(Enterprise) Linked Data: What’s missing?

Axel Polleres, Sabrina Kirrane, Javier D. Fernández
The PROPEL project:

- 15 January 2015 year's Semantic Web meetup:
- Business Semantics & Enterprise Linked Data
- Various companies in Austria already making business & products with Linked Data, e.g.:

- But: Still a niche market... What's missing?
  - Addressing Technology gaps (security, time, efficient interchange, standard tool chain)

… Can we learn from industry? Shall we?
PROPEL
Propelling the Potential of Enterprise Linked Data in Austria

Goals:

- How big is the market? Market Analysis, requirements and use cases
- How ready is Linked Data as an integration paradigm for the Enterprise?
- Research and Development Roadmap, Exploratory Study, Technology Gaps
- Eventually: Models for an Enterprise Linked Data value chain
So, how to approach this?

- First, we better know what we can offer…
  - What is Enterprise Linked Data?
  - What is the Semantic Web?
  - Who are we?
Some common misunderstandings...

What's the difference between Enterprise Linked Data and... ?
Linked Data vs. “The” Semantic Web

Overlaps:

• "Linked Data is the Semantic Web done right" (Tim Berners-Lee)
• The actual Semantic Web is made up of Linked Data.
• Linked Data is based on Semantic web standards.

Key Differences:

• Semantic Web was all about "semantifying" the Web, Linked Data is based on Web standards (URIs, http), but doesn't center around Web pages.
• LD is a more pragmatic "bottom-up" approach.

• "Linked Data is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inference."

M. Hausenblas "Exploiting Linked Data For Building Web Applications"
IEEE Internet Computing, 2009
Linked Data vs. Open Data

Overlaps:

- Openness is a core principle in the design of LD
- Many Linked Data sets published under an open license → Linked Open Data and LD are often used interchangeably

Key differences:

- Linked Data technologies can be used without publishing data – e.g., for internal and external data integration.
- Probably not all open data will ever be linked (the majority will remain in formats such as csv, txt, json, xml etc.)
Linked Data vs Big Data

Overlaps:

- LD as a whole is big (*38,606,408,854 triples and counting!*)
- No rigid up-front (e.g., relational) data model
- Big Data technologies (e.g., Hadoop) are used to handle LD
- LD can represent knowledge extracted from big unstructured data

Key Differences:

- Individual linked data sets are typically not "big" per se (e.g., English DBpedia dump currently < 5 GB)
- LD is structured and semantically explicit, single format (RDF) "big data lakes" are typically neither - RDMBS, NoSQL/“Polyglott persistence”, non-core formats, unstructured textual+mmedia data…
- Big data based on distributed data infrastructures within an organization (e.g., Hadoop clusters), LD creates a decentralized, globally distributed data infrastructure
ELD vs. LED

Enterprise Linked Data (intra-enterprise):
Internal use of LD technologies within organizations, e.g.,
• to integrate heterogeneous systems at the data level
• for advanced content/knowledge/… management
• as a basis for innovative products and services

Linked Enterprise Data (inter-enterprise):
• Cross-organizational data integration
• Data markets and data ecosystems
• Decentralized infrastructure for a networked economy
Let’s take a step back...

- What are the expectations/requirements?
- What can we offer as a community?
Let’s take a step back…

- What are the expectations/requirements?
- What can we offer as a community?

Taking a business/application centric view
Interviews

23 interviews:

- **Domains**
  - Consulting, Engineering, Environment, Finance and Insurance, Government, Healthcare, ICT, IT, Media, Pharmaceutical, Professional Services, Real Estate, Research, Startup, Tourism, Transports & Logistics

- **Roles**
  - Business Intelligence, CEO, Chief Engineer, Data and Systems Architect, Data Scientist, Director Information Management, Enterprise Architect, Founder, General Secretary, Governance, Risk & Compliance Manager, Head of Communications and Media, Head of Development, Head of HR, Head of R&D, Innovation Manager, Information Architect, IT Project Manager, Management, Managing director, Marketing Analyst, Principle System Analyst, Project Coordinator, Researcher, Technical Specialist

Note: Instead of explaining them what ELD is, we gathered their technology/research expectations from a more general SW perspective
Technologies in need…

- Analytics
- Computational linguistics & NLP
- Concept tagging & annotation
- Data integration
- Data management
- Dynamic data / streaming
- Extraction, data mining, text mining, entity extraction
- Logic, formal languages & reasoning
- Human-Computer Interaction & visualization
- Knowledge representation
- Machine learning
- Ontology/thesaurus /taxonomy management
- Quality & Provenance
- Recommendations
- Robustness, scalability, optimization and performance
- Searching, browsing & exploration
- Security and privacy
- System engineering

We pretty much ended up in all areas that SW touches upon!
Standards

https://www.w3.org/standards/semanticweb/

In addition to the classic "Web of documents" W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.

