Stratifying Semantic Data for Citation and Trust: an Introduction to RDFDF

Dario De Nart | Dante Degl'Innocenti | Marco Peressotti | and Carlo Tasso

CITING DATA

- Data is becoming an increasingly critical asset in research.
  - Big Data
  - Data Science
  - Linked Open Data
- The quality of research depends on the quality of data.
- A precise indication of what data was used and how it was collected can improve research quality.
  - Reproducibility

DATA CITATION REQUIREMENTS

- **Who and How**: who authored the data and what process generated them (e.g.: field test, crowdsourcing, bootstrapping, …).
- **What:** what data was used (what data set, which subset of a dataset, …).
- **When:** which version/revision of the data was used.

Plenty of metadata is needed!

DATA CITATION REQUIREMENTS

One more, fundamental feature:

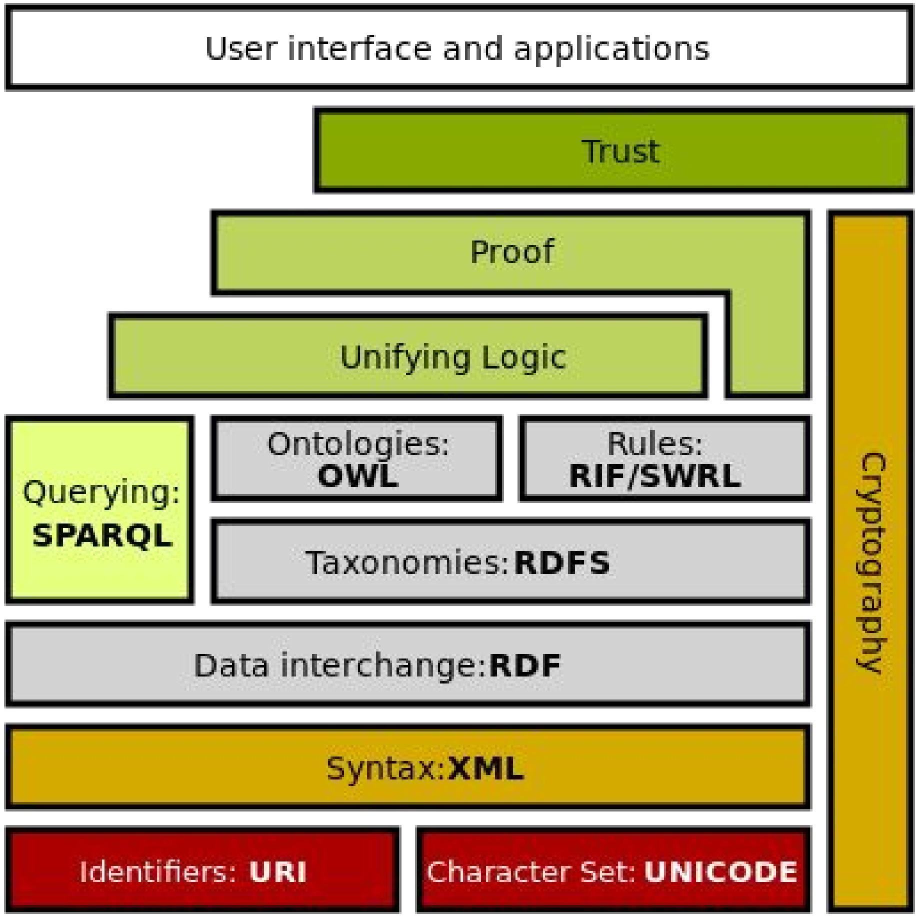
- **Feasibility**: data citation must be resolved in practical time.

In other words, getting the cited data should be computable and reasonably efficient.

LINKED DATA

- Linked Data are becoming the standard for open, machine readable data.
  - The Web of Data is constantly expanding.
- Relying on Semantic Web technologies
  - URIs
  - RDF
  - RDFS
  - OWL
  - ...

