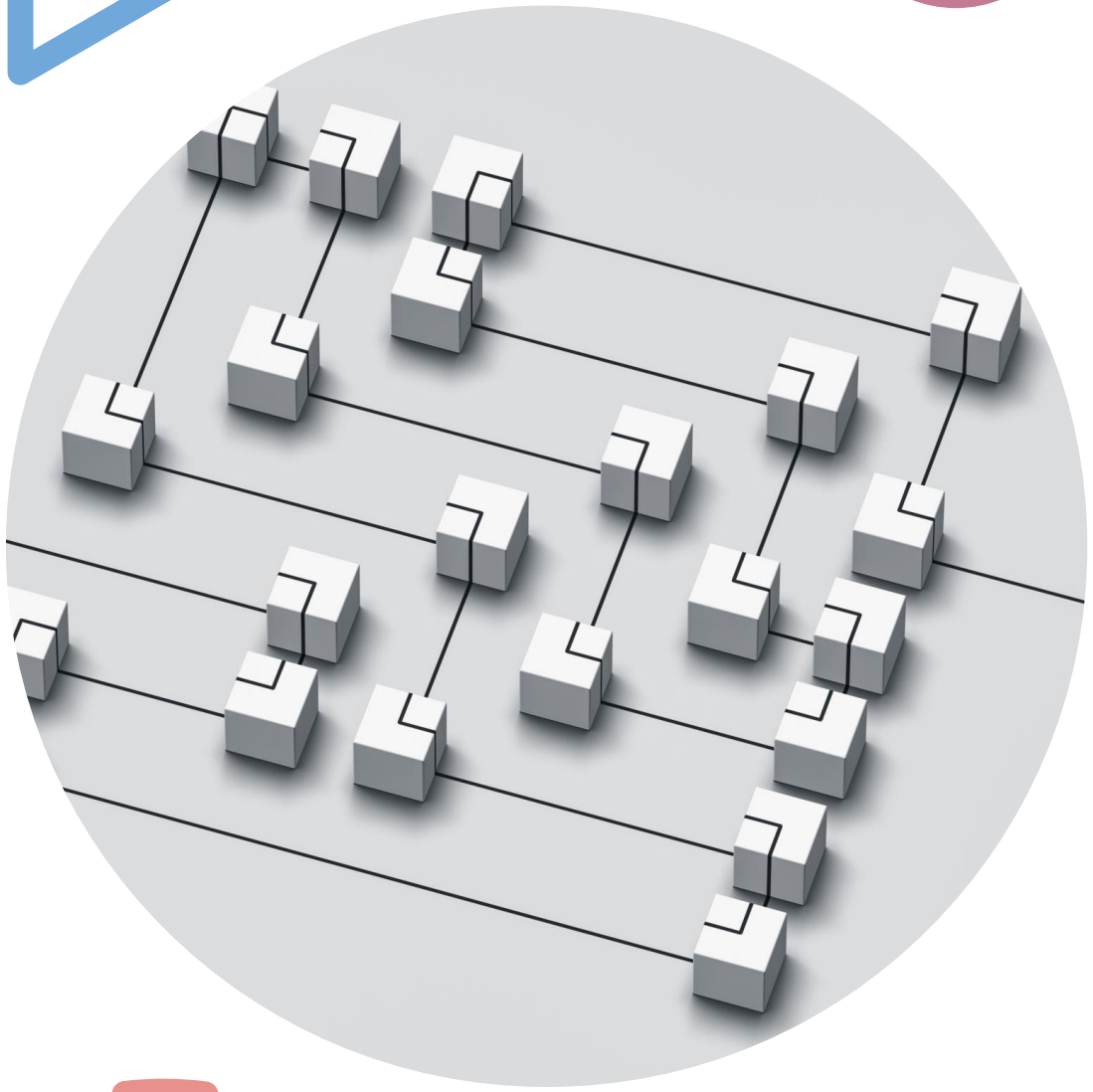
Representing provenance and track changes of cultural heritage metadata in RDF: a survey of existing approaches

Arcangelo Massari [1, 2], Silvio Peroni [1, 2], Francesca Tomasi [2] and Ivan Heibi [1, 2]

