low non-English share	378	40.5 [35.0, 48.0]	-39.0 [-46.8, -34.0]
Large institutions	350	39.2 [35.0, 46.0]	5.5 [-1.8, 14.0]
Smaller institutions	350	42.5 [38.0, 47.5]	0.0 [-7.0, 5.0]

4.4 Layer-specific decomposition

Coverage was the dominant source of explained OpenAlex-Web of Science discordance, accounting for 65.4% [63.7%, 67.2%] of the reduction obtained by fully harmonizing the four layers. Residual noise contributed 24.6%, and citation scaling contributed 11.7%. Affiliation produced a small negative Shapley share of -1.7%. This negative contribution is not a computational error. In the specified system, affiliation differences partially offset larger coverage differences; harmonizing affiliation in isolation therefore widened the gap slightly. Across all coalitions, affiliation acted as a suppressor.

The decomposition changes the interpretation of a rank gap. It separates a systematic difference in what becomes visible from instability introduced by residual data error. The coverage share is conditional on the declared parameters and cannot be generalized as a current estimate of vendor quality. Its value lies in showing how an auditable decomposition can be performed and how interactions can be retained instead of forced into exclusively positive attributions.

Table 8 Shapley decomposition of OpenAlex-Web of Science discordance

Layer	Mean rank-gap reduction	Share of total reduction
Coverage	31.48 [30.53, 32.41]	65.4% [63.7%, 67.2%]
Affiliation	-0.82 [-0.89, -0.74]	-1.7% [-1.8%, -1.6%]
Citation scaling	5.63 [5.46, 5.80]	11.7% [11.3%, 12.1%]
Residual noise	11.85 [10.55, 13.21]	24.6% [22.6%, 26.6%]

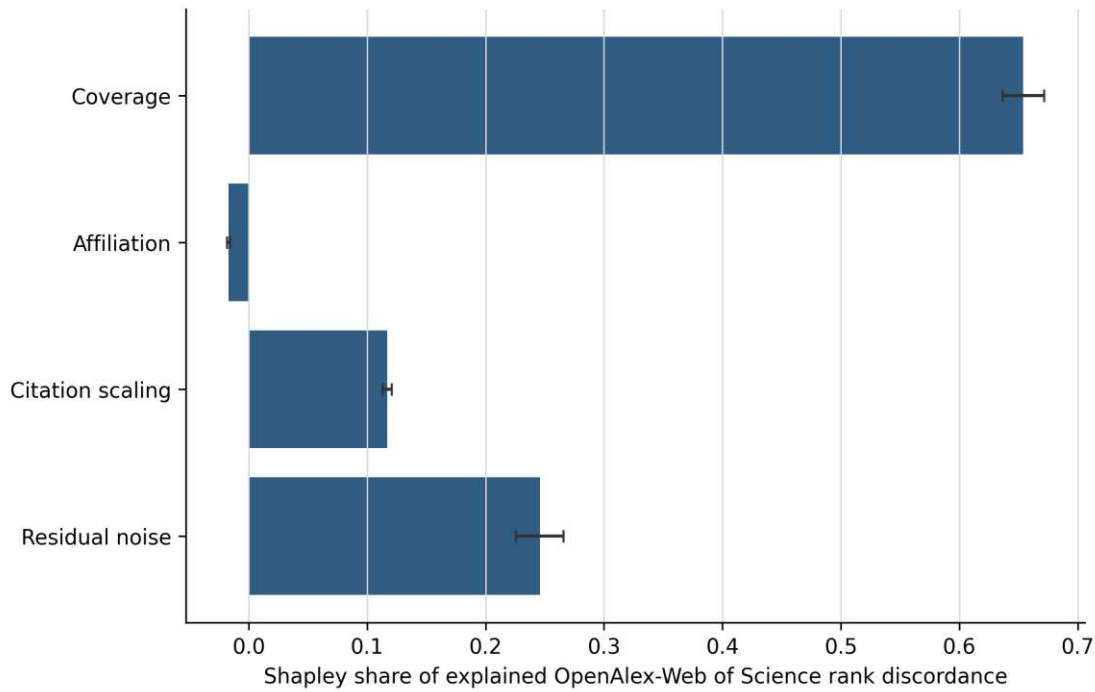


Fig. 3 Shapley shares of explained OpenAlex-Web of Science rank discordance. Error bars show 95% Monte Carlo intervals across 1,000 decompositions.

4.5 Sensitivity analysis

The principal conclusion survived all scenarios: correlation remained high while institution-level disagreement persisted. Its magnitude, however, was strongly conditional on the distributional assumptions. Removing language and geography penalties reduced the mean median gap from 40.8 to 17.9 places and increased top-100 overlap from 81.7% to 89.7%. Halving coverage differences reduced the gap to 24.7 and raised overlap to 88.2%. These were the largest changes.

