Representing provenance and track changes of cultural heritage metadata in RDF: a survey of existing approaches

Massari, Arcangelo
arcangelo.massari@unibo.it
Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy; Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Peroni, Silvio
silvio.peroni@unibo.it
Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy; Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Tomasi, Francesca
francesca.tomasi@unibo.it
Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Heibi, Ivan
ivan.heibi2@unibo.it
Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy; Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Recent years have seen the proliferation of many digital collections from all disciplinary fields that fall under the “big tent” (Svensson, 2012) of Digital Humanities. The data within them must be managed to be trustworthy, a goal usually achieved through the addition of provenance information, i.e., contextual metadata, primarily related to the identification of entities, such as the agent for the production, the date for the action and the reference sources (Gil et al., 2010). Moreover, in many humanities disciplines, truth loses its meaning without provenance since “truth” is defined as a statement with sufficient supporting sources, direct or indirect. As if this were not enough, sources may be at odds with each other, and it is essential to keep track of such conjectures (Barabucci et al., 2021; Daquino et al., 2022).

However, storing provenance information is not enough: having mechanisms to track how the metadata related to cultural objects change is crucial to promote the trustworthiness of the data. Indeed, such data evolves due to either the natural evolution of concepts or the correction of mistakes, and the latest version of knowledge may not be the most accurate. Representing such information in RDF is still an open challenge. In fact, the founding technologies of the Semantic Web – namely SPARQL, OWL, and RDF – did not initially provide an effective mechanism to annotate statements with metadata information. This limitation led to the introduction of numerous metadata representation models, none of which succeeded in establishing itself over the others and becoming a widely accepted standard to track both provenance and changes of RDF entities.

In this work, we present a systematic review of the representation models for provenance in RDF. The aim of our analysis is not to suggest a particular model, but rather to present the landscape of possible models to guide an informed choice. In order to conduct the review, we adopted a citation-based strategy (Figure 1), also known as “snowballing” (Wohlin, 2014), which consists of exploding the bibliography from a seed paper (Lecy & Beatty, 2012), which was (Sikos & Philp, 2020).

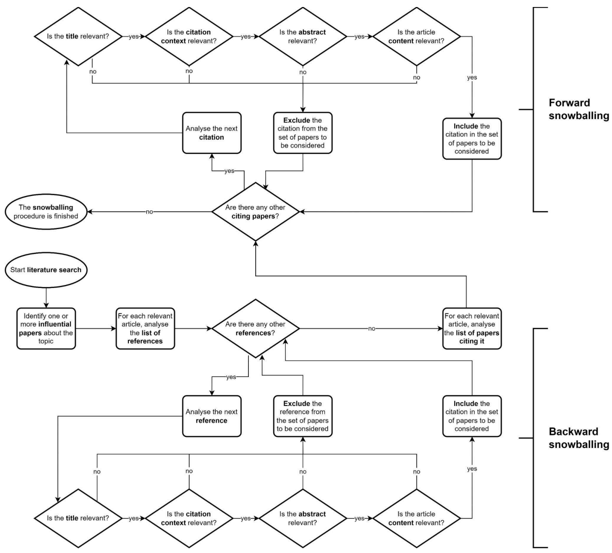


Figure 1. The snowballing procedure for conducting our systematic review

To date, the only W3C standard syntax to annotate the provenance is RDF reification (Manola & Miller, 2004) and is the only one compatible with all RDF-based systems, i.e., with every triplestore and software to process RDF data. However, there are several deprecation proposals for this syntax (Beckett, 2010), due to its poor scalability.

Different approaches have been proposed since 2005, varying in semantics, tuple typology, standard compliance, dependence on external vocabulary, blank node management, granularity, and scalability (Sikos & Philp, 2020). Three categories of solutions can be identified:

- Encapsulating provenance in RDF triples: n-ary relations (W3C, 2006), PaCE (Sahoo et al., 2010) and singleton properties (Nguyen et al., 2014).
- Associating provenance to the triple through RDF quadruples: named graphs (Carroll et al., 2005), RDF/S graphsets (Pediaditis et al., 2009), RDF triple coloring (Flouris et al., 2009), nanopublications (Groth et al., 2010), and conjectural graphs (Daquino et al., 2022).
- Extending the RDF data model: Notation 3 Logic (Berners-Lee, 2005), RDF+ (Dividino et al., 2009), SPOTL(X) (Hoffart et al., 2013), annotated RDF (aRDF) (Udrea et

al., 2010; Zimmermann et al., 2012), and RDF* (Hartig & Thompson, 2019).