[1] Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

[2] Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA

RESEARCH CENTRE
FOR OPEN SCHOLARLY METADATA

The Rise of Digital Collections in Humanities

- Proliferation of many digital collections across all disciplinary fields in the Digital Humanities
- The data within these collections needs careful management to maintain **trustworthiness**

The Role of Provenance Information

- **Trustworthiness** is typically achieved through the addition of **provenance information**
- Provenance information includes contextual metadata such as the **responsible agent**, the **generation time**, and the **primary sources**

The Concept of "Truth" in Humanities

- In many humanities disciplines, "truth" is defined as a statement with sufficient supporting sources
- Without provenance, "truth" loses its meaning
- There is a need to keep track of contradictory sources

Need for Mechanisms to Track Changes

Storing provenance information alone is **not enough**

Mechanisms to **track** how the metadata of cultural objects **change** are crucial

Data evolves due to the natural **evolution** of **concepts** or the **correction** of **mistakes**, and the latest versions of knowledge may not be the most accurate

Challenges in Representing Information in RDF

REPRESENTING PROVENANCE AND CHANGE INFORMATION IN **RDF** REMAINS AN OPEN CHALLENGE

FOUNDING TECHNOLOGIES OF THE SEMANTIC WEB (SPARQL, OWL, AND RDF) **INITIALLY LACKED** AN EFFECTIVE MECHANISM FOR **ANNOTATING STATEMENTS** WITH METADATA

THIS LED TO THE INTRODUCTION OF **NUMEROUS** METADATA REPRESENTATION **MODELS**, BUT **NONE** HAS BECOME A WIDELY ACCEPTED **STANDARD** TO TRACK BOTH PROVENANCE AND CHANGES OF RDF ENTITIES

Review of RDF Provenance Representation Models

Objective: Present a systematic review of provenance representation models in RDF

Not to advocate for a specific model, but to provide an **overview** of available models to help with an informed decision

Review Methodology: we adopted a citation-based approach, also known as "snowballing" (Wohlin, 2014). This method involves exploring the bibliography from a seed paper

Seed Paper: *Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs* (Sikos & Philp, 2020) was used as the starting point for the review

Current Standard and Issues

- RDF reification is the only W3C **standard** syntax for annotating provenance
- **Compatibility** with all RDF-based systems
- However, there are several **deprecation** proposals due to its **poor scalability**
- RDF reification leads to **triple bloat**: four triples must be added to add at least one piece of provenance information

Statement

meta:br/86766 dcterms:title "Open access and online publishing: a new frontier in nursing?".

Reification

statements:triple12345 **rdf:type rdf:Statement**.

statements:triple12345 **rdf:subject** meta:br/86766.

statements:triple12345 **rdf:predicate** dcterms:title.

statements:triple12345 **rdf:object** "Open access and online publishing: a new frontier in nursing?".

Provenance

statements:triple12345 **prov:hadPrimarySource** <https://api.crossref.org>.

From RDF Reification to N-ary Relations

- Recommended by W3C (2006) as an **alternative** approach to express provenance
- Properties are not only binary relationships but can connect a URI to multiple URIs or value
- Both RDF Reification and N-ary relations can reify relationships - RDF Reification reifies the statement, while N-ary relations **reify** the **predicate**
- **Advantage** of N-ary relations: avoids repeating all triple elements, **only** the **predicate** is **repeated**
- **Disadvantage**: it introduces **blank nodes**, which can't be globally dereferenced

Statement

meta:br/86766 dcterms:title _**:Title**.

N-ary relation

_**:Title** dcterms:title "Open access and online publishing: a new frontier in nursing?".

Provenance

_:Title **prov:hadPrimarySource** <https://api.crossref.org>.