Linked Data
The Semantic Web is a Web of data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. RDF provides the foundation for publishing and linking your data. Various technologies allow you to embed data in documents (RDFa, GRDDL) or expose what you have in SQL databases, or make it available as RDF files.

Vocabularies
At times it may be important or valuable to organize data. Using OWL (to build vocabularies, or "ontologies") and SKOS (for designing knowledge organization systems) it is possible to enrich data with additional meaning, which allows more people (and more machines) to do more with the data.

Query
Query languages go hand-in-hand with databases. If the Semantic Web is viewed as a global database, then it is easy to understand why one would need a query language for that data. SPARQL is the query language for the Semantic Web.

Inference
Near the top of the Semantic Web stack one finds inference — reasoning over data through rules. W3C work on rules, primarily through RIF and OWL, is focused on translating between rule languages and exchanging rules among different systems.

Vertical Applications
W3C is working with different industries — for example in Health Care and Life Sciences, eGovernment, and Energy — to improve collaboration, research and development, and innovation adoption through Semantic Web technology. For instance, by aiding decision-making in clinical research, Semantic Web technologies will bridge many forms of biological and medical information.
Standards Toolbox

- Constraint language
- Shapes Constraint Language (SHACL)
- Machine readable data
- RDF Quads, RDF N-Triples Trig ...
- Meta data
- PROV, PROV-O, PROV-N
- RDF Data Cube
- RDF Quads, RDF N-Triples Trig ...

Query language
- SPARQL

Update language
- SPARQL UPDATE

Rule language
- RIF

Transformation language
- RDB2RDF Direct Mapping R2RML

Vocabulary
- The Organization Vocabulary
- Time ontology in OWL
Standards Toolbox

(incl. W3C member submissions)

- Constraint language
- Shapes Constraint Language (SHACL)
- RDF Quads, RDF N-Triples
- Trig …
- HDT
- RDF Schema
- OWL
- SKOS
- PROV, PROV-O, PROV-N
- DCAT
- RDF Data Cube
- Multidimensional data
- Machine readable data
- Meta data
- Query language
- Update language
- SPARQL
- SPARQL UPDATE
- RIF
- N3
- SWRL
- WSML
- RDB2RDF
- Direct Mapping
- R2RML
- Rule language
- Transformation language
- Translation language
- GRDDL
- XSPARQL
- The Organization Ontology
- Time ontology in OWL
- SIOC
- …

Web Services Description
- SA-WSDL
- WSMO
- OWL-S
- METEOR-S

Schema/ Ontology/ Concepts/ Taxonomy
- RDF Schema
- OWL
- SKOS

Web Services Description
- SA-WSDL
- WSMO
- OWL-S
- METEOR-S

Multidimensional data
- Machine readable data
- Meta data

The Organization Ontology
- Time ontology in OWL
- SIOC
- …

Vocabulary

SPARQL

The Organization Ontology
- Time ontology in OWL
- SIOC
- …
Technologies in need…

- Analytics
  - Computational linguistics & NLP
  - Concept tagging & annotation
- Data integration
  - Data management
- Dynamic data / streaming
  - Extraction, data mining, text mining, entity extraction
  - Logic, formal languages & reasoning
- Human-Computer Interaction & visualization
- Knowledge representation
  - RDB2RDF
  - GRDDL Semantic Annotations
  - PROV DCAT
  - RDF
- Machine learning
  - Ontology/thesaurus/taxonomy management
  - Quality & Provenance
  - Recommendations
  - Robustness, scalability, optimization and performance
- Machine learning
  - Ontology/thesaurus/taxonomy management
  - Quality & Provenance
  - Recommendations
  - Robustness, scalability, optimization and performance
- Machine learning
  - Ontology/thesaurus/taxonomy management
  - Quality & Provenance
  - Recommendations
  - Robustness, scalability, optimization and performance
- Machine learning
  - Ontology/thesaurus/taxonomy management
  - Quality & Provenance
  - Recommendations
  - Robustness, scalability, optimization and performance
- Machine learning
  - Ontology/thesaurus/taxonomy management
  - Quality & Provenance
  - Recommendations
  - Robustness, scalability, optimization and performance
Technologies in need...

