MapAffil Tempo 2018 dataset: A temporal model of authors in Author-ity based on MapAffil linked to NSF/NIH grants and ORCID

Dongin Nam[1], Tzu-Kun Hsiao[1], Deepanshu Malhotra[1], Liri Fang[1], Malik Oyewale Salami[1], Xiaoliang Jiang[1], Vetle I. Torvik[1]

[1] School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign IL, USA

Abstract

We present the MapAffil Tempo 2018 dataset, which integrates MapAffil 2018 dataset (linked with NSF/NIH and ORCID) with Author-ity 2018 dataset to build a timestamped, name-disambiguated author-affiliation dataset covering nearly 120 million records. Based on an exponential smoothing technique, MapAffil Tempo 2018 extends its predecessor (MapAffil 2018) by inferring each author's three primary affiliation dimensions: place (city/state/country), institution (identifier/type), and discipline for each year. The dataset supports studies of mobility across institutions, geographic regions, and disciplines, as well as macro-level trends such as brain drain and knowledge diffusion, making a powerful foundation for Science of Science and Policy studies of geography, collaboration, and scientific impact. MapAffil Tempo 2018 is freely available online: *https://databank.illinois.edu/datasets/IDB-4273770?code=XZGjDaTB_4P2_oA3ycXN8UMKtvg4OG3vRiP_wWFK40U*

1. Introduction

Many bibliographic systems such as Dimensions, OpenAlex, Lens, and Web of Science provide standardized institution identifiers (e.g., Global Research Identifier Database (GRID), Research Organization Registry (ROR)) for authors' affiliations. However, the temporal information is often not provided as a ready-made, per-year author affiliation timeline in these repositories. For instance, OpenAlex provides the institution history in their author object, but only for the past decade [1]. Prior studies underscore this gap: Thelwall & Fairclough [2] operationalized "U.S. researchers" via first and last publication affiliations, without considering the possible changes in-between. In contrast, a recent study on researcher mobility [3] does identify year-by-year affiliation histories, but fills missing years by only considering the most recent publication's affiliation, which can lead to misdated information. These approaches can obscure temporal dynamics and introduce systematic error, biasing downstream analyses. To address this issue, we present MapAffil Tempo 2018, a dataset that provides robust timestamped geocoded affiliation trajectories, including year, institution, city and discipline, for name-disambiguated authors in PubMed.

MapAffil Tempo 2018 links MapAffil 2018 (a dataset that maps affiliations to city, country, and disciplines) with Author-ity 2018 (a dataset that disambiguates author instances from PubMed). Records are organized by year for each author, resulting in a simple per-author timeline of affiliations. In addition to PMIDs, external identifiers are added from linked sources such as the National Science Foundation (NSF) grants, National Institutes of Health (NIH) grants, and Open Researcher & Contributor IDentifiers (ORCIDs). Since the National Library of Medicine (NLM) did not systematically record first author affiliations until 1988 and non-first author affiliations until 2014 for PubMed articles, MapAffil 2018 contains a substantial amount of missing affiliation fields for earlier years [4, 5]. Using a robust exponential smoothing technique, MapAffil Tempo 2018 covers nearly 120 million disambiguated author name instances across PubMed, NSF, NIH, and ORCID. Figure 1 shows an illustrative example of how MapAffil Tempo 2018 is designed.

