A document-inspired way for tracking changes of RDF data

The case of the OpenCitations Corpus

Silvio Peroni[1], David Shotton[2], and Fabio Vitali[1]

[1] DASPLab, DISI, University of Bologna, Bologna, Italy
[2] Oxford e-Research Centre, University of Oxford, Oxford, UK
silvio.peroni@unibo.it, david.shotton@oerc.oc.ac.uk, fabio.vitali@unibo.it

Abstract. There are several distinct ways to represent data drift the in Linked Open Data world. In this paper we introduce the approach for tracking data changes that has been used in the context of the OpenCitations Project. Such approach has been inspired by existing works on change tracking mechanisms in documents created through word-processors such as Microsoft Word and OpenOffice Writer.
RASH: https://w3id.org/oc/paper/occ-driftalod2016.html

Keywords: OpenCitations, OpenCitations Corpus, PROV-O, RDF changes, change tracking, document changes

1 Introduction

Data change in time, and the reason for this change can be manyfold. On the one hand, they can have mistakes that are corrected once they are identified, even after the publication date. On the other hand, information (or, better, representations of certain actual situations, like the composition of the government of a country) naturally evolves in time.

RDF technologies (RDF, OWL, SPARQL, etc.) were not originally thought to keep track of such changes natively. Thus, alternative approaches have been proposed in the past years so as to extend such formalisms with a mechanism for adding such additional endeavour, so as to keep track how certain data have changed. The introduction of Named Graphs [1] and the Provenance Ontology (PROV-O) [4] are among the most used and appropriate ways for enabling the description of time-dependent (or, more generally, context-dependent) data. However, there can still exist different ways of keeping track of such changes in time.

In this paper we introduce the approach for tracking changes in RDF data that has been used in the context of the OpenCitations Project[3] [7] [10]. The main aim of this project is the creation of an open repository of scholarly citation data – the OpenCitations Corpus (OCC) – made available under a Creative Commons public domain dedication[4] to provide in RDF accurate citation information (bibliographic references) harvested from the scholarly literature. All the entities in the OCC have associated metadata describing its provenance, so as to keep track of the curatorial activities related to each OCC entity, the curatorial

[3] http://opencitations.net/
[4] https://creativecommons.org/publicdomain/zero/1.0/legalcode

agents involved, their roles, and the sources used for retrieving such data. These provenance data are organised in a way that allows us to reconstruct a particular status (or snapshot) of any entity in the OCC at a specified time by using a mechanism inspired by existing works on change tracking mechanisms in documents created through word-processors such as Microsoft Word and OpenOffice Writer.

The rest of the paper is organised as follows. In Section 2 we briefly introduce some possible approaches to keep track of changes of RDF data. In Section 3 we describe the approach we propose for addressing such issue, while in Section 4 we discuss its application in the context of the OCC. Finally, in Section 5 we conclude the paper sketching out some future works.

2 Approaches to changes

Two approaches can be used for representing how a particular dataset has evolved in time. On the one hand, we have *statement-centric* approaches, that basically provide mechanisms to record how the set of statements in a dataset have evolved by means of simple operation such as addition and deletion. On the other hand, we have *resource-centric* approaches, that mainly allow one to say when an instance of a time-dependent class or property (traditionally called *anti-rigid* concepts [3]) change its status somehow.

There are at least two possible approaches belonging to the first of the aforementioned categories: *physical snapshots* and *massive statement reification*.

A *physical snapshot* of a given LOD dataset is a particular record of all the statements in such dataset at a given time. Using this technique, the tracking of all the changes of the dataset is stored every time one think is appropriate, e.g. every time a statement has been added/modified, after a certain amount of modification to the dataset, after a particular time interval (every week, every month, etc.), and so on. This is a quite common strategy for several LOD datasets available online (such as DBPedia [5], which makes available versioned datasets as described at `http://wiki.dbpedia.org/datasets`), it is quite easy to implement, but one would need of an extraordinary amount of space and time for keeping track of how a dataset has changed, since every snapshot would record the entire dataset at a certain date.

