Monday, March 26, 2012 4:58 PM

Contents

[1. Current State 1](#_Toc320544766)

[Ontology might help big data, but why this usually fails 1](#_Toc320544767)

[2. Current Uses/Examples 1](#_Toc320544768)

[3. Areas of Use (both current and future)/Areas of non-use 2](#_Toc320544769)

[4. Accessibility (i.e., ease of use)/Impediments 3](#_Toc320544770)

[5. Recommendations 3](#_Toc320544771)

The goal of "Meeting Big Data Challenges through Ontology" is to identify challenges that will advance ontology and semantic web technologies, increase applications, and accelerate adoption.

# Current State

Ontology may tame big data, drive innovation, facilitate the rapid exploitation of information, contribute to long-lived and sustainable software, and improve Complicated Systems Modeling.

## Ontology might help big data, but why this usually fails

1. easy to create ontologies that myriad incompatible ontologies are being created in *ad hoc* ways leading to the creation of new, semantic silos
2. The Semantic Web framework as currently conceived and governed by the W3C (modeled on html) yields minimal standardization
3. The **more semantic technology is** successful**; they more we fail to achieve our goals**

* Just as it’s easier to build a new database, so it’s easier to build a new ontology for each new project
* You will not get paid for reusing existing ontologies (Let a million ontologies bloom)
* There are no ‘good’ ontologies, anyway (just arbitrary choices of terms and relations …)
* Information technology (hardware) changes constantly, not worth the effort of getting things right

Linked data to lower the costs of reusing data more than anything. In addition, government data is used quite widely already, so we feel there are huge opportunities in promoting this in the Federal space.

# Current Uses/Examples

Systems Engineering Modeling Languages and Ontology Languages

**Drive Innovation**

Federation and Integration of Systems

***Driving Innovation with Open Data - Creating a Data Ecosystem***

1. **Gather data**
   * from many places and give it freely to developers, scientists, and citizens
2. **Connect the community**
   * in finding solutions to allow collaboration through social media, events, platforms
3. **Provide an infrastructure** 
   * built on standards
4. **Encourage** **technology** **developers**
   * to create apps, maps, and visualizations of data that empower people’s choices
5. **Gather more data**
   * and connect more people
6. Energy.Data.gov connects works with challenges across the nation to integrate federal data and bring government personnel to code-a-thons
7. ***Data Drives Decisions***
   * Apps transform data in understandable ways to help people make decisions

Rapid exploitation of information

1. In this world, the benefit is derived from the rapid pace at which new data and new data sources can be combined and exploited.
2. **High level reasoning over curated information** In this world, the benefit is derived from non-trivial inferences drawn over highly vetted data.
3. **Many times people try to have both expressivity and scale. This is very expensive**
   * **Don’t be seduced by expressivity**
   * Just because you CAN say it doesn’t mean you SHOULD say it. Stick to things that are strictly useful to building your big data application.
   * **Computationally expensive**
   * Expressivity is not free. It must be paid for either with load throughput or query latency, or both.
   * **Not easily partitioned**
   * Higher expressivity often involves more than one piece of information from the abox – meaning you have to cross server boundaries. With lower expressivity you can replicate the ontology everywhere on the cluster and answer questions LOCALLY.
   * **A little ontology goes a long way**
   * There can be a lot of value just getting the data federated and semantically aligned.
4. Unfortunately it is now so easy to create ontologies that myriad incompatible ontologies are being created in *ad hoc* ways leading to the creation of new, semantic silos
5. The Semantic Web framework as currently conceived and governed by the W3C (modeled on html) yields minimal standardization
6. **The more semantic technology is successful, they more we fail to achieve our goals**

# Areas of Use (both current and future)/Areas of non-use

Ontology Design Patterns for Systems Engineering

Ontology for Software Production - Instantiating the ontology describes design of a particular system

* Decisions considered, rejected, made, changed
  + Rationale
* Formal software artifacts
  + Source and executable code; specifications; machine-readable models
* Structured informal artifacts
  + Pseudo-code, requirements, graphical models, test plans, email addressing info, subject
* Unstructured artifacts
  + Email body, notes, code comments, etc.