ID-type (ID)	au_order	Year	MapAffil 2018			MapAffil Tempo 2018		
			City	Disciplines	GRID	City-temporal	Discipline-temporal	GRID-temporal
PMID (4546360)	2	1974	None	None	None	Cambridge, MA, USA	biology	grid.116068.8
...	...	...	...	...	...	...	...	...
PMID (215328)	1	1978	Cambridge, MA, USA	biology|cancer	grid.116068.8	Cambridge, MA, USA	biology	grid.116068.8
...	...	...	...	...	...	...	...	...
PMID (6494891)	1	1984	None	None	None	Cambridge, MA, USA	biology	grid.38142.3c
NIH (1R01GM034028-01)	1	1984	Cambridge, MA, USA	None	grid.38142.3c	Cambridge, MA, USA	biology	grid.38142.3c
NSF (NSF8414045)	0	1985	Cambridge, MA, USA	None	grid.38142.3c	Cambridge, MA, USA	biology	grid.38142.3c
...	...	...	...	...	...	...	...	...
PMID (2565854)	2	1989	Boston, MA, USA	biology|genetics|molecular biology	grid.38142.3c	Cambridge, MA, USA	cell biology	grid.38142.3c
PMID (2583479)	4	1989	Cambridge, Cambridgeshire, UK	molecular biology	None	Cambridge, MA, USA	cell biology	grid.38142.3c
...	...	...	...	...	...	...	...	...
PMID (8252621)	3	1993	Hanover, NH, USA	None	grid.254880.3	Hanover, NH, USA	biology	grid.254880.3
...	...	...	...	...	...	...	...	...
PMID (11444045)	1	2001	Hanover, NH, USA	biology	grid.254880.3	Hanover, NH, USA	genetics	grid.254880.3
PMID (11486075)	1	2001	Hanover, NH, USA	genetics	grid.254880.3	Hanover, NH, USA	genetics	grid.254880.3
...	...	...	...	...	...	...	...	...
PMID (18559475)	4	2008	Hanover, NH, USA	genetics	grid.254880.3	Worcester, MA, USA	molecular medicine	grid.168645.8
...	...	...	...	...	...	...	...	...
PMID (29967059)	1	2018	Worcester, MA, USA	molecular medicine	grid.168645.8	Worcester, MA, USA	molecular medicine	grid.168645.8

1 Sources of data (PMID: from author affiliation string, NSF/NIH grant: from PI affiliation string)
2 Different city appearance (vs. previous record)
3 Different discipline appearance (vs. previous record)
4 Different institution (GRID) appearance (vs. previous record)
5 Fills in missing data via exponential smoothing
6 Infers primary affiliation information via exponential smoothing
7 Correction of mislabeled affiliation via exponential smoothing

Figure 1. An illustrative snippet of MapAffil Tempo 2018 records for Victor Ambros (Nobel laureate in Physiology or Medicine, AUID = 4546360_2) showing mobility and missing data ("None") inferred. Note that continuous records have been truncated with an ellipsis (...). Also, the instances are ordered in a time-sequence. Red boxed cells show the sources of the data. Highlighted cells indicate the appearance of a different affiliation information compared to the previous record (green: city, pink: discipline, yellow: institution). The target author's

(Victor Ambros) first paper in 1974 was with MIT (grid.116068.8), Cambridge, MA in biology, followed by a paper in 1984 with Harvard University (grid.38142.3c) still in Cambridge, MA in biology. Then, in 1989, his discipline focused on cell biology. In 1993, the paper with Dartmouth College (grid.254880.3), Hanover, NH was in biology. Another paper with Dartmouth College in 2001 was in genetics. In the year 2008, a paper was published in molecular medicine with University of Massachusetts (grid.168645.8), Worcester, MA. Additionally, inferring primary affiliation is noteworthy (orange, green, blue boxed cells), especially for correction of mislabeled affiliation. For example, for the year 1989 record (PMID (2583479)), MapAffil Tempo 2018 properly inferred Cambridge, MA, USA while MapAffil 2018 wrongly identified Cambridge, Cambridgeshire, UK, which is an artifact derived from PMC's indexing error.

2. Methodology

In this section, the two core datasets (MapAffil 2018 & Author-ity 2018) that serve as the backbone of our proposed dataset are first described. Then, two supplementary datasets linking authors to NSF/NIH grants and ORCIDs were incorporated when PMIDs were unavailable. Finally, based on the information of these source datasets, the construction pipeline of the MapAffil Tempo 2018 is provided.

