

Reverse-Engineering Radiation Oncology: A Reproducible Pipeline for Affiliation Disambiguation and Author Localisation in the Medical Sciences

David Kaul

david.kaul@hmu-potsdam.de

HMU Potsdam – Health and Medical University Potsdam

Research Article

Keywords: OpenAlex, affiliation disambiguation, author name disambiguation, bibliometric methods, reproducibility, radiation oncology

Posted Date: July 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10180340/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

aims

institutional tags are distorted by consortium over-attribution and by author-profile merge/split errors, so naïve location and seniority assignments are biased. Using German-speaking (DACH) radiation oncology as a worked example (17,270 articles, 2000–2025), we present an openly reproducible pipeline that derives locations and seniority from the raw affiliation strings rather than from the disambiguated institution tags. Its core components are: a robust, checkpointed OpenAlex client; a union corpus combining a topic with eight core journals through a dual-stream OR; a stem-based department filter separating radiation oncology from excluded specialties; a curated, corpus-grounded city gazetteer with word-boundary matching; a single-string consortium-clause guard that removes phantom locations from network-grant (“DKTK”) strings; and a person-gated fallback that recovers genuine senior locations while excluding non-radiation-oncology applicants. We define three transparent location measures (strict, gated, presence), report the data-quality failure modes encountered (P1–P7) with their corrections, and emit a full machine-readable provenance manifest. On a manually coded random sample, the raw-string measures localise last authors far more accurately than a naïve institution-tag baseline ($F1 \approx 0.92$ vs. 0.64). The pipeline is field-agnostic and transferable to any discipline-and-region slice of OpenAlex.

1. Introduction

Open bibliographic databases have changed what is possible in field-level scientometrics. OpenAlex in particular provides, for hundreds of millions of works, structured metadata on authorship, affiliation, venue and topic under a permissive licence, and, uniquely among the large free indices, exposes the raw affiliation strings exactly as they appeared on each publication (Priem et al., 2022). Systematic comparisons of large bibliographic sources have shown both their promise and the heterogeneity of their data quality (Visser et al., 2021). In principle the metadata allow an entire medical specialty to be “reverse-engineered” from public data: one can ask which centres lead a field, how its collaboration networks are structured, how researchers move between institutions, and how the gender composition of authorship evolves over a career.

In practice, however, the convenient *disambiguated* layer of such graphs, the institution objects attached to each authorship, cannot be used naïvely for any claim about *people* or *places*. Two distinct classes of error intervene. First, consortium over-attribution: when an affiliation line names a multi-site research network, the network’s headquarters institution is propagated to authors at every partner site, so a researcher who merely belongs to a consortium appears to be located at its centre. Second, author-profile errors: frequent names accrue foreign works and spurious career stations through profile *merges*, while single individuals fragment into several author identifiers through profile *splits*, a well-documented and only partially solved problem in bibliometric data (Tekles & Bornmann, 2020). A pipeline that reads location from the institution tags inherits all of these distortions.

We argue that reliable field reconstruction requires an explicit cleaning layer that works on the raw affiliation strings, with disambiguation rules that are stated, auditable and reproducible. This paper describes such a layer. We use German-speaking radiation oncology (Germany, Austria, German-speaking Switzerland; 17,270 articles, 2000–2025) as a worked example, but the components are field-agnostic. Our contribution is threefold: (i) a robust, resumable retrieval and union-corpus construction; (ii) a raw-string location-and-seniority layer built around a department filter, a corpus-grounded gazetteer, a consortium-clause guard and a person-gated fallback; and (iii) an explicit catalogue of the seven data-quality failure modes (P1–P7) we encountered, with their corrections. All code, the gazetteer, the term lists and a machine-readable provenance manifest are openly available.

2. Data source and robust retrieval

Records were retrieved from the OpenAlex /works endpoint using cursor pagination (`per_page = 200`) and the polite pool (a contact mailto supplied both as a query parameter and in the User-Agent header). Large pulls were made robust by a streaming-append client that writes each page to disk immediately, retries on HTTP 429/5xx responses with exponential backoff, and records a per-stream cursor checkpoint, so that an interrupted download resumes exactly where it stopped rather than restarting. This matters at field scale: the two streams below together comprise tens of thousands of works retrieved over many minutes.

