Ontology Summit 2014
Big Data and Semantic Web Meet Applied Ontology

Track A-Semantic Content Reuse:
Synthesis

MikeBennett, GaryBergCross, AndreaWesterinen

April 28, 2014

Summary

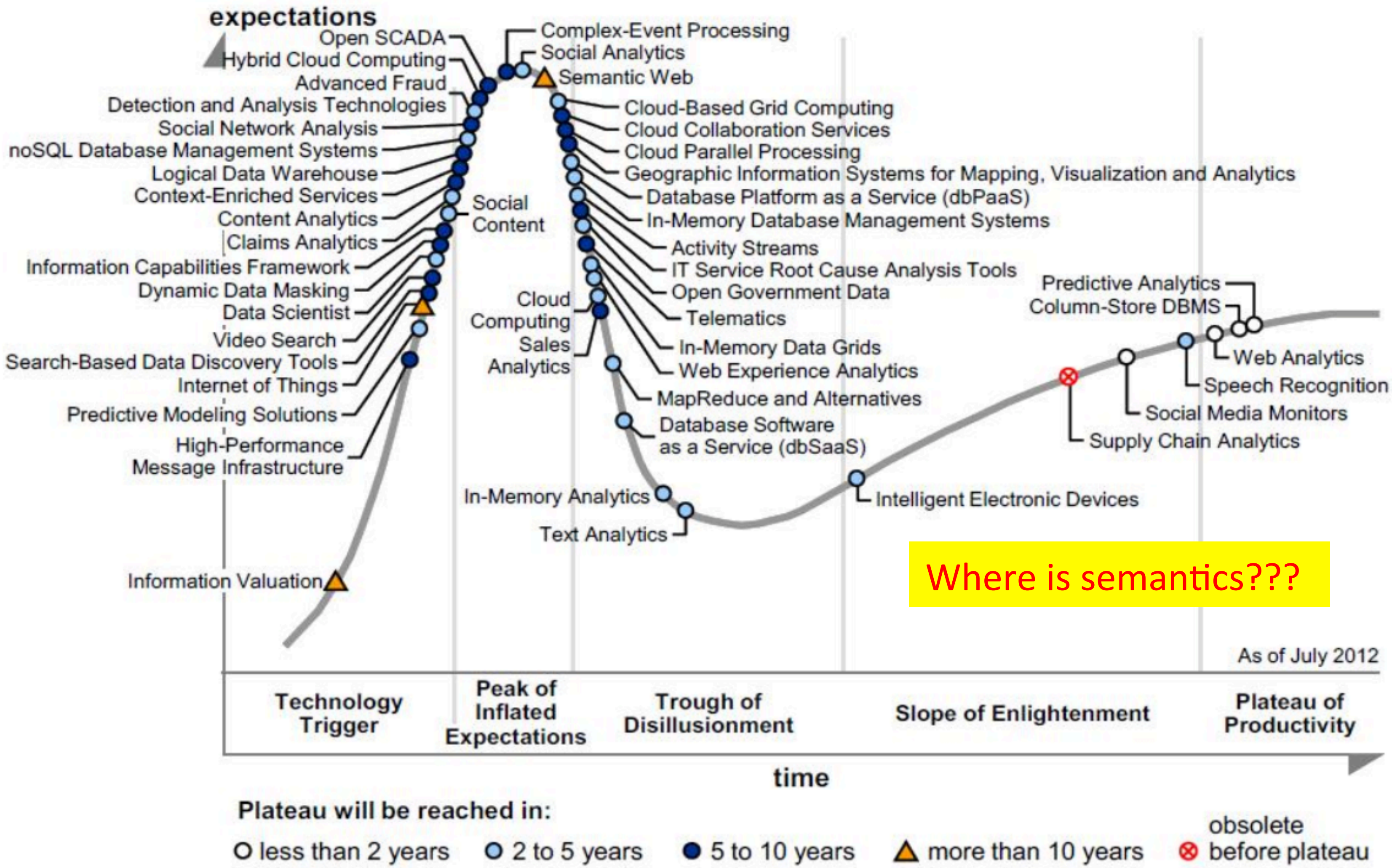
- Inputs:
  - 2 Track A presentations Sessions, Jan. & March 2014
  - Email dialogs
  - Co-Champion discussions & community page
- Track A Goals:
  - Define/document:
    - Explicit conditions for and issues with reuse
    - Concepts/meta-ontology
    - Approaches to modularization and best practices
    - Specific design patterns and exemplary content
      - For content reuse in applied ontologies and semantic web/linked data, and for reasoning and big data
  - Expand tooling, such as OOR, to enable defining and finding reusable content

Introduction, Mission and Scope of Track A

- Semantic technologies such as ontologies and related reasoning play a major role in the Semantic Web and are increasingly being considered to help process and understand information expressed in digital formats.
- The mission of Track A is to leverage common semantic content to reduce the burden of new, quality ontology creation while avoiding silos of different ontologies.
- The range of semantic content reuse being used on the Web and Big Data is broad.
  - We've considered content are whole or partial ontologies, ontology modules, ontological patterns and archetypes, and common, conceptual theories related to ontologies.
  - The role and relation of methods, bottlenecks and tools has also been discussed.

