BRAINIACS
BRAIN IMAGING AND COMPUTING SCIENCES

2023 Volume 4
Issue 1 Edoc X934EF46B

# Managing Bibliometrics with Contributor Roles and Literature Provenance for NPDS Metadata Records[*]

Adam Craig and Carl Taswell[†]

## Abstract

Tracking the provenance of information and knowledge with sources, concepts, ideas and the contributions of creators, inventors, and scholars remains essential to the reproducibility of scientific results, the reliability of engineering methods, the integrity of clinical trials, and the fair allocation of research and development resources. Since the first version, the PORTAL-DOORS Project (PDP) has provided schemas for the Nexus-PORTAL-DOORS-Scribe (NPDS) cyberinfrastructure which have always supported inclusion of provenance data as part of the metadata describing the entity corresponding to an NPDS record. However, PDP has not yet provided alternative templated formats for the provenance infosubset (aka, provenance facet) of the entire infoset that documents the entity for the identified NPDS record. Therefore, this report introduces a template for the provenance facet suitable for a generic entity type, along with alternatives appropriate for various specific entity types, including a template for authored publications that enables descriptions of contributor roles compatible with the CRediT taxonomy. By extending the PDP-DREAM ontology and creating a provenance subontology for the NPDS cyberinfrastructure with semantic classes and properties for these contributor roles, semantic search in NPDS repositories can be facilitated to support more effective discovery of resources with consideration of entities as same, similar, related, or different.

## Keywords

Bibliometrics, contributor roles, literature provenance, scholarly communications, data repositories, knowledge engineering, semantic web.

## Contents

## Introduction

The Nexus-PORTAL-DOORS-Scribe (NPDS) Cyberinfrastructure specifies a message-level protocol and web API for decentralized management of metadata for heterogeneous resources according to the hierarchically distributed mobile metadata (HDMM) architectural style also implemented by the IRIS-DNS system (Taswell 2010). NPDS facilitates grouping metadata records by problem domain in any of the three repository types: lexical metadata-oriented Problem-Oriented Registries of Tags And Labels (PORTAL), semantic description-oriented Domain Ontology-Oriented Resource System (DOORS) directories, and hybrid Nexus diristries (Craig, Bae, et al. 2016). Whereas these three server types provide read-only access for both manual and automated retrieval of records, Scribe registrars provide read-write access for curation of records (Craig, Bae, et al. 2016).

Since its earliest version, NPDS has supported inclusion of provenance among the semantic facets of a record (Taswell 2007). In the interest of supporting a wide variety of users with diverse needs, we have not previously specified a format for it. However, in order to encourage user uptake and promote interoperability, we here provide a recommendation for how to format record provenance for works of scholarly literature using the provenance module of the Portal Doors Project Discoverable Data with Reproducible Results for Equivalent Entities with Accessible Attributes and Manageable Metadata (PDP-DREAM) formal ontology. The PDP-DREAM Ontology provides a formal ontology that includes classes and properties useful for creating semantic markup to be embedded in a DOORS or Nexus record (Craig and Taswell 2021). The DREAM part of the name refers to the DREAM Principles that guide the PORTAL-DOORS project (Dutta et al. 2019).

For our purposes, *provenance* does not refer to the relatively narrow sense of why, how, and where-from input data in a database influenced the output of a query as reviewed in (Cheney et al. 2009), but rather to the broader concept of who (or what automated agent) produced which version of an artifact by what process, as defined in (Moreau et al. 2010). In prior work, we discussed how such a notion of provenance could apply to cultural artifacts, for which the emphasis is on physical artifacts rather than digital ones (Athreya et al. 2021). Here, we focus on representing the provenance of scholarly literature, where the key artifacts are the specific versions of a document, digital or physical, and the roles people and software agents created in producing it.

In the context of a written work reporting scholarly research, this concept of provenance includes how individual coauthors and others contributed to the origination of ideas, generation and analysis of data, drafting of the manuscript, and other steps that shaped the ultimate

[*] Document received 2023-05-14, published 2023-06-25.

