Acquiring Experience with Ontology and Vocabularies

Walt Melo
Risa Mayan
Jean Stanford

The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the authors.

MITRE

About MITRE

- Context of MITRE work:
  - Independent
  - FFRDC
  - No vendor affiliations

Background

- The agency receives a large amount of information through free text documents and images created by external entities.

- The submitted information drives many areas of the agency's work:
  - Enforcement activities
  - Policy development
  - Product approval
  - Public health research

- Regulators and scientists need to:
  - Analyze the submitted documentation
  - Research internal and external documentation related to domain
  - Formulate policies based on document analysis

MITRE

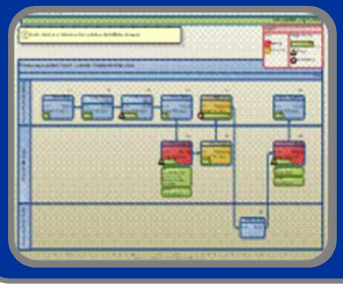
Agency Problem

- Difficult to find relevant documents. Need better ways to support:
  - Naming conventions and data standards
  - Document classifications or taxonomy
  - Document traceability
  - Standard terminology for components described in the documents
- Difficult to query across internal and external information sources: federated search
- Difficult to perform queries using both structured and non-structured information

Take Ways

- This presentation describes work done to:
  - Create an ontology from publicly-available resources
  - Improve search of unstructured data via semantic text mining
  - Enhance data quality by leveraging structured terminologies and data standards

Approach for Applying Ontology and SOA to Improve Federated Search of Documents

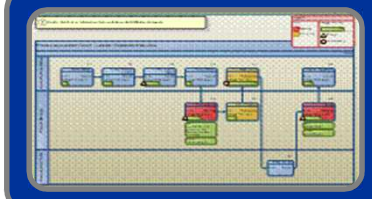


1) Define Business Architecture

2) Define Service & Technical Architecture

3) Define Ontology & Vocabularies

MITRE



Business Architecture

- **How**
  - **Prioritized relevant business functions as defined by Enterprise Architecture**
  - **Developed business processes**
  - **Used business modeling (BPMN)**
    - **Conducted Workshops with SMEs and key stakeholders**
    - **Leveraged best practices from the field to define business processes**
- **Why**
  - **Align the enterprise business processes to its business vision and strategic goals.**
  - **Identify needed capabilities (what *type of* business services are needed to support ontological search)**
  - **Define roles & responsibilities**
    - **"Who will be responsible for doing what"**
  - **Obtain buy-in**
  - **Disseminate ideas**
  - **Inputs to Requirements**

Service & Technical Architecture

- How
  - Gathered requirements
    - What capabilities are needed?
  - Technology inventory / survey
  - Identified business service blueprint
    - What are the major business & system services?
    - What capabilities they must provide?
  - Mapped business services to business processes
    - What business & system services are need to support the prioritized business processes?
- Why
  - Define an architecture blueprint for business services and technology
  - Identify capabilities gaps
    - What services are currently available?
    - What services are needed?
  - Prioritize procurement process
    - What *type* of COTS can provide the needed capabilities? Such as:
      - ontology engineering tools
      - semantic text mining
      - vocabulary management

MITRE

Ontology & Vocabulary

- **How**
  - **Used an approach based on semantic web technology and ontology for:**
    - **Defining data standards**
    - **Creating a conceptual data model**
    - **Addressing federated data integration**
    - **Improving document classification**
    - **Reusing existing ontologies relevant to domain**
- **Why**
  - **Improve search capabilities by scientists**
  - **Define a common vocabulary**
  - **Allow easy integration of structure and unstructured federated data sources**
  - **Allow automatic document classification**
  - **Enhance collaboration among inter- and intra- research groups**