Big Data Landscape

Figure 1. Hype Cycle for Big Data, 2012



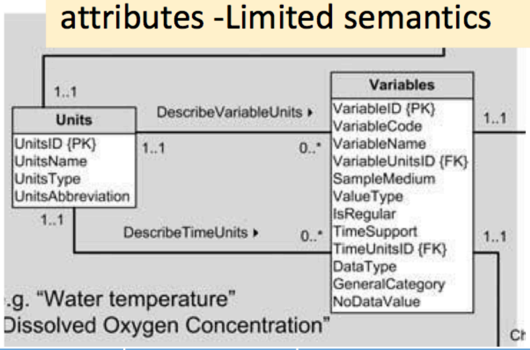
Where is semantics???

Source: Gartner (July 2012)

Big Data Vocabularies

Example fro Big Data Domain- Hydrology- Variables, Tags & “Ontology” Concepts

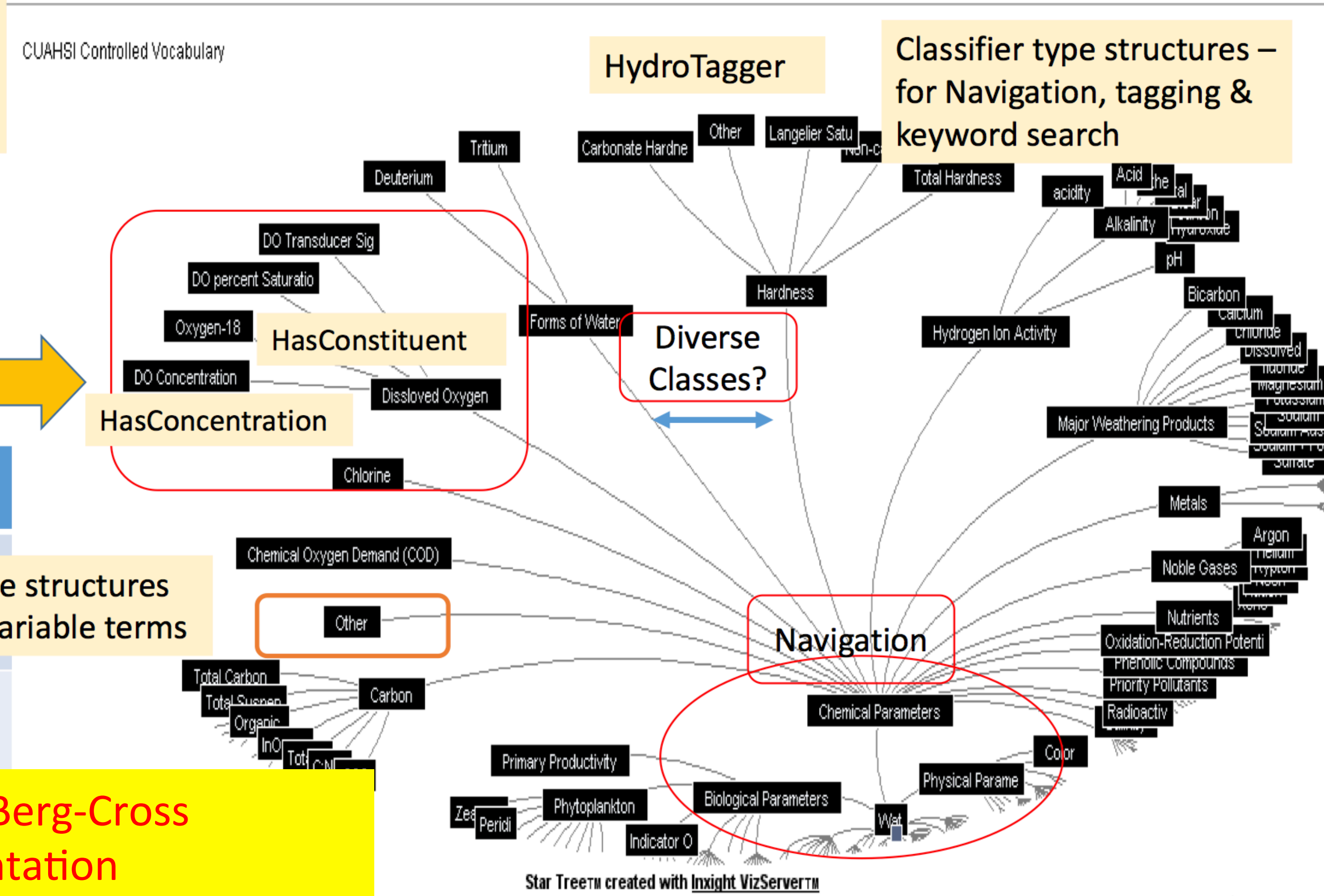
Observation DM uses RDB structure to
Integrate files & handle heterogeneity, Good MD attributes -Limited semantics