[†] Authors affiliated with Brain Health Alliance Virtual Institute, Ladera Ranch, CA 92694 USA; correspondence to CTaswell at Brain Health Alliance.

form of the document. With this in mind, we also set out to support representation of contributor roles and to make our approach compatible with the Contributor Roles Taxonomy (CRediT), which provides a controlled vocabulary for describing contributions of both authors and non-authors (Holcombe 2019; Brand et al. 2015). CRediT is one of several Contributor Role Ontologies and Taxonomies (CROTs), along with the Taxonomy of Digital Research Activities in the Humanities, but is notable for having attracted attention from a broad spectrum of disciplines and has become a National Information Standards Organization standard (at credit.niso.org) (Hosseini, Colomb, et al. 2022). One panel of experts with the National Science Foundation has called for journals to require that all contributing authors affirm their contributions as described using this taxonomy in a statement to be published along with the article metadata in order to discourage ghost writing, gift authorship, orphan authorship, and forged authorship (McNutt et al. 2018).

## Methods

We modeled the core classes for representing provenance in the PDP-DREAM Provenance Ontology on the Open Provenance Model (Moreau et al. 2011). This model includes three core classes of entity: artifacts, agents, and processes (Moreau et al. 2011). An instance of a class can relate to an instance of another class in one of five ways: A process can have used an artifact. An artifact can have been generated by a process. An process can have been controlled by an agent. A process can have been triggered by a prior process. An artifact can have been derived from another artifact. By design, these relationships are all in the past tense to indicate that the model is only for describing the past provenance of an artifact, not some potential future events (Moreau et al. 2011). We incorporated this model directly into the PDP-DREAM Provenance ontology by including OWL classes *Artifact*, *Agent*, and *Proccess*, and OWL object properties *used*, *wasGeneratedBy*, *wasControlledBy*, *wasTriggeredBy*, and *wasDerivedFrom* directly from the OPM specification (v 1.1 (Moreau et al. 2011)). We then added subclasses to make it clearer how this model can apply to scholarly literature, making *DocumentVersion* a subclass of *Artifact*, *Person* a subclass of *Agent*, and *Editing* a subclass of *Process*.

We intentionally made *DocumentVersion* a class separate from the preexisting class *Document*. In the OPM specification, an artifact must be an "immutable piece of state, which may have a physical embodiment in a physical object, or a digital representation in a computer system" (Moreau et al. 2011). A *DocumentVersion*, whether embodied as the exact sequence of bits in a file or the exact arrangement of ink on a page, satisfies this definition, while a *Document*, which may be a living document encompassing multiple versions does not.

As an additional tool for tracking document state, we extended the PDP-DREAM Provenance ontology to allow for recording of a hash text for each document version, along with the hashing algorithm used. To do this, we added the following: *Hash* and *HashSalt* as subclasses of *Artifact*, *HashingAlgorithm* as a subclass of *Agent*, and *Hashing* as a subclass of *Process*. Ideally, the user could specify an arbitrarily long list of documents to concatenate together before computing a hash, but representing lists in RDF tends to add significant complexity (Daga et al. 2019). Since our target use case is inclusion of three items, the clear text of the current version, the hash of the previous version, and a salt, we selected the simple, satisficing approach of creating three sub-properties of *used*: *usedFirst*, *usedSecond*, and *usedThird*. Including the hash of the hash from the previous version allows us to create a blockchain-like sequence of hashes associated with the document version data, helping to identify cases of incorrect tracking of document versions. Use of blockchain to ensure document integrity has received much attention in the biomedical field but mostly in the context of tracking Electronic Health Records and Personal Health Records (Hasselgren et al. 2020). We contend that it can also help publishers, editors, and authors to maintain the accuracy of the scientific record.