3. Corpus definition: a union via two streams

A field corpus must capture both the *technical/physics* literature and the *clinical* literature of radiation oncology. A pure topic filter (OpenAlex topic T10358, “Advanced Radiotherapy Techniques”) captures the former but severely under-counts clinical radiation oncologists; a set of core journals captures the latter. The corpus is therefore the union of a topic and eight core journals.

This union cannot be expressed in a single query, because OpenAlex combines distinct filter categories with logical AND: a query naming both a topic and a source set returns their *intersection*, not their union. We therefore ran two separate streams, a topic stream and a journal stream (Strahlentherapie und Onkologie; Radiotherapy and Oncology; International Journal of Radiation Oncology Biology Physics; Radiation Oncology; Clinical and Translational Radiation Oncology; Practical Radiation Oncology; Advances in Radiation Oncology; Physics and Imaging in Radiation Oncology), each into its own checkpointed raw file, and merged them locally with de-duplication by work identifier. We restricted to journal articles published 2000–2025 (2026 excluded from trends as a partial year) with at least one author affiliated in Germany, Austria or Switzerland. Because OpenAlex offers no language or sub-national region filter, the non-German-speaking Swiss share (French- and Italian-speaking centres such as Geneva, Lausanne and Ticino) was removed in a cleaning layer operating on the raw affiliation strings, retaining a Swiss paper only when a German-speaking Swiss city appears.

The construction flow was: 12,760 topic records and 8,682 journal records; minus 3,112 overlapping records = 18,330 unique; minus 1,060 purely non-German-speaking Swiss = 17,270 articles (Germany 13,752; Austria 2,206; German-speaking Switzerland 3,065; counts overlap for multi-country papers). A machine-readable provenance manifest (exact queries, parameters and per-stage counts, with a generation timestamp) is emitted automatically so the corpus can be reconstructed independently.

4. Location and seniority from raw strings

4.1 Why not the institution tags

The disambiguated institution tags are unsafe for localisation for three reasons that we encountered directly. Consortium strings (e.g. “German Cancer Consortium / DKTK, partner site X”) are expanded across the network, attaching the consortium headquarters to partner-site authors (failure mode P1). Frequent names accrue foreign works and false career stations through profile merges (P2). Single persons fragment into several author identifiers through profile splits (P3). Our principle is therefore: derive the city from the raw affiliation strings, never from authorships.institutions.

4.2 Department filter

A stem-based inclusion/exclusion list decides whether a single raw affiliation line belongs to radiation oncology. The inclusion list (DEPT) covers bilingual variants, radiation oncolog, the English compound radiooncolog (omitting which silently drops major chairs), strahlenther, strahlenheilkunde, radioonkolog, medical/radiation physics, proton/particle therap, and named centres such as OncoRay, MedAustron and Paul Scherrer. The exclusion list (EXCL), which overrides inclusion, blocks urolog, neuroradiolog, nuklearmedizin and diagnostic radiolog / department of radiology. Only lines passing this filter can contribute a city under the strict measure.

4.3 City gazetteer with word-boundary matching

A curated gazetteer maps each canonical city to its spelling variants (German/English; e.g. munich/München, berne/Bern, and “paul scherrer” → Villigen). Matching is anchored to word boundaries, so that “Heidelberglaan” (a street in Utrecht) or “Freiburgstrasse” (a street in Bern) do not match Heidelberg or Freiburg (P5). Where Mannheim and Heidelberg co-occur in one string, a Mannheim-over-Heidelberg priority resolves the University Medical Centre Mannheim ambiguity (the Mannheim faculty is formally part of Heidelberg University).

4.4 Consortium-clause guard

Within a single radiation-oncology line that contains a consortium marker (dktk, german cancer consortium, partner site) and several cities, only the first-named city, the author’s own department, is retained; the consortium headquarters, typically appended afterwards, is dropped. A bare “DKFZ, Heidelberg” (the genuine institution) is preserved, because the German Cancer Research Center really is

in Heidelberg. This single rule removed 95 phantom-Heidelberg attributions (the gated Heidelberg count fell from 1,408 to 1,313), correcting the dominant location artefact in the field.