Concept ID	Concept Name	Ontology Layer
41	Chemical	1
42	Organic	2
43	PCBs	3
1001	Homolog Groups	4

1220 Deca_Chloro_PCB 5

Classifier type structures Connect to variable terms



HydroTagger

Classifier type structures – for Navigation, tagging & keyword search

HasConstituent

HasConcentration

Diverse Classes?

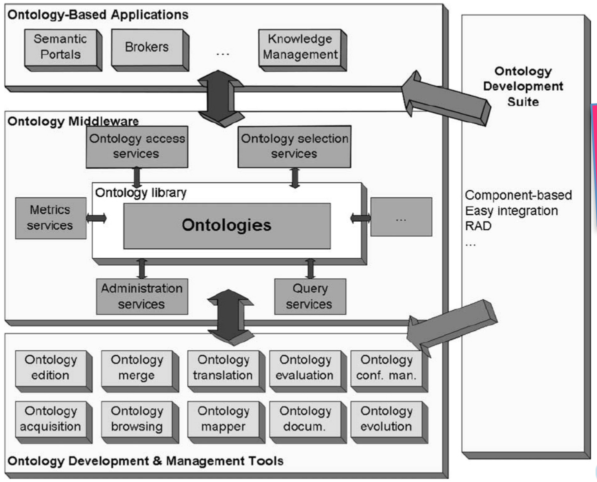
Navigation

From Berg-Cross Presentation

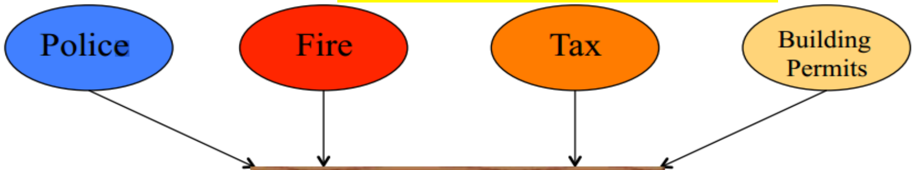
Vocabularies need semantics!

Summit's Track Themes Are Highly Related

Tools, Methods, Services



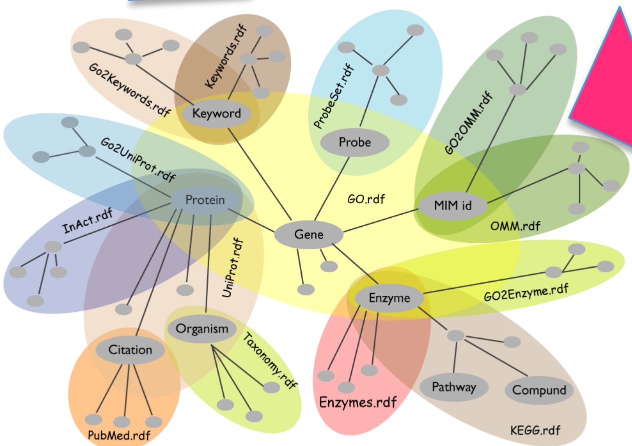
Data Variety



New York's Analytics team created a dataset combining data from 19 agencies:

- Buildings, preservation, police, fire, tax,

- By combining data analysis with expertise gleaned from inspectors, able to rate complaints such that in **70%** of inspector visits



Reuse is a bottleneck and is itself impeded by lack of tooling

Bottlenecks include:
Modeling axioms or knowledge representation language fragments that cause difficulties in terms of an increase in reasoning complexity or reducing the reusability of ontologies

Sharing and Reuse

- Reuse versus sharing …
  - Re-use: What does it take to make use of the work of others instead of having to re-invent?
  - Share-ability: How do you create an artefact in order for someone else to be able to re-use it?
- *How to* re-use versus What *makes* something re-usable?
- Reuse issues are not unique to ontologies/schemas
  - Parallels and differences with software reuse
  - Requires that the concepts (+ relationships, axioms and rules), assumptions and expression(s) of the included content meet a need, and can fit into the re-user's implementation
- Why reuse?
  - Reduce the development effort (by developing less)
  - Expand the benefit (improve the ROI) of the original content
  - Improve the quality of the original content (by identifying and eliminating errors)