The *massive statement reification* mechanism requires the creation of additional identifiers (one for each statement), and all of them are, in some way, marked when they have been created/removed and by whom. This kind of approach can be coupled easily with existing models, such as PROV-O [4], so as to keep track of how a statement has been modified in time – similarly to what Wikidata[5] [11] implements. In this case, the dataset is continuously increasing its size – since deleted statements are not really removed from the dataset, rather they are marked as deleted – but it also allows to track every change and to index it according to the time it happened. This is a quite huge advantage, since it would allow to restore any possible status of the dataset by discarding all the modifications happened after a certain date.

Among the *resource-centric* mechanisms, it is worth mentioning the *provenance-centric* and the *by-design* approaches, that allow one to record changes of a certain

[5] `https://www.wikidata.org/`

resource, e.g. a particular class or an individual, by means of re-using existing models and without explicitly referring to the set of statements they are involved in.

An ontology that can be used for addressing the former category is PROV-DC [2], which can be used to for expressing how entities change in time by means of additional classes and properties added to PROV-O, which now allows the specification of activities such as *prov:Create*, *prov:Modify*, etc. While this is a valuable and simple approach, it is not easy to understand in a formal way which particular aspect of an entity has actually changed.

The alternative approaches, i.e. those compliant with the *by-design* mechanism, oblige the dataset creator of including, from the very beginning, a finest conceptualisation of the (*anti-rigid*) entities that can change in time in the actual ontology she is using for representing the data. A good option here is to use particular ontology design patterns, such as the time-indexed situation pattern[6] (that is enough flexible to be reused with different kind of data), or the 4D Fluent OWL ontology [12]. However, if something that now can be modified was not considered as such at the very beginning, it would be possible that part of the ontology used for representing the data (and consequently the data themselves) could be modified accordingly – wasting time and, potentially, changing the current organisation of the data, thus limiting their reusability in the long term.

Both the aforementioned resource-centric mechanisms would allow not to delete permanently any information, rather they would oblige to include the entire history of each entity in the dataset, since they use particular ontological constructs to tell the user when an entity has been created/invalidated, by whom, and so on.

3 A document-inspired approach to data drift

The approach we propose readapts a well-known structure for keeping track of changes in word-processor documents, in particular OpenOffice Writer (OOW herein). When an author activate the change tracking plugin in OOW, every insertion and deletion into the document are tracked by using two different mechanisms proper to *overlapping markup* theories, called *milestone* (for insertions) and *stand-off markup* (for deletions) [8]. Milestones allows one to add the new content directly within the existing text, marking it in some way that can be recognisable. Contrarily, stand-off markup removes explicitly a piece of text from the actual content of the document, and places it in an auxiliary space for easy retrieving and, if needed, restoration.

Following the same principles, we developed a mechanism that allows us to either add or remove new statements directly to the current set of data related to an entity (i.e. the RDF triples that have such entity as subject), while preserving provenance information of such addition/deletion actions in an appropriate contextual space, i.e. the provenance graph associated to such entity. For doing that we leverage the PROV-O [4] ontology, and extend it by adding an additional data property called *hasUpdateQuery*, which allows us to record insertions and deletions as SPARQL INSERT and SPARQL DELETE queries.

[6] `http://ontologydesignpatterns.org/wiki/Submissions:TimeIndexedSituation`

The main idea of our approach is that each entity in a dataset (i.e. an instance *e* of the class *prov:Entity*) is represented by one or more snapshots (other instances *e1*, *e2*, *e3*, ... of *prov:Entity*, each intended as specialisation of *e* via *prov:specializationOf*), each recording the composition of such entity *e* (i.e. the statements that have the entity as subject) at a fixed point in time. In addition, each snapshot is linked to the others according to their temporal creation/invalidation by means of the property *prov:wasDerivedFrom*.