4.5 Three location measures

We report three transparent measures rather than a single opaque assignment (each city counted at most once per paper):

- Strict, last author, radiation-oncology lines only: pure departmental leadership (coverage 44.4%).
- Gated, strict, plus a fallback that, *only for known radiation-oncology seniors*, accepts a DACH city from any non-excluded line (coverage 49.0%).
- Presence, any author at the centre: participation/visibility (coverage $\approx 62\%$).

The gated fallback is the key robustness device. If the last author's radiation-oncology lines yield no city, a DACH city is taken from any non-excluded line, but only if the last author belongs to the senior set, defined as individuals with at least two strict last-author radiation-oncology papers, identified robustly by author-id combined with a (surname, first-initial) name-key to resist profile splits (724 individuals). This recovers genuine chairs who list only their hospital without a department word, while keeping non-field seniors (radiologists, surgeons, agency staff) out of the leadership measures. When a last author legitimately names two real cities (e.g. Gießen–Marburg, or MedAustron/Vienna), both are counted in full; city sums therefore exceed paper counts by design.

4.6 Handling out-of-field authors

Two layers protect against contamination. A hard exclusion ensures that urology, neuroradiology, nuclear medicine and diagnostic radiology never yield a city, even in the fallback. A soft exclusion means that a neutral out-of-field line without a radiation-oncology term simply contributes nothing. An out-of-field last author thus drops out of the leadership measures rather than corrupting a city; the conservative price is that a paper led by an out-of-field author with no radiation-oncology-tagged co-author falls out of the city counts entirely.

5. Validation

Quantitative accuracy against a manual gold standard. To quantify localisation accuracy we drew a reproducible random sample (seed 42) of 100 papers and, for each, manually coded the *true* German-speaking radiation-oncology city or cities of the last author by reading the raw affiliation strings (coded as “none” where the last author was foreign, in non-German-speaking Switzerland, or out of field). We then compared three methods against this gold standard: the strict and gated raw-string measures, and a naïve baseline that applies the same city gazetteer to the OpenAlex *institution tags* (without the department filter, the consortium-clause guard, or the Mannheim priority). Treating each (paper, city) pair as the unit, we computed micro-averaged precision, recall and F1 (Table 1).

Table 1
Localisation accuracy on a manually coded random sample of 100 papers (44 with no true DACH radiation-oncology city). Micro-averaged over (paper, city) assignments.

Method	Precision	Recall	F1	TP / FP / FN
Strict (raw strings)	0.980	0.862	0.917	50 / 1 / 8
Gated (raw strings + senior fallback)	0.914	0.914	0.914	53 / 5 / 5
Naïve (institution tags)	0.609	0.672	0.639	39 / 25 / 19

The two raw-string measures outperform the naïve institution-tag baseline by a wide margin ($F1 \approx 0.92$ vs. 0.64). The naïve baseline’s low precision (0.61; 25 false positives) is driven exactly by the failure modes this pipeline targets: phantom cities from consortium (DKTK) over-attribution and from the institution tags of out-of-field departments (e.g. a diagnostic-radiology last author wrongly localised to a radiation-oncology centre). The strict measure maximises precision (0.98), while the senior-gated fallback trades a little precision for higher recall, yielding the most balanced assignment (0.914 / 0.914).

Several further checks support the result. Coverage audits confirm that the residual unassigned last authors are overwhelmingly foreign or non-German-speaking Swiss (correctly unassigned) rather than missed German centres. A live OpenAlex count reproduced the documented corpus size to within rounding. The top-ranked seniors recovered by the gated measure are, on manual inspection, all genuine radiation-oncology or medical-physics figures, and detected institutional moves reproduce known chair appointments. Together these indicate that the raw-string layer corrects the institution-tag artefacts without introducing systematic new bias.

6. The data-quality catalogue (P1–P7)

The seven failure modes we encountered, and their corrections, are summarised below. We propose this catalogue as a checklist for any field-level reconstruction from OpenAlex.

