A document-inspired way for tracking changes of RDF data

The case of the OpenCitations Corpus

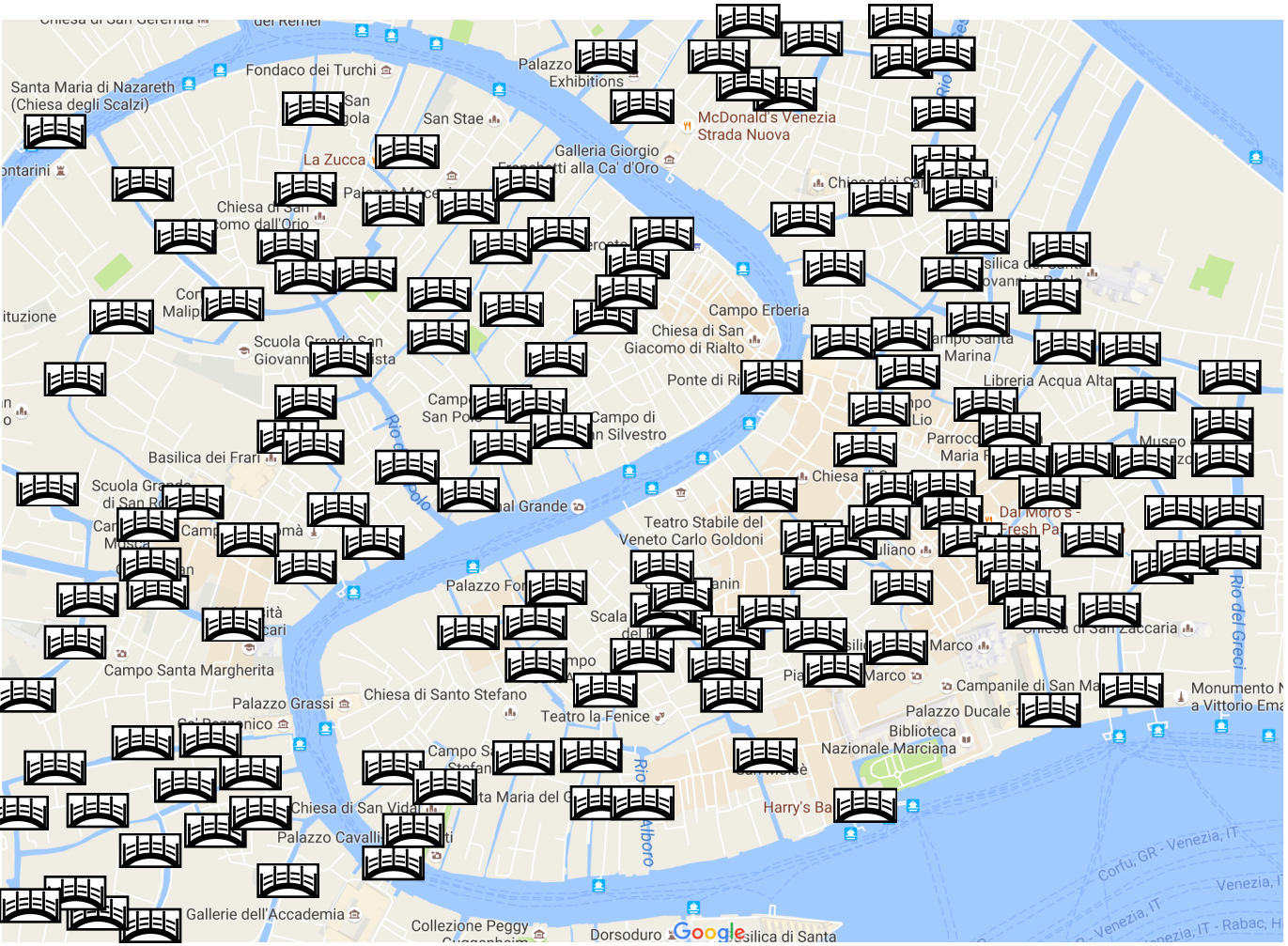
Paper: `https://w3id.org/oc/paper/occ-driftalod2016.html`

