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

When Crossref License Filters Match Different Entries on the Same DOI

Curtis Taylor · Leonard Harris · Kevin Wilson

Published: November 26, 2026 | Article 1242

Abstract

When an article has several deposited licenses, the URL and content role selected by a query may come from different entries. We tested five URL strings crossed with two roles in 1,593 Crossref records. Every combined search returned exactly the intersection of its two single-filter searches. The API returned 2,108 DOI–query matches, of which 1,673 had a matching license entry. The remaining 435 came from one journal's deposits, where the two filters could match different entries. Counting unique DOIs hid the discrepancy because every affected DOI also matched at least one valid URL–role pair.

When Crossref License Filters Match Different Entries on the Same DOI

Abstract

When an article has several deposited licenses, the URL and content role selected by a query may come from different entries. We tested five URL strings crossed with two roles in 1,593 Crossref records. Every combined search returned exactly the intersection of its two single-filter searches. The API returned 2,108 DOI–query matches, of which 1,673 had a matching license entry. The remaining 435 came from one journal's deposits, where the two filters could match different entries. Counting unique DOIs hid the discrepancy because every affected DOI also matched at least one valid URL–role pair.

A DOI can carry several license statements

A cataloguer reviewing a batch of articles might want the DOI records whose version of record is associated with a specific license URL. Crossref's works endpoint documents one filter for the URL and a second for the content version. Using both looks like a direct translation of the cataloguer's question. The translation depends on the domain over which the conjunction is evaluated. The record could list the requested URL in one entry and the requested role in another. The API might include the article in a search because both values exist somewhere on that record. A report that assigns the URL to the requested role would then make a statement for which there is no matching entry in the metadata.

The Crossref record for Yi et al. (2024), DOI 10.1016/j.esmoop.2024.102239, shows the practical difference. It supplies three license entries. Two give different Elsevier URLs with the role code tdm, and one gives a Creative Commons URL with the code vor. An exact DOI restriction allows us to check the six possible combinations of the three URLs and two roles individually. Each combination returns the record. Only three of them appear in individual entries. This article uses the record solely to illustrate how the public deposit represents those fields.

We ask whether a combined URL and role search can be reproduced by requiring both values on the same license entry, or by requiring each anywhere on the work record. We also ask which counts are affected by the choice of rule, how concentrated the differences are across journal frames, and whether the pattern survives aggregation to unique DOIs. The outcome is search membership and its correspondence to the public metadata. The target quantity is the count of DOI–query matches with an observed object carrying the two corresponding fields.

What the metadata represents

Crossref depends on deposits that provide persistent scholarly identifiers together with structured descriptions of works (Hendricks et al., 2020). The DOI anchors the work regardless of how many license references the deposit includes. This is the identity function described by Paskin (2010). Within a Crossref deposit an individual license reference can specify the object or content to which it applies, including a version of record or text and data mining, and the REST API publishes distinct filters for license URL and license version (Crossref, n.d.-a, n.d.-b). Field-level descriptions establish which values can be declared or queried. The empirical question here is what happens when those queries are applied jointly.

The National Information Standards Organization treats free-to-read status and a reference to terms as different indicators, and its Access and License Indicators recommendation defines an applicability qualifier for a license reference. Its JATS4R Permissions recommendation also locates permissions at the appropriate level of an article or an associated object (National Information Standards Organization,

2021a, 2021b). In another metadata model, the ODRL Information Model organizes rights statements around explicit relations among such elements as rules and targets (Iannella & Villata, 2018). The DataCite Metadata Working Group (2026) specifies a repeatable Rights property with optional identifiers and other qualifiers. These are different systems, but each makes it possible to represent more detail than a single work-level yes-or-no column.

