Linked Data as Part of the Big Data Landscape

Pascal Hitzler

Data Semantics Laboratory (DaSe Lab)

Data Science and Security Cluster (DSSC)

Wright State University

http://www.pascal-hitzler.de

Big Data Vs

Volume	-	High Performance Computing
Velocity	-	Internet of Things
Veracity	-	Social sciences, Humanities
Variety	-	Semantic Web

Variety

Core Semantic Web mostly concerned with Variety issues.

It is a generalization of the term “semantic heterogeneity”.

Variety can also be syntactic (e.g., different representation languages).

Linked Data Vs

Volume: not a major concern at the moment, but important in some corners, e.g. efficient query processing.

Velocity: In term of dealing, e.g., with sensor data. Efficiency of systems is an issue, but research also addresses the semantics of data with velocity (aka, streams).

Veracity: An aspect of Linked Data Quality research, and of research into the representation of uncertainty.

Variety: Linked Data overcomes syntactic heterogeneity issues. But it’s not so good on the semantic heterogeneity side.

Linked Data as Laboratory

Linked Data reduced Big Data variability by some of the scientifically less interesting dimensions.

- Syntactic issues vanish.
- Based on a relatively small set of conventions, e.g. the use of vocabularies (from SKOS, OWL, etc.)
- Can be accessed, stored, linked, queried, etc. by a set of (largely) compatible free and open source tools and systems on regular hardware.
- Linked Data is object-centric (this reduces the multi-mediality aspect).

Hence: Linked Data is a bit like *Big Data in a laboratory setting*.

I.e., studying linked data means addressing some central Big Data issues.

Linked Data Variety

Linked Data Variety challenges are still very substantial.

See e.g.

Integrated querying over multiple Linked Datasets.

- **If the Linked Datasets are previously unknown, then there are currently no systems which do a good job on this.**
- **Manual intervention (integration!) is still required.**

- **In fact, it's even troublesome to simply formulate SPARQL queries over a (new, large) Linked Dataset.**

Linked Data Quality

Difficulties in making practical use of Linked Data are sometimes attributed to *poor quality*.

Some quality dimensions fall under Veracity.

Most fall under Variety.

There is a large variance in perspectives and underlying data models

- Minimalistic agreed-upon requirements for Linked Data
- Community-generated, grassroots, datasets.

Linked Data Quality

In addition, there is a significant amount of genuine low-level quality issues:

- Erroneous data
- Missing data
- Triplification errors
- Misleading owl:sameAs statements (and other abuse of vocabulary with formal semantics)
- Faulty syntax
- Unavailable SPARQL endpoints.

We need to

We need to:

- Find technical and methodological solutions which perform well even under low-level quality issues.
- Start creating a culture of best practices in data publishing.
- Develop a way to communicate, illustrate, and document linked datasets effectively (for all stakeholders).

And as a very neglected issue:

- We need to establish a culture in which our claims are properly evaluated, e.g. claims that “proper” linked data should be structured in a specific way.
  For example, is there any work which shows that the links in Linked Data actually provide added value?

Semantic Heterogeneity

Linked Data as it currently exists does hardly address semantic heterogeneity.

Case in point: Take two new, large linked datasets (with highly overlapping domain) and try to integrate them. This is very hard, due to many factors which Linked Data does not address.

- Representational choices (subgraph shapes).
- Vocabulary with informal semantics (skos, foaf, rdf:label).
- Ambiguities (co-references need resolving).
- Ad-hoc schema (i.e., poor quality ontologies, if at all).
- Lack of links on the schema level.
- Etc.

Semantic Heterogeneity

Addressing Semantic Heterogeneity is in fact not the purpose of Linked Data (it seems).

Rather, Semantic Heterogeneity is about using (good!) ontologies.

Good ontologies (as schema which underlies Linked Data) helps with:

- Understanding (by a human) of the Linked Graph.
- Aligning different representational choices.
- Resolving naming ambiguities.
- Linking on the schema level
- Etc.

Call to Arms

Linked Data is part of the Big Data Landscape.

Linked Data addresses mostly syntactic variety issues.

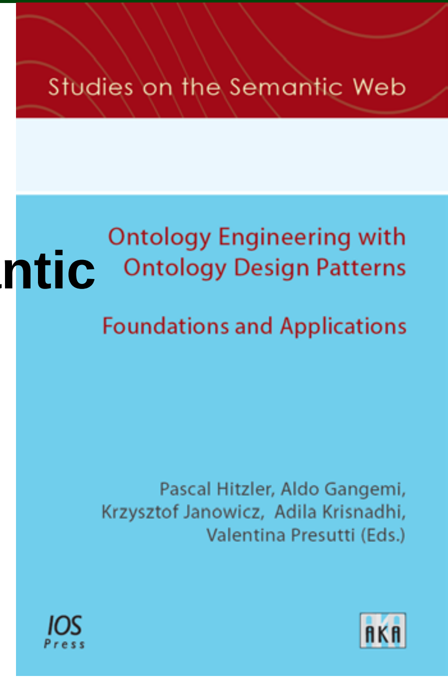
Ontologies (with Linked Data) address semantic variety issues.

Looking forward, we need to establish *evaluated* best practices for data sharing, integration, and reuse, based on both pillars of the current semantic web landscape.

Thanks!

References

- Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, Valentina Presutti (eds.), Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web. IOS Press/AKA Verlag, 2016. To appear.
- Krzysztof Janowicz, Frank van Harmelen, James A. Hendler, Pascal Hitzler, Why the Data Train Needs Semantic Rails. AI Magazine 26 (1), 2015, 5-14.
- Pascal Hitzler, Krzysztof Janowicz, Linked Data, Big Data, and the 4th Paradigm. Semantic Web 4 (2), 2013, 233-235.
- Victor Rodriguez-Doncel, Adila A. Krisnadhi, Pascal Hitzler, Michelle Cheatham, Nazifa Karima, Reihaneh Amini, Pattern-Based Linked Data Publication: The Linked Chess Dataset Case. In: Proceedings CoLD2015 at ISWC2015, Bethlehem, PA, 2015.



Studies on the Semantic Web

Ontology Engineering with
Ontology Design Patterns

Foundations and Applications

Pascal Hitzler, Aldo Gangemi,
Krzysztof Janowicz, Adila Krisnadhi,
Valentina Presutti (Eds.)

IOS Press

AKA

References

- Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, Charles Vardeman II, Five Stars of Linked Data Vocabulary Use. Semantic Web 5 (3), 2014, 173-176.
- Krzysztof Janowicz, Pascal Hitzler, Thoughts on the Complex Relation Between Linked Data, Semantic Annotations, and Ontologies. In: Proc. ESAIR 2013, ACM, San Francisco, 2013.