2.1 MapAffil: Mapping affiliation strings to cities, institutions, and disciplines

The MapAffil 2018 dataset [6] was curated based on the information of the PubMed author affiliations mapped to geographical locations. The MapAffil bibliographic tool [7] was used to build this dataset, which takes advantage of a rule-based algorithm that is based on a dictionary consisting of 24,000 city/state/country names and 1.1 million word n-grams indicating a unique country/US state for disambiguation. The algorithm extracts the cities, GRID identifiers, and disciplines from the raw affiliation string. After applying this algorithm to the December 2018 PubMed snapshot, which is an annual baseline file of all PubMed citation records provided by NLM, the MapAffil 2018 dataset was built. As a result, this dataset consists of 52,931,957 author-paper rows with cleaned affiliation strings and 18 feature columns such as *PMID, au_order* (i.e., order of the author in a given paper)*, lastname, firstname, affiliation, grid, city, state, country, lat* (i.e., latitude)*,* and *lon* (i.e., longitude). For more information, refer to MapAffil 2018 dataset [6].

2.2 Author-ity: Disambiguating author names in PubMed

The Author-ity 2018 dataset [8] is a PubMed author name disambiguated dataset. This dataset was derived from the Author-ity software [9, 10], which utilizes a pairwise probabilistic model across different attributes in MEDLINE. It disambiguated author identities collected on the December 2018 PubMed snapshot (same as the MapAffil 2018), consisting of a total of 29.1 million articles and 114.2 million author name instances. Each instance row is uniquely indexed by PMID and author position information (e.g., 10786286_3 is the third author name on PMID 10786286). Records were grouped based on the last name and first name initial (including some close variants) and then inferred to a cluster of predicted individuals. Each cluster has a unique Author ID (the PMID of the earliest paper in the cluster and the author name position) and includes 19 more features such as *name variants, email addresses, ORCIDs, top 20 most frequent Medical Subject Headings (MeSH), co-author names, grant IDs,* and *citation counts*. For more details, refer to Author-ity 2018 dataset [8].

2.3 Linked Datasets: NSF, NIH, and ORCID

While most of our data is based on PMIDs, there were still instances that could not collect relevant author affiliation information from PubMed. In this case, the missing affiliation data had to be harvested from external resources such as NSF/NIH grants and ORCIDs. To do this, the Author-linked data for

Author-ity 2009 dataset [11] was used, which links PubMed author IDs from Author-ity 2009 dataset with principal investigators (i.e., PI) on NSF/NIH grants and inventors on USPTO patents to connect researchers beyond PubMed articles. Records were each matched based on a probability model, and only the ones that had more than a 0.5 probability score were retained. For our new dataset, the PIs on NSF/NIH grants were only considered. In addition, ORCIDs Mapped to PubMed Authors dataset [12] was used, which is also a linked dataset that bridges PubMed author IDs with ORCID's employment and education records based on a rule-based matching system.

These additional linked datasets were essential in the process of linking external sources (in this case, NSF, NIH, and ORCID) beyond PubMed information. This enabled our dataset to be more robust by preventing valuable information from being lost.

2.4 MapAffil Tempo 2018

Each row in our proposed dataset uses the Author-ity 2018 dataset to expand the MapAffil 2018 dataset in a temporal fashion for each author. Also, it includes the linked data as records alongside PubMed records. However, when there was missing affiliation information in the PubMed data, relevant information from external sources were collected based on the aforementioned linked datasets.

The joining process of the two datasets was based on a shared key (i.e., PMID and author position). While Author-ity 2018 provided the disambiguated author name cluster, MapAffil 2018 provided the geocoded affiliation information for an author. Considering this, for a given disambiguated author, all cities, disciplines, and institutions were collected, respectively. To infer the value for a specific author *A* in a target year *T*, we designed an exponential smoothing technique that assigns exponentially decaying weights to observations from surrounding years *Y*. Thus, observations in the target year *T* receive the maximum weight (i.e., 1), and the weighted halves for each yearly span (0.5 for *T±1,* 0.25 for *T±2, etc.*). For example, assuming that an author instance has *Palo Alto* for its city attribute in the year *2010* (*T*), *Palo Alto* will receive 1 as the weight; if the same city is observed in 2011, the weight is 0.5; in 2012, 0.25; and so on. These weights are summed per candidate, and the highest score is selected. For missing data (e.g., non-first authors in PubMed prior to 2014), the first author's affiliation was considered as a candidate, however, with less weight. Weights also vary depending on the quality of the information and were adjusted through qualitative assessment. For instance, PubMed Central identifiers (PMCIDs) receive less weight as they tend to be less reliably assigned to the correct author. Excessively long affiliation strings were also downweighted because they typically are a result of multiple institutions conflated as one chunk.