- Analytics
- Computational linguistics & NLP
- Concept tagging & annotation
- Data integration
- Data management
- Dynamic data / streaming
- Extraction, data mining, text mining, entity extraction
- Logic, formal languages & reasoning
- Human-computer interaction & visualization
- Knowledge representation
- Machine learning
- Ontology/thesaurus/taxonomy management
- Quality & provenance
- Recommendations
- Robustness, scalability, optimization and performance
- Searching, browsing & exploration
- Security and privacy

Potentially interesting: exchange of ML models, e.g. learnt regression models, decision trees, etc. extension of PROV,…? W3C CGs offer a forum for these things!
The XML Binary Characterization Working Group was created as a result of the Binary Interchange Workshop.

The XML Binary Characterization Working Group was tasked with gathering information about uses cases where the overhead of generating, parsing, application, characterizing the properties that XML provides as well as those that are required by the use cases, and establishing objective, shared properties.

User Stories

4 out of 60 user stories we collected in the interviews:

- **Horizontals**
  - Business processes (e.g. product logistics and supply chain management)
  - Human resources (e.g. expert and resource management)

- **Verticals**
  - Media & Publishing
  - Healthcare & Pharma
“**I would like to be able to exchange information and coordinate production and logistics with suppliers and customers…**”

“…**so that I can improve efficiency, effectiveness and flexibility of my inventory management and operations**”
Human Resources

“I would like identify expertise within our large organisation and be able to pinpoint the relevant experts…”

“…so that I can identify top trends within the organisation and expertise for the organisation as a whole”
I would like to display personalized content as precise as possible so that my readers stay as long as possible on my website.
I would like to integrate disparate systems that are:
- Hard to integrate
- Widespread
- Contain the same data that contradicts each other

So that I can gain insights from other clinical trials
User Stories – Bottomline…

4 out of 60 user stories we collected in the interviews:

- **Horizontals**
  - Business processes (e.g. product logistics and supply chain management)
  - Human resources (e.g. expert and resource management)

- **Verticals**
  - Media & Publishing
  - Healthcare & Pharma

According to our interviews best fit to what we can offer technologywise!
Let’s take a step back…

- What are the expectations/requirements?
- What can we offer as a community?

What technologies and standards do we have available?
Technology Toolbox

I offer Linked Data /Semantic Web technologies for... you name it!
Let's take a step back…

- What can we offer as a community?

Taking an introspective view
Let’s check back with our research results...

- **3 seminal papers:**
  - [Scientific American](https://www.scientificamerican.com/article/the-semantic-web/)
  - [Scientific American](https://www.scientificamerican.com/article/the-semantic-web-in-action/)
  - Viewpoint - [A New Look at the Semantic Web](https://www.scientificamerican.com/article/a-new-look-at-the-semantic-web/)

- **Monitoring SW community’s major venues:**
  - ISWC (since 2006), ESWC (since 2006), SEMANTiCS (since 2007), JWS (since 2006), SWJ (since 2010)
The Semantic Web

A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities

By Tim Berners-Lee, James Hendler and Ora Lassila

Subtopics:

Expressing Meaning
Knowledge Representation
Ontologies
Agents
Evolution of Knowledge
Knowledge Representation

“The challenge of the Semantic Web, therefore, is to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web.”

“Adding logic to the Web—the means to use rules to make inferences, choose courses of action and answer questions”

“Ideally, the program must have a way to discover such common meanings for whatever databases it encounters.”

How has knowledge representation based research evolved?

The importance of Knowledge Representation & Reasoning:

Tendency of 'knowledge representation' in 2006-2015

![Graph showing the trend of 'knowledge representation' from 2006 to 2015.](image-url)
The importance of Knowledge Representation & Reasoning:

Tendency of 'knowledge representation' in 2006-2015

Year

Ocurrences

Plot Area

ESWC
ISWC
SEMANTICS
JWS
SWJ
Agents

“The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine-readable Web content and automated services (including other agents) become available.”

Are agents still a hot topic?

