Semantic Interoperability for Big Data

Ken Baclawski
College of Computer
and Information Science
Northeastern University

Emergence of Big Data

- The phenomenon was studied in the early 1980s
  - Originally called the “information onslaught” [1]
- Different fields experience the problem at different times.
- The transition from information scarcity to information abundance in a field appears to occur abruptly.
- Organizations have difficulty coping with the transition.
- Many organizations are still based on the information scarcity model.

Characteristics of Big Data

- Large amounts of data (volume)
- Rapid pace of data acquisition (velocity)
- Complexity of the data (variety)
  - Complexity of individual data items
  - Both structured and unstructured data
  - Complexity of data from one source
  - Multiple sources and time periods

Technology for Big Data

- Development of new storage and indexing strategies for handling volume and velocity
  - “Map Reduce” was developed in 1994. [2]
- Development of techniques for handling variety
  - Schema mapping
  - Controlled vocabularies
  - Knowledge representations
  - Semantic technologies

Semantic Interoperability

- The ability of computer systems to exchange data with unambiguous, shared meaning.[6]
- This can be accomplished by adding data about the data (metadata), linking each data element to a controlled, shared vocabulary.
  - What does it mean for a vocabulary to be controlled?

Why is semantics important?

- Misunderstanding the data can result in invalid or misrepresented analyses.
- A recognized problem well before Big Data.
- John P. A. Ioannidis claimed that most published scientific research findings are false. [3] [4]
  - Many forms of bias can invalidate results.
  - Bias and observer effects are difficult to recognize.
  - Standard experimental procedures and statistical techniques do not always eliminate bias.
- Could Big Data exacerbate this problem?

Current Problems

- Big Data practitioners focus on volume and velocity
  - Assume someone else will handle variety.
- Big Data curricula are inadequate
  - Assume that specifying the schema is all that is required for the variety problem.
- Incompatibility between statistical and logical techniques (hybrid reasoning gap)

Examples

- The track *Tackling the Variety Problem in Big Data* uncovered many examples of semantic interoperability problems and potential solutions. [5]
- The results of the track were summarized with a synthesis and use cases.
- The following Smart Cities example is due to Mark Fox and Rosario Uceda-Sosa.

Smart Cities: Successes

- Cities are collecting large amounts of data.
  - Each agency collects its own data.
- Significant social problems can be addressed.
- In a NYC project, data from 19 agencies was combined to find illegal rooming houses using data analytics. The productivity of inspectors was increased from 13% to 70%.
- However, finding the necessary data is difficult, and it is difficult to combine data from multiple sources.

Smart Cities: Failures

- The indicators used by cities are not consistent
  - In a pilot study of 9 cities, out of 1100 indicators, only 2 were comparable.
  - Conclusions based on one city need not apply to other cities in spite of using the same terms.
  - Merging data from several cities may not be meaningful.
- For example, student-faculty ratios currently cannot be meaningfully compared, yet they are often compared and decisions are based on the comparisons.

Smart Cities: Conclusions

- Using a shared vocabulary term is not sufficient to ensure semantic interoperability.
  - Most ontologies are little more than just vocabularies with informal, limited definitions.
- Inconsistencies occur at many levels
  - Different cities often collect and present data differently for the same metric.
  - Over time, a single city may change how it collects data for the same metric.
  - Derived indicators often combine data from different regions and time periods.

Smart Cities: Solution

- Ensuring consistency of indicators is possible but requires development of standards with detailed rules.
- Maintain provenance
- Conform to the standard
- Establish trust
- Develop bridges between cities, agencies, and previous standards.

Semantics and Standards

- There is a close relationship between standards and semantic interoperability.
- A standard is a formal agreement of the meaning of a collection of concepts among communities that share a common interest.
- Conformance to a standard ensures, in principle, semantic interoperability within the scope of the standard.

More Use Cases

- Use modular development techniques
- Take advantage of existing models
- Develop reusable frameworks with explicit extension points
- Build a semantic ecosystem for collaborative development

Bibliography

1. D. Kerr, K. Braithwaite, N. Metropolis, D. Sharp, G.-C. Rota (Ed). Science, Computers and the Information Onslaught. Academic Press. (1984)
2. K. Baclawski and J.E. Smith. High-performance, distributed information retrieval. Northeastern University, College of Computer Science. (1994)
3. Ioannidis JPA, Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124 (2005)
4. Critical Data Conference: Secondary Use of Big Data from Critical Care http://criticaldata.mit.edu/ (2014)
5. Tackling the Variety Problem for Big Data. (2014) http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2014_Tackling_Variety_In_BigData_Synthesis
6. NCOIC, "SCOPE", Network Centric Operations Industry Consortium (2008)