Besides, a wide range of ontologies and vocabularies represent provenance information, either upper ontologies, domain ontologies, or provenance-related ontologies.

- Upper ontologies: Proof Markup Language (da Silva et al., 2006), Provenance Ontology (Gil et al., 2010), and the Open Provenance Model (Moreau et al., 2011).
- Domain ontologies: SWAN Ontology for Neuromedicine (Ciccarese et al., 2008), Provenir Ontology for eScience (Sahoo & Sheth, 2009), and PREMIS for archived digital objects (Caplan, 2017).
- Provenance-related ontologies: Dublin Core Metadata Terms (Board, 2020), and the OpenCitations Data Model (Daquino, Peroni, et al., 2020).

Table 1 compares the metadata representation models mentioned above. Most of these solutions do not comply with RDF 1.1 (i.e., RDF/S graphsets, N3Logic, aRDF, RDF+, SPOTL(X), and RDF*), are domain-specific (i.e., Provenir, SWAN, and PREMIS ontologies), rely on blank nodes (n-ary relations), or suffer from scalability issues (singleton properties, PaCE).

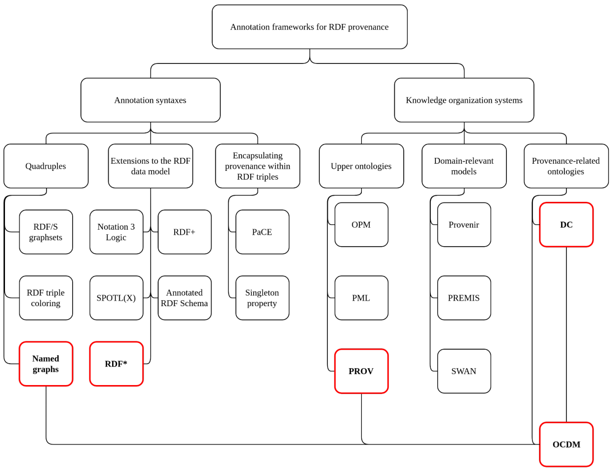
Despite being incompatible with RDF 1.1, it is worth mentioning that a W3C working group has recently published the first draft to make RDF* a standard (Gschwend & Lassila, 2022). RDF* embeds triples into triples as the subject or object. Its main goal is to replace RDF Reification through less verbose and redundant semantics. Since there is no serialization to represent such syntax, Turtle*, an extension of Turtle to include triples in other triples within << and >>, was also introduced.

Table 1. Advantages and disadvantages of metadata representations models for RDF. A glossary of the acronyms can be consulted on Zenodo (Massari, 2023). The table was expanded from the one in (Sikos & Philp, 2020)

Approach	Tuple type	RDF compliant	SPARQL compliant	RDF serialisations	External vocabulary	Scalable	Conjectures
Named graphs	Quadruple	yes	yes	TriG, TriX, N Quads	no	yes	no
RDF/S graphsets	Quadruple	no	no	TriG, TriX, N Quads	no	yes	no
RDF triple coloring	Quadruple	yes	yes	TriG, TriX, N Quads	no	yes	no
N3Logic	Triple (in N3)	no	yes	N3	N3 Logic Vocabulary	yes	yes
aRDF & Annotated RDF Schema	Non-standard	no	no	no	no	yes	no
RDF+	Quintuple	no	no	no	no	yes	no
SPOTL(X)	Quintuple/sextuple	no	no	no	no	Depends on implementation	no
RDF*	Non-standard	no	no	Turtle* (non-standard)	no	yes	yes
PaCE	Triple	yes	yes	Turtle, N-Triples, RDF-JSON, JSON-LD, RDFa, HTML5 Microdata	Provenir ontology	no	no
Singleton property	Triple	yes	yes	RDF/XML, N3, Turtle, N-Triples, RDF-JSON, JSON-LD, RDFa, HTML5 Microdata	Singleton property	no	no
Conjectural graph	Quadruple	yes	yes	TriG, TriX, N Quads	Conjectural property	no	yes