Please let us introduce a working example for discussing the approach proposed. For instance, let us consider the entity *sp* as composed by the following two statements:

```
:sp a foaf:Person ;
  foaf:name "Silvio Peroni" .
```

The addition of these statements also generates, at least, the following provenance statements, so as to set *sp* as a provenance entity, where its statements are implicitly encoded in a specific snapshot:

```
:entity a prov:Entity .

:entity-snapshot-1 a prov:Entity ;
  prov:specializationOf :entity .
```

Then suppose the curator of such data will decide to split the full name of *sp* using two distinct properties, i.e. *foaf:givenName* and *foaf:familyName*, so as to remove the more generic *foaf:name*:

```
:sp a foaf:Person ;
  foaf:givenName "Silvio" ;
  foaf:familyName "Peroni" .
```

In this case, a new snapshot of the entity will be generated, which specifies which statements have been added/deleted (by means of the property *new:hasUpdateQuery*) starting from the previous snapshot linked through the property *prov:wasDerivedFrom*, as shown as follows:

```
:entity-snapshot-2 a prov:Entity ;
  prov:specializationOf :entity ;
  prov:wasDerivedFrom :entity-snapshot-1 ;
  new:hasUpdateQuery "INSERT DATA { :sp foaf:givenName 'Silvio' ; foaf:familyName '
       Peroni' } ; DELETE DATA { :sp foaf:name 'Silvio Peroni' }" .
```

Using such snapshot-oriented structure, which clearly indicates how a previous snapshot of an entity has been modified to reach the set of statements currently available, makes easier to:

- retrieve the current statements of the entity, since they are those currently available in the dataset;
- restore the entity to a certain snapshot s_i by applying the inverse operations (i.e. deletions instead of insertions and vice versa) of all the update queries from the most recent snapshot s_n to s_{i+1}.

For instance, to get back to the status recorded by the first snapshot of the aforementioned example, we can just to run all the inverse operations of the update query indicated in the second snapshot, i.e.:

```
INSERT DATA { :sp foaf:name 'Silvio Peroni' } ;
DELETE DATA { :sp foaf:givenName 'Silvio' ; foaf:familyName 'Peroni' }
```

4 A real application: the OpenCitations Corpus

The OCC has been accompanied by a formal metadata model [9] which is strictly followed by all the data in the corpus. The metadata model is explicitly aligned with the SPAR Ontologies [6] for expressing the data and to other standard vocabularies, e.g. PROV-O [4] and PROV-DC [2], for expressing contextual information of entities, such as provenance data. All the ontological entities introduced by the metadata model are conveniently grouped together in the OpenCitations Ontology (OCO)[7], which also implements the *oco:hasUpdateQuery* for keeping track of changes as described in Section 3. The entities included in the corpus can have one of the following types:

- **bibliographic resource** (br), class `fabio:Expression` – a resource that either cites or is cited by other bibliographic resources (e.g. journal articles), or that contains such citing/cited resources (e.g. journals);
- **resource embodiment** (re), class `fabio:Manifestation` – details of the physical or digital form in which the bibliographic resource is made available by their publishers;
- **bibliographic entry** (be), class `biro:BibliographicReference` – the literal textual bibliographic entry occurring in the reference lists within the bibliographic resource, that references another bibliographic resource;
- **responsible agent** (ra), class `foaf:Agent` – an agent having certain roles with respect to the bibliographic resource;
- **agent role** (ar), class `pro:RoleInTime` – a role held by an agent with respect to the bibliographic resource (e.g. author, editor, publisher);
- **identifiers** (id) (class `datacite:Identifier`) – an external identifier (e.g. DOI, ORCID, PubMedID) associated with the bibliographic entity.