May 17, 2001

The Semantic Web
A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities

By Tim Berners-Lee, James Hendler and Ora Lassila

Agents vs Ontologies

![Graph showing occurrences of agents and ontologies from 2006 to 2015.](image-url)
Evolution of Knowledge

“temporal semantics/reasoning” “evolving/evolution”

… overall lower than than the other areas?
Early adopters:
MITRE
Chevron
British Telecom
Boeing
Ordnance Survey
Eli Lily
Pfizer
Agfa
Food and Drug Administration
National Institutes of Health

Software adopters/products:
Oracle
Adobe
Altova
OpenLink
TopQuadrant
Software AG
Aduna Software
Protége
SAPHIRE
“Other **companies** are improving the back-end operations of consumer services.”

*Did companies sustainably adopt SW technologies?*

*Which verticals/domains?*

*Who sponsors us?*

*Which conference sponsors also appear in papers?*

*The semantic web in action L Feigenbaum, I Herman, T Hongsermeier, E Neumann, S Stephens Scientific American 297 (6), 90-97, 2007*
Companies

Conference Sponsors that appear in papers 2006-2015

<table>
<thead>
<tr>
<th>Companies</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>1400</td>
</tr>
<tr>
<td>Oracle</td>
<td>500</td>
</tr>
<tr>
<td>Yahoo</td>
<td>500</td>
</tr>
<tr>
<td>SAP</td>
<td>200</td>
</tr>
<tr>
<td>IEEE Intelligent Systems</td>
<td>200</td>
</tr>
<tr>
<td>Franz</td>
<td>100</td>
</tr>
<tr>
<td>Bing</td>
<td>100</td>
</tr>
<tr>
<td>Expert System</td>
<td>100</td>
</tr>
<tr>
<td>IBM Research</td>
<td>100</td>
</tr>
<tr>
<td>Poolparty</td>
<td>100</td>
</tr>
</tbody>
</table>
Domains

“some of the most advanced progress is taking place in the life sciences and health care fields”

What are the primary research domains within our community?

The semantic web in action L Feigenbaum, I Herman, T Hongsermeier, E Neumann, S Stephens Scientific American 297 (6), 90-97, 2007
Domains

Topics grouped by domain 2006-2015

Well, they publish in other venues…
Applications

“And like an iceberg, the tip of this large body of work is emerging in direct consumer applications, too.”

Can we find evidence of this in our conference and journal paper corpus?

The semantic web in action L Feigenbaum, I Herman, T Hongsermeier, E Neumann, S Stephens Scientific American 297 (6), 90-97, 2007
End user applications?

applications, apps, tools, systems, toolkits” (purple)
“end users, usability” (grey) 2006-2015
Viewpoint
A New Look at the Semantic Web

Seeking to make Web data “smarter” by utilizing a new kind of semantics.
"As the early research has transitioned into these larger, more applied systems, today’s Semantic Web research is changing: It builds on the earlier foundations but it has generated a more diverse set of pursuits."

"the representations that they used became less formal and precise than many early Semantic Web researchers had envisioned."

"As the semantics, in a sense, becomes more “shallow,” it could be more widely applicable”
More applied/lightweight/big data?

Trends in topics 2006-2015

- Blue: large/big data
- Orange: applied
- Gray: formalism/reasoning
- Yellow: precise
- Purple: lightweight
The Next 10 Years

“We believe the objective of the next decade of Semantic Web research is to make this vast heterogeneous multilingual data provide the fuel for truly intelligent applications.”

“relies less on logic-based approaches and more on evidence-based ones.”
Less logic-based, more evidence-based?

Trends in topics 2006-2015

- heterogeneous
- multilingual
- intelligent applications
- logic-based
- evidence-based

Year

Ocurrences

Representation and lightweight semantics

- How do we leverage these diverse representations?
- How do we coordinate the diverse components of structured knowledge that are defined by various parties and that must interact in order to achieve increasingly intelligent behavior?
- How do we define lightweight, needs-based, “pay-as-you-go” approaches for describing knowledge?
- What are the languages and architectures that will provide this knowledge to the increasingly mobile and application-based Web?