Cyber-Physical Social Data Cloud Infrastructure

NIST & NICT Collaboration Project R&D of a cloud platform specialized for collecting, archiving, organizing, manipulating, and sharing very large (big) cyber-physical social data

Use case 1 - Healthcare data publishing & sharing

Use case 2 – Location Aware -based Service (e.g., disaster)

Globally monitoring and locally fencing (safe and rapid evacuation)

Information and Communication Technology (ict)

* Too much data
* Too much speed
* Too much complexity

Why a Materials Genome Initiative? Materials Are Complicated Systems Modeling is a Challenge

The Materials Genome Initiative is a new, multi-stakeholder effort to develop an infrastructure to accelerate advanced materials discovery and deployment in the United States. Over the last several decades there has been significant Federal investment in new experimental processes and techniques for designing advanced materials. This new focused initiative will better leverage existing Federal investments through the use of computational capabilities, data management, and an integrated approach to materials science and engineering.

Next steps

* File repository for first principles calculations
  + File repository for CALPHAD calculations
  + General data repository Prototype repository for data used in Calphad assessments
* Evaluation of data storage formats (e.g. markup language, hierarchical data format)

# Accessibility (i.e., ease of use)/Impediments

Ontology Quality for Large-Scale Systems

Ontology Tools and Training for Systems Engineers

# Recommendations

Some big systems and systems engineering needs and desires of ontology are:

* Fastintegration of data
* Integrated heterogeneous data, linked data, and structured data
* Easy exploitation of data
* Fine‐grained provenance of federated data.
* An Open, Transparent Platform for Everyone
* More opportunities for social, economic and political participation
* Open platform for everyone, new public good
* Non-expert system
* Crowd sourcing, citizen science
* Establish new information ecosystem to create new opportunities, services and jobs
* Benefit from cultural diversity
* Value-sensitive design

The European FuturICT (Information and Communication Technology) Paradigm is:

* Create a Big Data Commons
* Ethical, value-sensitive, culturally fitting ICT (responsive + responsible)
* Privacy-respecting data-mining
* Platforms for collective awareness
* Participatory platforms, new opportunities for everyone
* A new information ecosystem
* Coevolution of ICT with society
* Democratic control
* Socio-inspired ICT (socially adaptive, self-organizing, self-regulating, etc.)
* A ‘trustable web’

Big data might benefit from ontology technology but why this usually fails

* How to do it right
  + how create an incremental, evolutionary process, where what is good survives, and what is bad fails
  + create a scenario in which people will find it profitable to reuse ontologies, terminologies and coding systems which have been tried and tested
  + silo effects will be avoided and results of investment in Semantic Technology will cumulate effectively
  + ontologies should mimic the methodology used by the GO (following the principles of the OBO Foundry: [http://obofoundry.org](http://obofoundry.org/))
  + ontologies in the same field should be developed in coordinated fashion to ensure that there is exactly one ontology for each subdomain
  + ontologies should be developed incrementally in a way that builds on successful user testing at every stage
* AmandaVizedom: I can envision a Grand Challenge like this:
* Create a tool, of the sort that would work with an ontology repository such as OOR, to support the following activities (make them relatively easy and make them reliable/repeatable):
  + (a) someone with an ontology registers it, and either adds it to the repository or provides sufficient information for the tool to access it remotely. The tool provides assistance identifying key properties of the ontology that are relevant to its suitability for various types of usage. This assistance includes some manual entry, some automated validation and metrics generation, and some semi-automated generation of information.
  + (b) Someone looking for ontologies comes to the tool and gets help finding ontologies that might meet their needs. The tool assists them in specifying their need, by entering their ontology-specific requirements to the extent that they know them, and by describing their aspects of the intended usage. The tool makes this process also semi-assisted. Key feature of this that makes it a Grand Challenge: It's not just building a tool; it requires the research and testing to establish some of the relationships between ontology characteristics and usage characteristics. It also requires not just implementation of known evaluation techniques, but also research to develop others. On the other hand, it need not be complete to be valuable. Increments of improvement could be high value advancements over the current state.

Need a science of multi-level complex systems!

Linked Open Data (LOD)

Linked Open Data (LOD) is hard to create

* Linked Open Data is hard to query (Natural language query systems a research goal)
* Two ongoing UMBC dissertations hope to make it easier
  + Varish Mulwad: Generating linked data from tables (Inferring the  
    Semantics of Tables)
  + Lushan Han: Querying linked data with a quasi-NL interface (Intuitive Query System  
    for Linked Data)
  + **Key idea:** Reduce problem complexity by having (1) User enter a simple graph, and (2) Annotate it words and phrases
* Both need statistics on large amounts of LOD data and/or text
* Linked Data is an emerging paradigm for sharing structured and semi-structured data
  + Backed by machine-understandable semantics
  + Based on successful Web languages and protocols
* Generating and exploring Linked Data resources can be challenging
  + Schemas are large, too many URIs
* New tools for mapping tables to Linked Data and translating structured natural language queries help reduce the barriers