Using a common citation model produced a smaller reduction, to 34.6 places. A common affiliation model increased the gap to 43.5, consistent with the negative Shapley contribution. Halving residual noise raised Spearman correlation to 0.962 but left the mean median gap near 41.1 because systematic coverage differences remained. This result illustrates why reducing random error does not necessarily remove source-conditioned bias. Equal weights changed the magnitude modestly but did not reverse the interpretation.

Table 9 OpenAlex-Web of Science sensitivity results

Scenario	Spearman rho	Median rank gap	Top-100 overlap
Reference	0.954 [0.951, 0.958]	40.8 [38.0, 44.0]	81.7% [78.0%, 85.0%]
Half coverage differences	0.979 [0.977, 0.982]	24.7 [22.0, 27.0]	88.2% [85.0%, 92.0%]
No language or geography penalties	0.988 [0.986, 0.989]	17.9 [16.0, 20.0]	89.7% [86.0%, 93.0%]
Common affiliation model	0.950 [0.946, 0.954]	43.5 [40.0, 46.5]	80.5% [77.0%, 84.0%]
Common citation model	0.965 [0.962, 0.968]	34.6 [32.0, 37.0]	84.6% [81.0%, 88.0%]
Half residual noise	0.962 [0.960, 0.964]	41.1 [39.0, 43.0]	83.1% [80.0%, 86.0%]
Equal component weights	0.961 [0.958, 0.964]	38.2 [35.0, 41.0]	82.1% [79.0%, 85.0%]

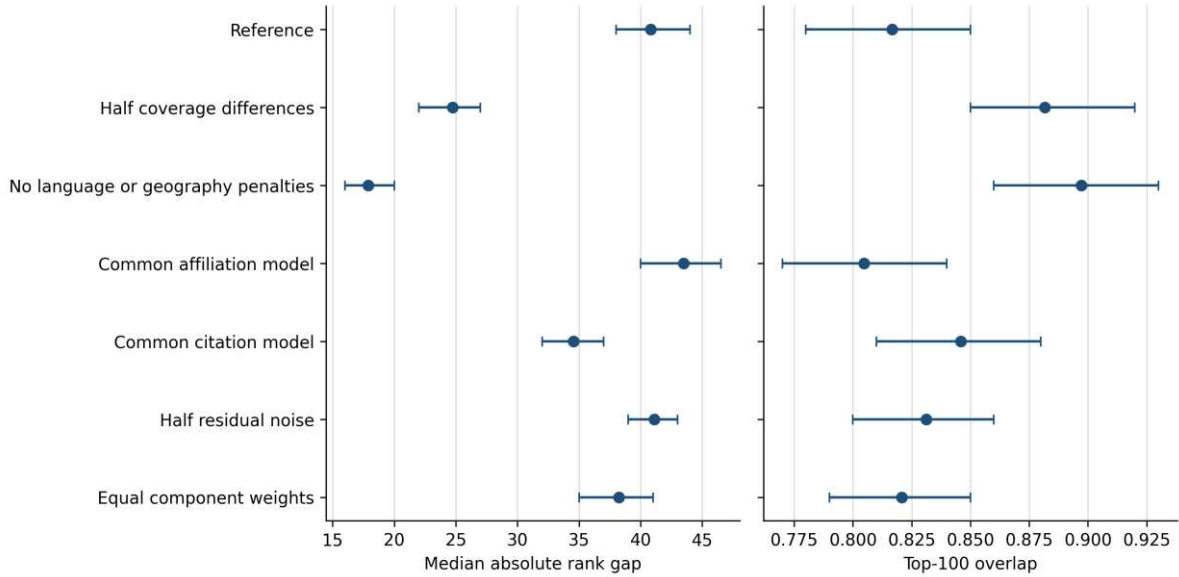


Fig. 4 Sensitivity of the OpenAlex-Web of Science median rank gap and top-100 overlap. Points are replication means; bars are 95% Monte Carlo intervals.

5 Discussion

5.1 What a high rank correlation does not establish

The central finding is structural: a ranking can preserve the broad hierarchy and still be unstable for individual universities and top-group membership. Correlation gives equal analytical weight to thousands of pairwise orderings that may have little relevance to a specific funding threshold, reputational boundary, or strategic benchmark. A university can move 80 places while the global coefficient barely changes. Consequently, the statement that two databases produce highly correlated rankings is not sufficient evidence that they are interchangeable.

This result aligns with empirical evidence showing stronger agreement at one analytical level than another. Scheidsteger et al. (2025) observed strong paper-level relationships but low university-level concordance for aggregated normalized impact. Abramo et al. (2026) likewise found that productivity was more stable than impact and that database choice could alter individual and institutional positions. The simulation adds a reproducible diagnostic framework: it links rank movement to declared observation layers, reports institutional intervals, and exposes boundary membership directly.