The contrast can be expressed using a relational representation. In Chen's (1976) entity–relationship terminology, the work has a one-to-many relationship with the license entries in its public deposit. A relation of observed statements can be formed with DOI, individual entry position, URL, and role. A same-entry filter selects rows by both URL and role before taking distinct DOIs. A document-level query intersects two sets of work identities after each field is selected independently. Codd's (1970) relational model provides the operations for both procedures. Dropping the entry position, then rebuilding the record by joining separate URL and role lists on DOI, can generate their cross-product. A work with several entries can therefore acquire apparent combinations without acquiring new information.

Different published comparisons motivate evaluating the association explicitly. Piwowar et al. (2018) operationalized several types of open access using an article-level access data source. Akbaritabar and Stahlschmidt (2019) noted that more than one license can occur in Crossref and reported disagreements over access status derived from different inventories. Jahn (2026) examined how data sources influence estimates of hybrid open access in an assessment of transformative agreements. Our comparison holds the public provider fixed. It examines the association encoded by two fields before they are interpreted or mapped to an open-access category.

Data and matched comparison

The study includes all Crossref works in four selected February 2024 journal frames satisfying the API's publication-date and journal-article filters. The frames are ESMO Open, Scientific Data, PLOS ONE, and PeerJ Computer Science, located by the four ISSNs in the receipt. The month includes the illustrative DOI. The journal sample gives a contrast between deposited records with one license entry and records with multiple entries. It is a descriptive evaluation of these frames, and the venue-specific structure prevents treating their relative sizes or deposit conventions as representative. Cross-provider comparison studies measure other questions, such as overlap in indexed publication sets or availability of fields given an indexed publication (Visser et al., 2021; Delgado-Quirós & Ortega, 2024). Full capture within each stated frame allows a direct comparison without separately estimating provider coverage.

We collected each journal frame through its DOI and license-array projection. We then retrieved a single pooled works frame using the same month, document type and four ISSN restrictions. The pooled set was checked against the union of the four journal lists. Results were paginated through the documented cursor and checked for complete retrieval to the displayed total and for unique DOI keys. DOI strings were compared case-insensitively. There was no inference of missing identities from titles, authors or edit distance. This decision avoids the uncertainty of probabilistic record linkage discussed by Christen (2012), although it necessarily constrains all measurements to identities exposed by the provider.

Each license object contributes its URL and content-version code as displayed by the API. The JSON form keeps these fields associated inside each object, consistent with the distinction between objects and arrays in the exchange format (Bray, 2017). The five URL strings include the three present on the illustrative DOI and both observed URI schemes for the standard Creative Commons Attribution 4.0 path. The two role strings are vor and tdm. We used exact field equality throughout, so a literal scheme, path or trailing-slash change can alter the query group. The general URI specification distinguishes scheme and path and provides normalization rules for other tasks (Berners-Lee et al., 2005). For this experiment an

HTTP-to-HTTPS rewrite would obscure whether the public filter matches the strings the API exposes. We supplied the strings as filter values without requesting any of the referenced license pages.

The combined filter, DOI-level intersection, and same-entry test are three parallel computations over the same work frame. For URL u, role v, and the license array L_d for DOI d, the document predicate accepts d if some item in L_d has URL u and some item in L_d has role v. The same-entry predicate accepts d if there exists one item with both. Every same-entry match meets the document predicate. A discrepancy can therefore be counted as a DOI returned by the API that lacks a corresponding pair in the original array. The API need not expose its internal query plan for these observable outcomes to be compared.

All five URL filters and both role filters were also run separately. These controls distinguish broadened matching in a single filter from ambiguity created when two valid work-level sets are intersected. Each independent control was compared with the analogous directly observed DOI set. The main unit is one match for a DOI for each tested pair of exact predicates. Repeated license rows cannot increase this count. A DOI returned for three distinct pair queries contributes three events, and the union of its query hits contributes only one work. We report both to show what aggregation changes. Lenzerini (2002) describes formally how the semantics of sources affect integrated queries, and Rahm and Bernstein (2001) distinguish matching levels that include record structure as well as the fields themselves. Here the same two fields are available to all methods, and their association is the structural distinction under test.