Broad Questions

- What range of semantic content is being shared and used on the Web?
  - From vocabularies to formally axiomatized ontologies
- What are the issues with reuse?
- What ontologies are available/being used/required?
- How can we find this content?
  - e.g. LOV, ontology repositories

Issues

Tools

Success Stories

Best Practices

More Specific Track Questions

1. How can we characterize or measure semantic content reuse, both between ontologies and by Big Data and Semantic Web communities?
2. What building blocks of common semantic content exists now to enable interoperability?
   – What additions are needed to move forward and how are these best achieved?
3. What is involved in reuse of Linked Data versus reuse of ontologies?
4. What is an example of a small set of semantic content that the community might propose for reuse?
   – Is there agreement on these or things like ODPs as building blocks?
5. What is an example of a large set that the community might propose for reuse?
6. Is it reasonable to expect reuse of an entire ontology like DOLCE and Semantic Sensor Network (SSN)?
   – If so under what conditions might this be reasonable?
   – Is it better to expect alignment rather than exact content reuse?
7. Is reuse about semantics alone or should it also address reasoning and data analytics?

Our Speakers & Their Talks

1. MikeBennett **(EDM Council)** Overview of the track
2. Dr. GaryBergCross (SOCoP) - "Use and Reuse of Semantic Content: The Problems and Efforts to Address Them - An Introduction"
3. Professor PascalHitzler (Wright State U) - "Towards ontology patterns for ocean science repository integration"
4. Ms. AndreaWesterinen (Nine Points Solutions) - "Reuse of Content from ISO 15926 and FIBO"
5. Ms. MeganKatsumi & Professor MichaelGruninger (U of Toronto) - "Reasoning about Events on the Semantic Web"
6. Dr. JohnSowa (VivoMind Intelligence) - "Historical Perspectives: On Problems of Knowledge Sharing"
7. Professor MichelDumontier (Stanford BMIR) - "Tactical Formalization of Linked Open Data"
8. Mr. KingsleyIdehen (OpenLink Software) - "Ontology Driven Data Integration & Big Linked Open Data"

Presentations: Common Themes

- Uses of ontologies: Annotation, query/search/retrieval, reasoning, integration etc.
- What makes an ontology meaningful?
  - Is it meaningful because you can reason over it? OR
  - Can you reason over it because it is meaningful?
- Ontology Design Patterns (ODPs)
  - Applicability
  - Abstraction levels
- Identifying what's reusable
  - Intended scope
  - Confidence (metadata, documentation etc.)
  - Overlaps, equivalences etc. across multiple reused ontologies
  - Theories, microtheories: There is no "one" ontology of the world
- When to use a whole ontology versus a snapshot or extract
- Tooling and interaction
- Knowledge Representation best practice
  - See Sowa's Historical Perspectives On Problems of Knowledge Sharing

Themes Deeper Dive

- Patterns (two kinds)
  - Ontological abstractions, e.g. Event (an Ontology Design Pattern)
  - Axiom patterns – Modelling approaches for concepts; String vs Thing
- Implementation and reuse strategies
  - Bottom Up: Derivation of meaning from example data
  - Top Down: Extension from common abstractions
  - E.g. Event extended to Trajectory; Trajectory extended to Cruise, Trip etc.
- Integrating ontologies
  - Spectrum of ontologies; Network of ontologies
  - Comparable to system design and integration
  - Tooling to support integration level view of new, extended and reused ontologies
- Common business language versus common logical expression
- Personnel: how to get the right mix and interaction?
  - Business domain experts; Knowledge representation specialists; Formal logic experts
  - Tooling to support the required level of information interchange

Reuse Issues

- For successful reuse of semantic content it is important to understand how content is being used, with what methods to coordinate reuse are available and what tools are helpful.
- Tooling for modularity, documentation, etc. is critical
  - Broader use by mainstream efforts including Big Data is bottlenecked in part by the paucity of semantic tools integrated into mainstream tools along with the inherent learning curve of understanding semantics.
- In practice reuse is dependent on both the availability of well-documented content AND tooling that supports finding and incorporating this range of content.

Perspectives for Reuse

- User Perspective
  - What do I look for in an ontology in order to confidently reuse it?
- Originator Perspective
  - What needs to be in place for someone to be able to reuse my ontology?