5.2 Coverage representation and fairness

The directional subgroup results should be read as a demonstration of distributive measurement, not as a factual ranking of regions. When one source is modeled as more inclusive of non-English and Global South portfolios, those institutions improve relative to a more selective source, and other institutions move downward because ranks are zero-sum. The mechanism is unsurprising; its scale and policy implications are not. A ranking that treats missingness as low performance can reproduce asymmetries while appearing procedurally uniform.

Fair evaluation does not require every database to index every output. It requires a ranking provider to state which scholarship is visible, quantify differential missingness, and avoid converting unobserved work into an unqualified performance judgment. Coverage audits should be stratified by language, region, field, document type, and publisher. Institutions should also have a correction process supported by record-level evidence. These practices follow the broader principles of transparent and context-sensitive research assessment articulated in the Leiden Manifesto (Hicks et al., 2015) and implemented in ranking methodologies that disclose collection and normalization decisions (Waltman et al., 2012).

5.3 Implications for ranking providers and universities

Ranking providers should treat database provenance as an output characteristic, not merely an input citation. Publishing a methodology document is insufficient when users cannot inspect the records, mappings, and exclusions that sustain a score. Providers using proprietary sources can still release institution-level diagnostics, shared and unmatched record counts, affiliation corrections, uncertainty intervals, extraction dates, and version histories. Open-data rankings should reciprocate by documenting cleaning decisions and resisting the assumption that openness guarantees validity.

Universities have a complementary responsibility. A durable affiliation audit should maintain persistent identifiers, historical names, translated variants, hospitals, institutes, and branch-campus relationships. The purpose is not to optimize a ranking through cosmetic metadata changes; it is to ensure that legitimate scholarship is consistently attributed. Institutions should also compare component indicators before reacting to an overall rank. If output is stable but normalized impact varies, the investigation differs from a case in which thousands of works are missing from one source.

For high-stakes use, the final position should be accompanied by a database robustness envelope, database-specific rank intervals, and top-group membership probabilities. A university with median positions of 47, 49, and 52 and narrow intervals is evaluated more confidently than one with median positions of 62, 114, and 181. The apparent precision of a single integer cannot substitute for evidence about the records and assumptions that produced it.

5.4 Minimum reporting standard

Table 10 Minimum reporting standard for database-dependent institutional rankings

Item	Minimum disclosure	Reason
Data snapshot	Database version and exact extraction date	Records and citations change continuously
Institution crosswalk	Persistent identifiers and parent-child rules	Prevents silent aggregation differences
Work matching	Shared, exclusive, DOI-less, and unresolved counts	Shows the evidentiary base of disagreement
Eligibility	Document types, years, and version rules	Defines the measured research portfolio
Indicator provenance	Formula, field map, window, counting method, and weights	Separates source effects from model effects
Rank uncertainty	Database-specific intervals and cross-database envelope	Avoids false positional precision
Boundary stability	Membership probability for policy-relevant top groups	Makes categorical consequences visible
Decomposition	Layer-specific contributions with interactions	Identifies plausible mechanisms
Reproducibility	Code, permissible intermediate tables, and correction log	Permits audit and revision

6 Limitations and empirical extension

This is a simulation study, not a measurement of current OpenAlex, Scopus, or Web of Science performance. The parameter directions are literature-informed, but their magnitudes were not estimated from synchronized vendor extracts. Numerical results are therefore conditional on the declared data-generating mechanism. They demonstrate possible consequences and an analytical workflow; they should not be quoted as observed database accuracy rates.

The model compresses publication-level processes into institution-level observation functions. It does not generate individual works, authors, journals, references, or time-varying citation histories. It also treats residual errors as independent across components and sources, whereas real databases inherit shared records and common upstream metadata. Field composition is represented by five broad domains. More granular taxonomies, multidisciplinary assignments, and database-specific normalization procedures could produce different interactions.

The ranking composite is illustrative. Alternative weights, transformations, and noncompensatory rules can alter score density and rank sensitivity. Equal weighting did not reverse the primary conclusions, but the analysis cannot exhaust the large space of ranking models. The top-100 boundary is similarly illustrative; empirical applications should use the thresholds that carry actual policy or reputational consequences.

A definitive empirical extension requires synchronized snapshots from all three databases, a fixed institutional population, a common citation census date, and sufficient access to audit records. Works should be matched primarily by normalized DOI, with conservative secondary linkage for DOI-less records. Affiliations should be aligned through persistent identifiers and explicit parent-child rules. A stratified manual sample should validate document type, affiliation, and version consolidation against publisher records. The empirical analysis should then reproduce the two estimands: database-native results to reveal representational differences and common-corpus results to isolate citation, metadata, and classification effects.

