Studies in Efficient Intelligence
Vol. 1 No. 1 (2026)

A Local XML Identifier Becomes Ambiguous When Article Fronts Are Merged

Martha Kelley · William Hart · Reed Walker

Published: November 26, 2026 | Article 1241

Abstract

A contributor pointer in journal XML identifies a target within its own document. To see what happens when that scope is dropped, we extracted affiliation and correspondence pointers from 60 exact PubMed Central records. All 629 pointers resolved uniquely in their source front matter. Joining the same tokens across the pooled documents yielded 9,532 candidate rows, including 8,903 from other articles. Preserving the declared pointer type produced the same result, while requiring the same journal still admitted 4,294 cross-document candidates. A first-match lookup using the pooled typed key chose a target outside the source on 567 of 629 pointers. A document key is therefore needed even when the XML identifier, the target class and the required number of output rows are all known.

A Local XML Identifier Becomes Ambiguous When Article Fronts Are Merged

Abstract

A contributor pointer in journal XML identifies a target within its own document. To see what happens when that scope is dropped, we extracted affiliation and correspondence pointers from 60 exact PubMed Central records. All 629 pointers resolved uniquely in their source front matter. Joining the same tokens across the pooled documents yielded 9,532 candidate rows, including 8,903 from other articles. Preserving the declared pointer type produced the same result, while requiring the same journal still admitted 4,294 cross-document candidates. A first-match lookup using the pooled typed key chose a target outside the source on 567 of 629 pointers. A document key is therefore needed even when the XML identifier, the target class and the required number of output rows are all known.

1. Why an exact key can give the wrong match

An article's structured front matter can point from a contributor to an affiliation element, a correspondence note, or another target in the same document. Rather than repeating the entire affiliation in each contributor element, the document assigns one identifier to the target and reuses a smaller pointer. When exported into a flat table, a pointer has a recognizable identifier string and a meaningful declared type. An analyst could join it to the table of extracted target attributes by exactly matching those values. If the document identifier was discarded when the tables were pooled, the join may associate the pointer with other articles whose front matter legitimately used the same identifier string.

The XML standard places uniqueness and reference constraints within a document (World Wide Web Consortium, 2008). The JATS tag library permits a reference attribute to identify one or several local targets and gives a vocabulary for specifying a cross-reference type (National Information Standards Organization, 2024a, 2024b). JATS4R (n.d.) describes how contributor and affiliation markup can maintain these links outside display text. Relational databases also support composite keys, which let several columns jointly identify a row (PostgreSQL Global Development Group, n.d.). The question for a metadata export is how much of the document context it must carry into that composite key. The nominal exactness of the identifier does not settle where it is unique.

A separate observation verifies the distinction in a real archive item. The PMC front matter for Yi et al. (2024), DOI 10.1016/j.esmoop.2024.102239, contains 24 affiliation pointer tokens that resolve to four distinct local affiliation elements, and two additional correspondence pointers. All 26 have unique typed targets when resolved inside the originating front. Several target IDs are also used in other observed front records. We use the matching archive and DOI identifiers to make the example reproducible. No names, institutions, correspondence content or publication content is needed for the comparison.

The empirical question is how the number of matching target rows changes when the source article is removed from the join key. We separately test whether keeping the declared target type, keeping only the journal as document context, or selecting one matching row can recover the same local association. The source of each valid target is known under the document-scoped interpretation, so all results are counted on a fixed set of pointers that initially succeed. This evaluation distinguishes ambiguity caused by pooling correct documents from failure to parse a malformed pointer in a source file.

2. Local and bibliographic identity serve different tasks

An article identifier is intended to remain interpretable across systems. The International DOI Foundation (n.d.) describes a persistent identifier for the registered object, and Lagoze et al. (2002) distinguish the item identifier from the representations a repository can offer. The PubMed Central protocol documentation specifies how an OAI identifier represents an archive item (PubMed Central, n.d.). A local XML identifier answers a different question by design. It gives a reusable target within an article under its XML document scope. Its text can be repeated across published items because each item carries its own identification space.

Institution identifiers provide another useful contrast. The Research Organization Registry (n.d.) describes persistent identifiers for organizations and their location in organization records. Crossref (2026) documents affiliation information and identifiers for contributors in registered metadata. Donner et al. (2020) compare the effects of different affiliation disambiguation systems on bibliometric indicators. Our experiment studies a step before organizational disambiguation. It tests whether an XML pointer is attached to the element the document intended. Two affiliation elements could accurately describe the same organization and still be different targets for a document's internal pointer. Conversely, an identical author-affiliation identifier can lead to unrelated elements across documents.