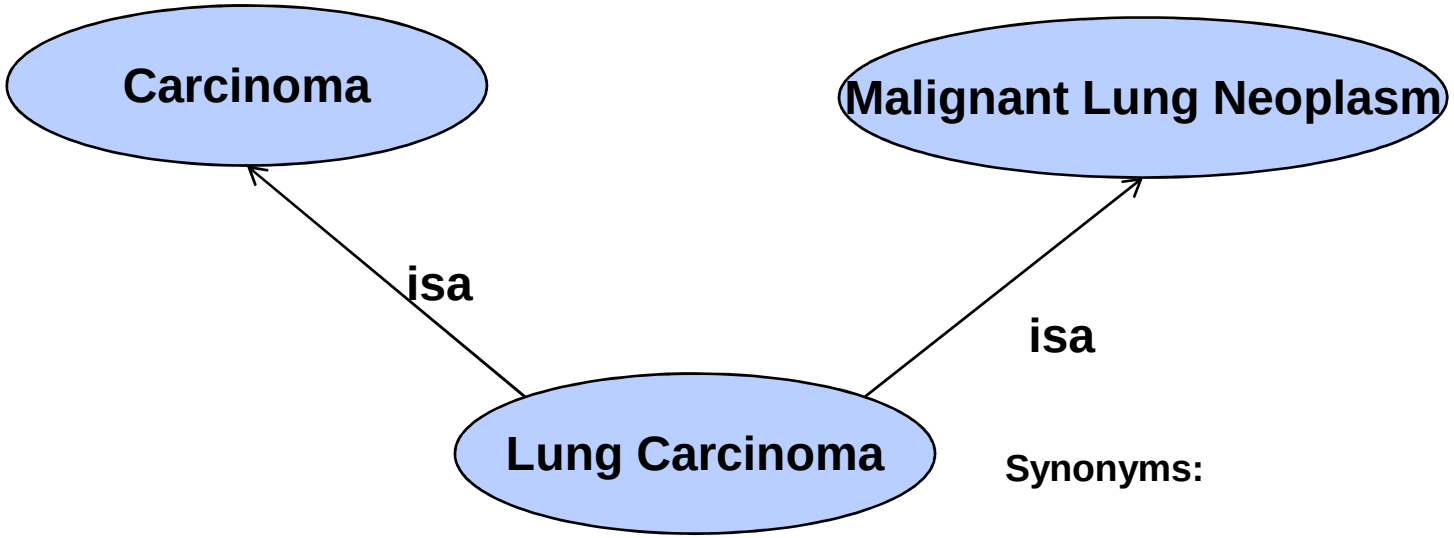
Definitions

- ***What is an Ontology?***
  - An ontology defines the terms used to describe and represent an area of knowledge. It includes computer-usable definitions of concepts in the domain and the relationships among them.
- ***Why should we care?***
  - Ontologies are used by people, databases, and applications that need to share domain information, such as the health care domain

Rationale

- Permit semantic interoperability - standard terminologies ensure the same words and phrases mean the same thing
- Provide consistent way to annotate industry submissions concerning products, ingredients, and constituents
- Enable translation of health-related information received from industry into a structured format that can be analyzed / mined
- Provide structured way to map the domain specific ontology with other relevant ontologies, such as diseases, toxicology, and so on
- Allow the semantic integration of data from different data sources – a semantic data warehouse