A new look at the semantic web A Bernstein, J Hendler, N Noy
Communications of the ACM 59 (9), 35-37
Heterogeneity, quality, and provenance

- How do we integrate **heterogeneous data** and particularly how can we understand which data can be integrated to what degree?
- How can we represent and assess **quality** and **provenance** of the data?
- How do we evaluate whether the quality of a particular source is sufficient for a given task?
Didn’t really find a trend for more lightweight/shallow approaches in the data yet…
Didn’t really find a trend for more lightweight/shallow approaches in the data yet… how about data quality & provenance topics?
Latent semantics

- How much of the semantics can we learn automatically and what is the **quality** of the resulting knowledge?
- As ontologies are learned or enhanced automatically, what is the very meaning of “**formal ontologies**”?
- How do we develop some notion of approximate **correctness**?
- Do similar or different reasoning mechanisms apply to the ontologies that are extracted in this way?
- How do **crowdsourcing** approaches allow us to capture semantics that may be less precise but more reflective of the collective wisdom?

A new look at the semantic web A Bernstein, J Hendler, N Noy
Communications of the ACM 59 (9), 35-37
More emphasis on data quality, less formal ontologies, more crowdsourcing?

**CrowdSourcing** becoming hugely popular! (logscale)
High volume and velocity data

- How do we triage the data in motion to determine what to keep and what we may choose, or need, to allow to be lost?

- How can our applications integrate constantly changing sensor data with fixed data of long duration and high quality semantic provenance?

A new look at the semantic web A Bernstein, J Hendler, N Noy
Communications of the ACM 59 (9), 35-37
Streams & Sensors

![Graph showing occurrences of sensors, streams, and events over years from 2006 to 2015.]

- **Sensors** (blue line)
- **Streams** (orange line)
- **Events** (gray line)

**Chart Area**

<table>
<thead>
<tr>
<th>Years</th>
<th>Sensors</th>
<th>Streams</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2007</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2008</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2011</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2012</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2013</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2014</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2015</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The Roadmap
Roadmap for Enterprise SW/LD?

- hmm, I think that I don't have enough data for that as of yet…
- … but gives me a chance to throw in my own personal taste/preferences!!! ;-) 

- What I believe to be hot/interesting SW topics (the no-brainers, if you want...)
  1. **Combining Open and Closed** Data (Data Security and Privacy)  
     - *Privacy & Policies* are the big thing in the age of “Big Linked Data”
  2. **Archiving** and storage of **temporal data**, efficient indexing and efficient updates (Data Management)
  3. Move into “not-quite-so-structured” data, don’t focus on just structured RDF data and non-structured data ...
    - Embrace and deal with de facto standards and formats: e.g. *schema.org, CSV, JSON*, ...
    - Syntax doesn’t matter, we’ve defined enough languages and syntaxes!
    - *Embrace a “Cognitive computing” approach* (bridge between purely symbolic and subsymbolic AI)

4. **Analysing our own research and its impact (...properly)!**
Giving back to the community
Where to find data about our community (for example...):

http://www.scholarlydata.org/

- Thanks to:
  - Abraham Bernstein (JWS)
  - Pascal Hitzler (SWJ)
  - Krzystof Janowisz (SWJ)
  - Annalisa Gentile, Andrea Nuzzolese (scholarlydata)
  - Francesco Osborne (Rexplore)
  - Knud Möller (data.semanticweb.org ...)
  - ...
We plan to play back our data into http://scholarlydata.org …

- What we plan* to add:
  - complete/link missing data from all major events
  - keyphrase extraction from fulltexts
  - keyphrase taxonomy
  - Company/Sponsorship data

- Still a lot TO DO:
  - respect *skews* such as:
    - higher importance of abstract/title keyphrases
    - No of papers per year (journals)
    - a curated SW keyphrase taxonomy
  - Doing the “usual NLP stuff”
    - improve stemming
    - disambiguate wordsenses

*) will try, licensing etc. e.g. needs to be clarified!
Key take-home:

• Stay tuned: PROPEL results will be collected at:
  • https://www.linked-data.at/
• Don’t think our standards are known by industry ;-)  
  • Nor that they cover “Semantic Linked” data…
  • … but standardization is still useful
• We could need some more research in understanding what we actually do and sell our successes!
  • Listen to the “wise old elves”!
  • … but maybe there is a more data-driven, data-analytics-driven way for this
• Feel free to go ahead and play with the data we collected (soon on http://scholarlydata.org)…
  • …improve our quick attempt!) … and let me know!!!! 😊

• We shouldn’t try to be