

Semantic Data

Chapter 11 : Application domains

Jean-Louis Binot

Sources and recommended readings

- The following case study documents are part of the course material :
 - Case study 2 : Montefiore (*Webinar by P. Mirhaji and other sources indicated in section 2*).
- Section 4 is not part of the material for the exam.
- Sources and useful additional readings :
 - The role of ontologies in bioinformatics is discussed in the review article *Managing, Analysing, and Integrating Big Data in Medical Bioinformatics* (Merelli et al. 2014).
 - The book *Ontology driven software development* (Pan & al. 2013) covers many aspects of the use of ontologies in software engineering.
 - The book *Handbook of Semantic Web Technologies* (Fensel et al. 2011) includes, i.e., a chapter on semantic web services.

Agenda

- 1 Bioinformatics: data deluge and ontologies
- 2 Medical informatics: a semantic data lake
- 3 Systems engineering: systems & models
- 4 Ontologies in software engineering
- 5 Thoughts on semantic data

The data deluge in biogenetics and medical sciences

- Omic sciences : multiple large-scale genome projects, among which :
 - [Genome 10K](#) (started in 2010, US) (genomes of 10,000 vertebrate species): 100 TB^(**) data.
 - [100K Genome](#) (started in 2012, UK) (100,000 genomes from around 85,000 NHS patients, reached in 2018).

- The rate is accelerating* :
 - In 2015, >2500 sequencing instruments in >1000 centers in 55 countries. Largest 20 store 100 PB^(**) of data.
 - The **Earth Biogenome Project**, created in 2018 by an international consortium (Smithsonian Institute, Beijing Genomics Institute...) plans to capture and catalog all the DNA from the world's flora and fauna (>1M species).
 - Countries are contemplating to sequence large portions of their populations : England ([Genomics England](#), 1M genomes), US ([All of Us](#), 1M), EU ([declaration of genomic cooperation](#), 1M+), China, others...

- Between 100 millions and 2 billions human genomes could be sequenced by 2025.
 - That would represent 2 and 40 exabytes^(**) of human genomic data.

*: *Big Data: Astronomical or Genomical?* (Stephens et al. 2015).

** : Terabyte : 10^{12} bytes; petabyte : 10^{15} bytes; exabyte : 10^{18} bytes.

Data deluge impacts – biomedical sciences

□ No time to cope

- “We have these giant piles of data and no way to connect them” said H. Steven Wiley, a biologist at the Pacific Northwest National Laboratory. “I’m sitting in front of a pile of data that we’ve been trying to analyze for the last year and a half.” ([*DNA Sequencing Caught in Deluge of Data*, A. Pollack, NY Times 2011](#)).
- Professor Brown of Michigan State said: “We are going to have to come up with really clever ways to **throw away data** so we can see new stuff.” (*same source*).
- “Right now, agencies like the National Institute of Health maintain public archives containing petabytes of genetic data. But without easy search methods, such databases are significantly underused, and **all that valuable data is essentially dead**.” ([*The DNA Data Deluge*, M. Schatz and B. Langmead, IEEE Spectrum 2013](#)).

□ Extremely energetic efforts, research, funding, to provide solutions.

Semantic solutions

□ Ontologies

- [Gene Ontology](#) : framework for the model of biology (OWL, OBO* - cf. next).
- [Cell ontology](#) : exhaustive organisation of cell types, excluding plants (OWL).
- [Human disease ontology \(DOID\)](#) : comprehensive hierarchical controlled vocabulary for human diseases (OBO).
- [UniProt](#) : protein sequence and functional information, integrated with GO annotations (RDF).
- [MGI](#) : Mouse Genome Informatics : international database resource for the laboratory mouse.
- Many others ...

□ Linked data

- Data that have been annotated using ontologies, such as UniProt or DOID and the GO, can be integrated with other community datasets, providing semantic support to perform rich queries.
- [European Bioinformatics Institute](#) (EBI) RDF platform : explicit links between datasets using Semantic Web technologies.


Semantic solutions ./.

- A large and growing number of biomedical ontologies
 - To the point that : *“Unfortunately, the very success of this approach has led to a proliferation of ontologies, which itself creates obstacles to integration.”* ([Smith et al. 2007](#)).

- The **O**pen **B**iomedical **O**ntologies project :
 - Aims to develop interoperable ontologies that are logically well-formed and scientifically accurate : non overlapping content, shared syntax and relations.
 - A resource of the US [National Center for Biomedical Ontology](#).
 - OBO format : flat file human-readable format convertible to OWL and vice-versa.


- The [BioPortal](#) :
 - Managed by the US National Center for Biomedical Ontology.
 - “The world’s most comprehensive repository of biomedical ontologies” (currently 859).

The BioPortal

 **BioPortal** [Ontologies](#) [Search](#) [Annotator](#) [Recommender](#) [Mappings](#) Login Support

Gene Ontology

Last updated: February 3, 2021



Summary Classes Properties Notes Mappings Widgets

Details

Acronym	GO
Visibility	Public
Description	Provides structured controlled vocabularies for the annotation of gene products with respect to their molecular function, cellular component, and biological role. The Gene Ontology consists of three Vocabularies.
Status	Production
Format	OBO
Contact	GO Helpdesk, https://github.com/geneontology/helpdesk/issues
Categories	Genomic and Proteomic
Groups	Cancer Biomedical Informatics Grid, Clinical and Translational Science Awards, OBO Foundry, Unified Medical Language System

Metrics

Classes	50,515
Individuals	0
Properties	9
Maximum depth	16
Maximum number of children	9,428
Average number of children	4
Classes with a single child	4,778
Classes with more than 25 children	361
Classes with no definition	3,305

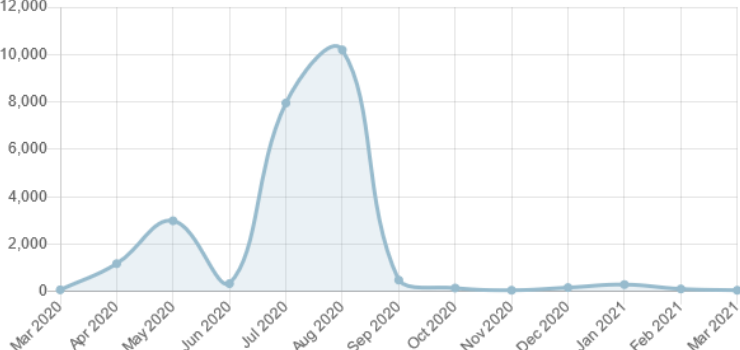
Submissions

Version	Released	Uploaded	Downloads
releases/2021-02-01 <small>(Parsed, Indexed, Metrics, Annotator)</small>	02/03/2021	02/03/2021	OBO CSV RDF/XML Diff
2021-01-01 <small>(Archived)</small>	01/13/2021	01/13/2021	OBO Diff
releases/2020-12-08 <small>(Archived)</small>	12/11/2020	12/11/2020	OBO Diff
releases/2020-10-09 <small>(Archived)</small>	11/20/2020	11/20/2020	OBO Diff
releases/2020-10-09 <small>(Archived)</small>	10/12/2020	10/12/2020	OBO Diff

[more...](#)

Views of GO

Visits



Month	Visits
Mar 2020	0
Apr 2020	1,000
May 2020	3,000
Jun 2020	500
Jul 2020	8,000
Aug 2020	10,500
Sep 2020	500
Oct 2020	200
Nov 2020	100
Dec 2020	100
Jan 2021	200
Feb 2021	100
Mar 2021	100

The Gene Ontology

- ❑ Founded to advance semantic standards for molecular biology.
- ❑ Controlled terminologies of terms in three dimensions : molecular function, biological process, and cellular location of gene products.
- ❑ Developed using the semantic web and biology ontology standards (OWL, OBO).
- ❑ The GO Consortium includes the major organism modeling database groups.
- ❑ The GO is also linked to a major sequence database resource (UniProt).
- ❑ It has established itself as the standard for [function annotations](#) (cf. later).