Conditions for Reuse

Critical

- Content is accessible, can be easily found and is documented
  - Documentation includes the basic details of the semantics, and the range of conditions, contexts and intended purposes for which the content was developed
- The re-user has defined their competency questions and overall micro-theories, and is motivated to find the content
  - Q: When it is best to reuse content within the development lifecycle?
- The content is in a form conducive to re-use or can be converted/transformed to a usable form
- The re-user knows how to do the conversion/transformation
- The content is logically consistent with the micro-theories of the re-user and this can be established
- The re-user trusts the content and its quality, and believes that this quality will be maintained

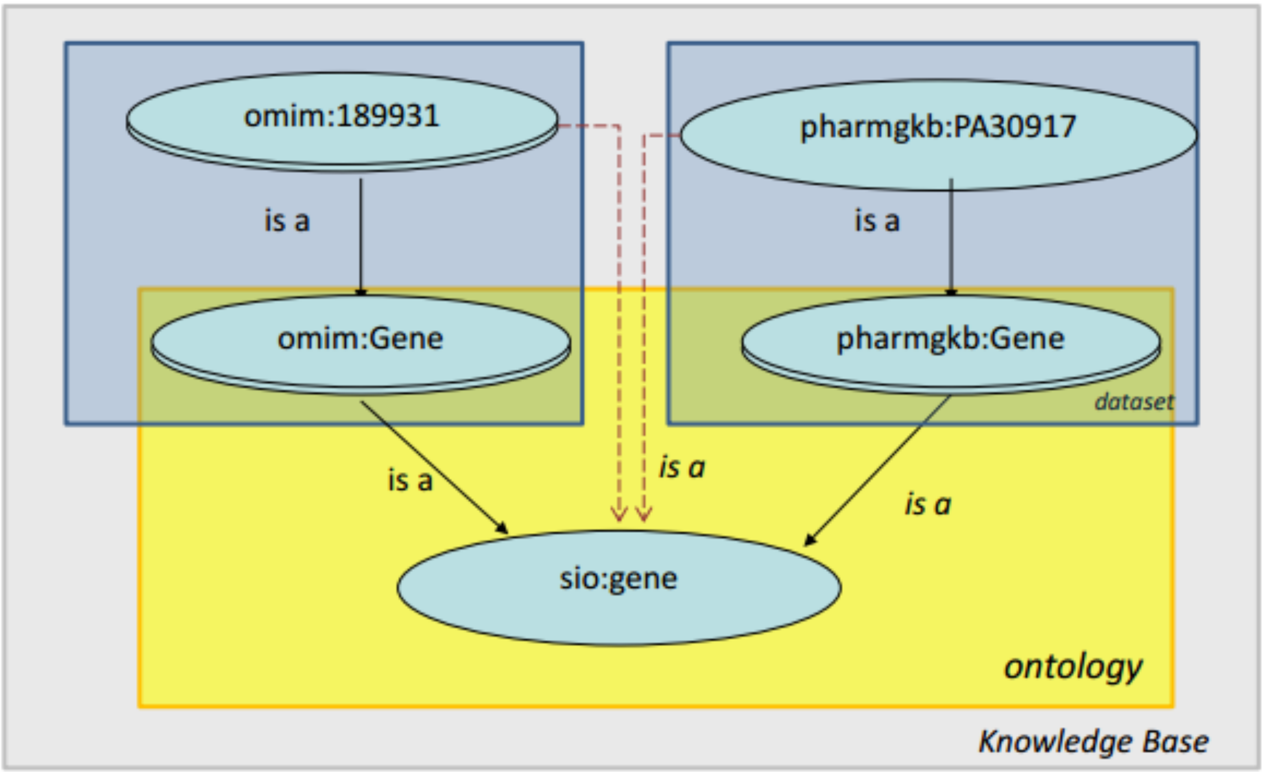
Need to broadly standardize metadata (annotations and object/data properties) for reuse and provide exemplars

schema.org

- Defines a widely used (and extensible) vocabulary for describing the contents of a web page
- Addresses the issues of:
  - Finding reusable content
  - Managing the size and complexity of the content
  - Integrating the various pieces and extensions together
  - Maintaining quality and trust
- Concepts are well documented including directions on how to use and extend the vocabulary
- Users are supported via blogs and discussion groups

SemanticScience **Integrated** Ontology Example

Semantic data integration, consistency checking and query answering over Bio2RDF with the Semanticscience Integrated Ontology (SIO)