Silvio Peroni, *David Shotton, Fabio Vitali*

1st Drift-a-LOD Workshop: Detection, Representation and Management of Concept Drift in Linked Open Data
Bologna, Italy, November 20, 2016

The Venice analogy

https://w3id.org/oc/paper/the-venice-analogy.html



- Island = scholarly publication
- Bridge = citation
- Current situation:
  - local travel to the next island is permitted
  - unrestricted travel over the entire network of bridges requires an expensive season ticket
  - general populace is excluded

Opening the bridges

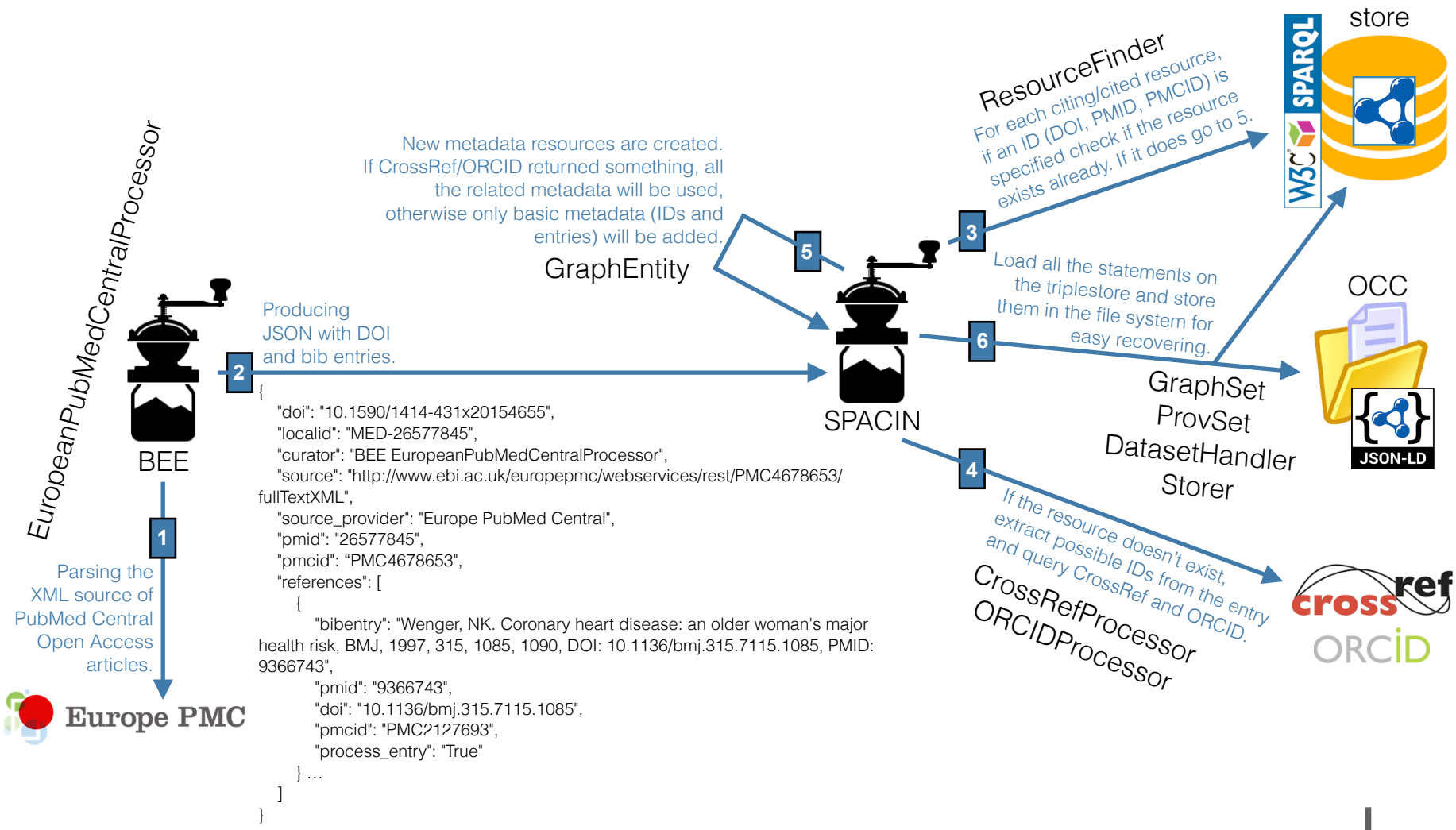
- **What** – Citation data are one of the main tools used by researchers to gain knowledge about particular topics, and they also serve institutional goals, for example in research assessment
- **Problem** – The most authoritative databases of citation data, Scopus and Web of Science, can only be accessed by paying significant annual access fees
  - The University of Bologna pays about 6,000,000 euros per year for accessing to digital bibliographic resources
- **Solution** – To create a citation database that freely and legally makes available citation data in an open repository to assist scholars with their academic studies and serve knowledge to the wider public