Each OCC entity is identified by a URL (e.g. `https://w3id.org/oc/corpus/br/525205`) that includes a two-letter short name for the class of such entity (e.g. "br" for bibliographic resources) and the number (e.g. "525205") that uniquely identifies it among the resources of the same type. Independently from the particular type assigned to entities, they have associated provenance information such as those introduced in Section 3. In particular, we record four different kinds of provenance entities, as indicated in [9]:

- *snapshot of entity metadata* (short: *se*) – a particular snapshot recording the metadata associated with an individual entity at a particular time;.
- *curatorial activity* (short: *ca*) – a curatorial activity relating to that entity, where possible activities are:
  1. creation, i.e. the activity of creating a new entity and of associating new metadata with it, within the corpus;
  2. modification, i.e. the activity of modifying (adding/removing) the metadata associated with an existing entity, or even of deprecating the entire entity;
  3. merging, i.e. the activity of unifying the metadata relating to two different OCC bibliographic entity descriptions, if they actually represent the same thing. This can result in the deprecation of one of the corpus entries in favour of the other one.

[7] `http://w3id.org/oc/ontology`

- *provenance agent* (short: *pa*) – the agent, such as a person, organisation or process, that creates or modifies entity metadata, or that is used as source provider of those metadata (e.g. Crossref);
- *curatorial role* (short: *cr*) – a particular role held by a provenance agent with respect to a curatorial activity (e.g. OCC curator, metadata source).

All these information are stored in the provenance graph related to the particular entity in consideration. The URL of such provenance graph is the URL of the entity in consideration plus "/prov/". The URL of all the aforementioned provenance entities (e.g. `https://w3id.org/oc/corpus/br/525205/prov/se/1`) is built using the provenance graph as base and adding two-letter short name for the class of such provenance entity (e.g. "se" for snapshot of entity metadata) and the number (e.g. "1") that uniquely identifies it among the resources of the same type in the context of that particular provenance graph. An exception to that URL template is provided for all the provenance agents, that are shared among the whole corpus and, thus, they have `https://w3id.org/oc/corpus/prov/pa/` as base URL (e.g. `https://w3id.org/oc/corpus/prov/pa/1`).

As an example, let us discuss the provenance statements added during the creation and modification of `https://w3id.org/oc/corpus/br/525205` – that are all accessible online. After the creation, the following statements are added to the corpus:

```
# Snapshot of entity metadata
<https://w3id.org/oc/corpus/br/525205/prov/se/1> a prov:Entity ;
  rdfs:label "snapshot of entity metadata 1 related to bibliographic resource
       525205 [se/1 -> br/525205]" ;
  prov:generatedAtTime "2016-08-08T22:25:48"^^xsd:dateTime ;
  prov:hadPrimarySource <http://api.crossref.org/works/10.2196/mhealth.5331> ;
  prov:specializationOf <https://w3id.org/oc/corpus/br/525205> ;
  prov:wasGeneratedBy <https://w3id.org/oc/corpus/br/525205/prov/ca/1> .

# Curatorial activity
<https://w3id.org/oc/corpus/br/525205/prov/ca/1> a prov:Activity, prov:Create ;
  rdfs:label "curatorial activity 1 related to bibliographic resource 525205 [ca/1
       -> br/525205]" ;
  dcterms:description "The entity 'https://w3id.org/oc/corpus/br/525205' has been
       created." ;
  prov:qualifiedAssociation
    <https://w3id.org/oc/corpus/br/525205/prov/cr/1> ,
    <https://w3id.org/oc/corpus/br/525205/prov/cr/2> .

# Curatorial roles
<https://w3id.org/oc/corpus/br/525205/prov/cr/1> a prov:Association ;
  rdfs:label "curatorial role 1 related to bibliographic resource 525205 [cr/1 ->
       br/525205]" ;
  prov:agent <https://w3id.org/oc/corpus/prov/pa/1> ;
  prov:hadRole oco:occ-curator .

<https://w3id.org/oc/corpus/br/525205/prov/cr/2> a prov:Association ;
  rdfs:label "curatorial role 2 related to bibliographic resource 525205 [cr/2 ->
       br/525205]" ;
  prov:agent <https://w3id.org/oc/corpus/prov/pa/2> ;
  prov:hadRole oco:source-metadata-provider .

# Provenance agents
<https://w3id.org/oc/corpus/prov/pa/1> a prov:Agent ;
  rdfs:label "provenance agent 1 [pa/1]" ;
  foaf:name "SPACIN CrossrefProcessor" .

<https://w3id.org/oc/corpus/prov/pa/2> a prov:Agent ;
  rdfs:label "provenance agent 2 [pa/2]" ;
  ns1:name "Crossref" .
```