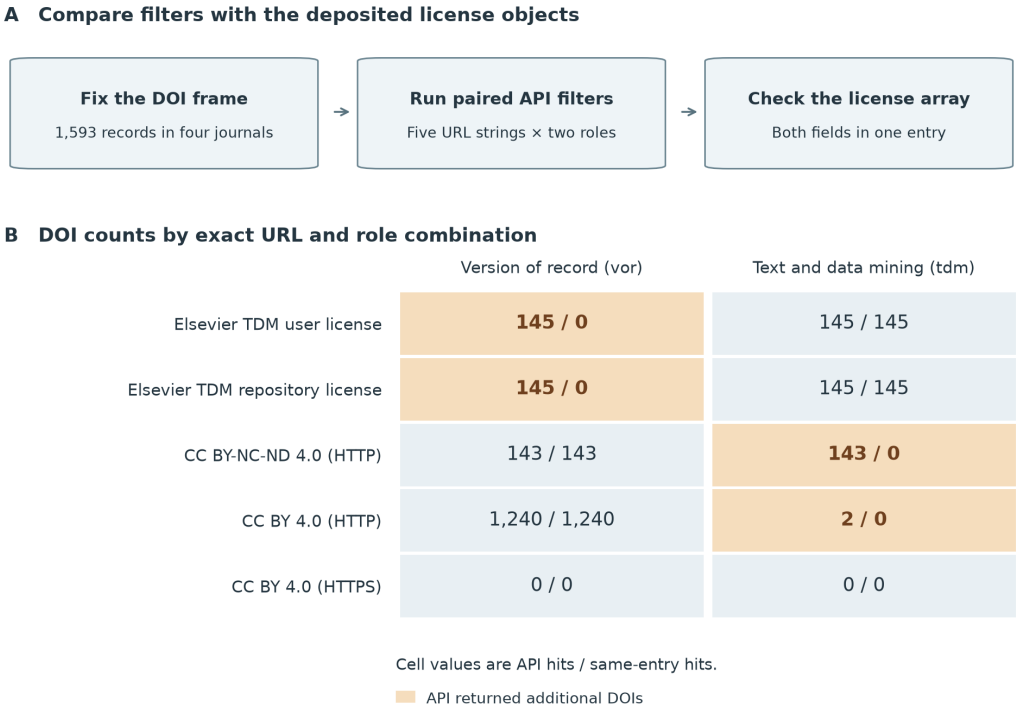

Figure 1. The study fixes a DOI set, observes ten paired searches, and checks the underlying license objects. The cells show numbers of Crossref API matches and same-entry matches, respectively. Highlighting marks a cell with additional API hits.

Results

The pooled base search and the four journal requests agreed on all 1,593 DOI records. The frame contained 145 ESMO Open, 108 Scientific Data, 1,275 PLOS ONE, and 65 PeerJ Computer Science records. Every DOI had at least one license entry. The first two journals supplied multiple entries for all their records, while the other two supplied one entry per record. The single-filter DOI sets agreed exactly with the sets reconstructed from the saved license arrays. In particular, 1,528 DOI records supplied a vor role, 253 supplied tdm, and 65 carried the tested HTTPS Creative Commons Attribution URL.

Table 1. DOI counts from the public API and direct same-entry verification

License URL label	Content role	API matches	Same-entry matches	Difference
Elsevier TDM user license	vor	145	0	145
Elsevier TDM user license	tdm	145	145	0
Elsevier TDM repository license	vor	145	0	145
Elsevier TDM repository license	tdm	145	145	0
CC BY-NC-ND 4.0 (HTTP)	vor	143	143	0
CC BY-NC-ND 4.0 (HTTP)	tdm	143	0	143
CC BY 4.0 (HTTP)	vor	1,240	1,240	0
CC BY 4.0 (HTTP)	tdm	2	0	2
CC BY 4.0 (HTTPS)	vor	0	0	0
CC BY 4.0 (HTTPS)	tdm	0	0	0
All tested DOI–query matches		2,108	1,673	435