7 Conclusion

One university can occupy three positions without changing its scholarship. The variation arises because bibliographic databases observe research through different coverage rules, affiliation systems, citation graphs, classification schemes, and residual errors. A high cross-database correlation does not neutralize those differences. In the reference simulation, OpenAlex and Web of Science correlated at 0.954, yet the typical replication retained a median absolute gap of roughly 41 positions and an average top-100 overlap of only 81.7%.

The appropriate response is neither to abandon rankings nor to crown one database by reputation. It is to make database dependence measurable. Component scores, Monte Carlo rank intervals, cross-database envelopes, membership probabilities, sensitivity scenarios, and layer-specific decomposition replace a solitary rank with an auditable result. The present study supplies the code and synthetic evidence needed to reproduce that logic. Its next test belongs in synchronized empirical data.

Statements and Declarations

Data availability

No new datasets were generated or analysed in this theoretical article. All literature used to develop the argument is identified in the reference list.

Ethics declaration

This conceptual study did not involve human participants, personal data, animals, interventions or primary data collection. Ethical approval and informed consent were therefore not required.

Author contributions

The author conceived the theoretical argument, selected and synthesised the relevant literature, developed the institutional-intelligence framework and propositions, designed the conceptual tables and figures, and drafted and revised the manuscript.

Competing interests

The author declares no competing interests.

Funding

No specific funding was received for this work. Authors should amend this statement if funding supported the final manuscript.

Generative AI Disclosure

Generative AI tools were used solely to assist with language editing, clarity, and the organization of the manuscript. The author reviewed, revised, and verified all AI-assisted output and assumes full responsibility for the accuracy, originality, interpretation, and integrity of the entire content.

References

- Abramo, G., Cicero, T., & D'Angelo, C. A. (2026). Data source effects in research performance assessment of individuals and institutions: Comparing OpenAlex with established bibliographic databases. *Scientometrics*, 131(5), 3411-3441. <https://doi.org/10.1007/s11192-026-05638-6>
- Alonso-Alvarez, P., & van Eck, N. J. (2025). Coverage and metadata completeness and accuracy of African research publications in OpenAlex: A comparative analysis. *Quantitative Science Studies*, 6, 1336-1357. <https://doi.org/10.1162/qss.a.396>
- Cespedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimi, S., Ayeni, P., Filali, S., Li, B., & Lariviere, V. (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*, 76(6), 884-895. <https://doi.org/10.1002/asi.24979>
- Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*, 130(4), 2475-2492. <https://doi.org/10.1007/s11192-025-05293-3>
- Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2), 787-804. <https://doi.org/10.1007/s11192-015-1798-9>
- Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2024). Analysis of the publication and document types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *arXiv*. <https://doi.org/10.48550/arXiv.2406.15154>
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429-431. <https://doi.org/10.1038/520429a>
- Maddi, A., Maisonobe, M., & Boukacem-Zeghmouri, C. (2025). Geographical and disciplinary coverage of open access journals: OpenAlex, Scopus, and WoS. *PLOS ONE*, 20(4), e0320347. <https://doi.org/10.1371/journal.pone.0320347>
- Martin-Martin, A., Thelwall, M., Orduna-Malea, E., & Delgado Lopez-Cozar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), 871-906. <https://doi.org/10.1007/s11192-020-03690-4>
- Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13), 2105-2125. <https://doi.org/10.1002/asi.20677>
- Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, 106(1), 213-228. <https://doi.org/10.1007/s11192-015-1765-5>
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074-2102. <https://doi.org/10.1002/sim.8086>
- Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*. <https://doi.org/10.48550/arXiv.2205.01833>
- Rodrigues, D., Lopes, A., & Batista, F. (2025). Detecting incoherent citation data among three bibliometric platforms: OpenAlex, Scopus and Web of Science. *Journal of Information Science*. <https://doi.org/10.1177/01655515251330579>

- Scheidsteger, T., Haunschild, R., & Bornmann, L. (2025). How similar are field-normalized citation impact scores obtained from OpenAlex and three popular commercial databases? An empirical comparison based on large German universities. *Scientometrics*, 130(7), 3537-3569. <https://doi.org/10.1007/s11192-025-05338-7>
- Vera-Baceta, M. A., Thelwall, M., & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics*, 121(3), 1803-1813. <https://doi.org/10.1007/s11192-019-03264-z>
- Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1), 20-41. https://doi.org/10.1162/qss_a_00112
- Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F. J., Visser, M. S., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-2432. <https://doi.org/10.1002/asi.22708>
- Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*, 129(10), 5869-5891. <https://doi.org/10.1007/s11192-023-04923-y>