THE SEMANTIC WEB STACK



User interface and applications

Trust

Proof

Unifying Logic

Querying: **SPARQL**

Ontologies: **OWL**

Rules: **RIF/SWRL**

Taxonomies: **RDFS**

Data interchange: **RDF**

Syntax: **XML**

Identifiers: **URI**

Character Set: **UNICODE**

Cryptography

CITING LINKED DATA

- To cite Linked Data we have to provide a precise reference to a portion of the Web of Data.
  - How to identify authorship?
  - How to identify the version of the data?
  - How to identify the data set?
- Data citation should take us to the cited data in practical time.
  - Dereferencing the citation must be computable.

WHO AND HOW

- Expressing and tracking **who and how** edited linked data is a well known problem.
- **Data Provenance** has been investigated by the community:
  - Many vocabularies available (e.g.: provONT)
  - Methodologies and best practices
- Once we have a URI identifying data we can easily express provenance information in RDF.

WHEN

- Data may change over time: we must provide a reference to a precise dataset **version** to ensure reproducibility.
- **Versioning** is a common problem in Engineering.
    - Versioning systems;
    - OWL provides out of the box versioning properties.
- Attaching versioning information to a data set identified by a URI is easy.

BUT WHAT DATA?

- Coarse grained citation
  - Citing a whole dataset (e.g.:dbpedia);
  - datasets as a whole can be identified by their base URI, easy to find and authoritative;
  - Seldom the whole data set is needed.
- Fine grained citation
  - Data subset or even Triple level;
  - More realistic scenario;
  - Non trivial problem: quad semantics is needed;
  - *Computability issues.*

QUAD SEMANTICS and REIFICATION

RDF allows assignment of identifiers to single triples

- **Quad Semantics**: shifting from triples made of subject, predicate, and object to quadruples made of subject, predicate, object, and *identifier.*
  - Officially part of the language since RDF 1.1 (2014).
- **Reification**: the technique that allows assignment of an identifier to a triple, thus allowing quad semantics.
  - Cumbersome, plus makes data hard to read.

QUAD SEMANTICS and REIFICATION

Reification

x type statement

x subject *a*

x predicate *b*

x object *c*

Quad Semantics

Subject	Predicate	Object	Identifier
a	b	c	x

x → a, b, c

To express a quadruple four triples are needed!

INFORMATION AND METAINFORMATION

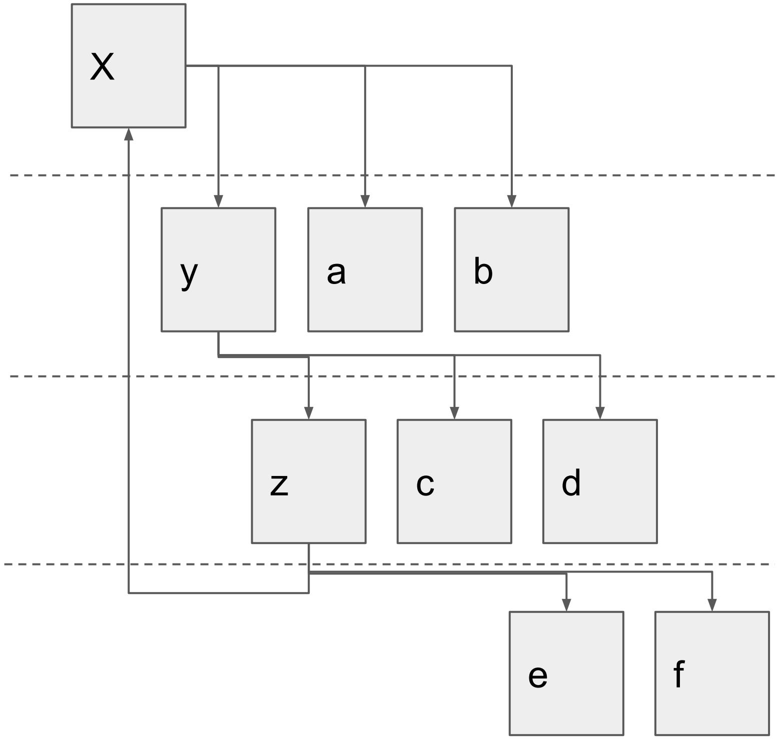
- Using quad semantics the *subject-predicate-object* triple is **information** and the *identifier* **metainformation**.
- Metainformation ideally stays at a higher level of abstraction.
  - We can identify an order relation.
- $x \rightarrow a, b, c$ implies that x is metainformation for a, b, c therefore $x > a, b, c$.
  - there can be several levels of metainformation.
- We assume that high level identifiers get cited.
  - Can we get to the **information** ?