#	Failure mode	Correction in the pipeline
P1	Consortium (DKTK) over-attribution of the network HQ to partner sites	City from raw strings; single-string consortium-clause guard; Mannheim-over-Heidelberg priority
P2	Homonym/profile merge (frequent names accrue foreign works)	No institution tags for location; paper-specific raw strings only
P3	Profile split (one person → several author-ids)	Senior set and de-duplication by (surname, first-initial) name-key alongside author-id
P4	Topic-only corpus under-counts the clinical slice	Union corpus (topic OR eight core journals) via two streams
P5	Substring / foreign-street false matches (Heidelberglaan, Freiburgstrasse)	Word-boundary regex; foreign and non-German-speaking-Swiss cleaning layer
P6	Filter-syntax traps (AND across categories; no author_position filter)	OR via the pipe operator across separate streams; last author determined client-side
P7	Left-censoring of careers at the window start	Window from 2000; a denser-data threshold for career modules

7. Discussion

The central methodological lesson is that the disambiguated metadata of a large open graph require an explicit, raw-string-checked cleaning layer before any person- or place-level claim is made. The convenience of pre-attached institution objects is a trap for field reconstruction: they encode exactly the consortium and profile artefacts that bias leadership, mobility and network measures. By contrast, reading location from the raw affiliation strings under stated rules, a department filter, a corpus-grounded gazetteer with word-boundary matching, a consortium-clause guard, and a person-gated fallback, yields measures that are transparent, auditable and reproducible.

None of these components is specific to radiation oncology or to the German-speaking region; each is a parameter (a department stem list, a city gazetteer, a consortium marker set, a seniority threshold) that can be re-instantiated for another discipline-and-region slice. We therefore argue for two routine practices in field-level bibliometrics: shipping a machine-readable provenance manifest with every corpus, and releasing the cleaning code rather than only the results.

8. Limitations

The keyword/topic-plus-journal corpus is deliberately conservative and will miss radiation-oncology work published outside the eight core journals and outside the topic. Name-key disambiguation has residual error. Location coverage is 44–62% depending on the measure; the remainder are predominantly foreign or non-German-speaking-Swiss last authors that are correctly left unassigned, plus a small residue of out-of-field or affiliation-less cases. The consortium-clause guard assumes that the first-named city in a consortium line is the author’s own department, which holds almost always but is not guaranteed.

Combination departments such as “Radiology and Radiotherapy” are currently excluded by the department of radiology term, a rare edge case that could be relaxed. Finally, OpenAlex grant coverage was empty for this field, so funding could not be derived from the graph itself.

9. Conclusion

A reproducible, openly available pipeline turns OpenAlex into a reliable instrument for reverse-engineering a medical specialty, provided that locations and seniority are read from the raw affiliation strings under an explicit disambiguation layer, and that the failure modes catalogued here (P1–P7) are corrected rather than ignored. The approach is field-agnostic and, with a re-instantiated gazetteer and department list, transferable to any region-and-discipline slice of the open scholarly graph.

Declarations

Declarations of generative AI use. During this work the author used an AI assistant (Anthropic Claude) to help debug the analysis code and for language editing and proofreading of the manuscript. The author verified all outputs and takes full responsibility for the content.

Funding.

No funding was received for conducting this study.

Competing interests. The author declares no competing interests. All analyses use only publicly available bibliographic metadata and are aggregate and non-individual.

Ethics approval. Not applicable. The study used only publicly available bibliographic metadata; no human participants or animals were involved.

Consent to participate / Consent for publication. Not applicable.

Data availability. The frozen corpus snapshot and the machine-readable provenance manifest are openly archived on Zenodo (Underlying records are from OpenAlex, openalex.org, CC0).

Code availability. The core pipeline (retrieval, corpus construction, the city gazetteer, the department and term lists, the localisation layer, and the validation scripts and gold standard) is openly available on GitHub (<https://github.com/davidpkaul-hue/radonc-pipeline>) and archived on Zenodo (version v1.1; <https://doi.org/10.5281/zenodo.20998589>) under the MIT License.

Author Contribution

D.K. is the sole author and is responsible for all aspects of the work, including conception, pipeline design and implementation, analysis, and writing.

References

1. Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. In: arXiv.
2. Tekles, A., & Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies*, 1(4), 1510–1528. https://doi.org/10.1162/qss_a_00081
3. 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