Advances in Laboratory Automation
Vol. 1 No. 1 (2026)

A Role Tag Can Describe an Editor or a Research Contribution

Ellen Cook · Nathan Hughes · Scott L. Parker

Published: November 26, 2026 | Article 1214

Abstract

An XML field called role can carry an editor's title or a standardized statement of contribution. We inspected the public PubMed Central front matter for 72 exact DOI records. Forty-nine records had at least one role node, and 32 carried CRediT activity terms. All 862 recognized activity nodes used the older content-type URI convention. A reader restricted to the newer vocabulary attributes would have recovered none. Canonical display text recovered the same 862 nodes in this sample, while counting every role would also have selected 62 editorial titles. Among 199 contributors with recorded CRediT activity, 176 carried several roles. Both the way an activity is recognized and the way its relationship is stored affect what an export can preserve.

A Role Tag Can Describe an Editor or a Research Contribution

Abstract

An XML field called role can carry an editor's title or a standardized statement of contribution. We inspected the public PubMed Central front matter for 72 exact DOI records. Forty-nine records had at least one role node, and 32 carried CRediT activity terms. All 862 recognized activity nodes used the older content-type URI convention. A reader restricted to the newer vocabulary attributes would have recovered none. Canonical display text recovered the same 862 nodes in this sample, while counting every role would also have selected 62 editorial titles. Among 199 contributors with recorded CRediT activity, 176 carried several roles. Both the way an activity is recognized and the way its relationship is stored affect what an export can preserve.

1. The ambiguity sits inside a familiar tag

A publisher can describe what contributors did, identify a reviewing editor, or record a publishing role in structured XML. An archive aggregator that translates all role elements into the same contributor-activity field may make these statements difficult to distinguish. This matters when downstream users count research contributions by category. If a role tag states that someone participated as a senior editor, that is useful administrative information, with different semantics from a standardized investigation or writing attribution. The tag name alone leaves the distinction unresolved.

The Contributor Roles Taxonomy, known as CRediT, supplies a controlled set of fourteen activities and lets a contributor receive several roles on one paper (National Information Standards Organization, 2022). Allen et al. (2014) and Brand et al. (2015) describe the need for more explicit recognition of contributions than a conventional author list can supply. The International Committee of Medical Journal Editors (n.d.) treats authorship and attribution as related but separate responsibilities. Public XML likewise can preserve author or editor status independently from the specific activities declared for that person. National Library of Medicine (2024) describes the role element as a place for either a function title or a term tied to a recognized vocabulary. An extractor therefore has a concrete semantic test to perform before aggregating contributions.

A publisher XML guide may also support more than one way to encode the same taxonomy. JATS4R (n.d.) documents both the older content-type convention and the newer vocab attributes. Taylor & Francis (2025) describes how free-text contribution statements and coded role tags can coexist in production. Crossref (2026) now documents CRediT support within its contributor schema, and DataCite (n.d.) describes separate enumerated contributor types in its own metadata model. These examples show that knowing a person is an author or editor and knowing which standardized activities were recorded are distinct questions. The results here concern public archive fronts under one acquisition profile. No attempt was made to derive missing activities from publication prose or infer them from who appears in the byline.

The exact public archive record of Yi et al. (2024), DOI 10.1016/j.esmoop.2024.102239, provides a separate administrative illustration. Its PMC front matter has eight explicitly typed author entries and no direct role elements on those entries. An export can retain all eight types while returning no structured author activity from the queried path. The record does not establish that an author contribution statement is absent elsewhere. We take it as an example of representation scope. It is outside the 72-record comparison rather than a count of absent activity in the main cohort.

We ask how often a generic role reader and a CRediT-aware reader differ at the article and role-node level, how much a parser restricted to modern vocabulary attributes retains, and what a reader would lose by storing only one activity per contributor. The same front records feed all of the compared rules. This controls for work identification and separates differences in field interpretation from differences in archive coverage.