An Example: Deceases

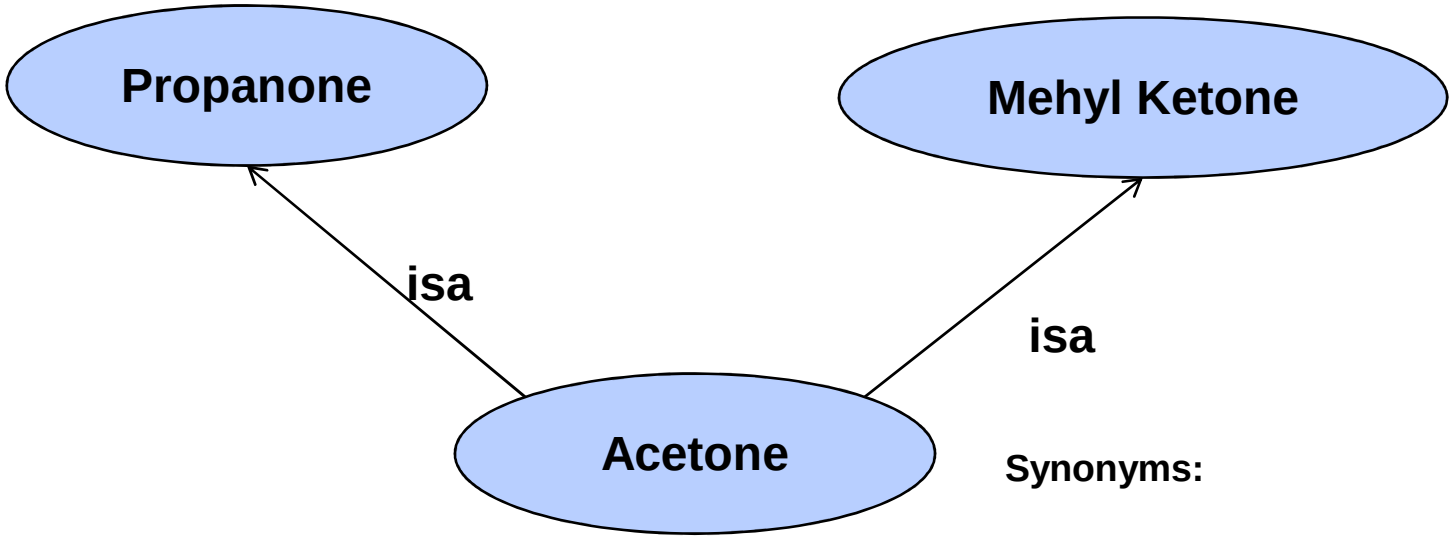


Synonyms:

- **Cancer of Lung**
- **Lung Cancer**
- **Cancer of the Lung**
- **Carcinoma of the Lung**

Disclaimer: this is not the actual ontology created by the agency

An Example: Chemical Substances

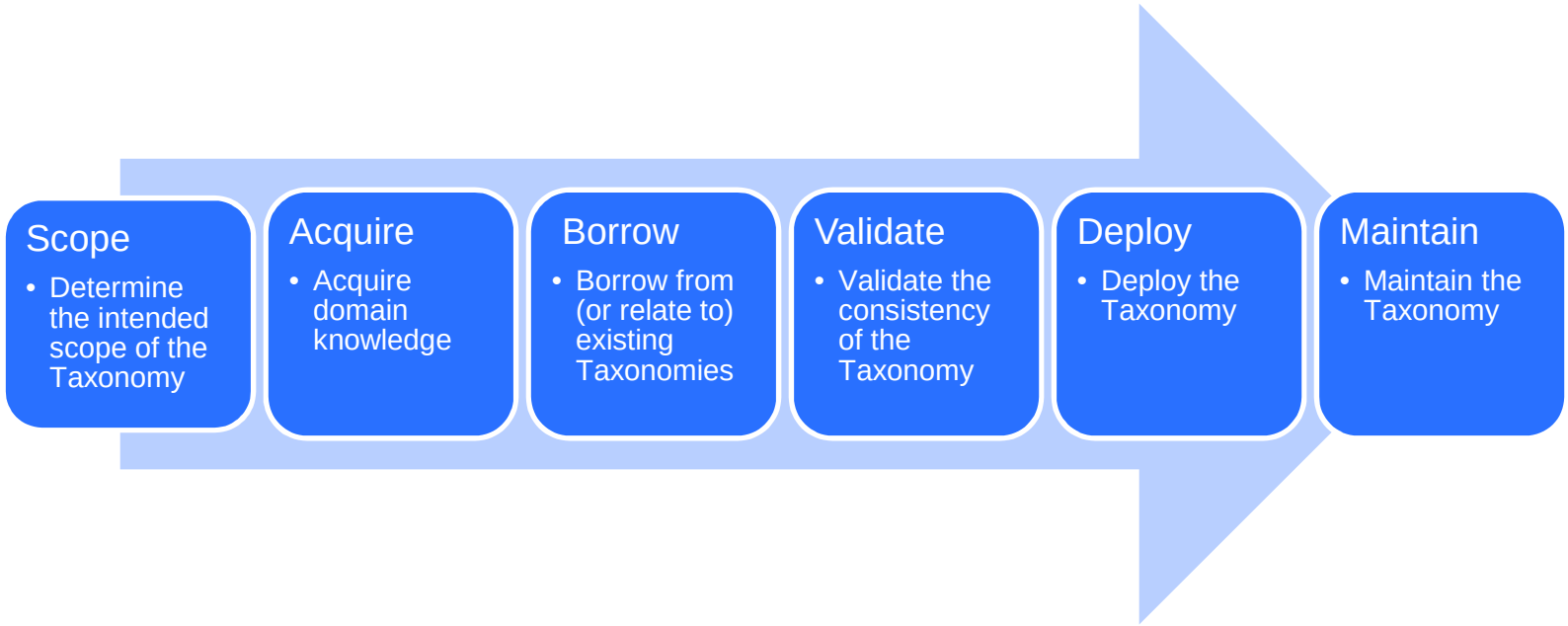


Synonyms:

