Transactions on Learning Architectures
Vol. 1 No. 1 (2026)

What an Initial Does to a Bibliographic Name Search

Ronald Boyd · Albert Simpson · Kevin Wilson

Published: November 26, 2026 | Article 1229

Abstract

Across 2,564 Crossref works, 11,276 of 24,298 author positions shared a surname-and-initial key with at least one other DOI. Retaining the complete given-name field reduced that count to 4,163. Forty-four groups still contained more than one distinct deposited ORCID under full-name matching. The reverse pattern also occurred. Fifty-six ORCID values were associated with more than one full-name key. The results describe candidate sets generated by different name rules and the limits of inferring author identity from those sets.

What an Initial Does to a Bibliographic Name Search

Abstract

Across 2,564 Crossref works, 11,276 of 24,298 author positions shared a surname-and-initial key with at least one other DOI. Retaining the complete given-name field reduced that count to 4,163. Forty-four groups still contained more than one distinct deposited ORCID under full-name matching. The reverse pattern also occurred. Fifty-six ORCID values were associated with more than one full-name key. The results describe candidate sets generated by different name rules and the limits of inferring author identity from those sets.

One character can change the work list

A bibliographic search often begins with a name as it appears in a formatted reference. Reducing a given name to one initial makes the search usable even when the original citation gives only one letter. The reduction also makes different given names look identical. A library or analytics team that uses such a key can gather a longer list of candidate records, even though the stored surname, individual DOIs, and their journal assignments have not changed. The size of that candidate list matters before any decision about which works belong to a particular author.

The citation by Yi et al. (2024), DOI 10.1038/s41392-024-02052-3, provides a way to inspect the change within a single byline. It carries 14 separate positions with a given name and a family name in Crossref. Thirteen positions have an abbreviated name key that returns at least one other DOI in our frame. Only one position does when the complete given-name field is retained. The first contributor's abbreviated key occurs on two distinct works in the pooled inventory. Its full-name key occurs on only the selected work. The byline supplies these observable comparisons without attributing the other DOI to that contributor.

We ask how the choice between the complete given name and a first initial changes the number of byline positions that share a key with another DOI. A second question is whether carrying the journal ISSN in the key changes that result. A third asks when the public identifier evidence points to observed conflicts. We compare the distinct deposited ORCID values collected under one name key and, in the reverse direction, the different full-name keys recorded under one ORCID. These are partly observed indicators. They describe what the metadata says about candidates, and identifying particular individuals would require additional evidence.

Candidate groups and authorship evidence

Record linkage separates generating candidate pairs from deciding their identity. Fellegi and Sunter (1969) formulate linkage decisions from agreement and disagreement on comparison evidence. Jaro (1989) describes grouping candidate records before matching in a large application, and Christen (2012) discusses how efficient candidate selection affects later resolution. Codd (1970) gives a relational model for separating keys from other recorded attributes. A group formed by equal name fields is an operational group. The inputs can be accurately copied from both publications while still giving limited evidence that they share a contributor.

In bibliographic disambiguation, title and source context can be used with names. Torvik and Smalheiser (2009) develop one such approach to organizing author occurrences, and Liu et al. (2014) describe a disambiguation system for PubMed. Tang et al. (2012) model the problem with combinations of evidence from digital library records. Ferreira et al. (2012) distinguish methods that organize publication occurrences in different ways. Our work stops at candidate generation. We use two exact string rules, with

or without the declared journal scope. No algorithm clusters the underlying people, and there is no demographic inference from a string.

Training and evaluation datasets can support more ambitious conclusions where independent identity labels are available. Müller et al. (2017) compare such datasets and present a new resource. Subramanian et al. (2021) harmonize several benchmarks with a consistent evaluation suite, and Zhang et al. (2023) develop labeled datasets for multiple views of the problem. Kim et al. (2023) show why retaining different representations of a name can matter to the output of a disambiguation system. We do not add a script conversion, rejoin name components, draw labels from personal profiles, or substitute direct human annotation. This keeps the empirical comparison narrow enough to replay from the deposited fields.