Basic concepts used in the Gene Ontology

- Gene* : contiguous region of DNA that encodes instructions for how the cell can make large (“macro”) molecules.
- Gene product* : macromolecule (protein or noncoding RNA) produced deterministically according to the instructions from a gene.
- Activity* : chemical action produced by a gene product acting as a **molecular machine**.
- Macromolecular complex* : combination of gene products from different genes into a larger molecular machine.

□ Three axes of classification.

A gene encodes a gene product, which :

- Carries out a molecular-level process or activity (**molecular function**),
- In a specific location relative to the cell (**cellular component**),
- Contributing to a larger biological objective (**biological process**).

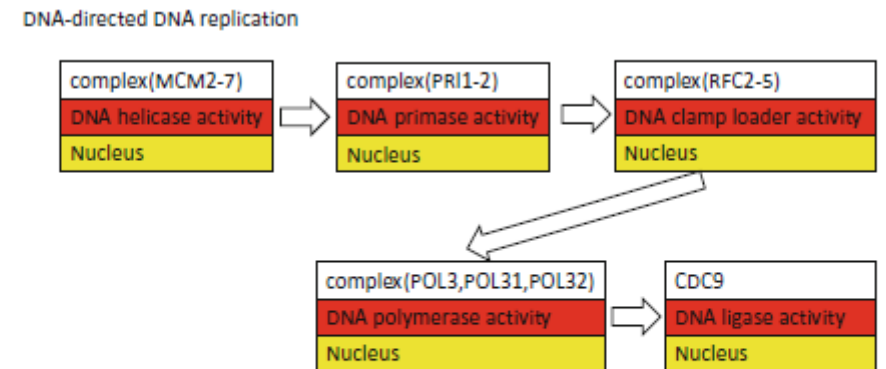
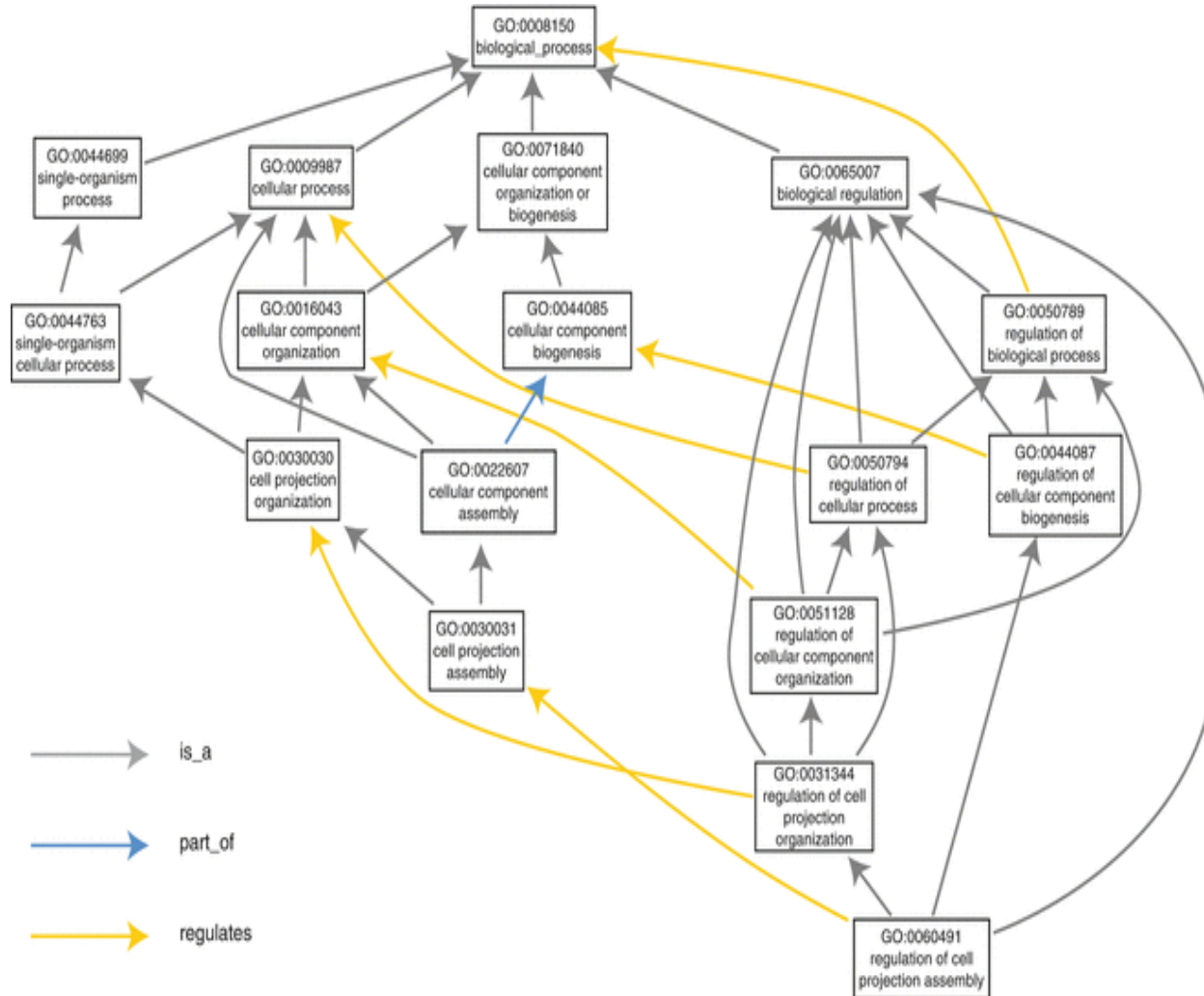


Fig. 1 DNA replication (in yeast) as modeled using the GO. Gene products/complexes (*white*) perform molecular processes (**molecular function**, *red*) in specific locations (**cellular component**, *yellow*), as part of larger biological objectives (**biological process**, specifically **DNA-directed DNA replication**)

(*Gene Ontology Handbook, Dessimoz and Škunca 2017*)