In order to express that a coauthor (Person) contributed in a particular way to a work (Publication), we define the object property *contributedTo* and a set of sub-properties, each corresponding to a role in the CRediT taxonomy (See 1). This differs from the approach taken in the CRO (2019-12-11 Release, retrieved from github.com/-data2health/contributor-role-ontology on 2023-03-14), another formal ontology designed to be compatible with the CRediT taxonomy, which represents roles as classes (Vasilevsky et al. 2021). This means that attributing a role to an author in the context of a particular publication in accordance with the Contributor Attribution Model (retrieved from contributor-attribution-model.readthedocs.io/en/latest/ on 2023-03-14) requires at least two triples to link the instance of a role to the contributor and to the document for which the contributor occupied that role. Representing roles as object properties as we have allows us to link a contributor to a work concisely via the role in a single RDF triple.

Furthermore, whereas the CRO has 32 sub-classes directly under *contributor role*, seemingly calling for a more fine-grained approach to parsing author roles than that found in the CRediT taxonomy, we considered possible cases where a work had a small number of contributors, each of whom participated in multiple capacities. In such situations, the roles described in the CRediT taxonomy may be more fine-grained than necessary. As a way of making contributor role statements more concise, we organized them into a hierarchy (See 1). In some cases, we observed that some roles in the taxonomy are specific components of other roles. For instance, formal analysis, not just wet-lab or other physical experimentation, is part of investigation. In other cases, we introduced a new term that could cover two closely related terms. Including a single item for writing to cover the case where one author wrote the original draft and all subsequent drafts was particularly straight-forward, and the comment in the subsection Results, Issue 1 of (Hosseini, Gordijn, et al. 2023) regarding randomized controlled trials in dermatology supports the utility of such a simplification. The choice to group together supervision and project administration under project management, conceptualization and methodology under project design, and both project management and project design under project leadership arose from the the authors' past experience in which all of these are common activities for a principal investigator.

## Results

One example of a paper published with a CRediT statement is (Craig, Yücel, et al. 2022). Its text, as published on the Elsevier website, reads, "Adam Craig: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Visualization. Mesut Yücel: Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft. Lev Muchnik: Conceptualization, Investigation, Writing – review & editing. Uri Hershberg: Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition." We can render this example CRediT statement as an RDF XML document using the PDP-DREAM Provenance ontology. This document is available as CRediTCraig2022IFSEAGFSDBN.xml in