2. The field as a unit of observation

Contributorship studies have used explicit disclosures to examine the division of labor. Larivière et al. (2016) analyze how activities are distributed among coauthors. Sauermann and Haeussler (2017) investigate what author contribution disclosures can reveal, and Haeussler and Sauermann (2020) examine team size and interdisciplinarity in relation to those divisions. Yang et al. (2025) use role categories in data papers to study the work behind data-intensive publishing. Lu et al. (2022) consider how contribution and byline order can relate. The measurements in these studies depend on the visibility and unit of the recorded contribution. An export that preserves only one role per contributor reduces the data before an analyst can decide how to interpret it.

Implementation studies offer complementary perspectives. Holcombe et al. (2020) describe recording contribution information in a collection tool that produces structured statements. Baum et al. (2023) consider ways to improve visibility and transparency in author credit. Godskesen et al. (2026) examine practical adoption and barriers, and Partin et al. (2026) describe the taxonomy's use as a shared vocabulary in assessing authorship disputes. The CRediT Matrix (n.d.) census distinguishes several markup and statement routes when tracking public adoption. This audit is narrower. It takes the same exact DOI front matter and tests how matched parsers would count nodes, contributors and works. It separately recognizes editorial titles under the same XML element. It does not estimate global adoption or attempt to reconstruct contribution roles across all document sections.

The analysis retains no contributor name, organization, contact information, prose or scientific content. A contributor slot is a single observed XML node within a designated article. More than one role can attach to the slot. The same label attached to different slots counts separately. Counting repeat roles within a slot is an operational issue with consequences for data portability. Focus on the node and its recognized vocabulary fields permits direct calculation of duplication and multiplicity without identifying who supplied any role.

3. Frame, parsing and outcomes

3.1. Fixing the article identity

We used four journal frames originally collected from PubMed with publication year 2024 and exact ISSN search predicates. The archived queries returned the first 100 records under the service's publication-date sort for each journal. These are bounded retrieval frames, and their selection is neither a census of 2024 nor a random sample of the full journal archives. PubMed E-utilities provide access to machine-readable metadata in the NCBI infrastructure described by Sayers et al. (2024). From each archived frame, we retained eligible pairs of a DOI and PMC accession. We ordered those pairs by a salted SHA-256 score of the DOI and took 18 from each of PLOS ONE, PLOS Biology, eLife and PNAS. PLOS (n.d.) documents an explicit contribution taxonomy in its author guidelines. The remaining journals support comparison with different observed front matter configurations. Equal source allocation bounds the workload and makes each stratum equally visible in the record-level comparison.

Each chosen PMC accession was queried through the public PMC Open Archives Initiative front-matter interface (National Library of Medicine, n.d.). Exact DOI equality was required between the stored

PubMed pair and the PMC article metadata. DOI comparison normalized letter case and surrounding whitespace. No title or person identifier was needed. Visser et al. (2021) compare the coverage of different scholarly databases, and Delgado-Quirós and Ortega (2024) distinguish publication presence from the availability of individual metadata fields. We likewise separate having a usable front record from having a role node on that record. The record and contributor denominators were fixed before parsing roles, making omission at one stage distinguishable from missing document access.

3.2. Testing readers on the same contributor paths

We retained direct contributor children of the front-matter contributor groups, recording declared type and a compact representation of the direct role elements on each node. We counted a group-level role separately if it existed rather than attributing it to individual descendants. The generic reader selects any such direct role element. A CRediT-aware reader applies three nested checks. It first recognizes a valid modern term URI or an explicitly declared CRediT vocabulary with a recognized term. It then permits the older content-type attribute if the value is a recognized CRediT role URI in an allowed namespace. Finally, it can recognize the exact canonical activity label after folding case, whitespace and punctuation. An exact term URI and a canonical display label can be compared independently. The validation uses the finite documented taxonomy, which means an arbitrary editorial function title remains outside the activity counts.

This design yields several distinct readouts. We count a work with any selected node, a work with at least one recognized CRediT activity, and a work that has direct role nodes but none recognized under the taxonomy. At the node level, we tally how many each rule selects and compare coded activity with canonical display text. A label reader can recognize a meaningful string without exposing its expressly declared vocabulary. Keeping label evidence separate avoids granting the same provenance to a string match and a URI match. We also distinguish the tests from changing the field. Other public places where contributor roles might be disclosed, including acknowledgments or written contribution sections, are outside the fixed front-matter path.