Semantic relations in the Gene Ontology



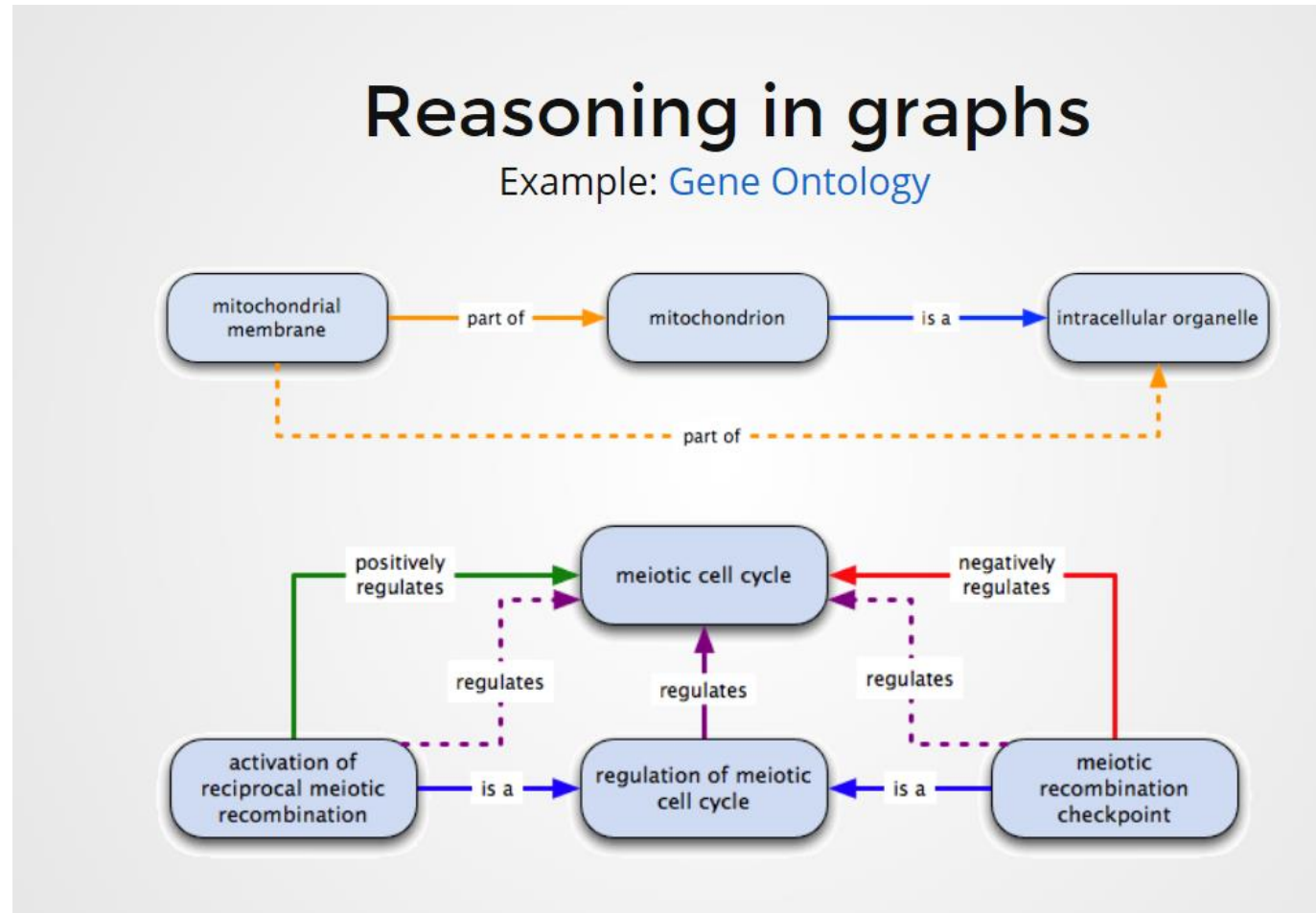
Gene Ontology Term: cell projection assembly

GO ID:	GO:0030031
Aspect:	Biological Process
Description:	Formation of a prolongation or process extending from a cell, e.g. a flagellum or axon.
Synonyms:	cell projection biogenesis, formation of a cell surface projection

□ The Gene Ontology uses 4 relations :

- **subsumption** (called **is-a**);
- **part_of**;
- **has_part** (inverse of **part_of**);
- **regulates** : one process directly affects the manifestation of another process or quality.
 - R₊ : positively regulates.
 - R₋ : negatively regulates.

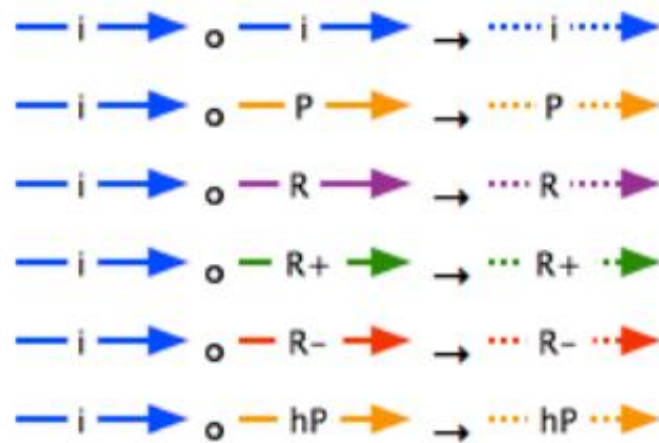
Semantic inferences in the Gene Ontology



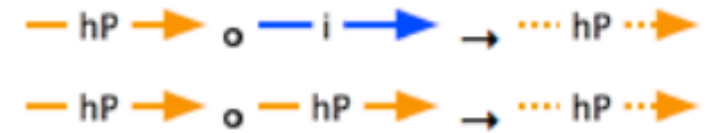
Semantic inferences in the Gene Ontology ./.

Gene Ontology inference rules explained graphically in the GO documentation, for biomedical scientists :

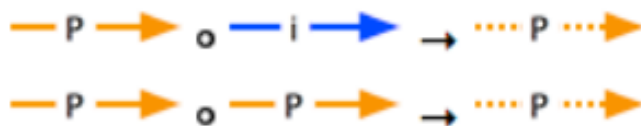
is_a



has_part



part_of



regulates



Gene Ontology annotations

- ❑ Genetics is still a science under development. **All formal representations lack precision in representing the full context surrounding the knowledge, as it is depicted in journals or articles.**
 - Detailed context of experiments, genesis and perspective of the study are important to interpret the results.
 - The required granularity of information is difficult to acquire in computable form.
- ❑ GO defines the possible functions a gene might have but **makes no claims** about the function of any particular gene. Information about those claims is captured as **annotations**.
- ❑ A **GO annotation** is a statement about **the function of a particular gene product**.
 - A GO annotation typically associates a single gene product to a term of the GO ontology.
The fundamental relation is an RDF triple **<gene product> <involved in> <GO term>**.
 - An annotation always refers to its supporting evidence, including the nature of evidence and the source publication.
“The human MSH2 gene product (represented by [UniProtKB:P43246](#)) is involved in ‘[GO:0006298 DNA mismatch repair](#)’ (a biological process), based on direct assay evidence published in PubMed [7923193](#). »

Gene Ontology annotations ./.

- GO has become a **community standard for annotations**.
 - Annotations can originate from the GO itself or from other ontologies
 - Annotations can be used by other ontologies.
 - Many genome annotation groups and bioinformatics centers use the GO annotation process.

- We will illustrate this with the interaction of two ontologies :
 - **Gene Ontology**.
 - **Mouse Genome Informatics**.

(the discussion of the example in the following slides is based on Blake and Bult 2006).

The Mouse Genome Informatics (MGI)

- The recognized community database for the laboratory mouse.
- Integrates genetic and genomic data for the mouse to support the use of the mouse as a model system for understanding human biology and disease processes.

The screenshot shows the MGI website interface. At the top left is the MGI logo with a mouse icon. To its right is the text "Mouse Genome Informatics". Below this is a navigation bar with links: Search, Download, More Resources, Submit Data, Find Mice (IMSR), Analysis Tools, Contact Us, and Browsers. The main content area is divided into two columns. The left column features a search box labeled "Keywords, Symbols, or IDs" with a "Quick Search" button. Below the search box is a section titled "Or use topic specific search and analysis tools:" followed by a list of categories: Genes, Phenotypes & Mutant Alleles, Human-Mouse: Disease Connection, Gene Expression Database (GXD), Recombinase (cre), Function, Strains, SNPs & Polymorphisms, Vertebrate Homology, Mouse Models of Human Cancer, Pathways, Batch Data and Analysis Tools, and Nomenclature. At the bottom of this column is a "Getting Started:" section with links to "Introduction to mouse genetics", "How to use MGI (Text & Video)", and "Cre Portal Tutorial". The right column contains a descriptive paragraph about MGI, social media icons for Facebook and Twitter, a "Community Feedback Needed: Minimal Information Standards for PDX models" banner with a mouse image, and a "What's new at MGI" section updated on October 20, 2017. This section lists several updates, including the release of the Alliance of Genome Resources (AGR) 1.0, enhancements to the Disease Ontology Browser, and the incorporation of the hierarchical Disease Ontology (DO) into MGI. At the bottom right of the right column are links for "MGI Statistics" and "More MGI news".