Record linkage research often considers how data values distinguish objects across systems. Fellegi and Sunter (1969) establish a framework for deciding when record observations support a match. Christen (2012) distinguishes the role of exact keys from approximate and probabilistic evidence. Elmagarmid et al. (2007) review duplicate-record detection as an integration problem, and Binette and Steorts (2022) cover entity resolution at scale. Here matching the string is the easy part. The ambiguity arises because the identifier comparison discards a required scope before it takes place. An algorithm that became more tolerant of spelling differences would have no bearing on whether an exact local string can map to a different article.

The distinction matters for exports made from several metadata formats or services. Haslhofer and Klas (2010) describe how schema mappings can preserve different quantities of context. Jackson et al. (2008) study the properties of metadata shared through OAI-PMH, and Han et al. (2009) analyze what can be lost when repository-specific fields are translated to simple shared fields. A flat target table operates like another mapping. If it keeps the label and local ID but drops the originating article, it preserves a display token and removes a condition for using that token. We measure that transformation on the same set of source documents rather than comparing different data providers.

3. Archived metadata and join rules

3.1. Select the records and verify article scope

The source frames were six existing 2024 PubMed queries by exact journal ISSN for PNAS, PLOS Biology, PLOS ONE, Nature Communications, Scientific Reports, and eLife. The recorded queries retrieved an initial set of up to 100 citations in publication-date order from each journal. We first restricted the candidates to records with the necessary PubMed ID, DOI and PMC accession. An ordered SHA-256 ranking of the DOI, with a fixed namespace string, then selected ten articles per journal. The finite frame tests six observed publishing configurations at equal sample sizes. It does not represent the volume of the journals or the whole archive.

Each selected PMC item was requested in the public OAI front-matter format. In the NCBI infrastructure described by Sayers et al. (2024), PubMed and PMC expose different projections of a work. We required the returned front to identify the expected DOI and archive item, and to agree with the source PMID

wherever that value was explicitly returned. We matched identifiers without author, title or organization comparisons. Visser et al. (2021) show that the scope of scholarly databases can vary, and Delgado-Quirós and Ortega (2024) distinguish catalog coverage from the completeness of specific metadata. Consequently, an item failing alignment would be reported separately from a locally unresolved reference within an aligned front. The separate C1 example was queried under the same rules and excluded from the 60-work comparison.

3.2. Define the comparison at the pointer token

We recorded direct contributor cross-reference elements explicitly typed as affiliation or correspondence within the selected article metadata front. The target candidates were affiliation and correspondence elements bearing local IDs in the corresponding administrative sections. If a pointer could carry multiple target IDs, the parsing rule tokenized the space-separated attribute values according to the documented syntax. In the observed sample every included element carried one token. The compact archived projection saves contributor positions, role labels, target elements identified by an ordinal and one-way equality-preserving encodings of the local values. It contains no names, affiliation text, emails or note content. Hashing supports exact equality within and across documents without storing the original markup strings.

The reference case resolves a typed pointer only inside the source article. Our primary comparison uses pointer tokens with exactly one matching local target of the expected type. Other tokens could be informative about the measured front scope, but their intended target is not established by these criteria. We compare three altered table joins on the fixed valid tokens. The first ignores article and type and matches the ID alone. The second keeps the target type with the ID but ignores the article. The third groups by journal and target type before matching the ID, replacing article-level context with serial-level context. Each candidate row retains its originating article. We can therefore count expansion within the source scope and expansion into a different source document separately.

A common data reduction could keep only one row per pointer. We simulate the first and last candidate under the unscoped typed key, ordered by the numeric portion of PMCID and then by target position within a document. The ordering is a deterministic convention for this sensitivity. It does not encode preference or certainty. Since the reference case already supplies the unique correct document, both rules can be assessed by how often their selected target originates elsewhere. We record these outcomes on the same pointers used for the full candidate expansion.

This is a structured data-quality audit. Stvilia et al. (2007) distinguish how a use context affects judgments of information quality. Wang and Strong (1996) describe dimensions of data quality from the consumer perspective, and Batini et al. (2009) compare data quality assessment methodologies. Park and Tosaka (2010) emphasize semantic quality controls for digital collections. The reference roles and accepted join keys are consequently defined before counting. Our result is an observed difference between projections of the same archive records, rather than a claim that any publicly visible target element has incorrect markup.