The final analysis counts distinct recognized activities within each contributor slot. A single-valued export could choose any one of the activities in a slot; the number of retained slot–activity pairs is the same no matter which chosen value it keeps. The contrast estimates multiplicity that would be erased under that storage rule. It does not rank the importance of activities or assume that lists supplied by publishers are complete. Park and Tosaka (2010) describe metadata quality controls for digital repositories, and Batini et al. (2009) discuss how data-quality assessments depend on explicit units and measurable criteria. Repeated or specialized values therefore belong in the outcome definition rather than being silently removed by a general export convention.

4. Results

4.1. Identical article frame, different activity inventory

All 72 requests supplied exact DOI-aligned front matter. The records had 590 contributor nodes, including 528 explicitly typed author nodes. Forty-nine records displayed at least one direct role element. Thirty-two of those records contained at least one recognized CRediT activity. Seventeen records had direct roles made up entirely of editorial titles. A generic tag-presence rule would thus identify 49 of 72 records as carriers of activity, while restricting the claim to recognized CRediT terms gives 32 of 72. Among records with role elements, 17 of 49, or 34.7%, would make that difference consequential. The roleless comparison frame and the title-only subset are retained in the denominator.

Table 1. *Direct contributor roles and recognized CRediT activities in the DOI-aligned PMC front matter.*

Journal	DOI-aligned records	Records with any direct role	Records with CRediT	Direct role nodes	CRediT nodes	Editorial title nodes	Slots with multiple CRediT activities
PLOS ONE	18	18	18	540	522	18	102
PLOS Biology	18	14	14	350	340	10	74
eLife	18	17	0	34	0	34	0
PNAS	18	0	0	0	0	0	0
All four journals	72	49	32	924	862	62	176

Note. The table counts roles attached directly to the listed contributor nodes at the captured PMC path. No group-level role nodes were observed. Records with editorial roles can also carry recognized CRediT, so the record columns should be read as nested rather than added together.

The 924 role nodes consisted of 862 CRediT activities and 62 editorial titles. All recognized activities belonged to explicitly typed author nodes and all editorial titles belonged to editor nodes. A filter that required authorship would therefore happen to exclude the 62 titles in this sample. Its success depends on the format and category mix. An author-type indicator alone cannot establish which research activity an author carried out, and in the two comparison strata with no author activity tags it supplies no activity details at all. The general vocabulary-aware parser retains and distinguishes both categories without requiring a title to serve as positive evidence of research activity.

4.2. Recognizing older metadata and preserving several activities

The modern-attribute reader recognized zero CRediT activities. The legacy content-type URI reader recognized all 862. Canonical display text likewise recognized 862, and its normalized activity agreed with the URI category on every one of those nodes. No additional label-only recognized activity was observed. The text-only parser therefore reproduces the activity counts here. It does not preserve the explicit URI evidence, and its agreement on this archive slice supplies no accuracy estimate for other publishers or noncanonical display labels. On these observations, fallback compatibility with older tagged metadata makes the difference between a fully empty and a fully recovered set of observed codes.

The recognized activity inventory spans all fourteen taxonomy categories. Within contributor slots, there were 862 distinct activity assignments because no duplicate recognized activity was found in the same slot. They were attached to 199 explicitly typed author slots. Of those, 176, or 88.4%, had more than one activity. A one-value storage rule would retain only 199 assignments and remove the other 663, or 76.9% of the observed slot-level assignments. The writing review and editing category occurred on 152 slots, while methodology appeared on 95. Because the same slots can participate in both categories, reporting their frequencies from a single-role export would introduce a dependency on how that export chooses among eligible values.