To date, named graphs (Carroll et al., 2005) and the Provenance Ontology (Moreau & Missier, 2013) are the most adopted approaches to attach provenance metadata to RDF triples. For example, mythLOD uses named graphs to map the provenance of artworks (Daquino, Pasqual, et al., 2020). On the one hand, Named Graphs are widespread because they are compliant with RDF 1.1 and can be queried with SPARQL 1.1; they have several serialization formats (i.e., TriX, TriG, and N-Quads), and are scalable, i.e., do not lead to triple bloat. On the other hand, the Provenance Ontology was published by the Provenance Working Group as a W3C Recommendation in 2013, meeting all the requirements for provenance on the Web and collecting existing ontologies into a single general model.

The OpenCitations Data Model (OCDM) (Daquino et al., 2020) represents provenance and tracks changes in a way that complies with RDF 1.1. It relies on well-known and widely adopted standards such as PROV-O (Lebo et al., 2013), named graphs, and Dublin Core (Figure 1). According to OCDM, one or more snapshots are linked to each entity, storing information about that resource at a specified time point. In particular, they record the validity dates, the primary data sources, the responsible agents, a human readable description, and a SPARQL update query summarizing the differences to the previous snapshot. Therefore, each snapshot is linked to the entity it describes and to the previous snapshot.



Annotation frameworks for RDF provenance. The edges represent a membership relationship, while the bold type and red circling highlight the syntaxes we focus on this contribution

We believe that understanding the complex landscape of models for adding metadata to RDF triples is essential for building digital collections that properly handle provenance and change-tracking. This aspect is crucial in order to build a reliable scholarly research in the Digital Humanities field.

Bibliography

Barabucci, G., Tomasi, F. & Vitali F. (2022), Modeling data complexity in public history and cultural heritage, in: Handbook of Digital Public History, Oldenbourg, De Gruyter, 459-474

Barabucci, G., Tomasi, F., & Vitali, F. (2021). Supporting Complexity and Conjectures in Cultural Heritage Descriptions. *CEUR Workshop Proceedings*, *2810*, 104–115.

Beckett, D. (2010). RDF Syntaxes 2.0. In *W3C*. https://www.w3.org/2009/12/rdf-ws/papers/ws11

Berners-Lee, T. (2005). Notation 3 Logic. In *W3C*. https://www.w3.org/DesignIssues/N3Logic

Board, D. U. (2020). DCMI Metadata Terms. In *Dublin Core Metadata Initiative*. http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/

Caplan, P. (2017). *Understanding PREMIS: an overview of the PREMIS Data Dictionary for Preservation Metadata. Library of Congress*. https://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf

Carroll, J. J., Bizer, C., Hayes, P., & Stickler, P. (2005). Named graphs, provenance and trust. *Proceedings of the 14th International Conference on World Wide Web - WWW '05*, 613. https://doi.org/10.1145/1060745.1060835

Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., & Clark, T. (2008). The SWAN biomedical discourse ontology. *Journal of Biomedical Informatics*, *41*(5), 739–751. https://doi.org/10.1016/j.jbi.2008.04.010

da Silva, P. P., McGuinness, D. L., & Fikes, R. (2006). A proof markup language for Semantic Web services. *Information Systems*, *31*(4–5), 381–395. https://doi.org/10.1016/j.is.2005.02.003

Daquino, M., Pasqual, V., & Tomasi, F. (2020). Knowledge Representation of digital Hermeneutics of archival and literary Sources. *Knowledge Representation of Digital Hermeneutics of Archival and Literary Sources*, 59–76.

Daquino, M., Pasqual, V., Tomasi, F., & Vitali, F. (2022). Expressing Without Asserting in the Arts. In G. M. D. Nunzio, B. Portelli, D. Redavid, & G. Silvello (Eds.), *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)* (Vol. 3160). CEUR-WS.org. http://ceur-ws.org/Vol-3160/short14.pdf

Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., Mayr, P., Romanello, M., & Zumstein, P. (2020). The OpenCitations Data Model. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), *The Semantic Web – ISWC 2020* (Vol. 12507, pp. 447–463). Springer International Publishing. https://doi.org/10.1007/978-3-030-62466-8_28