The combined filters agreed with the DOI-level intersection in all ten comparisons. They returned 2,108 events, 435 of which lacked any entry with the requested URL and role together. That is 20.6% of returned events. Each Elsevier URL yielded 145 additional DOI matches when requested with vor. The CC BY-NC-ND URL yielded 143 additional matches when requested with tdm. Two further records accounted for the difference for the HTTP form of the CC BY URL with tdm. The two tested HTTPS pairs returned zero records because none of the 65 DOIs with that string had either tested role. Absence under a combined filter can therefore coexist with presence under the URL filter.

All 435 extra events came from the 145 ESMO Open records. Here the two URLs associated with tdm can each be crossed with the vor role, while the URL assigned to vor can be crossed with tdm. Each DOI therefore generated six hits but only three recorded associations for this grid. Independent single-DOI queries reproduced the six-versus-three split for the illustrative work. Of the 1,240 same-entry results for the HTTP CC BY URL and vor, 1,238 belonged to PLOS ONE and two to ESMO Open. Scientific Data supplied multi-entry records yet no hits for the five tested URL strings. Multiple entries make divergence possible. Its appearance in an observed query grid also requires the selected strings and roles to be present in a configuration that produces new crossed combinations.

For the work-level union, the exact distinction disappears. Across the grid, the API identified 1,383 DOIs with at least one combined-filter hit. Every one of those DOIs also had a verified same-entry match for at least one pair, including each of the 145 DOIs with three additional matches. Comparing joint-coverage

rates based only on unique DOIs would therefore miss the full 435-event gap. Equal numbers of unique works are compatible with differing statements about which roles were associated with a particular URL.

What requires a preserved association

Completeness assessment typically starts by checking whether a value has been supplied. Margaritopoulos et al. (2012) formalize metadata completeness and measures for its comparison. Park and Tosaka (2010) distinguish metadata quality criteria from the mechanisms that institutions use to support them, and de Jonge and Kramer (2026) examine how publisher submission workflows are associated with Crossref metadata completeness. The contributing records in our discrepancy supplied both pieces of evidence, so a basic presence check would pass. Bruce and Hillmann (2004) describe a continuum of metadata quality that includes how effectively information can be used. This comparison supplies an example at that level. Whether any code or URL exists is a different question from whether the published fields support their association.

For a reader selecting candidate articles, a query that gathers all records containing each requested property is a coherent work-level policy. For a role-specific evidence table, a row should retain the object from which both fields were extracted. Wang and Strong (1996) ground data quality in the needs of data consumers, and Stvilia et al. (2007) similarly relate assessment to activities and contexts. Search relevance and evidence of a URL–role binding are therefore distinct criteria for evaluating the same result. The public API view demonstrates consistent candidate retrieval across the selected pair queries and the separate predicates. The same hits support fewer direct assertions when each reported association needs the original object.

A downstream representation can preserve this distinction by storing a license entry as one row with its DOI and sourced properties, then taking a distinct projection after the compound test. If the evidence is exported as two independently deduplicated lists, a later analyst can still ask whether a work includes a given URL or role, but can no longer reconstruct their association from those lists alone. A retained response identifier or a reproducible receipt makes the row traceable to the source data. Moreau and Missier (2013) provide a conceptual basis for recording such derivation, and Leo et al. (2024) give a practical model for recording computational workflow provenance. The FAIR principles emphasize recording provenance and representing information for machine reuse (Wilkinson et al., 2016). In the narrow setting here, retaining an observed tuple is the part that makes a paired claim auditable.

The date, venues, tested roles and observed exact strings delimit the result. A different publisher deposit, additional role codes, or pooled years may alter the proportion of returned DOI–query matches lacking an object witness. Field normalization could answer a separate question about URL families. The observed exact-match behavior cannot settle that question and does not establish how future deposits or the search index will behave. We can characterize the public result during the capture, while the search implementation and deposit history remain unobserved. The array establishes which fields accompanied each other in the displayed metadata. The study measures neither access to any work nor the legal implications of license terms.