An ORCID iD is intended to support attribution of scholarly activities to a persistent identifier (Haak et al., 2012). Practical research evaluation can use person identifiers together with identifiers for other objects (Haak et al., 2018). Porter (2022) documents that how identifiers enter a record depends on participation and integration practices. Kim and Owen-Smith (2021) show the usefulness of linking ORCID evidence for a larger disambiguation evaluation when that linking has been designed as an explicit source. The present evidence is more limited. An ORCID asserted on a Crossref author slot is one public bibliographic observation. We check its syntax and check digit. We treat separately populated values as conflicts for a one-person interpretation of the whole name group, and the absence of a value leaves that slot unclassified.

Four journal frames and one unit of observation

The empirical frame includes the journal-article records returned by Crossref during the 2024 publication window for Signal Transduction and Targeted Therapy, Nature Neuroscience, Scientific Data, and Communications Physics. The frame queries use the ISSN of each journal. Titles were selected to retain the demonstration record and observe exact name-key behavior across several different byline collections. They are a purposive group and cannot support a discipline-wide estimate. Crossref's infrastructure aggregates metadata deposited by participating publishers (Hendricks et al., 2020). Its public REST interface supports the source retrieval (Crossref, n.d.-b). We requested only each DOI and its author array. All results were recovered through cursor pages with a reported count check. The captured period spans 19:01:43 to 19:02:04 UTC on 23 November 2026.

An observation is one author position on one DOI. Paskin (2010) explains how a DOI identifies an object under registration. Here its stable role is to distinguish the underlying work records for each key. It does not represent one identifiable person. The Crossref contributor guidance describes how the given and family name fields can inform display and lookup (Crossref, n.d.-a). We process only positions with usable letters in both fields. Where a listed article has no eligible author position it can remain in the frame count but has no eligible slot in the position denominator. If the same key appears twice in one byline, counting distinct DOI names prevents that case alone from giving two candidate works.

For the full-name key we normalize each supplied given-name and family-name field with NFKC, case folding, and internal whitespace collapse. Unicode normalization guidance distinguishes compatibility equivalents from literal character sequences (Unicode Consortium, 2025). We preserve the order of the two fields and keep punctuation and diacritics. The complete given-name key refers to the available field even if a depositor already supplied a shortened value. The initial rule keeps the same family name and selects the first alphabetic character of the normalized given field. The journal-restricted versions pair the respective name key with the ISSN under which the record was acquired. A name-key group associated with more than one DOI can therefore be counted under four precisely specified rules.

The data archive retains deterministic digests for both equality keys. We also check the ORCID string after removing the resolver prefix and canonicalizing it. The ORCID identifier documentation specifies its check digit under the MOD 11-2 algorithm (ORCID, n.d.-b). The use of an authenticated integration has a different meaning in ORCID practice (ORCID, n.d.-a). We neither require nor measure authentication in this audit. The check digit rejects malformed values. It does not establish an authenticated, correctly assigned, or independently verified researcher identity. The captured equality label remains the value deposited on the source position and is not joined to person profiles.

The main reported outcome counts eligible positions whose group has two or more distinct work DOIs. We also report the number of different keys, the number of groups with multiple works, the largest number of DOIs per group, and groups with more than one checksum-valid ORCID value. The latter measure requires at least two distinct observed assertions, so it is a conservative diagnostic of conflicts among identifier values. It does not estimate how many groups really contain different people. In the reverse direction we group the same observed slots by their ORCID label and count how often more than one full-name key is present. Figure 1 shows the grouping design and how the pooled name rules change each journal's candidate load.