OpenCitations

http://opencitations.net

- The OpenCitations Project aims at creating an open repository of scholarly citation data – the OpenCitations Corpus (OCC) – made available under a Creative Commons public domain dedication to provide in RDF accurate citation information (bibliographic references) harvested from the scholarly literature
  - All scripts are released with Open Source ISC Licence and available on GitHub at `http://github.com/essepuntato/opencitations`
- Currently processing papers available in the PubMedCentral Open Access subset
- As of November 20, 2016 the OCC contains **2,076,645 citation links**
- Six distinct kinds of bibliographic entities
  - **bibliographic resources** (citing/cited articles, journals, books, proceedings, etc.)
  - **resource embodiments** (format information about bibliographic resources)
  - **bibliographic entries** (literal textual entries occurring in the reference lists)
  - **responsible agents** (agents having certain roles with respect to the bibliographic resources)
  - **agent roles** (author, editor, publisher);
  - **identifiers** (DOI, ORCID, PubMedID, URL, etc.)

Ingestion workflow



Issues with data

- Automatic workflow built upon external services
  - Efficient, but no human check of the data extracted
  - Some errors could be propagated
- Data do change in time
  - Information can be incomplete (e.g. citations added in another iteration of the ingestion workflow)
  - Information can be wrong (e.g. circular citations – paper A cites paper A)
  - Information can be ambiguous (e.g. author disambiguation)

Document-inspired data drift

- Inspiration from the Document Engineering domain
  - Well-known structure for keeping track of changes in word-processor documents, e.g. OpenOffice Writer
    - New content added directly within the existing text and marked in some way
    - Removed context moved out from the actual content of the document and placed in an auxiliary space for easy retrieving and restoration
  - Two basic operations (add & remove) are enough for keeping track of document changes
- Solution for RDF data: using PROV-O + SPARQL UPDATE (INSERT DATA and DELETE DATA only) for keeping track of the way entities change in time