Proposed Approaches and their Categories

Due to the limitations of both RDF Reification and N-ary relations, various new approaches have been proposed since 2005

Three categories of solutions identified:

- Encapsulating provenance in RDF triples (e.g., n-ary relations, PaCE, singleton properties)
- Associating provenance to the triple through RDF quadruples (e.g., **named graphs**, RDF/S graphsets, RDF triple coloring, **nanopublications**, and conjectural graphs)
- Extending the RDF data model (e.g., Notation 3 Logic, RDF+, SPOTL(X), annotated RDF, **RDF-star**)

Ontologies and Vocabularies for Provenance Information

Range of ontologies and vocabularies to represent provenance information:

- Upper ontologies (e.g., Proof Markup Language, **Provenance Ontology**, Open Provenance Model)
- Domain ontologies (e.g., SWAN Ontology, Provenir Ontology, **PREMIS**)
- Provenance-related ontologies (e.g., **Dublin Core Metadata Terms**, **OpenCitations Data Model**)

Issues with Existing Solutions

Most solutions:

- Do not comply with RDF 1.1 (i.e., RDF/S graphsets, N3Logic, aRDF, RDF+, SPOTL(X), and RDF-star),
- Are domain-specific (i.e., Provenir, SWAN, and PREMIS ontologies)
- Rely on blank nodes (n-ary relations)
- Have scalability issues (singleton properties, PaCE)

- **Domain-relevant models** are specifically suited for provenance handling within that domain, but usually lack the generality for other contexts
  - Example: **PREMIS** model, which focuses on preserving **archived digital objects** - files, bitstreams, and aggregations
- The use of a provenance and change tracking model that doesn't comply with RDF 1.1 implies a **prescriptive choice** at the **technological** level

Case Study: Wikidata - Provenance

- Wikidata provides an interesting example of **provenance** and **change tracking**
- Regarding **provenance**, Wikidata uses a **proprietary RDF extension** for context information about statements
- The **qualifiers** (e.g., start and end date of a statement) are **associated** with the **predicate**, forming an **n-ary relation**
- SPARQL queries can be made on this provenance information

Query: “Select the cities with a population greater than 1,000,000 inhabitants and return the year in which the minimum population was recorded for each city”

```
SELECT ?city (min(?time) as ?year) WHERE {
        ?city wdt:P31/wdt:P279* wd:Q515.
        ?city wdt:P17 wd:Q38 .
        ?city p:P1082 ?statement .
        ?statement
<http://www.wikidata.org/prop/statement/value/P1082> ?value ;
<http://www.wikidata.org/prop/qualifier/P585> ?time .
?value <http://wikiba.se/ontology#quantityAmount>
?population .
        FILTER (?population > 1000000 )
 } GROUP BY ?city
```

https://w.wiki/4rs

Wikidata – change tracking

- Wikidata adopts a **non-RDF**, **backup-based policy**, creating a revision every time an entity-related page is modified
- **Provenance metadata** such as timestamp, contributor's username, ID, and summary of modifications are also saved
- Each revision contains a complete **copy** of the page post-change, stored in **compressed XML files**
- This data is available for on the Wikidata website for **download**
- The content of the text field is in JSON format with non-ASCII characters escaped
- Users can explore single revisions and **compute the delta** between versions via the user interface
- However, it's **not possible** to perform **SPARQL queries** on revisions

```
<page>
    <title>Q78189694</title>
    <ns>0</ns>
    <id>77644210</id>
    <revision>
        <id>1467205756</id>
        <parentid>1233484847</parentid>
        <timestamp>2021-07-26T18:45:13Z</timestamp>
        <contributor>
            <username>Twofivesixbot</username>
            <id>2691515</id>
        </contributor>
        <comment>/* wbeditentity-update-languages-short:0||bn */ KOI</comment>
        <model>wikibase-item</model>
        <format>application/json</format>
        <text bytes="19449" xml:space="preserve">{&amp;quot;type&amp;quot;:&amp;quot[...]}
        </text>
        <sha1>jm79xfec7qbv4o5adf7umx1r94wblh4</sha1>
    </revision>
</page>
```

RDF-star and Turtle-star

- Despite incompatibility, a W3C working group has published a draft to make RDF-star a **standard**
- RDF-star **embeds** triples into triples as the subject or object
- Goal to replace RDF Reification through **less** verbose and **redundant** semantics
- **Turtle-star**, an extension of Turtle, was introduced to represent such syntax
- RDF-star is already compatible with many RDF-based systems (e.g., GraphDB, rdflib)

RDF Reification

meta:br/86766 dcterms:title "Open access and online publishing: a new frontier in nursing?".

statements:triple12345 rdf:type rdf:Statement.

statements:triple12345 rdf:subject meta:br/86766.

statements:triple12345 rdf:predicate dcterms:title.

statements:triple12345 rdf:object "Open access and online publishing: a new frontier in nursing?".

statements:triple12345 prov:hadPrimarySource <https://api.crossref.org>.