NOPE...

x→y, a, b

y→z, c, d

z→x, e, f

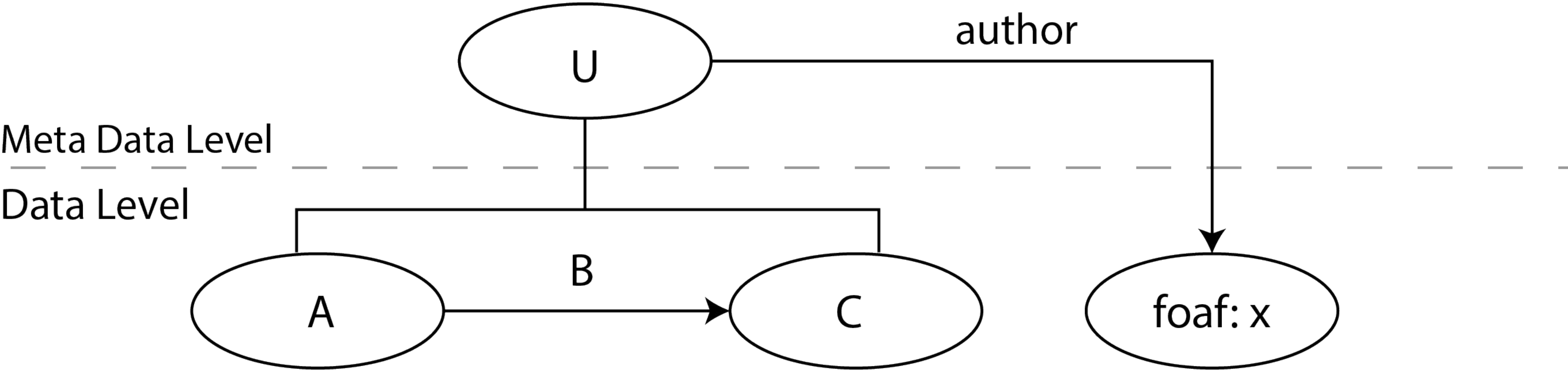


- Is *x* information or metainformation ?
- Infinite chains prevent discrimination of information and metainformation.

WELL STRATIFIED DATA

- To retrieve the information there must be no infinite chains.
- We call data with no such chains **well stratified.**
- Well stratification makes data citations computable.
- To make data citable well-stratification must be guaranteed.

WELL STRATIFIED DATA



- It is possible to draw a line that separates information from metainformation.
- There can be multiple levels of metainformation.

HOW TO GUARANTEE WELL STRATIFICATION?

- RDF, RDFS, and OWL provide no means.
- Good knowledge engineering practices may help, but are not 100% failproof.
    - plus assessing the actual absence of loops is still an issue.
- We propose an extension of the RDF language that allows efficient checking of well-stratification.

