Provenance and Annotations for Linked Data

Dublin Core Conference 2013, Lisbon
Kai Eckert
Overview

1. Motivating Example: Digitised Manuscripts to Europeana (DM2E)
2. Linked Data Publishing
3. Linked Data Provenance
4. Triple Identity → Provenance Context
5. DCAM and Linked Data
6. Linked Data Annotations
Linked Data Provenance
Digitised Manuscripts to Europeana

WP 1

WP 2

WP 3
The Linked Data Gap

- Linked Data publication is often one-way.
- Linked Data as an export from the „real“ data.
- Linked Data as a source for new data.
- The connection easily gets lost!
Linked Data Publishing
Linked Data Provenance

Various options and best practices:

– Reification to provide provenance of single statements,
– URL of RDF data (Web documents) as subject for provenance statements,
– Link Web documents to a larger dataset (VoID vocabulary), provide provenance for this dataset.

→ It is easy to provide some provenance.

But how is the provenance preserved in client systems?
Too many options?

- Web document URLs can be preserved as named graphs in a local triple store.
- So can VoID datasets.
- VoID datasets can be nested...
Triple Identity

• Several sources can make the same statement. No distinction within RDF.

• Statements (Triples) can be retrieved and become part of a new dataset.

→ A statement has **no identity**.
Can we establish triple identity?
What determines identity?

• Philosophical Question.
• Proposed Answer:

The provenance of a resource determines its identity.

„If we want to preserve the identity of the statements in our data, we have to preserve their provenance.“
Provenance Context

• One of our nested graph boundaries (hopefully) was created to provide provenance information.

• To enable preservation of triple identity, we indicate the Provenance Context.
“**Provenance Context** is a set of RDF triples that share the same provenance, identified by a URI.”

- Web documents (foaf:Document)
- Named Graphs
- ORE Resource Maps (ore:ResourceMap)
- VoID Datasets (void:Dataset)
- ...
1. Per default, the Provenance Context of a triple is the document identified by the URL it is retrieved from or the Named Graph that contains the statement.

2. If the document or the Named Graph is related to a void:Dataset via void:inDataset, the Provenance Context is the void:Dataset.

3. The Provenance Context can be stated explicitly using the property dm2e:inProvenanceContext.
Consequences

→ There must always be one and only one Provenance Context for each statement.

→ Every RDF graph either is a Provenance Context or it is contained completely within one Provenance Context.

→ The Provenance Context determines the maximum permissible set of RDF statements that are published together.
The Provenance Context in DCAM

- Description Set: Logical Boundary that creates identity.
- Record: Physical embodiment of a Description Set.
DCAM and Linked Data

• Any RDF publication is a Record containing a Description Set.
• These Description Sets are part of a larger Description Set, the Provenance Context.
Linked Data Annotations
Annotating Triples

Idea: Use an XPointer-style way to point to statements within a Provenance Context.

<scheme name>:<hierarchical part>[?<query>][#<fragment>]

Fragment: spo=subject,predicate,object
Example


→

Statement:

<http://example.org/data/doc1>
  <http://purl.org/dc/terms/creator>
  <http://example.org/persons/kai>.

within the Provenance Context:

<http://example.org/provcontext1>
What does this mean?

- The fragment URIs can be created and interpreted on the fly.
- But semantics in the URI are an anti-pattern.
- So let’s explain what the URI represents.

No semantics in the URL!
Contextual Reification

Dereferencing the URI explains the meaning. It is a Statement (Reification), connected to a specific Provenance Context.
Conclusion

- Provenance-tracking for data requires data identity.
- For the preservation of data identity, we need guidance.
- The Provenance Context abstracts from technical details and indicates the boundary that defines data identity.
- Furthermore, we can use it to connect statements about statements (annotations) to a concrete context.
- Technical issues (length!) with the fragment URIs still have to be investigated.
→ Provenance-tracking for data requires data identity.

→ For the preservation of data identity, we need guidance.

→ The Provenance Context abstracts from technical details and indicates the boundary that defines data identity.

→ Furthermore, we can use it to connect statements about statements (annotations) to a concrete context.

→ Technical issues (length!) with the fragment URIs still have to be investigated.

Thank you.