The Meaningful Use of Big Data: Four Perspectives – Four Challenges

Christian Bizer[1], Peter Boncz[2], Michael L. Brodie[3], Orri Erling[4]
[1]Web-based Systems Group, Freie Universität Berlin; [2]Centrum Wiskunde & Informatica, Amsterdam; [3]Verizon Communications, USA; [4]OpenLink Software, Utrecht
christian.bizer@fu-berlin.de, P.Boncz@cwi.nl, michael.brodie@verizon.com, oerling@openlinksw.com

Abstract

Twenty-five Semantic Web and Database researchers met at the 2011 STI Semantic Summit in Riga, Latvia July 6-8, 2011[1] to discuss the opportunities and challenges posed by Big Data for the Semantic Web, Semantic Technologies, and Database communities. The unanimous conclusion was that the greatest shared challenge was not only engineering Big Data, but also doing so meaningfully. The following are four expressions of that challenge from different perspectives.

Michael's Challenge: Big Data Integration is Multi-disciplinary

The exploding world of Big Data poses, more than ever, two challenge classes: engineering - efficiently managing data at unimaginable scale; and semantics – finding and meaningfully combining information that is relevant to your concern. Without the meaningful use of data, data engineering is just a bunch of cool tricks. Since every computer science discipline and every application domain has a vested interest, Big Data becomes a use case for multi-disciplinary problem solving[2]. The challenge posed here is of the meaningful use of Big Data regardless of the implementation technology or the application domain.

Emerging data-driven approaches in the US Healthcare Big Data World[3] involves over *50 million* patient databases distributed US-wide for which the US Government defines *Meaningful Use* and the medical community has identified challenges [4] across queries such as: "For every 54-year-old white, female high school dropout with a baseline blood pressure of 150 over 80 in the beta blocker group who had these *two* concurrent conditions and took these *three* mediations. Magid matched her to another 54-year-old female high school dropout with a baseline blood pressure of 150 over 80 in the ACE inhibitor group, who had the same drugs."[5]

In this Big Data World information is unbelievably large in scale, scope, distribution, heterogeneity, and supporting technologies. Regardless of the daunting engineering challenges, meaningful data integration takes the following form (step order can vary):

- **Define** the concern – the problem to be solved - the query to be answered, e.g., efficacy of a drug for 54-year-old hypertensive women.
- **Search** the Big Data space for candidate data elements that map to the concern; e.g., all hypertensive 54-year-old women.
- **Transform** Extract, Transform, and Load (ETL) the relevant parts of the candidate data elements into appropriate formats and stores for processing for processing.
- **Entity Resolution**: Verify that data elements are unique, relevant, and comprehensive, e.g., all hypertensive 54-year-old women. Since unique identification is practically and technically infeasible, not all candidate data elements will refer to the entity of concern. More challenging are data elements that describe aspects of the entity of concern at different level of abstraction and from different perspectives, e.g., data elements on myriad details of hypertensive 54-year-old women, e.g., physiology, social network membership, salary, education.
- **Answer the query/solve the problem**: Having selected the data elements relevant to the entity of concern, compute the answer using domain-specific computations, e.g., efficacy of the drug.

It is hard to conceive of the scope and scale of data elements in the Big Data World. The above method has worked amazing well for more than 30 years in the $27 billion per year relational database world with blinding efficiency over ever expanding database sizes from gigabytes, to terabytes, to petabytes, and now exabytes. Data elements that are genuinely relational constitute less than 10% of the Big Data World and that share is falling rapidly.

The rare properties of single value of truth, global schema, and view update of semantically homogeneous relational databases are often underlying assumptions of relational database integration. However, few relational databases are semantically homogeneous and like most data stores, they lack these properties. Hence, meaningful data integration solutions cannot be based on these properties without supporting evidence that must be derived manually. Since the real world involves multiple truths over every concern, relational data

integration has semantic (correctness) and engineering (efficiency) limits.

My challenge is meaningful data integration in the real, messy, often schema-less, and complex Big Data World of databases and the (Semantic) Web using multi-disciplinary, multi-technology methods.

Chris' Challenge: The Billion Triple Challenge

Over the past few years, an increasing number of web sites have started to publish structured data on the Web according to the Linked Data principles. This trend has led to the extension of the Web with a global data space – the Web of Data [6].