After this whole process, the final dataset contains 14 attributes for each author instance. Records are grouped by Author-ity Author ID and ordered by publication year to form temporal timelines for each author. The original city information (i.e., *City*) from MapAffil 2018 dataset alongside inferred temporalized fields (e.g., *grid-temporal, discipline-temporal, City-temporal, Country-temporal*), and consistency indicators (*CitySame, CountrySame, Moved*) were retained.

3. Dataset Summary

MapAffil Tempo 2018 contains city, institution, and discipline for 119,966,689 name instances across 30,638,055 unique records from PubMed, ORCID, NSF grants, and NIH grants. The dataset has been deposited into the Illinois Data Bank, along with descriptions of the 14 attributes: https://databank.illinois.edu/datasets/IDB-4273770?code=XZGjDaTB_4P2_oA3ycXN8UMKtvg4OG3vRiP_wWFK40U.

The majority of name instances are from PubMed (114,445,384 instances, 95.4%), followed by ORCID (2,922,037 instances, 2.4%). Note that an author may occur multiple times in the dataset if the author appears on more than one PubMed, ORCID, or NSF/NIH record. Table 1 presents a breakdown of the number of name instances and the number of unique authors by source.

Table 1. Distribution of MapAffil Tempo 2018 records across sources

Source	Number of Name Instances (%)	Number of Unique Authors (%)
PubMed	114,445,384* (95.4)	13,702,526** (100)
ORCID	2,922,037 (2.4)	518,648 (3.8)
NIH grant	1,825,538 (1.5)	138,925 (1.0)
NSF grant	773,730 (0.6)	49,420 (0.4)
Total	119,966,689 (100)	Not Applicable

*The number of instances is slightly larger than Author-ity 2018 (114.2 million) because MapAffil Tempo 2018 has some duplicate authors.
**The total number of unique authors is slightly smaller than Author-ity 2018 (13,722,485) because some authors have no affiliation information whatsoever.

The records range from the years 1786 to 2030. The dataset is based on a PubMed snapshot collected in December 2018, and very few records go beyond 2018 due to errors (e.g., ORCID typos or anticipated dates, prepublications, grant extensions, etc.). For example, grants' amendment dates may go beyond 2018 (e.g., NSF1333564), and ORCID records may list future dates such as expected graduation years. *ORCID ID: 0000-0002-1816-1721* shows that the person started their medical scientist training program in 2018 and anticipates completing the program in 2026.

As addressed in the Methodology section, the MapAffil Tempo 2018 dataset extends MapAffil 2018. In MapAffil 2018, data is available only for first authors of articles published before 2014, except for data harvested beyond PubMed, including PubMed Central (PMC) and Microsoft Academic Graph (MAG). This limitation is addressed in MapAffil Tempo 2018 by inferring cities, institutions, and disciplines for all authors based on their linked-records. Of the 119,966,689 name instances in MapAffil Tempo 2018, 44% (52,811,275 instances) originated from MapAffil 2018, while 56% (67,155,414 instances) were newly inferred. As shown in Figure 2, MapAffil Tempo 2018 substantially supplements data for instances at years prior to 2014: Between 1867 and 2013, the number of newly inferred name instances increased from 311 to 3,178,477 (left panel in Figure 2). However, the proportion inferred varied systematically by time periods: 1946-1987: 77-92% inference; 1988-2013: 58-70%; after 2013, it goes gradually down to 7% in 2018 (right panel in Figure 2).