(<http://www.informatics.jax.org/>)

The Mouse Genome Informatics (MGI)

- Data integration (identifying disparate data that describe the same biological entity (e.g., gene, transcript, protein, etc.) is a primary focus of MGI.

- **Many difficulties:**
 - Knowledge about each genome or cellular component is complex and incomplete.
 - **Tens of thousands of scientific papers** are published each year, correcting old information and adding new details about millions of distinct biological entities.
 - The same gene can be referred to by multiple different names in the literature.
 - The same name can be applied to more than one gene.
 - Different data about the same gene can be obtained from various sources.
 - Sequence data from different sources may include overlapping sequences identified by different identifiers.
 - Data may be incomplete; inconsistent; conflicts in data appear over time.

- Strategy : use of controlled terminologies and ontologies for **semantic normalization**, **relying on curated evidence** with links to the literature.

Example : functional annotations of HEXA gene



Hexa Gene Detail

Summary

Symbol Hexa

Name hexosaminidase A
Synonyms Hex-1

Feature Type protein coding gene

IDs MGI:96073
NCBI Gene: 15211

Gene Overview MyGene.info: [HEXA](#)

Alliance [gene page](#)

- Functional annotation of mouse [hexosaminidase A](#) gene in MGI (HEXA, MGI:96073).
- MGI shows the relevant links towards the Gene Ontology classification.
- Gene product of Hexa gene is involved (a.o.) in the molecular function [hydrolase activity](#).

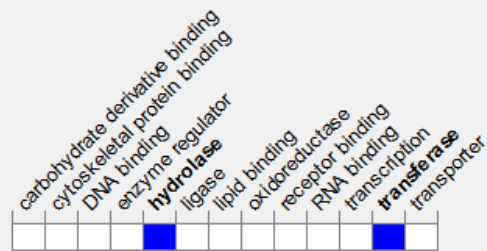
Gene Ontology (GO) Classifications

less ▾

All GO Annotations 35

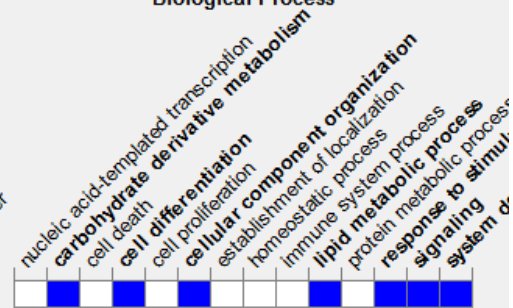
GO References 21

Molecular Function



Click cells to view annotations.

Biological Process



Hexosaminidase : an enzyme involved in the hydrolysis of terminal N-acetyl-D-hexosamine residues (amine sugars).

Hydrolase activity : catalysis of the hydrolysis of various bonds, e.g. C-O, C-N, C-C, phosphoric anhydride bonds, etc.

GO annotations for hydrolase in MGI

Gene Ontology (GO) annotations for hydrolase					
Filter annotations by: Aspect Category Evidence << f					
Export: Text File Excel File					
Aspect	Category	Classification Term	Context	Evidence	Reference(s)
Molecular Function	hydrolase	beta-N-acetylhexosaminidase activity	happens in liver .	IDA	J:45044 [PMID:9417048]
Molecular Function	hydrolase	beta-N-acetylhexosaminidase activity	happens in liver . happens in spleen . happens in heart . happens in metanephros . happens in gastrocnemius muscle . happens in brain .	IMP	J:54553 [PMID:10196372]
Molecular Function	hydrolase	beta-N-acetylhexosaminidase activity	happens in liver .	IMP	J:21008 [PMID:7937929]
Molecular Function	hydrolase	beta-N-acetylhexosaminidase activity	happens in spleen . happens in liver . happens in brain .	IDA	J:30435 [PMID:8747922]
Molecular Function	hydrolase	beta-N-acetylhexosaminidase activity		IDA	J:441 [PMID:1914521] , J:88531 [PMID:11854359]
Molecular Function	hydrolase	beta-N-acetylhexosaminidase activity	happens in brain . happens in liver .	IMP	J:30899 [PMID:8789434]
Molecular Function	hydrolase	beta-N-acetylhexosaminidase activity	happens in liver .	IGI	J:36305 [PMID:8896570]
Molecular Function	hydrolase	beta-N-acetylhexosaminidase activity	happens in spleen . happens in liver .	IMP	J:30435 [PMID:8747922]
Molecular Function	hydrolase	hydrolase activity		IEA	J:60000
Molecular Function	hydrolase	hydrolase activity, acting on glycosyl bonds		IEA	J:60000
Molecular Function	hydrolase	hydrolase activity, hydrolyzing O-glycosyl compounds		IEA	J:72247

Reference papers

- Of of the GO molecular functions associated to this gene in MGI is [beta-N-acetylhexosaminidase activity](#), a subclass of [hydrolase activity](#).
- Many papers support this functional annotation, but the nature of evidence and the wordings varies in each paper.
- GO is used as a reference to enforce standards.

[beta-N-acetylhexosaminidase activity](#) : catalysis of the hydrolysis of N-acetyl-D-hexosamine residues in N-acetyl-beta-D-hexosaminides

Examples of queries and analyses based on the GO

- ❑ Use of biomedical ontologies and **careful mappings of data** across them provide a robust mechanism to retrieve relevant set of data in answer to complex queries typically asked in biological research.
- ❑ A typical research example : GO enrichment analysis in cancer research :
 - A set of 1000 genes expressed at a higher level in a cancer sample than in a matched healthy tissue sample.
 - Are there any functions (terms from the GO molecular function, cellular component, or biological process aspects) unusually common among these 1000 overexpressed genes ?
 - The functions present in the set of 1000 genes need to be compared to the functions of the 20000 human protein-coding genes. GO is used to retrieve all functions performed by the 20000 human genes and create all possible groupings by functional class.
 - Each grouping is tested for statistical enrichment(*), and the small number of enriched functional classes enables the researcher to identify candidate biological processes.

(*) : Statistical enrichment : a method to identify classes of genes or proteins that are over-represented in a large set of genes.