Topology of the Web of Data Like the classic document Web, the Web of Data covers a wide variety of topics ranging from data describing people, organizations and events, over products and reviews to statistical data provided by governments as well as research data from various scientific disciplines.
W3C Linking Open Data (LOD) community effort has started to catalog known Linked Data sources in the CKAN data catalog and regularly generates statistics about the content of the data space [7]. According to these statistics, the Web of Data currently contains around 31 billion RDF triples. A total of 466 million of these triples are RDF links which connect data between different data sources. Major topical areas are government data (13 billion triples), geographic data (6 billion triples), publication and media (4.6 billion triples), life science (3 billion triples).

Characteristics of the Web of Data The Web of Data has several unique characteristics which make it an interesting use case for research on data integration as well as the Big Data processing:

- **Widely-used vs. proprietary vocabularies.** Many Linked Data sources reuse terms from widely-used vocabularies to represent data about common types of entities such as people, products, reviews, publications, and other creative works. In addition, they use their own, proprietary terms for representing aspects that are not covered by the widely used vocabularies. This partial agreement on terms makes it easier for applications to understand data from different data sources and is a valuable starting point for mining additional correspondences.
- **Identity and vocabulary links.** Many Linked Data sources set identity links (owl:sameAs) pointing at data about the same entity within other data sources. In addition data sources as well as vocabulary maintainers publish vocabulary links that represent correspondences between terms from different vocabularies (owl:equivalentClass, owl:equivalentProp-erty, rdfs:subClassOf, rdfs:subPropertyOf). Applications can treat these links as integration hints which help them to translate data into their target schema as well as to fuse data from different sources describing the same entity.
- **Data Quality:** The Web is an open medium in which everybody can publish data on the Web. As the classic document Web, the Web of Data contains data that is outdated, conflicting, or intentionally wrong (SPAM). Thus, one of the main challenges that Linked Data applications need to handle is to assess the quality of Web data and determine the subset of the available data that should be treated as trustworthy.

Pre-Crawled Data Sets One approach to obtain a corpus of Linked Data is to use publicly available software, such as LDSpider, to crawl the Web of Data. However, there exist already a number of publicly available data sets that have been crawled from the Web of Data and can be promptly used for evaluation and experimentation.

- **BTC 2011.** The Billion Triple Challenge 2011 data set (BTC 2011) has been crawled in May/June 2011 and consists of 2 billion RDF triples from Linked Data sources. There are also two older versions of the data set available which have been crawled in 2009 and 2010. The BTC data sets are employed in the Semantic Web Challenge, an academic competition that is part of the International Semantic Web Conference. The BTC data sets can be downloaded from the Semantic Web Challenge website [8].
- **Sindice 2011.** The Sindice 2011 data set has been crawled from the Web by the Sindice search engine. The data set consists of 11 billion RDF triples which (1) originate from Linked Data sources and (2) have been extracted from 230 million Web documents containing RDFa and Microformats markup. The data set contains descriptions of about 1.7 billion entities and can be downloaded from [9].

Now then, the task A concrete task, which touches all challenges around data integration, large-scale RDF processing, and data quality assessment that arise in the context of the Web of Data, is to (1) find all data that describes people (in whatever role) as well as creative works produced by these people (ranging from books, films, musical works to scientific publications) in the BTC 2011 or the Sindice 2011 data set; (2) translate this data from the

different vocabularies that are used on the Web into a single target vocabulary, (3) discover all resources that describe the same real-world entity (identity resolution), and (4) fuse these descriptions into an integrated representation of all data that is available about the entity using a general or several domain-specific trust heuristics.

Success metrics: Success metrics for this task are the number of people and creative works discovered in the data set and on the other hand the completeness and consistency of the integrated data.

Peter's Challenge: The LOD Ripper

Motivation. For broader adoption of semantic web techniques, two main challenges arguably exist: (I) lack of good use cases (ii) ever existing data integration troubles that makes creating links so hard. The LOD Ripper idea originates from the thought that the best window of opportunity is linked open government data. If it became easy for people and companies to earn money and reap value from this high-quality & free information out there, linked open data might break through in this domain. If this fails to catch on soon, linked open government data investment in the early adaptor countries might drop, and might altogether fail to take off in the rest. Use cases outside government or academic data are much harder to find as one then faces the issue of an economic model for LOD production. So, better to succeed here.