Dividino, R., Sizov, S., Staab, S., & Schueler, B. (2009). Querying for provenance, trust, uncertainty and other meta knowledge in RDF. *Journal of Web Semantics*, *7*(3), 204–219. https://doi.org/10.1016/j.websem.2009.07.004

Flouris, G., Fundulaki, I., Pediaditis, P., Theoharis, Y., & Christophides, V. (2009). Coloring RDF Triples to Capture Provenance. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. Bernstein, D. R. Karger, T. Heath, L. Feigenbaum, … K. Thirunarayan (Eds.), *The Semantic Web—ISWC 2009* (Vol. 5823, pp. 196–212). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04930-9_13

Gil, Y., Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L., & Silva, P. (2010). Provenance XG Final Report. In *W3C*. http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. *Information Services & Use*, *30*(1–2), 51–56. https://doi.org/10.3233/ISU-2010-0613

Gschwend, A., & Lassila, O. (2022). *PROPOSED RDF-star Working Group Charter*. W3C. https://w3c.github.io/rdf-star-wg-charter/

Hartig, O., & Thompson, B. (2019). Foundations of an Alternative Approach to Reification in RDF. *ArXiv:1406.3399 [Cs]*. http://arxiv.org/abs/1406.3399

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, *194*, 28–61. https://doi.org/10.1016/j.artint.2012.06.001

Lebo, T., Sahoo, S., & McGuinness, D. (2013). PROV-O: The PROV Ontology. In *W3C*. http://www.w3.org/TR/2013/REC-prov-o-20130430/

Lecy, J. D., & Beatty, K. E. (2012). Representative Literature Reviews Using Constrained Snowball Sampling and Citation Network Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1992601

Manola, F., & Miller, E. (2004). RDF Primer. In *W3C*. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., & den Bussche, J. V. (2011). The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, *27*(6), 743–756. https://doi.org/10.1016/j.future.2010.07.005

Moreau, L., & Missier, P. (Eds.). (2013). *PROV-DM: The PROV Data Model*. W3C. http://www.w3.org/TR/2013/REC-prov-dm-20130430/

Nguyen, V., Bodenreider, O., & Sheth, A. (2014). Don't like RDF reification?: Making statements about statements using singleton property. *Proceedings of the 23rd International Conference on World Wide Web - WWW '14*, 759–770. https://doi.org/10.1145/2566486.2567973

Pediaditis, P., Flouris, G., Fundulaki, I., & Christophides, V. (2009). On Explicit Provenance Management in RDF/S Graphs. In *First Workshop on the Theory and Practice of Provenance*. USENIX. https://www.usenix.org/legacy/event/tapp09/tech/full_papers/pediaditis/pediaditis.pdf

Sahoo, S. S., Bodenreider, O., Hitzler, P., Sheth, A., & Thirunarayan, K. (2010). Provenance Context Entity (PaCE): Scalable Provenance Tracking for Scientific RDF Data. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. Gertz, & B. Ludäscher (Eds.), *Scientific and Statistical Database Management* (Vol. 6187, pp. 461–470). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13818-8_32

Sahoo, S. S., & Sheth, A. P. (2009). *Provenir Ontology: Towards a Framework for eScience Provenance Management*. https://corescholar.libraries.wright.edu/knoesis/80

Sikos, L. F., & Philp, D. (2020). Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs. *Data Science and Engineering*, *5*(3), 293–316. https://doi.org/10.1007/s41019-020-00118-0

Svensson, P. (2012). Beyond the Big Tent. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (NED-New edition, pp. 36–72). University of Minnesota Press. http://www.jstor.org/stable/10.5749/j.ctttv8hq.7

Udrea, O., Recupero, D. R., & Subrahmanian, V. S. (2010). Annotated RDF. *ACM Transactions on Computational Logic*, *11*(2), 1–41. https://doi.org/10.1145/1656242.1656245

W3C. (2006). *Defining N-ary Relations on the Semantic Web*. http://www.w3.org/TR/2006/NOTE-swbp-n-aryRelations-20060412/

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 1–10. https://doi.org/10.1145/2601248.2601268

Zimmermann, A., Lopes, N., Polleres, A., & Straccia, U. (2012). A general framework for representing, reasoning and querying with annotated Semantic Web data. *Journal of Web Semantics*, *11*, 72–95. https://doi.org/10.1016/j.websem.2011.08.006