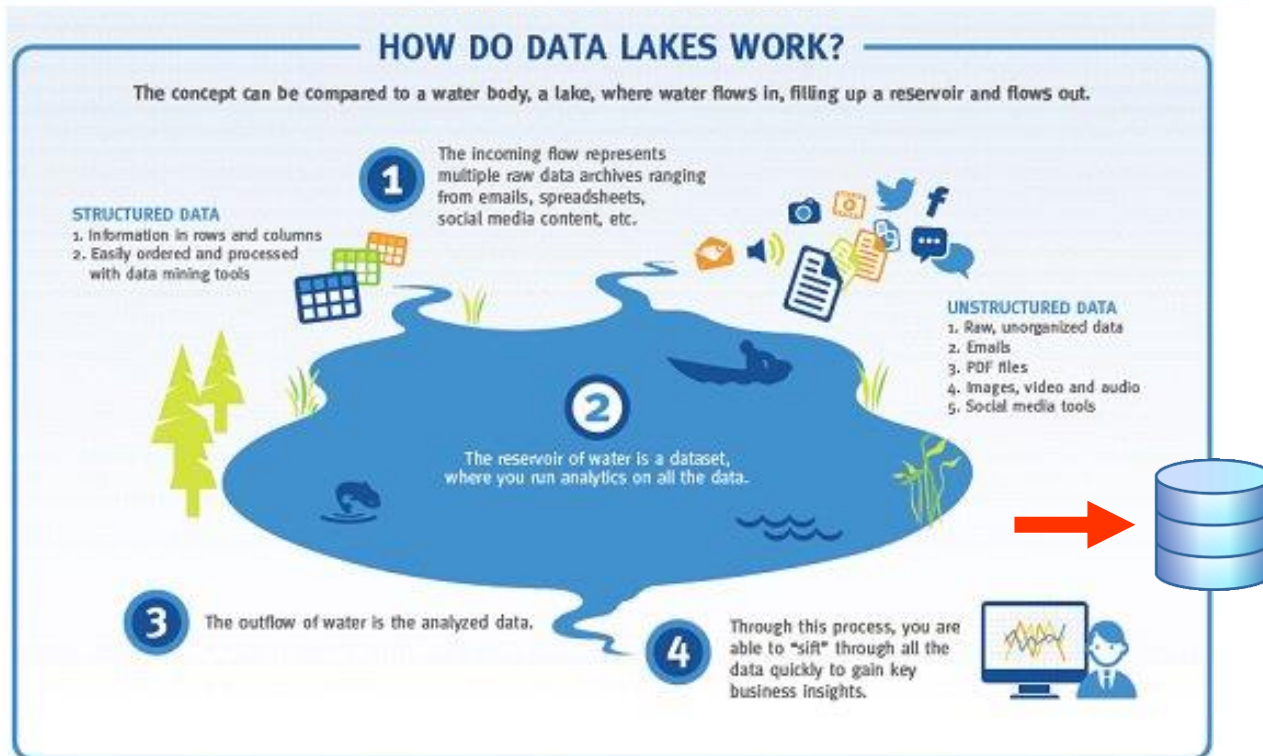
(source *Gene Ontology Consortium 2019*)

Agenda

- 1 Bioinformatics: data deluge and ontologies
- 2 Medical informatics: a semantic data lake
- 3 Systems engineering: systems & models
- 4 Ontologies in software engineering
- 5 Thoughts on semantic data

Data integration : data lake

AS THEY ARE ENVISIONED TODAY...






Source: <http://www.tangerine.co.th/tag/how-do-data-lake-work/>

Intended for Knowledge Sharing only



- ❑ Data is gathered from various sources
 - Both structured and unstructured.
- ❑ Distributed file system architecture.
 - Typically, Hadoop HDFS, HBase.
 - Cloud services: Amazon S3, Microsoft Azure...
- ❑ There is no effort to structure the data at the time of capture (**schema on read**):
 - Data is stored in its initial raw format.
 - Data consumers will set up their analysis applications to perform specific data exploration.
 - Less up-front costs than RDBMS (**schema on write**), more flexibility, less optimization.
 - The data lake can feed a data warehouse.

Data lakes in current ICT market (2021)

59 offre(s) Pertinence / Date

-  **Junior - Senior Data Engineer**
Smals
30/03 Bruxelles
Data Engineer
-  **BI Consultant**
Business & Decision
04/04 Bruxelles
PowerBI, Qlikview, SSIS, ETL, SAS
-  **Data Analyst Senior**
Contraste Europe
22/03 Bruxelles
Data Analyst Senior

[Créer une Job Alert](#)

-  **Data Platform Consultant**
Visser & Van Baars
08/03 Bruxelles
data lake, DWH, Data Vault, data
-  **Data Visualisation Consultant**
Business & Decision
31/03 Bruxelles

(source: Job openings on ICTjob.be (search done in April 2021))

Microsoft Azure

Présentation Solutions **Produits** Documentation Tarifs

Data Lake Analytics

Service de travaux d'analyse à la demande pour agir en connaissance de cause

[Démarrer gratuitement](#)

[Vous utilisez déjà Azure ? Essayez Data Lake Analytics maintenant >](#)

Data Lake Storage on AWS

The most secure, durable, and scalable storage to build your data lake

[Build your data lake on Amazon S3](#)

IBM

Analytics Products Solutions Learn **Explore More**

Data lake

Drive smarter, data-driven decisions by capitalizing on a broader variety of data

[See data lake products](#) [Talk to a data lake expert](#)

[Learn about the IBM-Cloudera partnership](#)

Data lake or data swamp ?

GARTNER : [beware the data lake fallacy!](#) (2014).



- ❑ Tries to tackle two hard problems :
 - Old: tries to get rid of information silos.
 - New : Not clear how and when data will be useful.

- ❑ Growing hype and significant risks :
 - **Lack of semantics and meta-data.**
 - Quality of data unknown.
 - Security still embryonic.
 - Business users lack skills and supporting tools.
 - Performance issues for tools and interfaces.
 - **Enterprise-wide data management is still an issue.**
 - So is data privacy.

- ❑ Badly managed data lake = data swamp !
 - Data lost, never accessed, business value not realized.

Case study: Montefiore semantic data lake

- ❑ The data deluge is impacting genomics, but also healthcare, in particular precision medicine : *an approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person (NIH).*
- ❑ Chosen case study : **Montefiore semantic data lake.**
 - It demonstrates a life application of semantic data in critical aspects of healthcare.
- ❑ Montefiore Medical Center:
 - Large US medical organisation in the Bronx and Westchester County, New-York state, U.S.A.
 - 11 hospitals, > 2000 beds.
 - (One of the) university hospital for Albert Einstein College of Medicine.

Case study: Montefiore semantic data lake

Sources

- ❑ Webinar by P. Mirhaji, 2016.
- ❑ Written sources:
 - ❑ [Montefiore Creates Data Analytics Platform to Advance Patient Care](#) – Intel solution brief
 - ❑ [Semantic Big Data Lakes Can Support Better Population Health](#), news article.
 - ❑ [Montefiore Semantic Data Lake Tackles Predictive Analytics](#), news article.
 - ❑ [Semantic data lake architecture in healthcare and beyond](#), news article.



The screenshot shows a video player interface. The main content is a title slide for a webinar. The title is "The Semantic Data Lake in Healthcare". Below the title, the speaker's name is listed as "Parsa Mirhaji MD, PhD", followed by his affiliations: "Associate Professor, Systems and Computational Biology, Albert Einstein College of Medicine" and "Director, Clinical Research Informatics, Montefiore Health System". At the bottom of the slide, it says "CTO, NYC-Clinical Data Research Network". The video player controls at the bottom show a play button, a progress bar at 0:00 / 1:11:09, and icons for settings, full screen, and a list.

Semantic Data Lake in Healthcare - Parsa Mirhaji, MD PhD - Yosemite Project Webinar

What is the problem ?

- Servicing a large patient community (Bronx & Westchester) :
 - Hundreds of thousands of patients, ethnically and socioeconomically diverse population.
 - Precision medicine : individualized patient treatment => expand breadth of patient data.
 - Complex array of factors : from genetics to home and work environment, nutrition, population health management and community care.

- Financial and regulatory pressures :
 - **Accountable health organization** : not *pay as service*, but *value-based* or performance-based payment. An accountable care organisation ties payments to quality metrics and the cost of care.
 - **EHR incentive program** : provides incentive payments to healthcare to use EHR (electronic health record) technology, inter-operability and standards in a way that impacts positively patient care. Program requirements are becoming more demanding at each stage.
 - Cost reduction pressure.