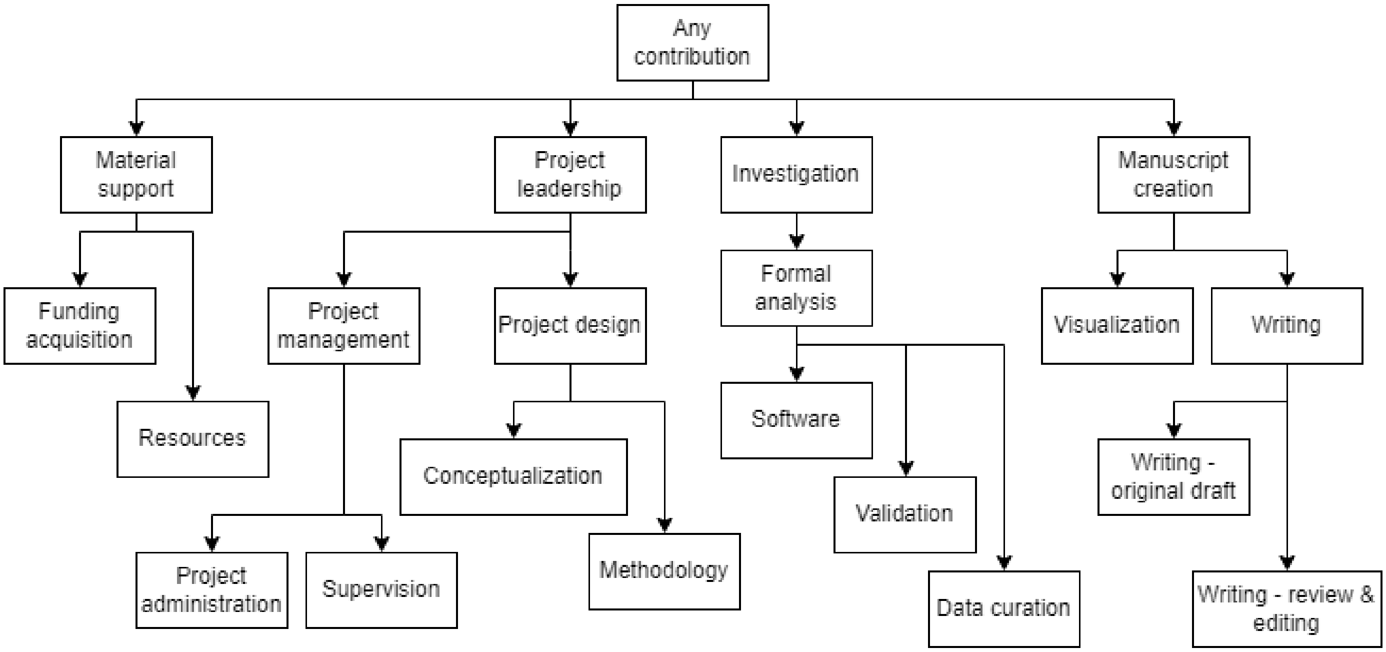

Figure 1: Hierarchy of contributor roles in the PDP-DREAM Provenance Ontology.

Table 1: Correspondence between roles in the CRediT taxonomy and object properties in the PDP-DREAM Provenance ontology

| CRediT role | PDP-DREAM Property |
|---|---|
| Conceptualization | contributedConceptualizationTo |
| Methodology | contributedMethodologyTo |
| Software | contributedSoftwareTo |
| Validation | contributedValidationTo |
| Formal Analysis | contributedFormalAnalysisTo |
| Investigation | contributedInvestigationTo |
| Resources | contributedResourcesTo |
| Data Curation | contributedDataCurationTo |
| Writing – Original Draft | contributedWritingOriginalDraftTo |
| Writing – Review & Editing | contributedWritingReviewEditingTo |
| Visualization | contributedVisualizationTo |
| Supervision | contributedSupervisionTo |
| Project Administration | contributedProjectAdministrationTo |
| Funding Acquisition | contributedFundingAcquisitionTo |

the online supplementary materials. This example demonstrates that, while it is straight-forward to use orcid.org URLs as identifiers for coauthors, other URLs or URIs can serve just as well when a coauthor does not have an ORCID. In this case, one of the coauthors went into industry after completing his graduate program and had no interest in obtaining one. Similarly, one can use DOIs as identifiers for publications or use any other URL or URI, including ones registered as entity labels with a Nexus diristry or PORTAL registry. The PDP Provenance ontology itself is available as an OWL 2.0 XML file in CRLP.owl in the supplementary files.

## Discussion

By implementing the provenance facet of NPDS records as an RDF description, we enable semantic search of resource records retrieved from Nexus diristries and DOORS directories. While the current implementation of PDP software does not implement semantic search, requiring a third-party SPARQL engine or similar utility, we hope to improve on this with a user-friendly semantic search interface soon.

Future editions of Brain Health Alliance's Brainiacs Journal will include CRediT roles for all contributors, though they will have the option to use other unique identifiers instead of ORCIDs and DOIs. In particular, because it is possible to register an ORCID or DOI as a cross-reference in a PORTAL or Nexus metadata record (Taswell 2008), making lookup of the ORCID or DOI using a resolvable PDP entity label straight-forward, using one provides the best of both worlds. This functionality will be especially useful when extending contributor roles to datasets, which do not necessarily have DOIs and may have contributors that are organizations rather than individuals and thus do not have ORCIDs, as discussed in (Hosseini, Gordijn, et al. 2023).