The approach

Time **Data** **Provenance**

T_1

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

```
:sp a prov:Entity .

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

T_2

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

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

A **snapshot** records the composition of the entity it specialises (i.e. the set of statements using such entity as subject) at a fixed point in time

Advantages

- Easy to retrieve the current statements of an entity, since they are those currently available in the dataset
- It is possible to restore the entity to a certain snapshot $\mathbf{s_i}$ by applying the inverse operations (i.e. deletions instead of insertions and vice versa) of all the update queries from the most recent snapshot $\mathbf{s_n}$ to $\mathbf{s_{i+1}}$
  - For instance, to get back to the status recorded by the first snapshot of the previous example, we have to run all the inverse operations of the update query specified in the second snapshot:

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

Implementation in the OCC

- We use:
  - PROV-O
  - PROV-DC, an extension of PROV-O mapping it with DC
  - OpenCitations Ontology (OCO), which defines **oco:hasUpdateQuery**
- Each entity in the OCC tracks provenance information about:
  - snapshot of entity metadata (**prov:Entity**), a particular snapshot recording the metadata associated with an individual entity at a particular time
  - curatorial activity (**prov:Activity**), a curatorial activity relating to that entity
    - creation (**prov:Create**), the activity of creating a new entity with statements
    - modification (**prov:Modify**), the activity of adding/removing statements of an entity
    - merging (**prov:Replace**), the activity of unifying the statements relating to two entites
  - provenance agent (**prov:Agent**), a person, organisation or process, that is involved in some way in the creation of an entity (e.g. Crossref)
  - curatorial role (**prov:Association**), a particular role held by a provenance agent with respect to a curatorial activity (e.g. OCC curator, metadata source)

An example

Time **Data** **Provenance**

T_1

```
br:525205
  a fabio:Expression , fabio:JournalArticle ;
  dcterms:title "The Electronic Patient ..." ;
  datacite:hasIdentifier
    id:816997 , id:816998 , ... ;
  fabio:hasPublicationYear "2016"^^xsd:gYear ;
  pro:isDocumentContextFor
    ar:1591190 , ar:1591191 , ... ;
  frbr:embodiment re:217773 .
```

```
se:1 a prov:Entity ;
  prov:generatedAtTime "2016-08-08T22:25:48"^^xsd:dateTime ;
  prov:hadPrimarySource
    <http://api.crossref.org/works/10.2196/mhealth.5331> ;
  prov:specializationOf br:525205 ;
  prov:wasGeneratedBy ca:1 .
ca:1 a prov:Activity, prov:Create ;
  dcterms:description
    "The entity 'https://w3id.org/oc/corpus/br/525205'
     has been created." ;
  prov:qualifiedAssociation cr:1 , cr:2 .
cr:1 a prov:Association ;
  prov:agent pa:1 ;
  prov:hadRole oco:occ-curator .
pa:1 a prov:Agent ;
  foaf:name "SPACIN CrossrefProcessor" . …
```

T_2

```
br:525205
  cites:cites br:1095420 , br:1095421 , ... ;
  frbr:part be:727446 , be:727447 , ... .
```

```
se:1
  prov:invalidatedAtTime "2016-08-29T22:42:06"^^xsd:dateTime ;
  prov:wasInvalidatedBy ca:2 .
se:2 a prov:Entity ;
  prov:generatedAtTime "2016-08-29T22:42:06"^^xsd:dateTime ;
  prov:hadPrimarySource <http://www.ebi.ac.uk/europepmc/
webservices/rest/PMC4911509/fullTextXML> ;
  prov:specializationOf br:525205 ;
  prov:wasDerivedFrom se:1 ;
  prov:wasGeneratedBy ca:2 ;
  oco:hasUpdateQuery "INSERT DATA {
    GRAPH <https://w3id.org/oc/corpus/br/> {
      br:525205
        cito:cites br:1095459 , br:525205 , ... ;
        frbr:part be:727491 , be:727452 , ... } }" . …
```

Conclusions

- Approach for keeping track of changes in RDF data inspired by existing implementations in the Document Engineering domain
  - PROV-O
  - SPARQL UPDATE (INSERT/DELETE DATA)
- Implementation: provenance data in the OpenCitations Corpus
  - PROV-DC
  - OpenCitations Ontology
- Future works
  - automatic tools for snapshot restoration
  - handling non-automatic modification to the OpenCitations Corpus (e.g. curation by humans)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Silvio Peroni

silvio.peroni@unibo.it - @essepuntato

Digital And Semantic Publishing Laboratory
Department of Computer Science and Engineering
Alma Mater Studiorum – Università di Bologna
Bologna, Italy

dasplab.cs.unibo.it – www.unibo.it