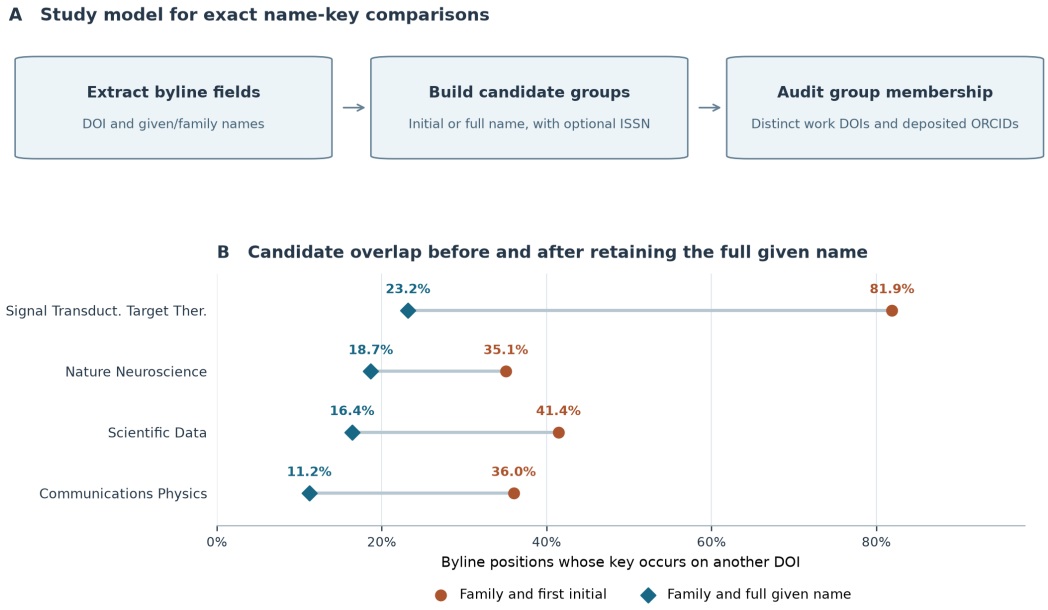

Figure 1. Each eligible slot is checked using the same four-journal inventory. The chart displays the initial-only and complete-given-name variations without filtering the candidate DOI list by journal.

How much the candidate groups change

Table 1. Observed key overlap and multiple asserted identifiers under four rules

Equality fields	Distinct key groups	Groups on several DOIs	Byline positions in those groups	Groups with several deposited ORCIDs	Largest DOI count per group
Family name and first given initial	15,258	2,261	11,276	398	110
Family name and full given name	21,886	1,758	4,163	44	11
ISSN, family name, and first given initial	17,099	2,483	9,653	316	61
ISSN, family name, and full given name	22,359	1,550	3,482	20	9

There were 2,564 distinct DOI records and 24,313 recorded author positions. Fifteen positions did not meet the name-field eligibility rule, leaving 24,298 in the analysis. Under the pooled initial key, 11,276 positions, or 46.4%, were associated with at least one other DOI. Under the complete given-name rule that number was 4,163, or 17.1%. The difference of 7,113 positions is the reduction in multiple-work candidacy reached by preserving the longer field on the observed rows. The largest group under the initial rule spans 110 different works, while the largest full-name group spans 11.

The effect differs across the four frames. Signal Transduction and Targeted Therapy had 3,232 of 3,946 eligible positions in pooled multi-work initial groups and 915 in pooled multi-work full-name groups. The corresponding counts were 1,181 and 629 for Nature Neuroscience, 5,713 and 2,261 for Scientific Data, and 1,150 and 358 for Communications Physics. The position-level comparison is affected by the number of author slots on each record and the mix of names. These journal differences cannot be used to establish what causes names to overlap.

Requiring the same journal frame reduces pooled initial-rule candidate positions by 1,623 and full-rule positions by 681. The number of multi-DOI initial keys nonetheless rises from 2,261 to 2,483. A single pooled group can split into several journal groups that each still contain several DOI values. Other slots lose all their other candidates. Thus a count of groups and a count of positions can move in opposite directions when the same matching rule is scoped to narrower collections. This property would be obscured by recording only the total number of repeated labels.