Success Metric. A side note on what success could be. Success is not only achieved when the IT world switches to semantic-everything technology. Given the value of installed base, this is unrealistic. Success is already achieved when people combine multiple LOD datasets, and link them to their own data, but then import the result e.g. as a flat relational table (via CSV, XML, etc.) for use in existing infrastructure and tools. Think of existing enterprise middleware, business logic, data warehouses, and OLAP and data mining tools: technology that has been invested heavily in, and which would profit from enrichment by linked open government data. The semantic success will be in the fact that semantic technology has made data integration easier and partially automatic. Data integration is one of the highest cost issues in IT, worth tens of billions of dollars per year. Therefore, my name for the project, the "**LOD Ripper**": a technology to rip valuable data out of LOD sources. Admittedly, this is intended to be provocative to the Semantic Web community and to emphasize practicality. But, you could also use the LOD ripper to extract data in triple form, of course. The LOD ripper could also search non-LOD open government datasets, just like CKAN. Mapping these together may trigger the incremental LOD-ification of such datasets.

Now then, the proposal: the LOD Ripper is a vision of a web portal, driven by goals similar to CKAN, however going way beyond CKAN in its practical support for an information engineer in finding and combining useful open government data, and integrating it with his own. The portal would do the maximum possible, given a vague information need on the part of the information engineer, to put him as quickly as possible into hands-on mode with real data (snippets) from the entire data collection. This means among other things that one of the main ways to interact with the system is keyword search, which would search in (1) ontologies/schemas (2) the data itself and (3) mappings/views provided by earlier users of the portal. The goal of the portal is to assist the information engineer in obtaining a useful mapping that allows him to retrieve ("rip") a derived dataset that is valuable for his problem space. Point (3) stresses that this portal should facilitate a pay-as-you-go process.

Mappings. Obtaining a mapping may happen by finding an existing mapping, by combining multiple existing ones into a new one, or by fresh composition. The resulting mapping should be made available again for future users. Mapping languages are hence an important aspect, and user interfaces to compose mappings and mapping systems, as well as entity resolution algorithms are part of such systems. Mappings are not only specifications, but in the end will also take the form of new data, new or better triples, that add meaning in and between existing dataset(s). Such new triples may be generated by a mapping system following a mapping specification automatically, but should be materializable as triple sets, because often these need to be manually curated as well. Note that mappings need provenance tracking, at least in the form of a simple version tracking system.

Ranking. As we search schema, mapping and data, we need also ways to usefully rank these. On the one hand, ranking could be based on precision of match with keywords, but on the other, should be based on usefulness/quality assessment by previous users of the datasets and dataset elements.

Visualization. To show results of a search, we need good snippets or summaries of what we find. In the case of ontologies, one would use dataset

summarization techniques, to visualize the most common structures in a dataset and where the keyword search matched in that. If we look for multiple types of data, one would also visualize the structure of any existing mappings between the hits, leaving out irrelevant details as much as possible. When searching for views/mappings, these should similarly be visually summarized. It should be one click to switch from looking at schema visualizations to see representative samples of underlying data occurring in the wild. There should be strong support for generating tabular data views out of the LOD sources. The ability to extract tables, using all the mapping machinery, is the prime output of the LOD Ripper portal.

Key Matching. The system should allow users to define keys, and upload possibly a large number of key values, which typically come from the users' own environment. Think for instance of a column containing city names as a potential key column. One purpose of such a key column in the LOD Ripper is to measure the overall effectiveness of finding useful data ("how many of my cities did I find info for?"). It also provides a concrete starting point for instance-driven data integration ("find me matching city properties anywhere!"). Note that this works on the instance level, and one needs algorithms to quickly search for similar and overlapping data distributions.

Snappiness. Visualizing results and creating mappings *interactively* is going to be very important. This means emphasis on cool GUI design as well as low-latency performance. This requires a solid LOD warehouse with advanced indexing performed in the background. A technique probably useful for instance-level data matching would be NGRAM indexing (to speed up partial string and distribution matching) as well as massive pre-computation of entity resolution methods.

Call to action: Can we organize such a portal? Do you have ideas and time, or even components available?