Existing situation

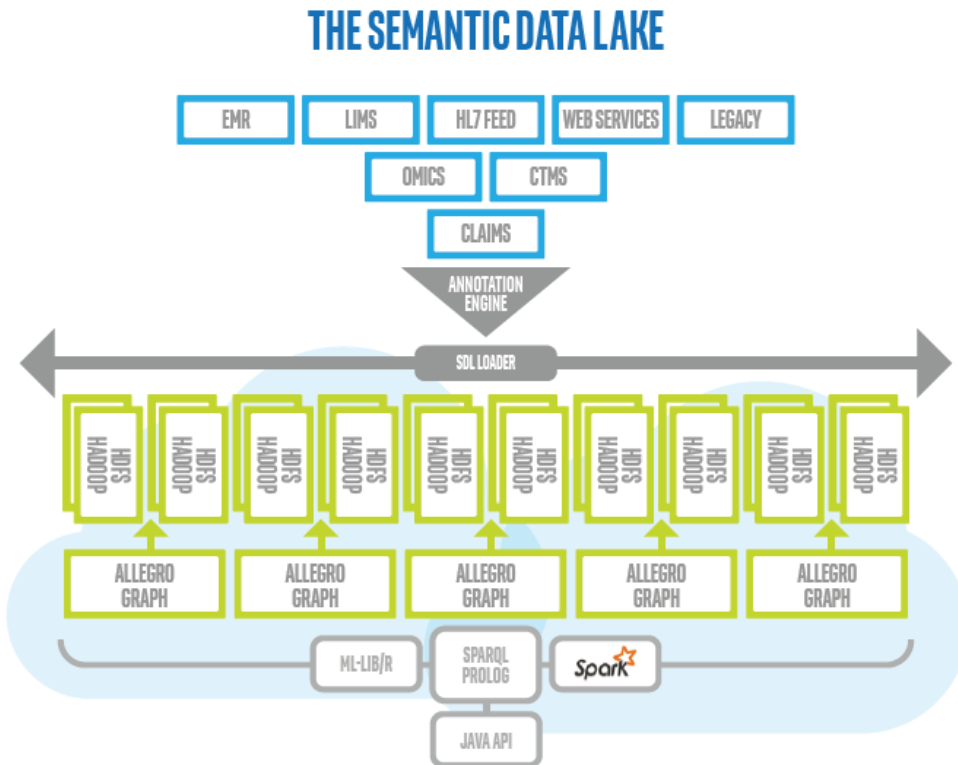
□ Data :

- Hundreds of thousands of patients impacting the institution at any given point.
- More than 100 data points per patient per day.
- Data includes research results, financial data, patient demographics. Ranges from unstructured free text information to images and waveforms to data from sensors and monitoring devices.

□ Architecture :

- Each type of data may be locked into its own individualized analytics architecture.
- Relational databases require to pre-define the problems. Predicting all future cases is impossible. Cost of change is huge and creates data silos.
- Data cannot be analyzed quickly, nor can hospital data be easily combined with external data sources such as those from pharmaceutical companies and researchers.

Solution proposed

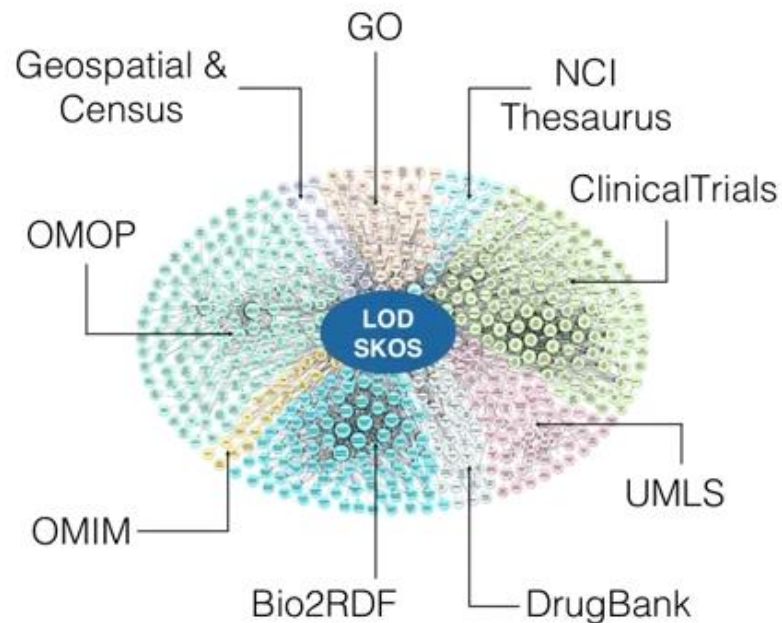


- ❑ Intel Xeon processors.
- ❑ Cloudera Hadoop (HDFS, Yarn, Spark).
- ❑ AllegroGraph semantic graph DB to integrate data.
 - Real time (HL7) feeds and large batch input (ETL process).
 - Transformed into triples through JAVA API.
- ❑ Ontological pipeline (annotation engine) integrating multiple ontologies :
 - Unified Medical Language System (UMLS) as upper ontology.
 - PharmGKB (correlates genetic variations and drug responses).
 - Online Mendelian Inheritance in Man (OMIM).
 - OMOP : common data model for observation data, ...
- ❑ SPARQL for queries.
 - Compact ! Graphical tool for visual query creation.
- ❑ “Analytic tapestry” tool (supports machine learning), interface to R Studio, Spark ...

Knowledge base and integration of ontologies

Google

The Knowledgebase



- ❑ Over half a billion triples.
- ❑ All ontologies integrated through Linked Open Data.
- ❑ UMLS (Unified Medical Language System) translated into W3C SKOS, used as upper ontology.

SKOS : Simple Knowledge Organization System

Simple models for expressing concept schemes of controlled vocabularies.

- ❑ Data curated by specialized team.

Semantic Data Lake in Healthcare - Parsa Mirhaji, MD PhD - Yosemite Project Webinar

930 vues

7 0 PARTAGER

Using SPARQL for 360° patient view

50v21r

Record Locator Service (aka 360° Patient View)

```
Select distinct ?type (count(?type) as ?types)
FROM <2d6e36f6a0e25c757d52780b4488>
WHERE {
  ?s ?p ?o.
  ?o a ?type}
group by ?type
```

?type	
Census Block Group	1
Census Tract	1
City	1
Contact Event	1
Date Of Birth	1
Demography Event	1
Diagnosis Source	4
Emergency Encounter	2
First Name	1
Gender	1
Outpatient Encounter	22
Patient Diagnosis	33
Patient Encounter End Date	2
Patient Encounter Seen Time	22
Patient Encounter Triage Time	2
Patient Encounter Type	24
Patient Procedure	52
Phone Number	1
Procedure Date	52
Race	1
State	1
System Level Person ID	1
Zipcode	

Evaluation

- ❑ Started in November 2015
- ❑ First pilot *“uses predictive analytics to flag any patient hospitalized at Montefiore Health System locations with respiratory problems who is at risk of death or in need of intubation within the following 48 hours, which is the window of opportunity to complete an effective intervention.” (Mirhaji)*
 - > 68000 admissions, over 28000 patients.
 - Over 77% detection of likelihood of a respiratory event, 48 hours in advance of a fatal episode or respiratory failure in the hospital.
 - False positives less than 1%.
- ❑ *“Semantic Data Lake ... is now being instituted throughout our critical care systems in all of the hospitals within the medical center.” (Racine, Chief Medical Officer).*
- ❑ System is live, positive customer testimony, plan to deploy.

Agenda

1

Bioinformatics: data deluge and ontologies

2

Medical informatics: a semantic data lake

3

Systems engineering: systems & models

4

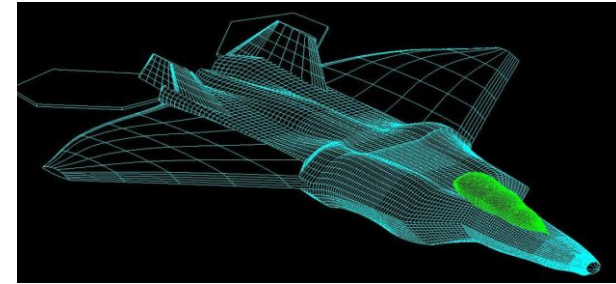
Ontologies in software engineering

5

Thoughts on semantic data

Data in Systems Engineering

- ❑ Systems engineering digitalization
 - CAD / CAM
 - Concurrent design
 - Virtual prototyping
 - Additive layer manufacturing (3D printing)
 - Smart products (autonomous car, airplane smart sky vision ..)
- ❑ Big data, but also big systems designed by large enterprises !