One possible future direction for journals is to address and formalize the relationship between contributorship and authorship. The CRediT taxonomy explicitly and by design side-steps the issue by taking an inclusive approach calling for the listing anyone who contributes to a research effort, regardless of whether they appear in the byline (McNutt et al. 2018). If those who submit manuscripts comply scrupulously, this may lead to a more accurate record of the knowledge creation process, but it will not by itself resolve the controversies surrounding authorship. As discussed in (Hosseini, Gordijn, et al. 2023), claims that CROTs will help to resolve disputes over authorship still lack empirical evidence, and the entrenched systems of incentives that reward, in some cases directly with cash, being first, corresponding, or senior author on an article published in a high-impact-factor journal will not disappear overnight merely because of the addition of a second, parallel way of describing involvement. While Brainiacs Journal is unlikely to be at the center of such a controversy any time soon, it can serve as a testbed and working example of new approaches, such as publishing a set of rules for which contributors qualify as coauthors and which should receive acknowledgments in the articles it publishes and encoding these rules in the PDP-DREAM Provenance ontology. While this is beyond the scope of the present work, we welcome feedback from the community on this subject.

## Conclusion

Accurately recording and communicating the contributions of researchers is important in order to maintain a just and efficient scholarly community. This feature added to the functionality of the NPDS Cyberinfrastructure enhances its usefulness in managing and disseminating such records.

## Citation

Brainiacs 2023 Volume 4 Issue 1 Edoc X934EF46B, Pages 1 – 4
Title: “Managing Bibliometrics with Contributor Roles and Literature Provenance for NPDS Metadata Records”
Authors: Adam Craig and Carl Taswell
Dates: created 2023-03-08, received 2023-05-14, updated 2023-06-25, published 2023-06-25, endorsed 2023-12-28
Copyright: © 2023 Brain Health Alliance
Contact: CTaswell at Brain Health Alliance
URL: www.BrainiacsJournal.org/arc/pub/Craig2023MBCRLP
PDP: /Nexus/Brainiacs/Craig2023MBCRLP
DOI: /10.48085/X934EF46B

## References

[1] A. Athreya, S. K. Taswell, and C. Taswell. “Management Software for Monitoring Related Versions of Cultural Heritage Artifacts for Libraries and Museums.” *Proceedings of the Association for Information Science and Technology* 58.1 (2021), pp. 676–678 (cited p. 1).

[2] A. Brand, L. Allen, M. Altman, M. Hlava, and J. Scott. “Beyond authorship: Attribution, contribution, collaboration, and credit.” *Learned Publishing* 28.2 (2015), pp. 151–155 (cited p. 2).

[3] J. Cheney, L. Chiticariu, W.-C. Tan, et al. “Provenance in databases: Why, how, and where.” *Foundations and Trends in Databases* 1.4 (2009), pp. 379–474 (cited p. 1).

[4] A. Craig, S. H. Bae, T. Veeramacheneni, S. K. Taswell, and C. Taswell. “Web Service APIs for Scribe Registrars, Nexus Diristries, PORTAL Registries and DOORS Directories in the NPD System.” In: *Proceedings 9th International SWAT4LS Conference*. Amsterdam, Netherlands, 2016. URL: https://ceur-ws.org/Vol-1795/paper4.pdf (cited p. 1).

[5] A. Craig and C. Taswell. “PDP-DREAM Software for Integrating Multimedia Data with Interoperable Repositories.” *Brainiacs Journal of Brain Imaging And Computing Sciences* 2.1, HA46280EF (1 Dec. 31, 2021), pp. 1–6. DOI: 10.48085/HA46280EF. URL: https://BrainiacsJournal.org/arc/pub/Craig2021SIMDIR (cited p. 1).

