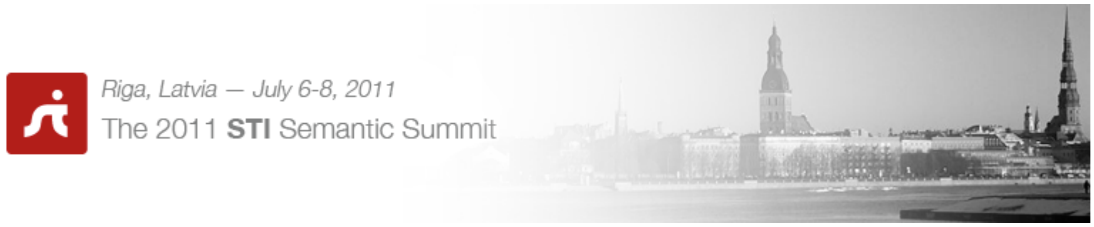
The 2011 STI Semantic Summit Closing Session



Key Summit issue: Search, access, and integration of real data in the Big Data World involving semantic and engineering aspects.

Observations

- Shockingly, **most Big Data integration is manual**, error prone, and costly.
- **Realistic solutions for meaningful data integration require**:
  - Realistic use cases that reflect semantic and engineering issues at realistic scope and scale.
  - Acceptable data quality hence engineering issues such as integrity, formatting, and cleaning of data.
  - All database **engineering tricks** for data cleansing (Extract, Transform, and Load) and for performance, e.g., optimizing RDF queries, – extended to all data not just relational databases plus data integration tricks – extended well beyond those of SQL to all data.
  - A **deep understanding of the problem domain** and the related data – thus relevant human expertise hence manual effort.
  - **Semantic technology approaches, tools, and tricks**.
- Understanding Big Data challenges including data integration should be driven by **real industry use cases.** Realistic use cases can be found on the web, e.g., data.gov and domain-specific sites. Data.gov sites are ideal as they are pubically available and industrial data is seldom available. Data.gov sites are emerging worldwide and are underutilized in part due to the challenges of meaningful data integration. If this is not addressed – if value is not derived, the window of opportunity for government data may close. Other industrial data is typically not available due to security, privacy, or confidentiality restrictions.
- **We need tools** such as a Peter Boncz's LOD ripper [see *The Meaningful Use of Big Data: Four Perspectives – Four Challenges]:* a portal for retrieving a relation table out of the linked data sets, such as Data.gov, with tools to summarize, visualize, explore, rank, annotate, and understand data to integrate and make available for reuse. It would be helpful for the community to develop, share, and extend tools to address Big Data challenges and publish a related Big Data tools catalogue.
- **Lightweight tools and approaches seem to offer the greatest opportunities**, including: schema-last, lightweight ontologies, and pay-as-you-go (data integration).
- While the Summit focused on the semantic web the key summit issue is applicable to all data in databases, on the web, and anywhere else. Hence, data integration solutions that combine database engineering tricks and semantic

technologies will apply not only to the semantic web, e.g., Linked data, but are **broadly applicable across the entire Big Data World**.