4. Results

4.1. Removing the document key multiplies target rows

All 60 sampled front records satisfied the identity checks. They contained 629 eligible pointer elements and 283 typed target nodes. The observed pointers included 592 affiliation tokens and 37 correspondence tokens. Each token resolved to exactly one typed target within the source article. One sampled record contributed no eligible pointer. Among the remaining 59, every record had at least one pointer that acquired a candidate from another document when the typed key was applied across the collection.

Matching only the element identifier returned 9,532 candidate rows. This is more than 15 times the number of pointers properly resolved within their document. Of those rows, 8,903, or 93.4%, attached the pointer to another article's target element. Keeping the type without keeping the article removed no candidates in this sample. The recognized IDs reused across documents generally occurred under the same target class already. There were 34 distinct typed key forms across the included target declarations, and 28 appeared in at least two articles. At the pointer level, 618 of 629 tokens had several candidates when the exact typed ID was applied to the full collection.

Table 1. *Candidate expansion from the same 629 locally resolved pointer tokens.*

Join scope	Candidate target rows	Rows from a different article	Pointers with several candidates
Source article, role and ID	629	0	0
Journal, role and ID	4,923	4,294	612
Role and ID across all fronts	9,532	8,903	618
ID across all fronts	9,532	8,903	618

Note. The reference interpretation treats ID strings as local to the source XML article and requires the declared target role. Pooled joins use the same input tokens and exact string equality. A row from another article describes a different target element regardless of organizational similarity, which was not measured.

Restricting the match to the same journal removed some candidates. It still attached 4,294 targets from other articles to the input pointer rows. This corresponds to nearly seven foreign candidates for each correctly scoped pointer on average. Affiliations accounted for 8,332 outside-article rows under the global typed merge and correspondence for 571. The smaller number of correspondence tokens makes their relative expansion visible. Every one of the 37 received several candidates under the global typed rule. Both target classes therefore contribute to the observed effect rather than a single type accounting for all of it.

4.2. A fixed one-row decision can conceal the ambiguity

The global typed key leaves the correct target somewhere among the candidates because it does not exclude the source document. Choosing only the first row by the prespecified order selected a target from a different article for 567 of the 629 pointer tokens, leaving 62 correct. Choosing the last left 61 correct and selected 568 outside the source article. Both decisions return a single row per pointer. A downstream record count based only on the resulting number of rows would therefore look unchanged. It would hide a large change in which element was selected.

The separate C1 example illustrates why a result verified on one item is insufficient to establish scope in a collection. In a table containing only its front, the typed and untyped exact keys both return the same 26 local candidates. Appending that work to the comparison set only for the illustration yields 396 global typed candidates for its pointer tokens. Of these, 370 originate in other documents. Twenty-five of its 26

tokens become ambiguous. The example shows that local validity and global confusion can coexist for precisely the same strings and declared roles. These extra rows and targets are not included in the aggregate counts in Table 1.