- **Dimethyl ketone**
- **2-Propanone**
- **C3H6O**
- **dimethylcetone**

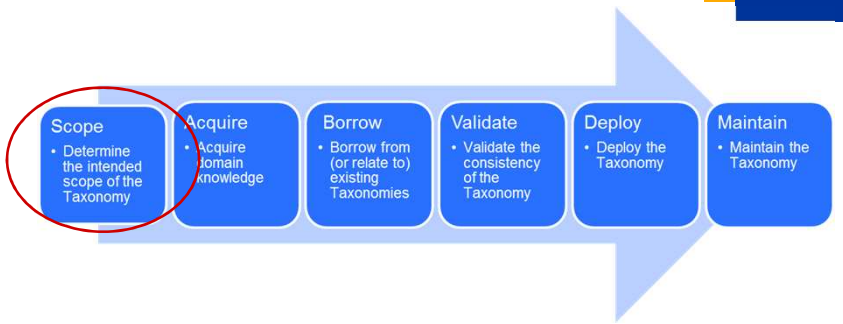
Disclaimer: this is not the actual ontology created by the agency

High Level View of the Ontology Development Cycle



Baclawski K, Niu T. Ontologies for bioinformatics. Cambridge: MIT Press; 2006.

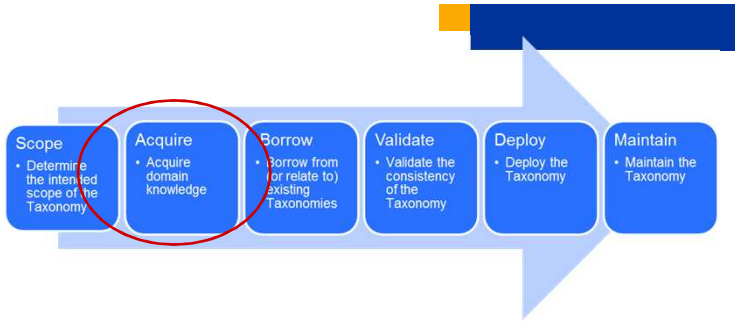
Scope



A bio-medical domain with a strong interrelationship with other domains

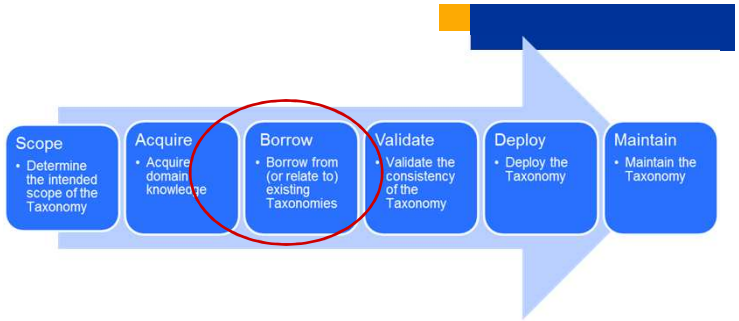
- **Medical**
- **Chemical**
- **Diseases**
- **Toxicology**
- **Legal entities**
- **Standard organizations**

Acquire Domain Knowledge



- Leverage SME from different domains:
  - Chemical
  - Medical
  - Toxicology
  - Manufacture
  - Public Health
- Utilize public domain publications
  - Specialized publications
  - Well known web sites maintained by non profit organizations
  - Standards
- Reuse existing efforts
  - National Cancer Institute (NCI)
  - National Library of Medicine