Data and materials

The data package records the exact Crossref receipt, the frame and probe lists, and the original URL–role pairs retained for analysis. Frozen data are in data/raw/license_pair_scope_1242/snapshot.json. Run scripts/analyse_license_pair_scope_1242.py and scripts/plot_license_pair_scope_1242.py for the results and the figure. These materials contain only public administrative license fields and DOI identities.

References

Akbaritabar, A.; Stahlschmidt, S. (2019). Merits and Limits: Applying open data to monitor open access publications in bibliometric databases. *arXiv*. https://arxiv.org/abs/1902.03937v2

Berners-Lee, T.; Fielding, R.; Masinter, L. (2005). Uniform Resource Identifier (URI): Generic Syntax. *RFC 3986*. https://doi.org/10.17487/rfc3986

Bray, T. (2017). The JavaScript Object Notation (JSON) Data Interchange Format. *RFC 8259*. https://doi.org/10.17487/rfc8259

Bruce, T. R.; Hillmann, D. I. (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. *Metadata in Practice*. https://hdl.handle.net/1813/7895

Chen, P. P. S. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems*, 1(1), 9–36. https://doi.org/10.1145/320434.320440

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). License information. *Crossref Documentation*. https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/license-information/

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

DataCite Metadata Working Group (2026). DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs, Version 4.7. *DataCite*. https://doi.org/10.14454/qdd3-ps68

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. https://doi.org/10.1016/j.esmoop.2024.102239

de Jonge, H.; Kramer, B. (2026). Manuscript submission systems and metadata completeness in Crossref: Patterns and associations. *PLOS One*, 21(3), e0345417. https://doi.org/10.1371/journal.pone.0345417

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

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

Iannella, R.; Villata, S. (2018). ODRL Information Model 2.2. *W3C Recommendation*. https://www.w3.org/TR/odrl-model/

Jahn, N. (2026). Estimating transformative agreement impact on hybrid open access: a comparative large-scale study using Scopus, Web of Science and open metadata. *Scientometrics*, 131(2), 901–924. https://doi.org/10.1007/s11192-025-05390-3

Lenzerini, M. (2002). Data Integration: A Theoretical Perspective. *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 233–246. https://doi.org/10.1145/543613.543644

Leo, S.; Crusoe, M. R.; Rodríguez-Navas, L.; Sirvent, R.; Kanitz, A.; De Geest, P.; Wittner, R.; Pireddu, L.; Garijo, D.; Fernández, J. M.; et al. (2024). Recording provenance of workflow runs with RO-Crate. *PLOS ONE*, 19(9), e0309210. https://doi.org/10.1371/journal.pone.0309210

Margaritopoulos, M.; Margaritopoulos, T.; Mavridis, I.; Manitsaris, A. (2012). Quantifying and measuring metadata completeness. *Journal of the American Society for Information Science and Technology*, 63(4), 724–737. https://doi.org/10.1002/asi.21706

Moreau, L.; Missier, P. (2013). PROV-DM: The PROV Data Model. *W3C Recommendation*. https://www.w3.org/TR/prov-dm/

National Information Standards Organization (2021a). Access & License Indicators (2021 Revision). *NISO RP-22-2021*. https://doi.org/10.3789/niso-rp-22-2021

National Information Standards Organization (2021b). JATS4R Permissions. *NISO RP-37-2021*. https://doi.org/10.3789/niso-rp-37-2021

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

Piwowar, H.; Priem, J.; Larivière, V.; Alperin, J. P.; Matthias, L.; Norlander, B.; Farley, A.; West, J.; Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6, e4375. https://doi.org/10.7717/peerj.4375

Rahm, E.; Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4), 334–350. https://doi.org/10.1007/s007780100057

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

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