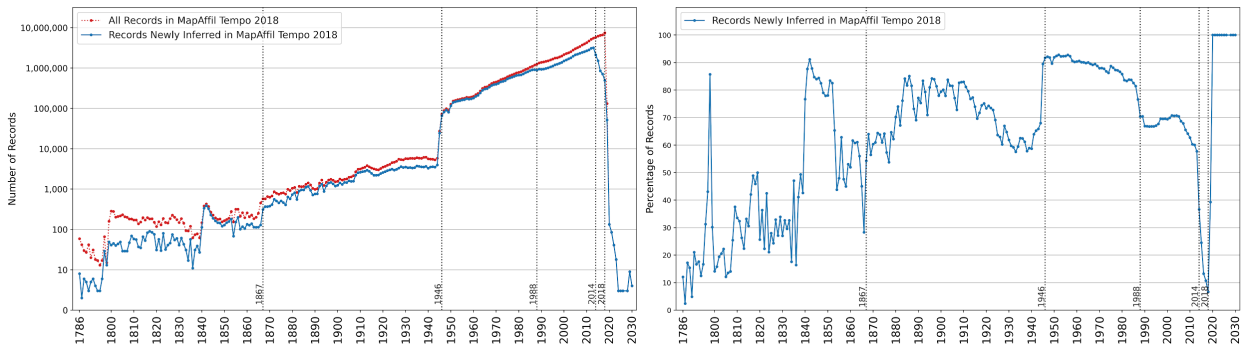


Figure 2. Number of Name Instances Newly Inferred in MapAffil Tempo 2018 (left) and Percentages of Name Instances Newly Inferred in MapAffil Tempo 2018 (right) over Time. PubMed

provides access to MEDLINE and OLDMEDLINE [13], dating back to 1966 and 1949, respectively [14, 15]. 1988 is the year that the US National Library of Medicine (NLM) started consistently recording first-author affiliations. 2014 is the year that NLM started recording affiliations for all authors when available. 2018 is the data collection year. Records with years beyond 2018 are due to preprints and future dates listed in grants and ORCID records.

4. Discussion

MapAffil Tempo 2018 dataset provides robust inferences of institution, city, and discipline for all authors in PubMed over time. The exponential smoothing technique incorporating linked MapAffil 2018 and linked data captures large portions of missing and noisy data. The advantage of having a temporal dimension has already been demonstrated in relevant studies. Jiang et al. [16] combined metadata-driven features measuring consistency between affiliations and candidate place names with embeddings. Further, the authors linked data to train a probabilistic model for geographic named entity recognition. This was achieved by incorporating temporal alignment and the model correctly accounted for author mobility, improving disambiguation accuracy and reducing false positives. Vaccario et al. [17], which analyzed temporal correlations in researcher mobility, also leveraged the temporal aspect by using both MapAffil and Author-ity software, which showed that mobility patterns depend not only on the current location but also on earlier paths. In addition, authors' temporal profiles can be used to trace intellectual lineage between coauthors and identify their mentors using modeled multidimensional metadata features such as age, author position, discipline, institution, and location. Tracking how these features change over time helps to understand how mentorship evolves, how mentees transition into mentors, and how patterns of academic genealogy emerge. Beyond these cases, MapAffil Tempo 2018 further enables broader and accurate investigations into researcher mobility across institutions, geographic regions, and disciplines, as well as macro-level trends such as brain drain and knowledge diffusion making a powerful foundation for Science of Science and policy studies of geography, collaboration, and scientific impact.

Accuracy of the inferences have yet not been evaluated, but we expect countries to be very robust and discipline to be the most variable, while city and institution are somewhere in the middle. The illustrative example of Victor Ambros (Figure 1) aligns well with the Wikipedia page [18], but we expect performance to be worse for people who authored few papers, or few first-authored papers before 2015. This is also the case for authors that were merged by Author-ity, and even more common for authors with very common names.