A checksum-valid ORCID was available on 7,661 positions, or 31.5% of eligible positions. With all journals pooled, the initial key gives 398 groups containing at least two distinct ORCID values and the full-name key gives 44. The 44 groups span 186 positions altogether, including 113 with an observed ORCID label. Adding journal scope leaves 20 such groups using the full-name rule. A different check starts with the 7,000 distinct ORCID values observed in the eligible data. Fifty-six occupy more than one full-name key. The result shows both kinds of disagreement between directly observable fields. It leaves room for several explanations, including differences in supplied name forms or errors in the deposited assertions.

Different observed coverage of identifiers also changes the value of the conflict counts across venues. Of the 3,946 eligible positions in the Signal frame, 639 carried a checksum-valid identifier. The comparable fractions were 1,424 of 3,369 in Nature Neuroscience, 4,084 of 13,788 in Scientific Data, and 1,514 of 3,195 in Communications Physics. A group can register several distinct deposited identifiers only when at least two of its members have them. With sparse coverage in one portion of the inventory, a grouping rule

has fewer opportunities to show conflicts there. The aggregate counts describe the evidence that can be observed under each rule. They do not serve as a denominator for classifying incomplete groups.

Interpreting a shorter or longer list

Bibliographic service comparisons report differences in scope and in the availability of optional fields (Visser et al., 2021 and Delgado-Quirós & Ortega, 2024). A comparison using one registry snapshot avoids combining those differences with the choice of name key. It also means an unmatched name on a different service says nothing about the candidate sets we observed. Wang and Strong (1996) frame information quality in terms of how data are used. Stvilia et al. (2007) propose a framework for evaluating that suitability, and Park and Tosaka (2010) describe the importance of metadata quality controls. A short name can be usable as a display form while giving a much longer candidate set in a data merge.

The same limitation applies in the other direction. When a full-name group contains only one DOI in this frame, a longer time window could reveal other records under the same name. Conversely, the group might never acquire another DOI under exactly that name because the same contributor's later record uses a different form. Looking within one ISSN gives a useful comparison of browsing scope, while any person whose work spans journals could be divided by that restriction. The observed proportions quantify what each rule does to database candidates, without measuring the rates of correct identity resolution.

A reusable candidate audit can preserve group definitions and source selection along with observed counts. Wilkinson et al. (2016) emphasize accessible methods and provenance for reuse of research data. Here the raw author strings are used locally to form equality labels and are outside the archived analytic table. The figures can be rebuilt from those labels and the stored DOI and position information. The main effect is substantial at the stage of choosing candidates. Retaining a complete given-name field reduced the number of positions with a second candidate, while the identifier comparison shows why even the remaining groups cannot be treated as ready-made author profiles.

Data and materials

The archived records contain public DOI and journal frame identifiers, contributor positions, and deterministic equality labels for name keys and observed ORCID values. Aggregate results, the derived position table, and both research figure formats are documented in the draft record.

References

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. *Springer*. https://doi.org/10.1007/978-3-642-31164-2

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), 377–387. https://doi.org/10.1145/362384.362685

Crossref (n.d.-a). Authors and editors. *Crossref Documentation*. https://support.crossref.org/hc/en-us/articles/214567746-Authors-and-editors

Crossref (n.d.-b). REST API. *Crossref Documentation*. https://www.crossref.org/documentation/retrieve-metadata/rest-api/

Delgado-Quirós, L.; Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 5(1), 31–49. https://doi.org/10.1162/qss_a_00286

Fellegi, I. P.; Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Ferreira, A. A.; Gonçalves, M. A.; Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record*, 41(2), 15–26. https://doi.org/10.1145/2350036.2350040

Haak, L. L.; Fenner, M.; Paglione, L.; Pentz, E.; Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, 25(4), 259–264. https://doi.org/10.1087/20120404

Haak, L. L.; Meadows, A.; Brown, J. (2018). Using ORCID, DOI, and Other Open Identifiers in Research Evaluation. *Frontiers in Research Metrics and Analytics*, 3, 28. https://doi.org/10.3389/frma.2018.00028