RDF Description Framework

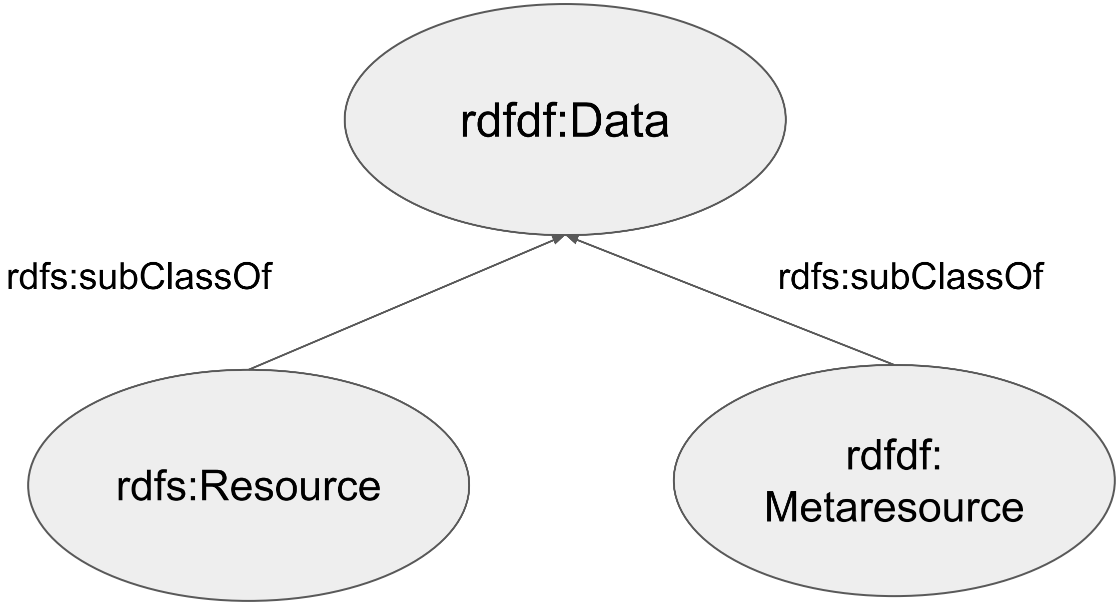
- Quadruples instead of triples.

  (subject, predicate, object, identifier)

- Has a **type system** with two types:
  - Well stratified data;
  - Ill stratified data.
- The fourth element is always of *rdf:type metaresource*.
  - Does not interfere with OWL and RDFS semantics
  - Considered for type checking

RESOURCES, METARESOURCES, AND DATA

- *Metaresource* is a sibling class of *rdfs:Resource*
- They are both subclasses of a more general class called *Data*.
- A URI can be both resource and metaresource.



IDENTIFIER AND REIFICATION

- The fourth element is an implicit reification.
  - a, b, c, x is syntactic sugar for:
  - x, type, Statement
  - x, type, Metaresource
  - x, subject, a
  - x, predicate, b
  - x, object, c.
- Data-level reification is allowed, but is not considered metainformation.

TYPE CHECKING

- We are not considering *rdf:type* properties for **typing** since they actually provide **classification**.
- If *x* is metainformation for *y*, then *Γ(x) > Γ(y).*
- Three typing rules (void dataset, triple, and union) to identify well stratified data:

$$\frac{}{\Gamma \vdash \varnothing : \checkmark}$$

$$\frac{\Gamma(x) > \Gamma(a) \quad \Gamma(x) > \Gamma(b) \quad \Gamma(x) > \Gamma(c)}{\Gamma \vdash x \mapsto (a, b, c) : \checkmark}$$

$$\frac{\Gamma_1 \vdash n_1 : \checkmark \quad \Gamma_2 \vdash n_2 : \checkmark \quad \Gamma = \Gamma_1 \sqcup \Gamma_2 \quad n = n_1 \sqcup n_2}{\Gamma \vdash n : \checkmark}$$

CONCLUSIONS

- We provided a theoretical framework for ensuring computability of Linked Data citation.
    - These ideas can be implemented leveraging on techniques well established in programming languages.
- RDFDF could be a new level in the Semantic Web stack.
    - Well stratification may be linked data’s sixth star.
- We are open to discussion and contribution.

THANKS FOR YOUR ATTENTION