[6] A. Craig, M. Yücel, L. Muchnik, and U. Hershberg. “Impact of finite size effect on applicability of generalized fractal and spectral dimensions to biological networks.” *Chaos, Solitons & Fractals* 164 (2022), p. 112707 (cited p. 2).

[7] E. Daga, A. Meroño-Peñuela, and E. Motta. “Modelling and querying lists in RDF. A pragmatic study.” In: *ISWC Workshops: QuWeDa*. 2019 (cited p. 2).

[8] S. Dutta, P. Kowshik, A. Ambati, S. Nori, S. K. Taswell, and C. Taswell. “Managing Scientific Literature with Software from the PORTAL-DOORS Project.” In: *2019 IEEE 15th International Conference on eScience (eScience)* (Sept. 24, 2019). San Diego, California: IEEE, Sept. 2019. DOI: 10.1109/eScience.2019.00081. URL: https://portaldoors.org/pub/docs/BCDC2019PdpDemo0806.pdf (cited p. 1).

[9] A. Hasselgren, K. Kralevska, D. Gligoroski, S. A. Pedersen, and A. Faxvaag. “Blockchain in healthcare and health sciences—a scoping review.” *International Journal of Medical Informatics* 134 (2020), p. 104040 (cited p. 2).

[10] A. O. Holcombe. “Contributorship, not authorship: Use CRediT to indicate who did what.” *Publications* 7.3 (2019), p. 48 (cited p. 2).

[11] M. Hosseini, J. Colomb, A. O. Holcombe, B. Kern, N. A. Vasilevsky, and K. L. Holmes. “Evolution and adoption of contributor role ontologies and taxonomies.” *Learned Publishing* (2022) (cited p. 2).

[12] M. Hosseini, B. Gordijn, Q. E. Wafford, and K. L. Holmes. “A systematic scoping review of the ethics of contributor role ontologies and taxonomies.” *Accountability in Research* (2023), pp. 1–28 (cited pp. 2–4).

[13] M. K. McNutt, M. Bradford, J. M. Drazen, B. Hanson, et al. “Transparency in authors’ contributions and responsibilities to promote integrity in scientific publication.” *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2557–2560 (cited pp. 2, 3).

[14] L. Moreau et al. “The foundations for provenance on the web.” *Foundations and Trends in Web Science* 2.2–3 (2010), pp. 99–241 (cited p. 1).

[15] L. Moreau, B. Clifford, J. Freire, J. Futrelle, et al. “The open provenance model core specification (v1.1).” *Future generation computer systems* 27.6 (2011), pp. 743–756 (cited p. 2).

[16] C. Taswell. “DOORS to the Semantic Web and Grid with a PORTAL for Biomedical Computing.” *IEEE Transactions on Information Technology in Biomedicine* 12.2 (2 Mar. 2007). In the Special Section on Bio-Grid published online 3 Aug. 2007, pp. 191–204. ISSN: 1089-7771. DOI: 10.1109/TITB.2007.905861 (cited p. 1).

[17] C. Taswell. “DOORS to the Semantic Web and Grid with a PORTAL for Biomedical Computing.” *IEEE Transactions on Information Technology in Biomedicine* 12.2 (2 Mar. 2008). In the Special Section on Bio-Grid published online 3 Aug. 2007, pp. 191–204. ISSN: 1089-7771. DOI: 10.1109/TITB.2007.905861 (cited p. 3).

[18] C. Taswell. “A Distributed Infrastructure for Metadata about Metadata: The HDMM Architectural Style and PORTAL-DOORS System.” *Future Internet* 2.2 (2010), pp. 156–189. ISSN: 1999-5903. DOI: 10.3390/FI2020156. URL: https://www.mdpi.com/1999-5903/2/2/156/ (cited p. 1).

[19] N. A. Vasilevsky, M. Hosseini, S. Teplitzky, V. Ilik, et al. “Is authorship sufficient for today’s collaborative research? A call for contributor roles.” *Accountability in Research* 28.1 (2021), pp. 23–43 (cited p. 2).