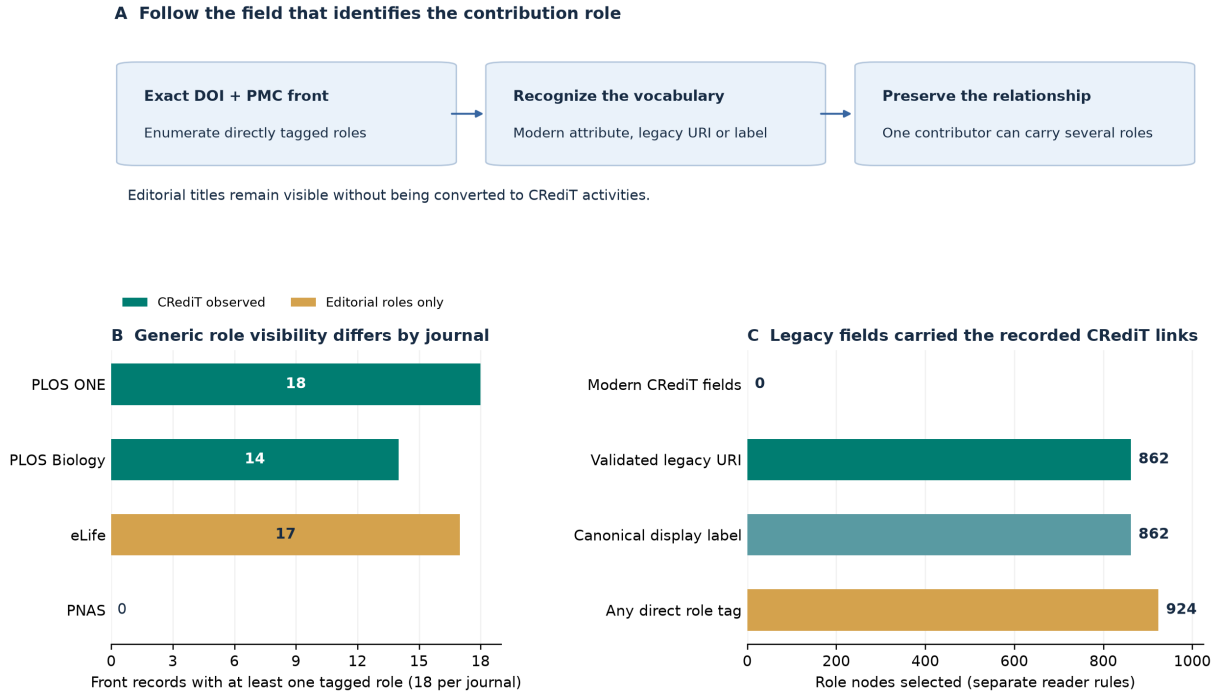

Figure 1. *Panel A shows the path from exact archive matching to activity recognition and multi-valued storage. Panel B distinguishes works with recognized CRediT from works carrying direct editorial role tags only. Panel C compares separate reader rules on the same role-node inventory.*

5. Interpretation for public contribution exports

This audit separates completeness of the document response from completeness of a contribution extract. Every chosen DOI had a usable archive front. Only some fronts had structured activity at the queried path. A reader choosing only newer vocabulary attributes would assign zeros even within two strata that explicitly coded hundreds of activity occurrences under an older convention. Conversely, a broad reader would associate editorial titles with the same field if it only checked that a role node existed. The two errors operate in opposite directions. Recovering more XML elements does not necessarily recover more observations of the desired scholarly activity.

The paired agreement with display text supports a focused distinction between reconstructing a category and retaining how that category was asserted. A recognized URI establishes that the deposited form pointed to a vocabulary term. A matching display string can recover an interpretable category for a reader even in an export stripped of attributes. A description of an activity in a local production vocabulary can be intelligible with no claim of standardization. The categories that the code recognizes, the places at which it looks for evidence, and whether it stores multiple values are all extract properties that can be documented independently of publication practice.

The record and slot contrasts are defined on the sampled archive configurations. They do not estimate whether journals use the taxonomy elsewhere, generalize to publisher practice over time, or rank implementations. Different records can expose other locations for legitimate activity statements. Both the CRediT census and the archive documentation identify routes beyond the one measured here. A companion study using additional structured formats could compare the new Crossref schema and full-text XML paths with the same linked DOI set. The current scope is enough to show how strict recognition and a one-role data model change the results even before any cross-platform or content comparison is undertaken.