RDF-star

<<meta:br/86766 dcterms:title "Open access and online publishing: a new frontier in nursing?"**>>** **prov:hadPrimarySource** <https://api.crossref.org>.

Most Adopted Approaches

- **Named graphs** and the **Provenance Ontology** are the most adopted approaches for attaching provenance metadata to RDF triples
- Reasons for adoption:
  - RDF 1.1 compliance
  - query capabilities
  - scalability
  - multiple serialization formats
  - meeting all requirements for provenance on the Web

Case study: MythLOD

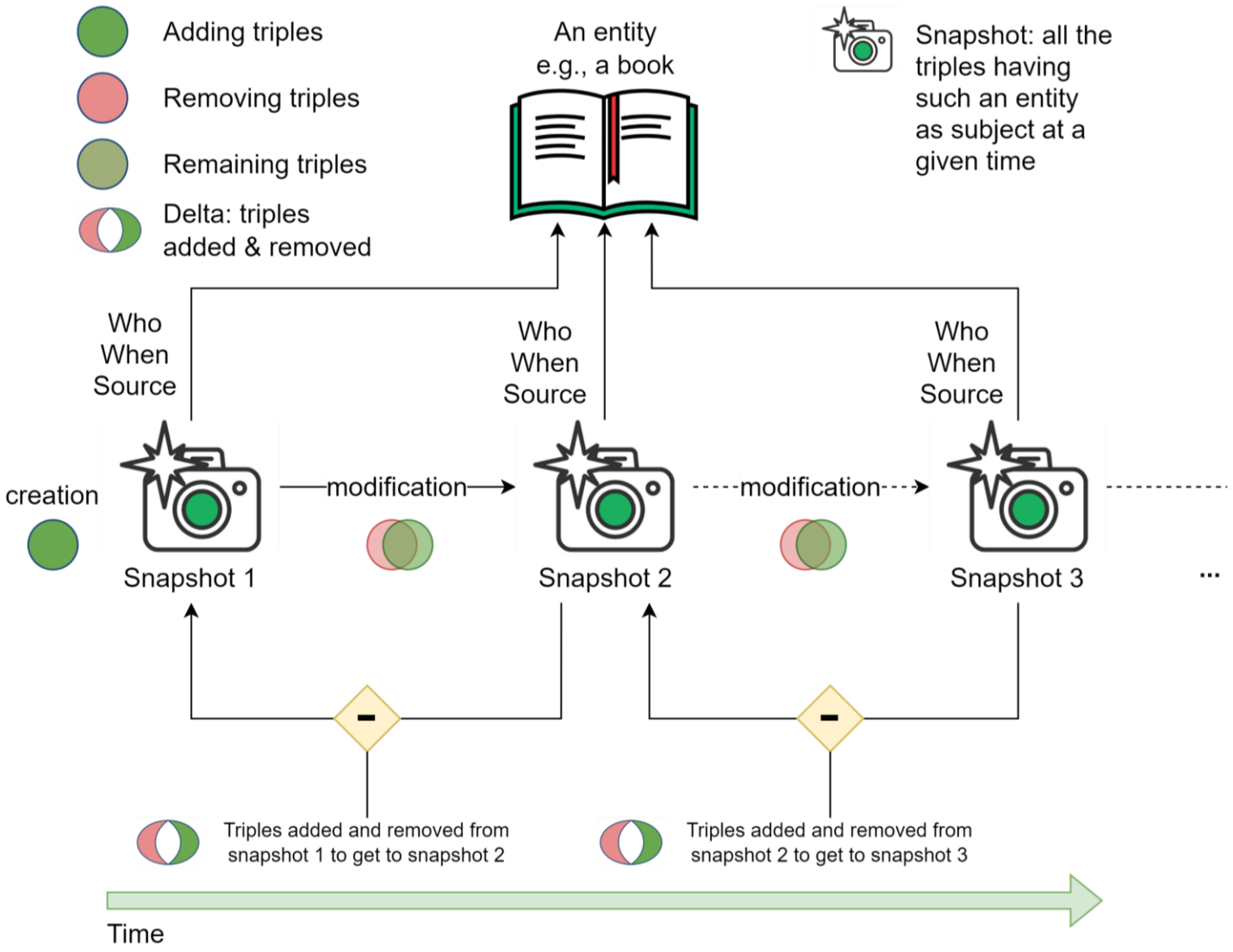
- MythLOD focuses over the formal representation of experts' analysis when **associating artworks** (and their interpretation) **to literary sources**
- mythLOD utilizes **named graphs** (**nanopublications**) and **PROV-O** for mapping the provenance of artworks