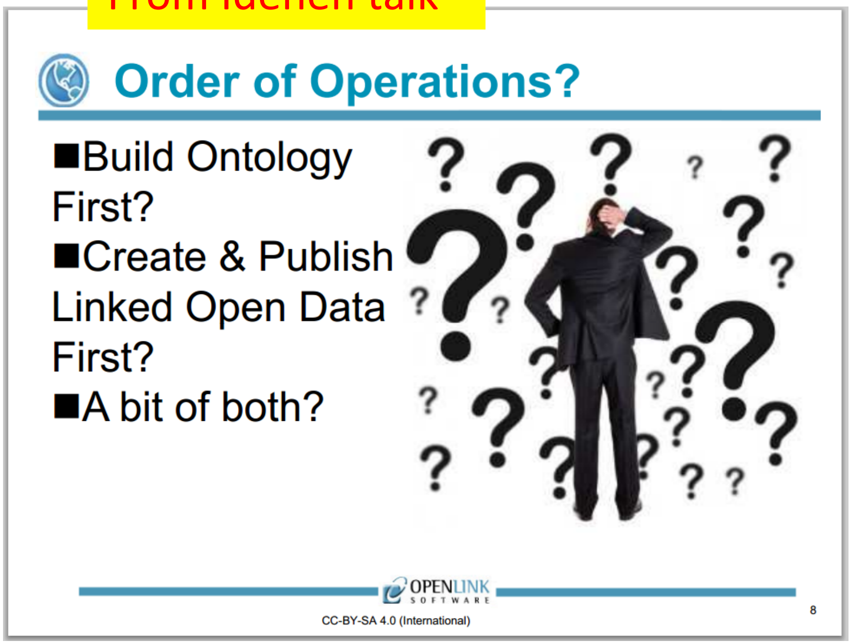


Querying Bio2RDF Linked Open Data with a Global Schema. Alison Callahan, José Cruz-Toledo and Michel Dumontier. Bio-ontologies 2012.

From Dumontier talk

Using Ontologies for Linked Data and Big Data

From Idehen talk



- Top down versus Bottom Up approach
  - Analysis of linked data to define ontology
  - Creation of ontology from common abstractions, patterns
  - Meet in the middle?

- Inference Rules
  - Enable / disable inference rules on available ontologies
  - Use inference rules, conditionally
- Don’t let the perfect be the enemy of the good

Reasoning - A Research Question

Objective

Research Questions

From Katsumi & Gruninger talk

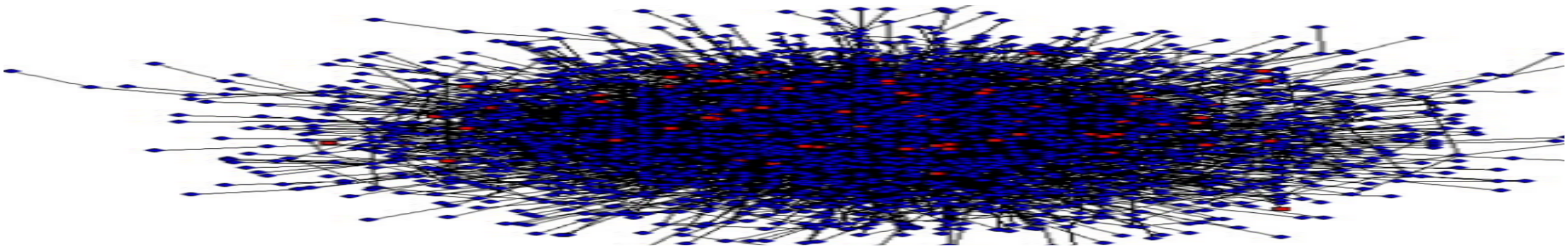
All of these observations raise the questions:

- Are Semantic Web ontologies able to support non-trivial reasoning problems?
- If not, why?
- Are the existing ontologies simply not designed with enough semantics to support these applications, or have they reached the limit of what Semantic Web languages can support?

Tooling – What Is Needed?

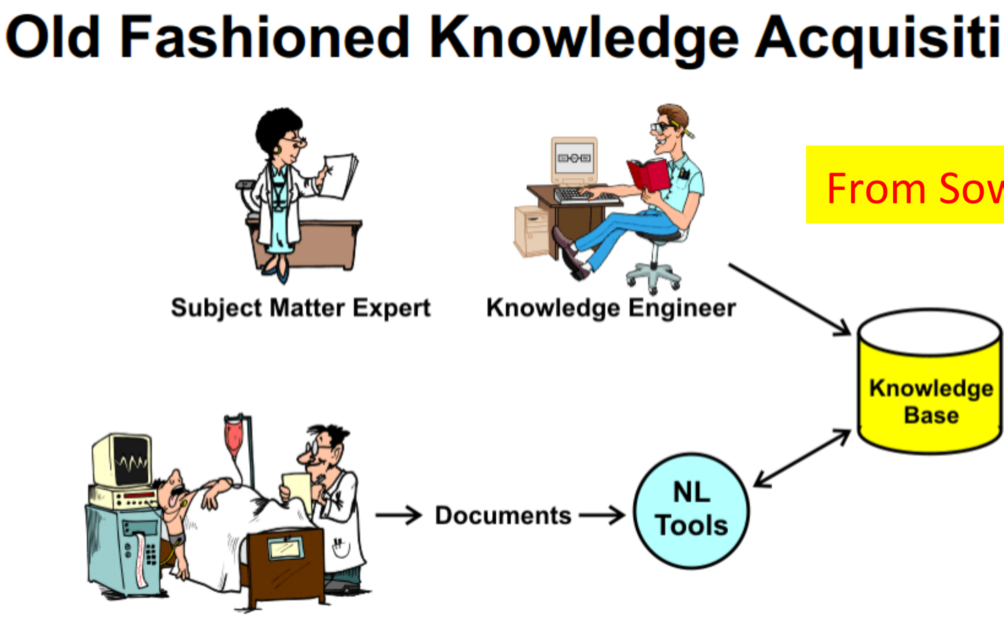
- Ontology repositories supporting modularity and governance, with good search capabilities
  - For finding relevant content
  - Including both topical ontologies and linked data schemes
  - Integrated with standard IT tooling
- Possibilities:
  - Open Ontology Repository
  - Linked Open Vocabulary
    - "In an increasingly linked data world, vocabularies rely more and more on each other through reusing, refining or extending, stating equivalences, declaring metadata … LOV provides a service to find relevant vocabularies."
- Q: What tools (and techniques) are needed to support the development of modular ontologies and schemas?
  - Are they different than the tooling to query/find/ reuse the modules?