The public extracts document the archive paths, retrieval times and sanitized field evidence. Wilkinson et al. (2016) emphasize metadata that can continue to be interpreted and reused over time. Peng (2011) distinguishes computational replay from a new study performed on a different sample. Gebru et al. (2021) formalize the benefit of describing collection and selection conditions for downstream users. For this audit a future reader can regenerate the counts from the fixed field projection. It can also recompute a different export rule and state clearly which sense of role survived that transformation.

Data and code availability

The reproduction record links the public front-matter receipts, coded role observations, analysis, figure and reference metadata. Author contributions, funding and competing-interest declarations await author confirmation.

References

Allen, Liz; Scott, Jo; Brand, Amy; Hlava, Marjorie; Altman, Micah (2014). Publishing: Credit where credit is due. *Nature, 508(7496), 312–313*. DOI.

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

Baum, Myriam A.; Braun, Moritz N.; Hart, Alexander; Huffer, Véronique I.; Meßmer, Julia A.; Weigl, Michael; Wennerhold, Lasse (2023). The first author takes it all? Solutions for crediting authors more visibly, transparently, and free of bias. *British Journal of Social Psychology, 62(4), 1605–1620*. DOI.

Brand, Amy; Allen, Liz; Altman, Micah; Hlava, Marjorie; Scott, Jo (2015). Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing, 28(2), 151–155*. DOI.

CRediT Matrix (n.d.). CRediT census. *CRediT Matrix*. Source.

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

DataCite (n.d.). Contributor. *DataCite Metadata Schema*. 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.

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.

Godskesen, Tove; Helgesson, Gert; Eriksson, Stefan (2026). Implementation, barriers, and improvement strategies for CRediT: A scoping review. *Accountability in Research, 33(4), 2528953*. DOI.

Haeussler, Carolin; Sauermann, Henry (2020). Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity. *Research Policy, 49(6), 103987*. DOI.

Holcombe, Alex O.; Kovacs, Marton; Aust, Frederik; Aczel, Balazs (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLOS ONE, 15(12), e0244611*. DOI.

International Committee of Medical Journal Editors (n.d.). Defining the Role of Authors and Contributors. *ICMJE Recommendations*. Source.

JATS4R (n.d.). CRediT taxonomy. *JATS for Reuse*. Source.

Larivière, Vincent; Desrochers, Nadine; Macaluso, Benoît; Mongeon, Philippe; Paul-Hus, Adèle; Sugimoto, Cassidy R (2016). Contributorship and division of labor in knowledge production. *Social Studies of Science, 46(3), 417–435*. DOI.

Lu, Chao; Zhang, Chenwei; Xiao, Chengrui; Ding, Ying (2022). Contributorship in scientific collaborations: The perspective of contribution-based byline orders. *Information Processing & Management, 59(3), 102944*. DOI.

National Information Standards Organization (2022). CRediT, Contributor Roles Taxonomy. *ANSI/NISO Z39.104-2022*. Source.

National Library of Medicine (2024). Role or Function Title of Contributor. *NISO JATS Journal Publishing Tag Library, version 1.4*. Source.

National Library of Medicine (n.d.). PMC Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) API. *PubMed Central documentation*. Source.

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.

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.

Partin, Kathryn; Hosseini, Mohammad (2026). Using the contributor role taxonomy (CRediT) as a tool in resolving authorship disputes at the NIH. *Accountability in Research, 33(5), 2596063*. DOI.

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

PLOS (n.d.). Authorship. *PLOS ONE author guidelines*. Source.

Sauermann, Henry; Haeussler, Carolin (2017). Authorship and contribution disclosures. *Science Advances, 3(11), e1700404*. DOI.

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.

Taylor & Francis (2025). Author Contributions and CRediT. *Taylor & Francis JATS Guide*. Source.

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.

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.

Yang, Heng; Yu, Yonglin; Liu, Fenghong (2025). How do data authors perform in data-intensive research activities? Evidence from author contribution statement in data papers. *Information Research an international electronic journal, 30(iConf), 219–239*. DOI.