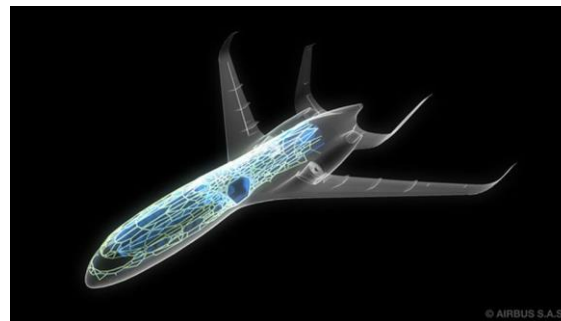
F22 Raptor



European space agency collaborative design



Google Waymo Chrysler Pacifica



Airbus 2050 smart plane vision



Ford immersive vehicle environment

Some key challenges

□ Product life cycle

- Product Life Cycle Management :

The activity of managing products all the way across their lifecycles from the very first idea through design, production, maintenance ... until it is retired and disposed of.

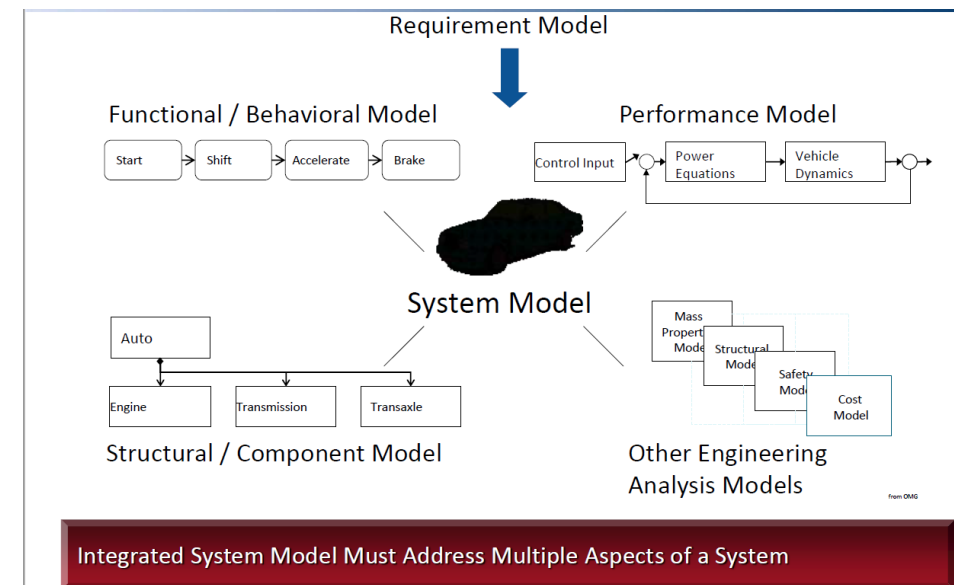
- Traceability of products may reach > 50 years (planes, space industry ...).

□ Models

- Model-based systems engineering.

□ Interoperability between systems and data !

- Huge interoperability costs.



Systems engineering is a world of standards

- Three main standards specific for system engineering :
 - Ansi 632 : *Processes for Engineering a System.*
 - IEEE 1280 : *Standard for application and Management of the Systems Engineering Process.*
 - ISO 15288 : *Systems Engineering – System Life-Cycle Processes* – the main reference.

- Other important standards from the point of view of systems engineering data :
 - CMMI : *Capability Maturity Model Integration* (intended more for software engineering).
 - ISO 9001 : *Quality management.*
 - OMG SysML : *Systems Modeling Language.*
 - ISO 10303 STEP : *Standard for the Exchange of Product Model Data.*
 - ISO 15926 : *Integration of life-cycle data for process plants.*

- System integration / federation / interoperability is key.

A detour through a well-known standard: ISO 9001

- ❑ Probably best-known ISO standard.
- ❑ Standard for quality management.
- ❑ Applies to many industries.
 - Software engineering: [ISO/IEC 90003](#).
- ❑ Certification by recognized organisms.
 - In 2013 > 1 million certifications over 187 countries.
 - Certification required explicitly by requests for proposals.
 - New version integrates “continual improvement” (**kaizen** ideas).



ISO 9001:2015

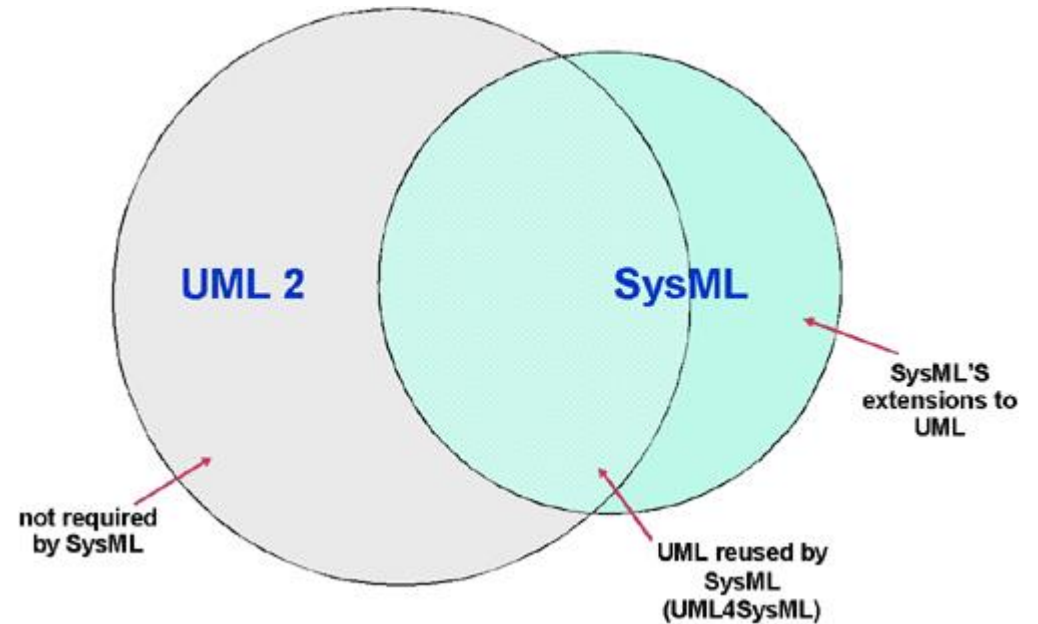
What is ISO 9001?

ISO 9001 is a standard that sets out the requirements for a quality management system. It helps businesses and organizations to be more efficient and improve customer satisfaction.

A new version of the standard, ISO 9001:2015, has just been launched, replacing the previous version (ISO 9001:2008).

OMG SysML

- ❑ **OMG standard 2006 :**
 - General-purpose graphical modeling language for specifying, analyzing, designing, and verifying complex systems (inc. hardware, software, information, personnel, procedures, and facilities).
 - Developed by OMG, Incose (International Council on Systems Engineering) and ISO.
 - Is for systems engineering what UML is for software engineering.
- ❑ Subset of **UML 2** with extensions needed for Systems Engineering.
- ❑ Compatible with ISO 10303 systems engineering data interchange standard.



SysML v2 RFP (2019)

SYSML V2: THE NEXT-GENERATION SYSTEMS MODELING LANGUAGE



The SysML® v2 RFP was issued on December 8, 2017. This culminated an 18-month effort to develop the requirements for the next-generation systems modeling language, which is intended to improve the precision, expressiveness, and usability over SysML v1. The requirements reflect lessons-learned from applying model-based systems engineering (MBSE) with SysML since its adoption more than 10 years ago.