Yi, L.; Pan, H.; Ning, Z.; Xu, L.; Zhang, H.; Peng, L.; Liu, Y.; Yang, Y.; Si, W.; Wang, Y.; et al. (2024). Clinical and biomarker analyses of SHR-1701 combined with famitinib in patients with previously treated advanced biliary tract cancer or pancreatic ductal adenocarcinoma: a phase II trial. *Signal Transduction and Targeted Therapy*, 9(1), 347. https://doi.org/10.1038/s41392-024-02052-3

Hendricks, G.; Tkaczyk, D.; Lin, J.; Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. https://doi.org/10.1162/qss_a_00022

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406), 414–420. https://doi.org/10.1080/01621459.1989.10478785

Kim, J.; Owen-Smith, J. (2021). ORCID-linked labeled data for evaluating author name disambiguation at scale. *Scientometrics*, 126(3), 2057–2083. https://doi.org/10.1007/s11192-020-03826-6

Kim, J.; Kim, J.; Kim, J. (2023). Effect of Chinese characters on machine learning for Chinese author name disambiguation: A counterfactual evaluation. *Journal of Information Science*, 49(3), 711–725. https://doi.org/10.1177/01655515211018171

Liu, W.; Islamaj Doğan, R.; Kim, S.; Comeau, D. C.; Kim, W.; Yeganova, L.; Lu, Z.; Wilbur, W. J. (2014). Author Name Disambiguation for PubMed. *Journal of the Association for Information Science and Technology*, 65(4), 765–781. https://doi.org/10.1002/asi.23063

Müller, M. C.; Reitz, F.; Roy, N. (2017). Data sets for author name disambiguation: an empirical analysis and a new resource. *Scientometrics*, 111(3), 1467–1500. https://doi.org/10.1007/s11192-017-2363-5

ORCID (n.d.-a). Research Information Systems Integration Best Practices. *ORCID*. https://info.orcid.org/research-information-systems/

ORCID (n.d.-b). Structure of the ORCID Identifier. *ORCID Support*. https://support.orcid.org/hc/en-us/articles/360006897674-Structure-of-the-ORCID-Identifier

Park, J. R.; Tosaka, Y. (2010). Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly*, 48(8), 696–715. https://doi.org/10.1080/01639374.2010.508711

Paskin, N. (2010). Digital Object Identifier (DOI®) System. *Encyclopedia of Library and Information Sciences, Third Edition*, 1586–1592. https://doi.org/10.1081/e-elis3-120044418

Porter, S. J. (2022). Measuring Research Information Citizenship Across ORCID Practice. *Frontiers in Research Metrics and Analytics*, 7, 779097. https://doi.org/10.3389/frma.2022.779097

Stvilia, B.; Gasser, L.; Twidale, M. B.; Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12), 1720–1733. https://doi.org/10.1002/asi.20652

Subramanian, S.; King, D.; Downey, D.; Feldman, S. (2021). S2AND: A Benchmark and Evaluation System for Author Name Disambiguation. *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 170–179. https://doi.org/10.1109/jcdl52503.2021.00029

Tang, J.; Fong, A. C. M.; Wang, B.; Zhang, J. (2012). A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *IEEE Transactions on Knowledge and Data Engineering*, 24(6), 975–987. https://doi.org/10.1109/tkde.2011.13

Torvik, V. I.; Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, 3(3), 1–29. https://doi.org/10.1145/1552303.1552304

Unicode Consortium (2025). Unicode Normalization Forms. *Unicode Standard Annex #15, Revision 57*. https://www.unicode.org/reports/tr15/tr15-57.html

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

Wang, R. Y.; Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J. W.; da Silva Santos, L. B.; Bourne, P. E.; et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zhang, L.; Lu, W.; Yang, J. (2023). LAGOS-AND: A large gold standard dataset for scholarly author name disambiguation. *Journal of the Association for Information Science and Technology*, 74(2), 168–185. https://doi.org/10.1002/asi.24720