Orri's Challenge: Demonstrate the Value of Semantics: Let Data Integration Drive DBMS Technology

Advances in database technology will continue to facilitate dealing with large volumes of heterogeneous data. Linked data and RDF have a place in this, as they are a schema-less model with global identifiers and a certain culture of, or at least wish for, reuse of modeling.

Systematic adoption of DBMS innovation into the semantic data field, backed by systematic benchmarking, will make the schema-less flexibility of these technologies increasingly affordable.

These developments set the stage for the real challenge:

- **Demonstrate the benefit of semantics for data integration**. The RDF/data world does not exist in a bubble, in any real life situation it will be compared to alternatives.
- **Meaningfully combine DBMS and reasoning functions**. Identify real-world problems where there is real benefit in having logics more expressive than SQL or SPARQL close to the data. We have talked extensively about smarter databases but the actual requirement remains vague. We do not think of OWL or RIF as such answers for data integration even if they may be a part of it.
- **Bring Linked Data and RDF into the regular data-engineering stack:** Use existing query and visualization tools against heterogeneous data. There are many interactive SPARQL builders but are these performs comparable to MS Query for SQL? Since data here is schema-less, data set summarization will have to play the role that the schema plays with relational tools. There are many RDF bound UI widgets but few bind to Excel for business graphics?

We know how to make DBMS's. To get to the next level we need use cases that represent real needs, e.g. data integration. This information is required to determine what ought to be optimized or in what way the existing query languages / logics / processing models fail to measure up to the challenge.

So, users / practitioners, does there exist functionality that belongs with the data but cannot be expressed in queries? What about entity resolution frameworks? What about inference? What kind of inference? What of the many things people do in map/reduce, is there a better way? How about Berkeley Orders of Magnitude (BOOM) work for declarative data centric engineering for big data? I envision expanding the Semdata benchmarks activity to include specific use cases that come from you. What did you always want to do with a DBMS but never dared ask?

This could result in a set of use cases with model solutions with different tools and techniques. We are not talking about fully formalized benchmarks but about samples of problems motivating DBMS advances beyond standard query languages.

This in turn would bring us closer to quantifying the benefits of semantic technology for real world problems, which is after all our value proposition. Of course, this involves also non-RDF approaches, as we do not believe that there ought to be a separate RDF enclave but that technologies should be appreciated according to their merits. It is no wonder the bulk of database research has been drawn to the performance aspect, as success in this is fairly unambiguous to define and the rationale needs no explaining. But when we move to a more diverse field like data integration, which indubitably is the core question of big data, we need more stakeholder involvement.

Tell us what you need and we'll see how this shapes the future of DBMS.

If you are struggling with doing things that DBMS' s ought to do but do not support, let us know. Chances are that these problems could be couched in terms of open government data even if your application domain is entirely different, thus alleviating processes confidentiality problems.

References

[1] 2011 STI Semantic Summit, Riga, Latvia, http://www.sti2.org/events/2011-sti-semantic-summit

[2] M.L. Brodie, M. Greaves and J.A. Hendler, Databases and AI: The Twain Just Met, 2011 STI Semantic Summit, Riga, Latvia, July 6-8, 2011

[3] Preliminary Observations on Information Technology Needs and Priorities at the Centers for Medicare and Medicaid Services: An Interim Report, Committee on Future Information Architectures, Processes, and Strategies for the Centers for Medicare and Medicaid Services; CSTB; National Research Council of the Academies of Science, 2010

[4] E.M. Borycki, A.W. Kushniruk, S. Kuwata, J. Kannry, Engineering the electronic health record for safety: a multi-level video-based approach to diagnosing and preventing technology-induced error arising from usability problems, Stud Health Technol Inform. 2011;166:197-205.

[5] S. Begley, The Best Medicine: The Quiet revolution in comparative effectiveness research just might save us from soaring medical costs, Scientific American **305**, 50 - 55 (2011)

[6] C. Bizer, T. Heath, and T. Berners-Lee: Linked Data - The Story So Far. International Journal on Semantic Web & Information Systems, Vol. 5, Issue 3, Pages 1-22, 2009.

[7] C. Bizer, A. Jentzsch, and R. Cyganiak: State of the LOD Cloud. http://www4.wiwiss.fu-berlin.de/lodcloud/state/

[8] Semantic Web Challenge website. http://challenge.semanticweb.org/

[9] Sindice-2011 Dataset for TREC Entity Track. http://data.sindice.com/trec2011/