item:3701 dct:title **"Venere di Milo con cassetti".**

myth:provenance3701 {
myth:assertion3701 **prov:wasGeneratedAtTime** "2017-05-24T14:43:00";
prov:wasGeneratedBy int-act:3701.
int-act:3701 a **prov:InterpretationAct** ;
hico:hasInterpretationCriterion myth:**hermeneutic-analysis** ;
hico:hasInterpretationType myth:**iconographic-approach** ;
prov:wasAttributedTo person:**gamba-hubert** .
}

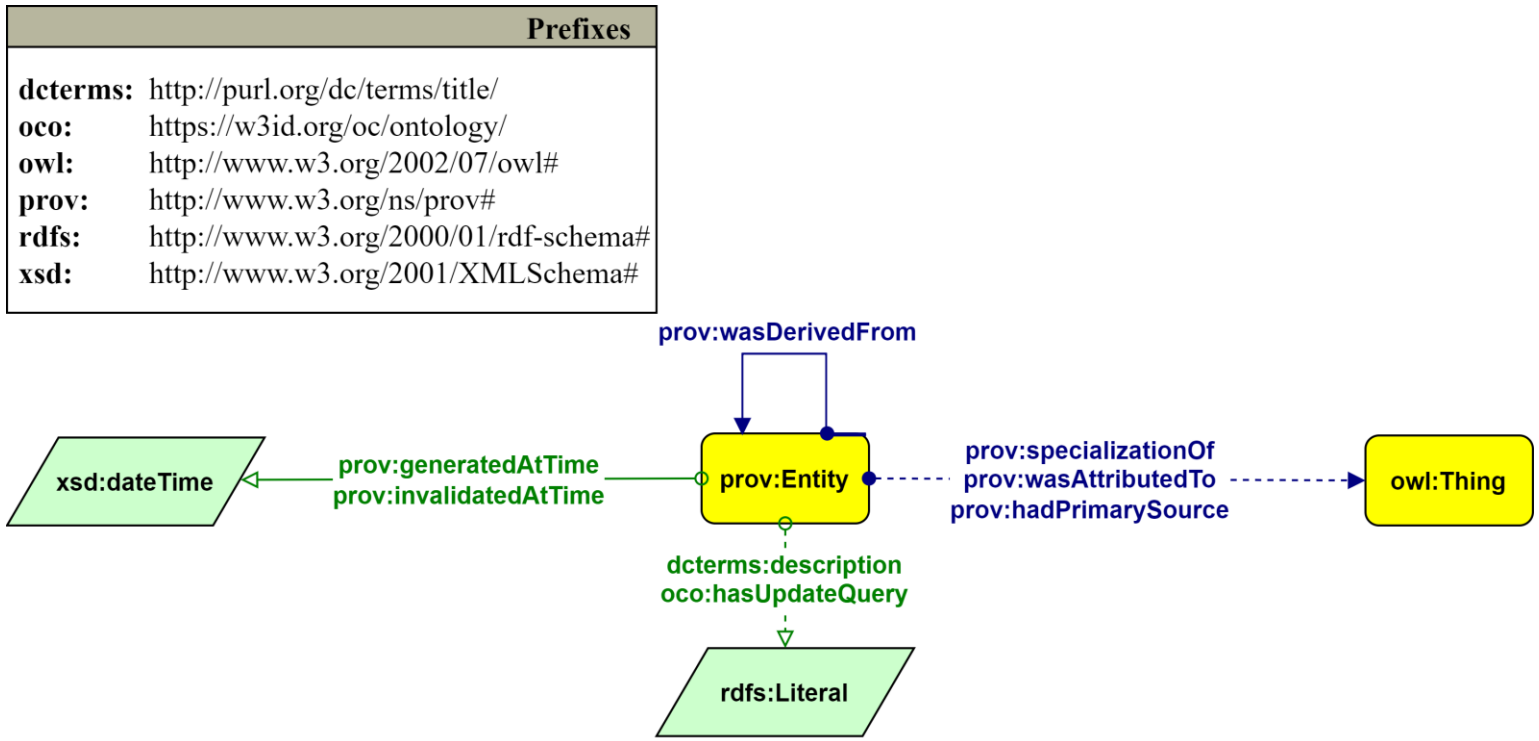
Overview of the OpenCitations Data Model (OCDM)

