The Role of Ontology in the Era of Big Data

T. Malyuta, Ph. D
New York City College of Technology, NY, NY

B. Smith, Ph. D
University at Buffalo, Buffalo, NY

R. Rudnicki
CUBRC, Buffalo, NY
Big Data Problem

• Wikipedia defines Big Data as “…a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools.”

• Gartner defines Big Data with three ‘V’s:
  – Volume
  – Velocity (of production and analysis)
  – Variety

• This means that Big Data are beyond our control (as opposed to those complex and big systems with diverse and changing data where the complexity is known)
Big Data Solution – Agility

• Dimensions of agility
  – Data storage and processing paradigms that can deal with massive volumes of heterogeneous data coming onstream
  – Dynamic data stores that can easily accommodate diverse and a priori unknown data types and semantics
  – Methods and tools that leverage dynamic and diverse content
Agile Data Utilization

• Today, the main problem of Big Data is *using* it

• Utilization of ‘Variety’ – diverse types and diverse semantics, in particular, for integration and interoperability

• Traditional utilization approaches fail

• Agile utilization paradigms are needed
Agile Semantic Utilization of Big Data

• A good solution has to be
  – Able to grow incrementally
  – Able to be developed in a distributed manner
  – Without losing consistency
  – Independent of particular implementations, and data producers and consumers
  – Applicable to data in an agile manner

• We call our solution: ‘semantic enhancement’ (SE) of data
SE

- SE is realized with the help of ontologies that are used to annotate (tag) data
  - Ontologies provide controlled vocabulary for annotations allowing agile integration and interoperability
  - Ontologies, by virtue of their nature and organization, provide semantic enhancement of data

<table>
<thead>
<tr>
<th>PersonID</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>Java</td>
<td>Programming</td>
</tr>
<tr>
<td>222</td>
<td>SQL</td>
<td>Database</td>
</tr>
</tbody>
</table>
The Meaning of ‘Enhancement’

• Semantic enhancement/enrichment of data = arm’s length approach (no change to data) – through simple annotation we associate an entire knowledge system with a database field
  – enables analytics to process data, e.g. about computer skills, “vertically” along the Skill hierarchy, as well as “horizontally” via relations between Skill and Education.
  – and further… while data in the database does not change, its analysis can be richer and richer as our understanding of the reality changes

• For this richness to be leveraged by different communities, persons, and applications it needs to have the properties mentioned above and be constructed in accordance with the principles of the SE
SE Principles

- Create a Shared Semantic Resource (SSR) of ontologies to be used for annotation
- Establish an agile strategy for building ontologies within this SSR, and apply and extend these ontologies to annotate new source data as they come onstream
  - Strategy pioneered in biomedical and other scientific fields: leaves data as they are, and incrementally tags data sources with terms from a growing, consistent, non-redundant set of ontologies
- Problem: Given the immense and growing variety of data sources, the development methodology must be applied by multiple different groups
  - How to manage collaboration?
Achieving the Goal

- Methodology of incremental distributed ontology development
- A common ontology architecture incorporating a common, domain-neutral, upper-level ontology (BFO)
- A shared governance and change management process
- A simple, repeatable process for ontology development
- An ontology registry
- A process of intelligence data capture through ‘annotation’ or ‘tagging’ of source data artifacts
Addressing Challenges

• Challenges
  – Too many lexicons
  – The scope of the domain: signal, sensor, image, ... intelligence about ... the whole world
  – Difficult to conduct governance and management of ontology development to ensure consistent evolution
  – Lack of expertise

• Preventing Failure
  – Realism
  – Arm’s length application
  – Architecture and methodology
  – Expertise
The SSR

Reference Ontologies (Shared Semantic Resource)

Agent
Geospatial Organization
Weapon Information Artifact

Application Ontologies:
• Agent-related
• Weapon-related
• ...

for purposes of
Event Reporting
Intelligence Analysis
Video Analysis
NLP
SSR and Interoperability/Integration
References

• Smith B. et al., Horizontal Integration of Warfighter Intelligence Data: A Shared Semantic Resource for the Intelligence Community, STIDS Conference, 2012.


• Salmen D. et al., Integration of Intelligence Data through Semantic Enhancement, STIDS Conference, 2011.