- **Core challenges remain** including unique identification and dynamic aspects of data reflecting the constantly changing world. Unfortunately, some powerful semantic tools, such as OWL fail to address these important data integration challenges just as database tools tend to ignore dynamic aspects of data.
- Due to the idiosyncratic nature of Big Data Challenges and the current need to develop expertise at solving such problems plus the need for domain specific knowledge and tools to automate problem solving at massive scope and scale, we should hold **Big Data Hackathons** - events for Big Data problem owners and Big Data solution developers to share and develop LOD tools and platforms.
- **Big Data Challenges are inherently pragmatic** – or at least our limited knowledge and expertise in Big Data requires deeper knowledge and experience with real data, real problems, and real solutions. For the moment, researchers cannot simply develop theories and architectures. Proposed solutions to Big Data Challenges must be validated against reality. Architectures must be validated and tested on real data sets under real constraints – meaning real projects with practical and financial consequences. Success will not be immediate. Solutions must be pursued empirically documenting what went well and what went wrong. That is **Big Data Challenges are inherently data driven** and must be pursued accordingly.
- **Big Data Integration is inherently multi-disciplinary** involving semantic and database technologies, amongst others. Either community alone will not solve the challenges. The semantic technology and database communities should partner to address Big Data Integration challenges thus opening the door to a myriad of other opportunities the are also inherently engineering-semantic challenges and opportunities including hot topics like social and mobile computing, energy, and sensors – all of which are data driven domains.
- In light of the discussions, Semantic Web experts suggested that the semantic web stack be revisited, e.g., focus on lightweight solutions. **Should there be a lightweight Semantic Web Stack?**
- There was optimism that progress is being made due to the growth of linked data, the recognition of the need for tools to integrate and better understand data largely driven by the potential value in the Big Data stores such as Data.gov.

Participant Quotations

Chris Bizer (invited speaker):

I am very optimistic that a year from now we will be in a much better position to address the data integration challenges using LOD. There continues to be a significant growth in the adoption of LOD. For example, many libraries are publishing their catalogs using LOD.

Industry has an important role to play in this development. As illustrated in the healthcare and Verizon examples, large enterprises have been facing Big Data challenges internally that are analogous to what the Web is facing as a whole. For example, how will the US Healthcare system uniquely identify their hundreds of millions of beneficiaries so that when Helen is in need of critical care for failed liver function they can select the correct records across the healthcare system that are relevant to Helens' liver that may be distributed longitudinally and geographically over hundreds of databases reflecting Helen's 60 year lifespan? Researchers and

Industry should work collaboratively on use cases like emergency healthcare for Helen’s failing liver.

Orri Erling (invited speaker):

As data continues to scale and in particular as Chris says, as LOD and RDF data continues to grow, engineering can assist to dramatically reduce the resources and time required for execution, for example, RDF query performance will increase significantly over the next year. There are a lot of engineering optimizations that can be applied to RDF data and we are only beginning to apply them. We also need significantly improved tools and engineering optimizations for the central problem of data integration. These tools will not only respond to real needs, they will reduce the cost of producing value from RDF data and drive the adoption of RDF data. Not only does OpenLink endorse the need for use cases, we claim that detailed, industrial scale use cases are central to progress on RDF data and hence to progress on the Semantic Web.

Dr. Jörg Wurzer (STI Partner)

As an industrial partner trying to deploy semantic technologies, I see the challenges of turning research results in Semantic Technologies, like RDF and LOD, into industrial value. I continue to look to the research community for collaboration to reduce research results to practice.

Valentin Zacharias (STI junior Member)

The benefits of semantics are still not clear. For example what specific benefits have arisen from publishing structured data in the many instances of Data.gov around the world?

Michael Brodie (STI Fellow)

The Summit unanimously concluded that meaningful use of data is a core problem of using Big Data in all its forms, e.g., text, files, databases, special purpose data sets, especially in highly distributed and connected contexts such as intra-nets, Internet, Web, and the Semantic Web. The core challenges are to ensure meaningful data integration, requiring semantics, at scale, requiring engineering. Due to the lack of semantic solutions Big Data integration requires the heavy involvement of human experts to understand and combine data meaningfully. But human involvement to resolve semantic challenges arise not only in the integration step; they arise in every step: Have we found the right data candidates? Have we extracted (ETL) the relevant data? Have we correctly identified those instances that pertain to our problem (entity resolution)? Have we correctly computed the result that we need? In general, full automation is feasible for an extremely limited set of data integration challenges, those that can be defined and executed precisely without human intervention.

Big Data Integration offers a necessary and wonderful opportunity for semantic technologies to collaborate with database technologies. The potential value of data in the Big Data World, the current human involvement required to derive that value, and the current growth rate of Big Data may soon bring us to the tipping point that will necessitate this collaboration.