- OCDM represents **provenance** and **tracks changes** in **compliance** with **RDF 1.1**
- It leverages widely adopted standards such as **PROV-O**, **named graphs**, and **Dublin Core**
- Provenance mechanism of OpenCitations encapsulates an initial creation **snapshot** for each stored **entity**
- The initial snapshot can be followed by others detailing **modification**, **merge**, or **deletion** of data, each marked with its snapshot number



Provenance Metadata in OCDM

- Each snapshot is **connected to the previous one** via the prov:wasDerivedFrom predicate
- Each snapshot is **linked to the entity** it describes via prov:specializationOf
- Each snapshot corresponds to a named graph with provenance metadata:
  - The **responsible agent** (prov:wasAttributedTo)
  - The **primary source** (prov:hadPrimarySource)
  - The **generation time** (prov:generatedAtTime)
  - The **invalidation time** (prov:invalidatedAtTime), following the generation of an additional snapshot
- Each snapshot can optionally include a natural language **description** (dcterms:description)



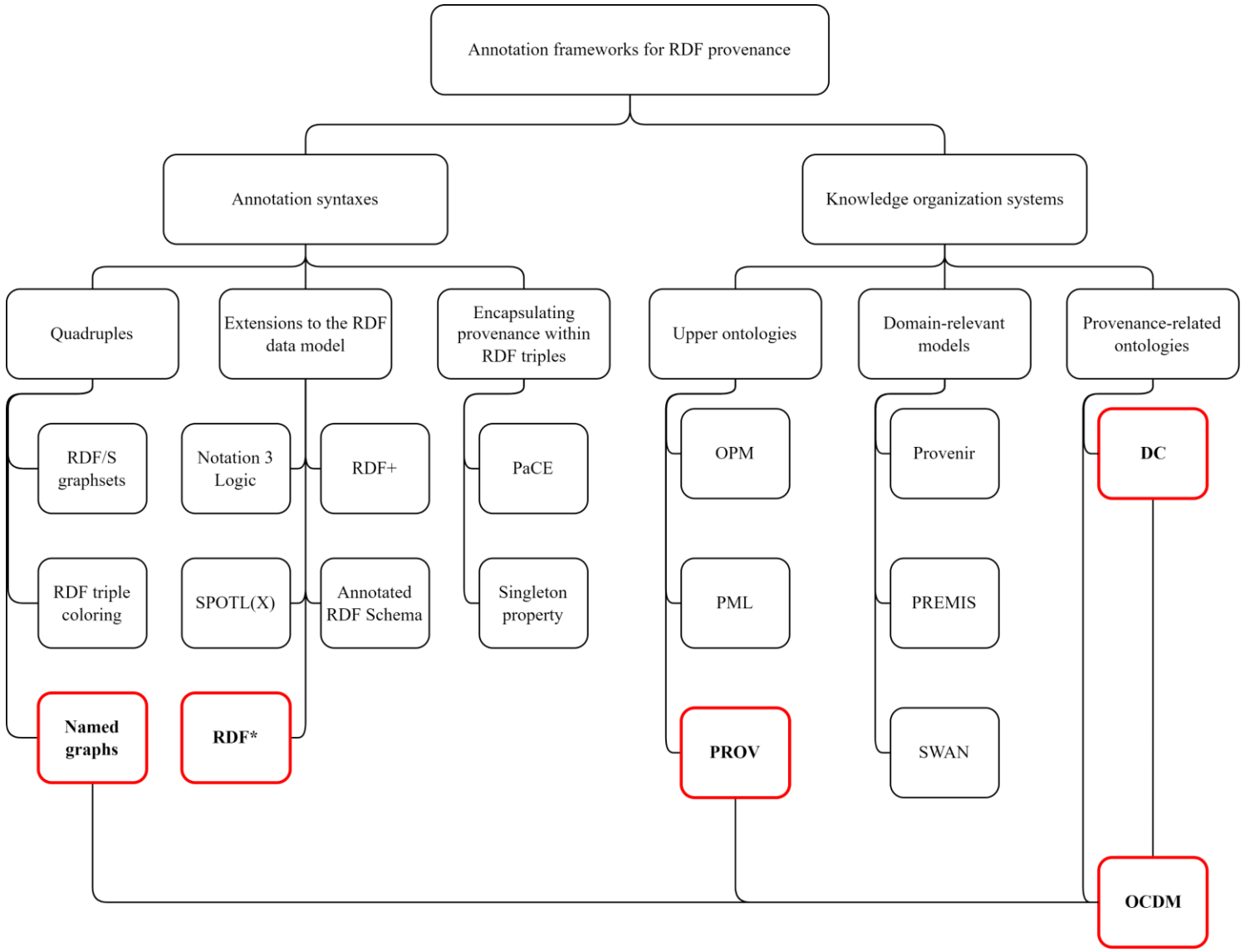
Change tracking in OCDM

- The OCDM provenance model introduces a **new predicate**, oco:hasUpdateQuery
- oco:hasUpdateQuery expresses the **delta** between two versions of an entity via a **SPARQL UPDATE query**

```
<br/86766> a <http://purl.org/spar/fabio/Expression>;
    dcterms:title "Open access and online publishing: a new frontier in
    ↪  nursing?"^^xsd:string;
    cito:cites <br/301102>, <br/301103>, <br/301104>, <br/301105>, <br/301106>;
    datacite:hasIdentifier <id/80178>.

<id/80178> a datacite:Identifier;
    datacite:usesIdentifierScheme datacite:doi;
    literal:hasLiteralValue "10.1111/j.1365-2648.2012.06023.x"^^xsd:string.

<id/80178/prov/se/2> a prov:Entity;
    oco:hasUpdateQuery "
        DELETE DATA {