Acknowledgments

Research reported in this publication was supported by the National Institutes of Health (Award Number P01AG039347). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] OpenAlex. Why is affiliation history metadata incomplete? *OpenAlex Support*. Retrieved from https://help.openalex.org/hc/en-us/articles/29664870327447-Why-is-affiliation-history-metadata-incomplete

[2] Mike Thelwall and Ruth Fairclough. 2020. All downhill from the PhD? The typical impact trajectory of U.S. academic careers. *Quantitative Science Studies* 1, 3 (August 2020), 1334–1348. https://doi.org/10.1162/qss_a_00072

[3] Silvia Dobre, Rachel Herbert, Alvin Shijie Ding, and Hans Pohl. 2024. A bibliometric analysis using a newly developed model and a customizable research tool: A case study of researcher mobility in Sweden. *PLOS ONE* 19, 12 (December 2024), 1–25. https://doi.org/10.1371/journal.pone.0308147

[4] National Library of Medicine. pubmed_250101.dtd: Affiliation element. *NLM DTD*. Retrieved from https://dtd.nlm.nih.gov/ncbi/pubmed/doc/out/250101/el-Affiliation.html

[5] National Library of Medicine. 2014. Author, Corporate Author, and Collaborator Affiliation Display Changes. *NLM Technical Bulletin*. Retrieved from https://www.nlm.nih.gov/pubs/techbull/nd13/nd13_author_affiliation_display.html

[6] Vetle I. Torvik. 2021. MapAffil 2018 dataset -- PubMed author affiliations mapped to cities and their geocodes worldwide with extracted disciplines, inferred GRIDs, and assigned ORCIDs. *University of Illinois Urbana-Champaign, Illinois Data Bank*. https://doi.org/10.13012/B2IDB-2556310_V1

[7] Vetle I. Torvik. 2015. MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide. *Dlib Mag* 21, 11–12 (December 2015). https://doi.org/10.1045/november2015-torvik

[8] Vetle I. Torvik and Neil R. Smalheiser. 2021. Author-ity 2018 - PubMed author name disambiguated dataset. *University of Illinois Urbana-Champaign, Illinois Data Bank.* https://doi.org/10.13012/B2IDB-2273402_V1

[9] Vetle I. Torvik, Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. 2005. A probabilistic similarity metric for Medline records: A model for author name disambiguation: Research Articles. *J. Am. Soc. Inf. Sci. Technol.* 56, 2 (January 2005), 140–158.

[10] Vetle I. Torvik and Neil R. Smalheiser. 2009. Author name disambiguation in MEDLINE. *ACM Trans. Knowl. Discov. Data* 3, 3 (July 2009). https://doi.org/10.1145/1552303.1552304

[11] Vetle I. Torvik. 2018. Author-Linked data for Author-ity 2009. *University of Illinois at Urbana-Champaign, Illinois Data Bank*. https://doi.org/10.13012/B2IDB-4370459_V1

[12] Vetle I. Torvik. 2021. ORCIDs mapped to PubMed authors. *University of Illinois at Urbana-Champaign, Illinois Data Bank.* https://doi.org/10.13012/B2IDB-9246015_V1

[13] National Library of Medicine. 2025. About PubMed. *PubMed*. Retrieved from https://pubmed.ncbi.nlm.nih.gov/about/

[14] National Library of Medicine. 2025. About MEDLINE. *NLM*. Retrieved from https://www.nlm.nih.gov/medline/medline_overview.html

[15] National Library of Medicine. 2024. OLDMEDLINE Data. *NLM*. Retrieved from https://www.nlm.nih.gov/databases/databases_oldmedline.html

[16] Xiaoliang Jiang, Nigel Bosch, and Vetle I. Torvik. 2025. Training a Geographic Entity Recognizer on Biomedical Abstracts with the Aid of Embeddings, Metadata, and Linked Data. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries (JCDL '24)*. Association for Computing Machinery, New York, NY, USA, Article 12, 1–5. https://doi.org/10.1145/3677389.3702515

[17] Giacomo Vaccario, Luca Verginer, and Frank Schweitzer. 2020. The mobility network of scientists: analyzing temporal correlations in scientific careers. *Applied Network Science* 5, 1 (July 2020), 36. https://doi.org/10.1007/s41109-020-00279-x

[18] Wikipedia. 2025. "Victor Ambros.". Last modified September 24, 2025. https://en.wikipedia.org/wiki/Victor_Ambros.