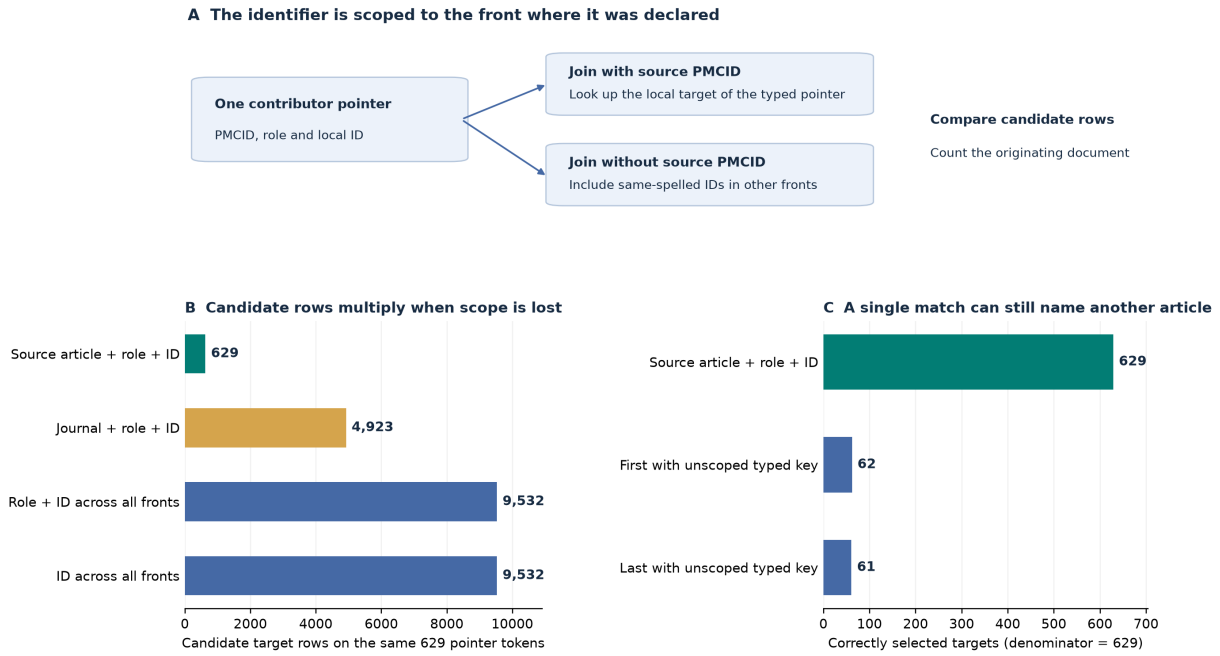

Figure 1. *Panel A traces how an article's contributor pointer retains or loses its source scope. Panel B counts target candidates for the same 629 input tokens. Panel C shows how many selected targets still originate in the correct article when the logic requires one row.*

5. Implications for local identifier exports

A table that has already lost its parent document can still display plausible contributor relationships. Both sides contain stable identifiers as far as their original XML context is concerned, and an exact value match can produce a well-shaped result. The substitution becomes visible only if the associated article is retained on both sides of the equality test or if the candidate list is audited. In these records, a journal key and declared role had insufficient selectivity to recover intended edges. The archive document identifier changed which target is addressable rather than simply improving the match rate of an approximate algorithm.

There are limits to interpreting these results across collections. We deliberately kept only two administrative pointer classes and required a definite local target for the comparative analysis. The observed range of reused IDs may vary by publisher template, metadata process, topic-free archive configuration and sample overlap. A distinctly patterned identifier format could make exact matches unique across a particular sample. It would still be a separate observation from the local validity constraint. This study does not assign an organization identity to any affiliation label, compare people across texts or diagnose archive deposits as erroneous. The simulated joins are the only thing that discards source scope.

The preservation of local context is also part of reproducibility. Wilkinson et al. (2016) place identifiers and interpretation inside reusable metadata principles, while Peng (2011) distinguishes rerunning a computation from repeating a data collection. Gebru et al. (2021) describe how documenting the composition and uses of a dataset supports later analysis. Leo et al. (2024) provide additional provenance methods for tracking computational inputs and outputs. The article-level receipts and anonymous local key representations in this audit make it possible to recompute candidate multiplicities without saving

contributor names or targets' textual values. A later collection could repeat the checks with another set of templates and report exactly which joins preserve the original pointer relationships.

Data and code availability

The reproduction record links the archive profiles, compact XML equality projections, analysis and figures. Author contributions, funding and competing-interest declarations await author confirmation.

References

Batini, Carlo; Cappiello, Cinzia; Francalanci, Chiara; Maurino, Andrea (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41(3), 1–52*. DOI.

Binette, Olivier; Steorts, Rebecca C. (2022). (Almost) all of entity resolution. *Science Advances, 8(12), eabi8021*. DOI.

Christen, Peter (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. *Springer Berlin Heidelberg*. DOI.

Crossref (2026). Contributors. *Crossref Metadata Documentation*. Source.

Delgado-Quirós, Lorena; Ortega, José Luis (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies, 5(1), 31–49*. DOI.

Donner, Paul; Rimmert, Christine; van Eck, Nees Jan (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies, 1(1), 150–170*. DOI.

Elmagarmid, Ahmed K.; Ipeirotis, Panagiotis G.; Verykios, Vassilios S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering, 19(1), 1–16*. DOI.

Fellegi, Ivan P.; Sunter, Alan B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association, 64(328), 1183–1210*. DOI.

Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; III, Hal Daumé; Crawford, Kate (2021). Datasheets for datasets. *Communications of the ACM, 64(12), 86–92*. DOI.