Tooling for Ontology Management



- As noted by MichelDumontier
  - "the lack of coordination makes Linked Open Data … quite chaotic and unwieldy … [there is a] Massive Proliferation of Ontologies / Vocabularies "
- Ontologies must include consistent, supporting metadata for query
  - Possible metadata includes context, use cases, labels, governance information, etc.
  - Possible definition is the Ontology Metadata Vocabulary
- Reuse enhanced by feedback and user input
  - Possibly include both a recommendation system and feedback mechanisms in the repository
- Governance needs a process and its enforcement
  - Process should include open consideration, comment, revision and acceptance

Tooling for Knowledge Acquisition and Verification



From Sowa talk

- Using Controlled Natural Language tools to generate candidate ontologies
  - May ease the KA bottleneck

- Verification
  - Similar to testing an algorithm against input values
  - Run pre-defined SPARQL queries (based on the competency questions) against imported data

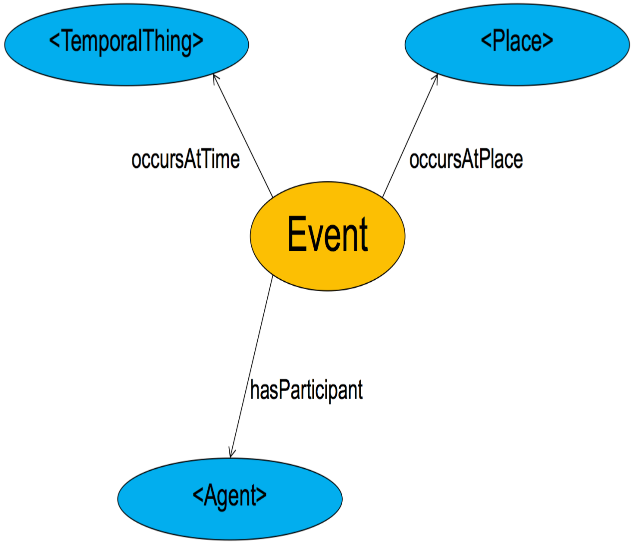
Best Practices – Upper Level Ontologies

- Upper-level ontologies, such as Aristotle's, Wilkins', or Kant's, show the broad patterns of how things fit together
  - For interoperability, upper level definitions must be underspecified with the barest minimum of axioms
- Be aware that different problems and different purposes require different representations and processing/algorithms
- For precise reasoning and problem solving, the details must be pushed down to the highly specialized, lower-level ontologies and design patterns

Best Practices - Modularization

- Create small, more **modular** ontologies and **tactical** schemes
  - More possibilities for reuse due to greater focus and cohesiveness, and likely less dependency on the original context
  - Modularity viewed from the perspective of the user, not the creator
- Collect and document approaches to modularization, best practices and specific patterns
- Dimensions of variability should be addressed
  - Variability across the contexts (for example, a certain concept or property may be present or absent in different contexts and uses)
  - Variability over time

From Hitzler talk



Ontology Design Patterns are "reusable successful solutions to recurrent modeling problems."