Basically, the first snapshot of the resource `br/525205` has been created on August 8, 2016, at 22:25:48 (property *prov:generatedAtTime*), starting from the

data retrieved in the source document `http://api.crossref.org/works/10.2196/mhealth.5331` (property *prov:hadPrimarySource*). The activity that generates the data of `br/525205` (property *prov:wasGeneratedBy*) was a creation activity (class *prov:Create*) that involved (property *prov:qualifiedAssociation*) two agents (referred by the property *prov:agent*), i.e. SPACIN CrossrefProcessor (that is one of the automatic scripts of OpenCitations responsible for the creation of RDF data) and Crossref, as OCC curator and source metadata provider respectively.

Then, few days after its creation, the resource `br/525205` has been extended with additional data concerning its citation links to other bibliographic resources, as well as the completion of the full textual references it includes. The following provenance statements have been, thus, generated:

```
# The old snapshot has been invalidated...
<https://w3id.org/oc/corpus/br/525205/prov/se/1>
  prov:invalidatedAtTime "2016-08-29T22:42:06"^^xsd:dateTime ;
  prov:wasInvalidatedBy <https://w3id.org/oc/corpus/br/525205/prov/ca/2> .

# ... and it was substituted by a new one
<https://w3id.org/oc/corpus/br/525205/prov/se/2> a prov:Entity ;
  rdfs:label "snapshot of entity metadata 2 related to bibliographic resource
       525205 [se/2 -> br/525205]" ;
  prov:generatedAtTime "2016-08-29T22:42:06"^^xsd:dateTime ;
  prov:hadPrimarySource <http://www.ebi.ac.uk/europepmc/webservices/rest/PMC4911509
       /fullTextXML> ;
  prov:specializationOf <https://w3id.org/oc/corpus/br/525205> ;
  prov:wasDerivedFrom <https://w3id.org/oc/corpus/br/525205/prov/se/1> ;
  prov:wasGeneratedBy <https://w3id.org/oc/corpus/br/525205/prov/ca/2> ;
  oco:hasUpdateQuery "INSERT DATA { GRAPH <https://w3id.org/oc/corpus/br/> { <https
       ://w3id.org/oc/corpus/br/525205> <http://purl.org/spar/cito/cites> <https://
       w3id.org/oc/corpus/br/1095459> . <https://w3id.org/oc/corpus/br/525205> <
       http://purl.org/vocab/frbr/core#part> <https://w3id.org/oc/corpus/be/727491>
        . <https://w3id.org/oc/corpus/br/525205> <http://purl.org/vocab/frbr/core#
       part> <https://w3id.org/oc/corpus/be/727452> ... } }" .

# Curatorial activity
<https://w3id.org/oc/corpus/br/525205/prov/ca/2> a prov:Activity, prov:Modify ;
  rdfs:label "curatorial activity 2 related to bibliographic resource 525205 [ca/2
       -> br/525205]" ;
  dcterms:description "The entity 'https://w3id.org/oc/corpus/br/525205' has been
       extended with citation data." ;
  prov:qualifiedAssociation
    <https://w3id.org/oc/corpus/br/525205/prov/cr/3> ,
    <https://w3id.org/oc/corpus/br/525205/prov/cr/4> .

# Curatorial roles
<https://w3id.org/oc/corpus/br/525205/prov/cr/3> a prov:Association ;
  rdfs:label "curatorial role 3 related to bibliographic resource 525205 [cr/3 ->
       br/525205]" ;
  prov:agent <https://w3id.org/oc/corpus/prov/pa/1> ;
  prov:hadRole oco:occ-curator .

<https://w3id.org/oc/corpus/br/525205/prov/cr/4> a prov:Association ;
  rdfs:label "curatorial role 4 related to bibliographic resource 525205 [cr/4 ->
       br/525205]" ;
  prov:agent <https://w3id.org/oc/corpus/prov/pa/2> ;
  prov:hadRole oco:source-metadata-provider .
```