Reuse: Ontologies Being Investigated

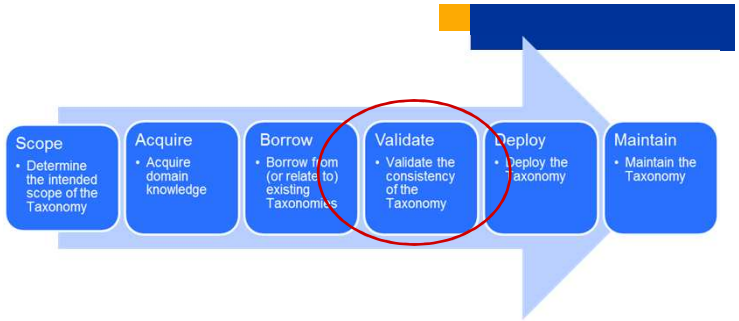


- Chemical Entities of Biological Interest ~ freely available dictionary of molecular entities focused on 'small' chemical compounds.
- National Cancer Institute (NCI) Thesaurus ~ widely recognized standard for biomedical coding and reference, used by a broad variety of public and private partners both nationally and internationally.
- SNOMED CT (Systematized Nomenclature of Medicine--Clinical Terms).
- And so on

Do not re-invent the wheel:

- Import existing ontologies
- Map created concepts to these ontologies

Validation



Method	Notes
Gold Standard	Compare to a gold standard. *In our case, there is no gold standard, but it is possible to develop metrics of quality (such as term coverage, lack of ambiguity) that can be applied to candidate taxonomies.*
Task Based	This approach involves testing the candidate taxonomies against a specific task (such as identifying all variations on a specific chemical component) and assessing the results.
Data or Corpus Driven	*This method compares the fit of a terminology to texts in a domain. In our case, this involves testing how many concepts in the terminology were found in a set of industry documents and evaluating how useful the terminology was in navigating through the documents.*
Manual Assessment Against A Set Of Pre-Defined Criteria	This would involve manual inspection of the terminology and development of specific related criteria, such as the number of synonyms that were correctly identified. *In our case, this validation was not yet done.*

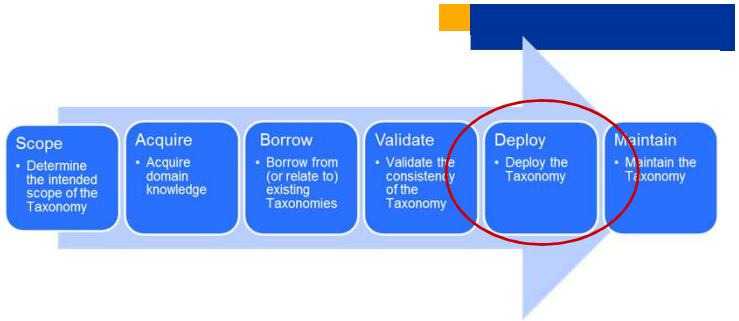
G MAIGA, DDEMBE WILLIAMS, "A Flexible Approach for User Evaluation of Biomedical Ontologies."

Int'l Journal of Computing and ICT Research, Vol. 2, No. 2, December 2008

Corpus Driven Validation: an example

Concept	Preferred Name	NIST	CAS#
CHEBI_15347	acetone	Yes	67-64-1
CHEBI_15366	acetic acid	Yes	64-19-7
CHEBI_22698	benzaldehydes		
CHEBI_31457	decanal	Yes	112-31-2
CHEBI_29309	methyl	Yes	2229-07-4
CHEBI_29805	glycolate		
CHEBI_35701	ester		
CHEBI_18346	vanillin	Yes	121-33-5
CHEBI_22695	base		
CHEBI_25627	octadecadienoic acid		
CHEBI_27542	methyl oleate		
CHEBI_32368	undecanoic acid	Yes	112-37-8
CHEBI_32544	nicotinate		
CHEBI_35209	label		
CHEBI_35366	fatty acid		
CHEBI_42504	pentadecanoic acid	Yes	1002-84-2
CHEBI_48408	ethyl vanillin	Yes	121-32-4