Han, Myung-Ja; Cho, Christine; Cole, Timothy W.; Jackson, Amy S. (2009). Metadata for Special Collections in CONTENTdm: How to Improve Interoperability of Unique Fields Through OAI-PMH. *Journal of Library Metadata, 9(3-4), 213–238*. DOI.

Haslhofer, Bernhard; Klas, Wolfgang (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys, 42(2), 1–37*. DOI.

International DOI Foundation (n.d.). DOI Handbook. *DOI Foundation*. Source.

Jackson, Amy S.; Han, Myung-Ja; Groetsch, Kurt; Mustafoff, Megan; Cole, Timothy W. (2008). Dublin Core Metadata Harvested Through OAI-PMH. *Journal of Library Metadata, 8(1), 5–21*. DOI.

JATS4R (n.d.). Authors and affiliations. *JATS for Reuse, version 2.0*. Source.

Lagoze, Carl; Van de Sompel, Herbert; Nelson, Michael L.; Warner, Simeon (2002). The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. *Open Archives Initiative Specification*. Source.

Leo, Simone; Crusoe, Michael R.; Rodríguez-Navas, Laura; Sirvent, Raül; Kanitz, Alexander; De Geest, Paul; Wittner, Rudolf; Pireddu, Luca; Garijo, Daniel; Fernández, José M.; Colonnelli, Iacopo; Gallo, Matej; Ohta, Tazro; Suetake, Hirotaka; Capella-Gutierrez, Salvador; de Wit, Renske; Kinoshita, Bruno P.; Soiland-Reyes, Stian (2024). Recording provenance of workflow runs with RO-Crate. *PLOS ONE, 19(9), e0309210*. DOI.

National Information Standards Organization (2024a). Reference to an Identifier. *Journal Publishing Tag Library, JATS version 1.4*. Source.

National Information Standards Organization (2024b). Type of Cross-Reference. *Journal Publishing Tag Library, JATS version 1.4*. Source.

Park, Jung-Ran; Tosaka, Yuji (2010). Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly, 48(8), 696–715*. DOI.

Peng, Roger D. (2011). Reproducible Research in Computational Science. *Science, 334(6060), 1226–1227*. DOI.

PostgreSQL Global Development Group (n.d.). Constraints. *PostgreSQL 18 Documentation*. Source.

PubMed Central (n.d.). PMC OAI-PMH API. *PubMed Central Documentation*. Source.

Research Organization Registry (n.d.). Registry. *Research Organization Registry*. Source.

Sayers, Eric W; Beck, Jeff; Bolton, Evan E; Brister, J Rodney; Chan, Jessica; Comeau, Donald C; Connor, Ryan; DiCuccio, Michael; Farrell, Catherine M; Feldgarden, Michael; Fine, Anna M; Funk, Kathryn; Hatcher, Eneida; Hoeppner, Marilu; Kane, Megan; Kannan, Sivakumar; Katz, Kenneth S; Kelly, Christopher; Klimke, William; Kim, Sunghwan; Kimchi, Avi; Landrum, Melissa; Lathrop, Stacy; Lu, Zhiyong; Malheiro, Adriana; Marchler-Bauer, Aron; Murphy, Terence D; Phan, Lon; Prasad, Arjun B; Pujar, Shashikant; Sawyer, Amanda; Schmieder, Erin; Schneider, Valerie A; Schoch, Conrad L; Sharma, Shobha; Thibaud-Nissen, Françoise; Trawick, Barton W; Venkatapathi, Thilakam; Wang, Jiyao; Pruitt, Kim D; Sherry, Stephen T (2024). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 52(D1), D33–D43*. DOI.

Yi, L.; Ning, Z.; Xu, L.; Shen, Y.; Zhu, X.; Yu, W.; Xie, J.; Meng, Z. (2024). The combination treatment of oncolytic adenovirus H101 with nivolumab for refractory advanced hepatocellular carcinoma: an open-label, single-arm, pilot study. *ESMO Open, 9(2), 102239*. DOI.

Stvilia, Besiki; Gasser, Les; Twidale, Michael B.; Smith, Linda C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58(12), 1720–1733*. DOI.

Visser, Martijn; van Eck, Nees Jan; Waltman, Ludo (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies, 2(1), 20–41*. DOI.

Wang, Richard Y.; Strong, Diane M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 12(4), 5–33*. DOI.

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3(1), 160018*. DOI.

World Wide Web Consortium (2008). Extensible Markup Language (XML) 1.0, Fifth Edition. *W3C Recommendation*. Source.