The new snapshot has substituted the previous one (properties *prov:invalidatedAtTime* and *prov:wasInvalidatedBy*) by updating the information about the resource `br/525205` with those indicated by the property *oco:hasUpdateQuery*. The new snapshot has been created by a particular modification activity (class *prov:Modify*) that involved the same agents with the same roles as before.

5 Conclusions

In this paper we have introduced a new approach for keeping track of changes in RDF data and, consequently, in LOD datasets. The method proposed is actually derived from existing techniques applied to the Document Engineering domain for addressing similar issues. We have also described as this approach has been implemented within the OpenCitations Project, and it is used as the main mechanism for providing a complete history of how the entities in the OpenCitations Corpus have evolved in time. In the future, we plan to develop automatic tools that allow us to restore a particular snapshot of an entity by looking at its provenance information only, so as to facilitate the restoration of entities at a particular time.

References

1. Carroll, J. J., Bizer, C., Hayes, P., & Stickler, P. (2005). Named graphs. Web Semantics: Science, Services and Agents on the World Wide Web, 3(4): 247–267. http://dx.doi.org/10.1016/j.websem.2005.09.001
2. Garijo, D., Eckert, K. (2013). Dublin Core to PROV Mapping. W3C Working Group Note, 30 April 2013. https://www.w3.org/TR/prov-dc/
3. Guarino, N., Welty, C. A. (2009). An Overview of OntoClean. In Handbook on Ontologies: 201–220. Berlin, Germany: Springer. ISBN: 978-3-540-70999-2
4. Lebo, T., Sahoo, S., McGuinness, D. (2013). PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. World Wide Web Consortium. http://www.w3.org/TR/prov-o/
5. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C. (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6 (2): 167–195. http://dx.doi.org/10.3233/SW-140134
6. Peroni, S. (2014). The Semantic Publishing and Referencing Ontologies. In Semantic Web Technologies and Legal Scholarly Publishing: 121–193. http://dx.doi.org/10.1007/978-3-319-04777-5_5
7. Peroni, S., Dutton, A., Gray, T., Shotton, D. (2015). Setting our bibliographic references free: towards open citation data. Journal of Documentation, 71 (2): 253–277. http://dx.doi.org/10.1108/JD-12-2013-0166
8. Peroni, S., Poggi, F., Vitali, F. (2014). Overlapproaches in documents: a definitive classification (in OWL, 2!). In Proceedings of Balisage 2014. http://dx.doi.org/10.4242/BalisageVol13.Peroni01
9. Peroni, S., Shotton, D. (2016). Metadata for the OpenCitations Corpus. Figshare. https://dx.doi.org/10.6084/m9.figshare.3443876
10. Peroni, S., Shotton, D., Vitali, F. (2016). Freedom for bibliographic references: OpenCitations arise. To appear in Proceedings of 2016 International Workshop on Linked Data for Information Extraction (LD4IE 2016). https://w3id.org/oc/paper/occ-lisc2016.html
11. Vrandecic, D., Krötzsch, M. (2014). Wikidata: a free collaborative knowledge base. Communication of the ACM, 57 (10): 78–85. http://dx.doi.org/10.1145/2629489
12. Welty, C. A., Fikes, R. (2006). A Reusable Ontology for Fluents in OWL. In Proceedings of FOIS 2006: 226–236.