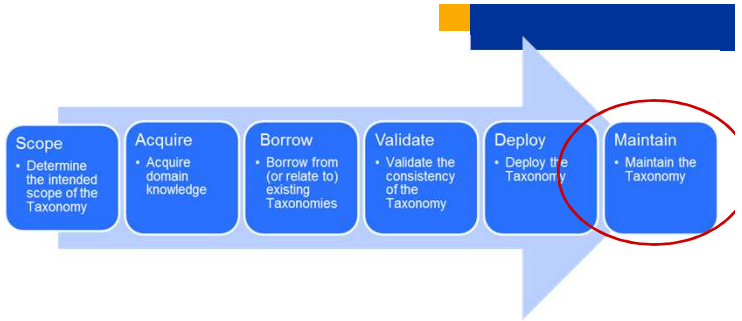
Deploy



- **Several alternatives**
  - **Internal**
    - **Only for inter agency systems , users, and authorized partners**
    - **Normative**
  - **External**
    - **Available to external communities of interest**
      - **Open Source ontology?**
      - **National Center for Biomedical Ontology?**
      - **NCI EVS (Enterprise Vocabulary Services (EVS)?**
    - **Guidance and reference**
  - **A combination of both**

Alternatives being evaluated

Maintain

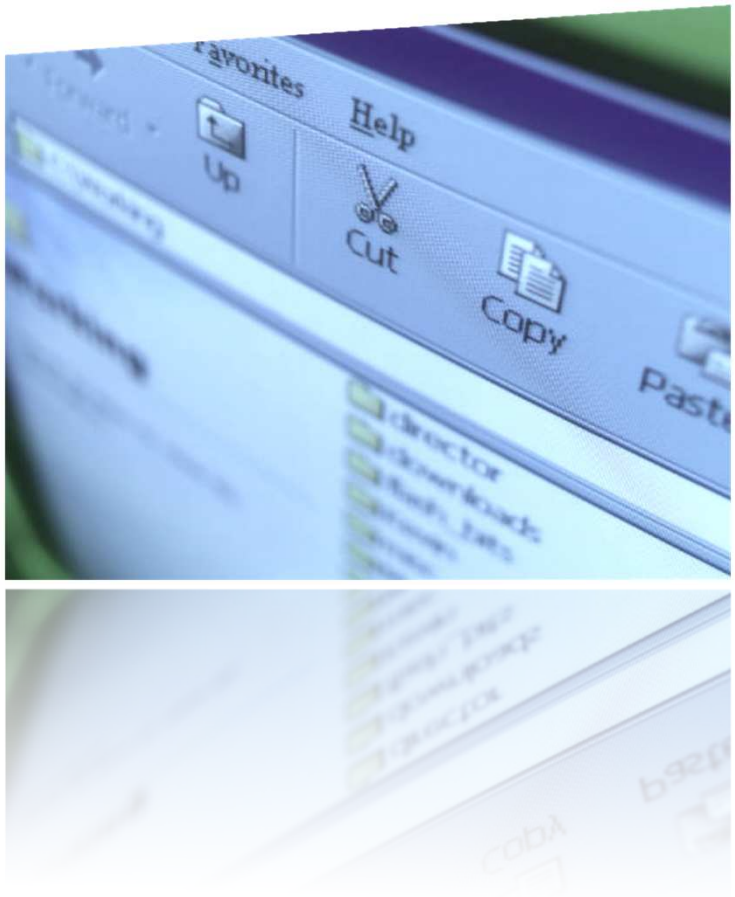


- Create a taxonomy management group
  - Unified team composed of key stakeholders
  - Data standard group

- Leverage collaborative development tools, such as Collaborative Protégé, supporting:
  - Discussion threads
  - Proposing and voting
  - Annotating ontology components
  - Support for user, groups, and access policies
  - Version and release control

Tools

- Ontology Engineering Tool
  - Protégé
  - TopBraid
- Semantic Text Mining
  - ODIE
  - NCBO Annotator
  - Open Calais
  - SmartLogic
- Ontology repository
  - Virtuoso
  - AllegroGraph



Market research is in progress:
This list is not exhaustive

Conclusion

- Semantic Web technology and products can be used integrate heterogeneous data sources in an architecture where semantics plays a pivotal role with metadata exchanges and ontology-based searches

- Semantic Text Mining tools allow analysts to:
  - execute queries which combine structured data with unstructured data
  - navigate unstructured data in a structured way

- Ontologies, taxonomies, and vocabularies can improve the productivity of those who need to review and evaluate large amount of documentation from different information sources