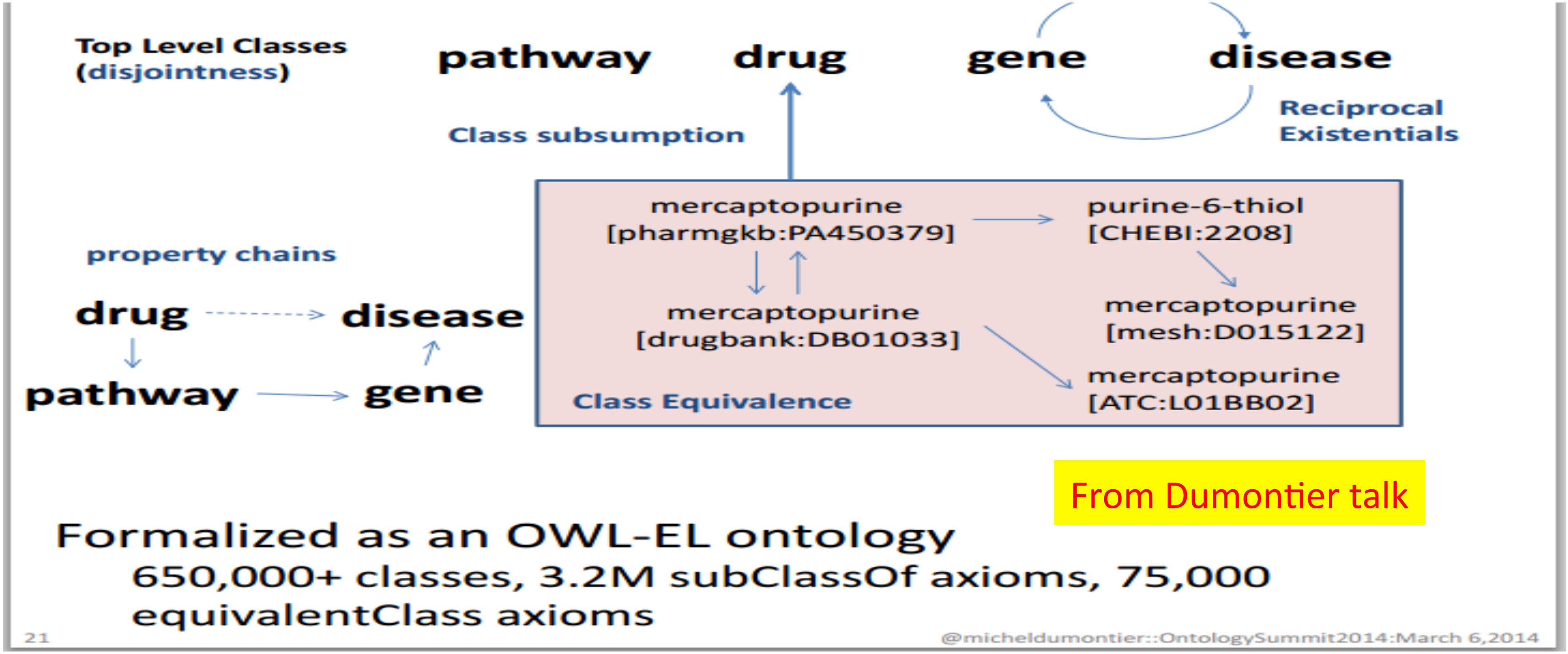
Best Practices – Modularization and Integration

- Create modules to separate classes/concepts, from properties, individuals, axioms and domain-specific usage, analysis, traversal and diagnosis information
  - Easier to target what is possible to reuse and reduces the amount of transformation and cleaning that is necessary
  - Take care to distinguish definitive entities ("defining" core concepts) from pragmatic ones (related to the business uses or a particular domain)
- Use ”integrating" modules
  - Employing owl:equivalentClass and axioms to map between the concepts, properties, etc. of all the ontologies that address an application/domain
- Use a consistent, explicit approach and partitioning scheme across all modules

"Tactical" Formalization

- Reuse what you need and represent it in a way that directly serves your objective

Applied to discovery of drug and disease pathway associations



Best Practices – Ontology Formalization

- Document everything, including identifying the application domain
  - Via well-established labels and predicates (SKOS, Dublin Core, etc.)
  - Describe why the ontology/schema was created, how it was tested, and how it may be used
- Where possible, describe reuse across multiple domains
  - Such that the ontologies and schemes represent "common needs” with no single domain focus
- Plan for evolution and document it when it occurs
- Correctly use domain and range properties
  - Take care to use domains and ranges where they clarify meaning and semantics
  - But avoid “lockdown” of meaning too early (where possible, “float” the domains and ranges high to enable later reuse)

Best Practices - Naming

- Names can be surrogate or human-readable identifiers
  - Both approaches have their advocates, and pros and cons
- Labels are valuable as documentation (such as from SKOS)
  - Regardless of the identifier scheme that is chosen
- Organize and name concepts distinguishing general from application-specific concepts
  - Avoid misapplication of very specific concepts as general entities

BACKUP

Approach

- Enlisted 6 speakers and the community to discuss reuse issues and problems, and present their efforts and experiences to address these.
- We referenced past Ontology Summits (for example, the Ontology Repositories discussions)
- Promoted discussion of track session topics on the Ontolog/Summit forum both before and after sessions
- Organized a Hackathon reuse and tool exercise.
- Distilled the virtual meeting topics to a useful summary and set of speakers for the face-to-face Symposium.