        GRAPH <https://github.com/opencitations/time-agnostic-library/id/> {
            <https://github.com/opencitations/time-agnostic-library/id/80178>
            <http://www.essepuntato.it/2010/06/literalreification/hasLiteralValue>
            '10.1111/j.1365-2648.2012.06023.x.' . } };
        INSERT DATA {
        GRAPH <https://github.com/opencitations/time-agnostic-library/id/> {
             <https://github.com/opencitations/time-agnostic-library/id/80178>
             <http://www.essepuntato.it/2010/06/literalreification/hasLiteralValue>
             '10.1111/j.1365-2648.2012.06023.x' . } }"^^xsd:string.
    dcterms:description "The entity
    ↪  'https://github.com/opencitations/time-agnostic-library/id/80178' has been
    ↪  modified."^^xsd:string;
    prov:generatedAtTime "2021-10-19T19:55:55"^^xsd:dateTime;
    prov:specializationOf <id/80178>;
    prov:wasAttributedTo <https://orcid.org/0000-0002-8420-0696>;
    prov:wasDerivedFrom <id/80178/prov/se/1>;

<id/80178/prov/se/1> a prov:Entity;
    dcterms:description "The entity
    ↪  'https://github.com/opencitations/time-agnostic-library/id/80178' has been
    ↪  created."^^xsd:string;
    prov:generatedAtTime "2021-10-10T23:44:45"^^xsd:dateTime;
    prov:hadPrimarySource <https://api.crossref.org/works/10.1007/s11192-019-03265-y>;
    prov:invalidatedAtTime "2021-10-19T19:55:55"^^xsd:dateTime;
    prov:specializationOf <id/80178>;
    prov:wasAttributedTo <https://orcid.org/0000-0002-8420-0696>.
```

Importance of Understanding Metadata Models

- Understanding the complex landscape of metadata models for RDF triples is crucial
- Essential for building digital collections that handle **provenance** and **change-tracking** properly
- This aspect is fundamental for building a **reliable scholarly research** for any Digital Humanities discipline



Thank you for your attention