The RFP requires the specification to include both a SysML profile of UML® and a SysML metamodel, and a mapping between them. In addition, submitters have the option to specify additional features that include model interchange and formal semantics.

The capabilities provided by SysML v2 should enable improved effectiveness and broader adoption of MBSE.

[READ RFP](#)

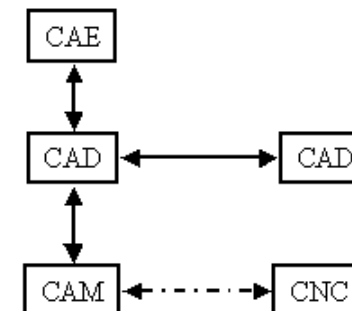
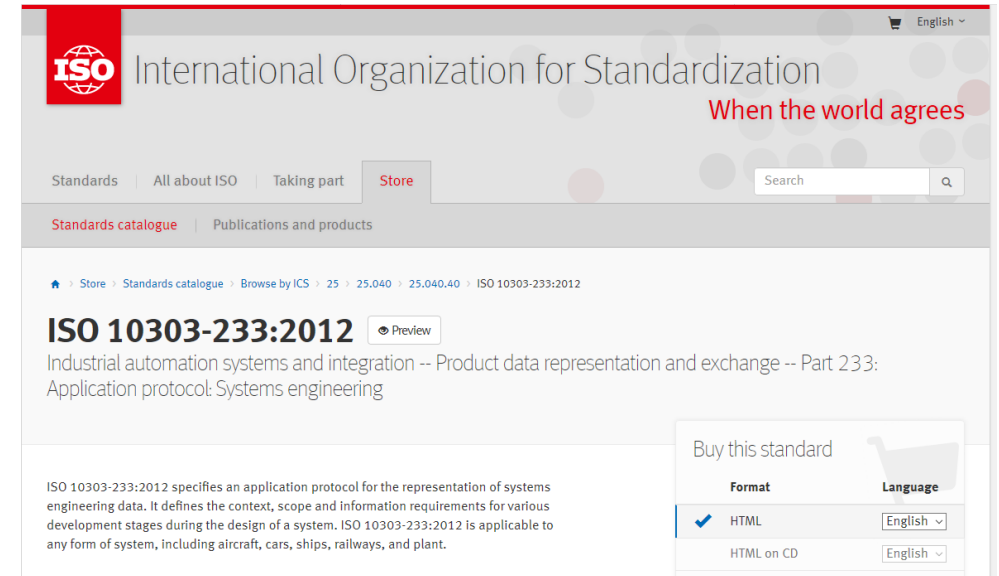
[WEBINAR](#)

[CONTACT US](#)

Emphasis on model (data) interchanges and formal semantics

ISO 10303: STEP

- ❑ **Standard for Exchange of Product Model Data.**
 - Application Protocol 239 for Product Life Cycle Support (PLCS).
 - Application Protocol 233 for Systems Engineering.
- ❑ Unified international standard for all aspects of technical product data.
- ❑ Describes designs and assemblies of solid models.
- ❑ Mainly geometric data.
- ❑ Nearly every CAD/CAE/CAM system has a module to read and write STEP data.



CAD (Computer Aided Design)
CAE (Computer Aided Engineering Analysis)
CAM (Computer Aided Manufacturing)
CNC (Computerized Numerical Control)

↔ STEP data exchange
↔ STEP-NC exchange

Evolution of STEP

□ Major requirements :

- Cover the whole system lifecycle.
- Support for data sharing.
- Semantics.

□ Current evolution :

- Include most of AP 233 (*Systems Engineering*) in AP 239 (*Product Life Cycle Support*)
- Normative XML schema for presenting EXPRESS.
- Add use of SysML alongside EXPRESS.

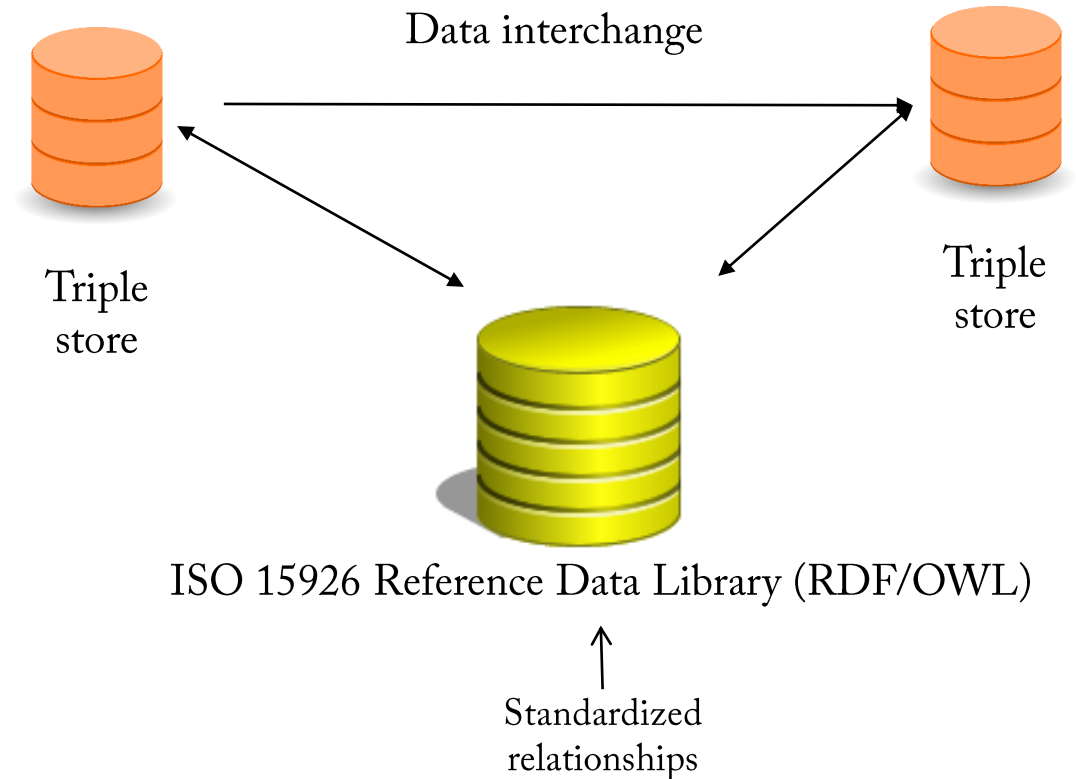
ISO 15926

- ❑ Standard for Integration of life-cycle data for process plants.
 - Standard for data integration, sharing, exchange, and hand-over between computer systems.
 - Developed initially for Oil and Gas but applicable to many sectors dealing with process material and their transport. (oil, gaz, electricity generation, steam generation, security and safety, buildings and accommodation)
- ❑ Implemented using W3C standards (RDF, OWL).
 - Includes an extensible upper ontology.
- ❑ Benefits :
 - Open standards.
 - Covers full life cycle.
 - Syntactic interoperability with XML.
 - Semantic interoperability with OWL.

Home > Store > Standards catalogue > Browse by ICS > 75 > 75.020 > ISO/TS 15926-8:2011

ISO/TS 15926-8:2011 [Preview](#)

Industrial automation systems and integration -- Integration of life-cycle data for process plants including oil and gas production facilities -- Part 8: Implementation methods for the integration of distributed systems: Web Ontology Language (OWL) implementation



Agenda

- 1 Bioinformatics: data deluge and ontologies
- 2 Medical informatics: a semantic data lake
- 3 Systems engineering: systems & models
- 4 Ontologies in software engineering
- 5 Thoughts on semantic data

Not part of the material
for the exam

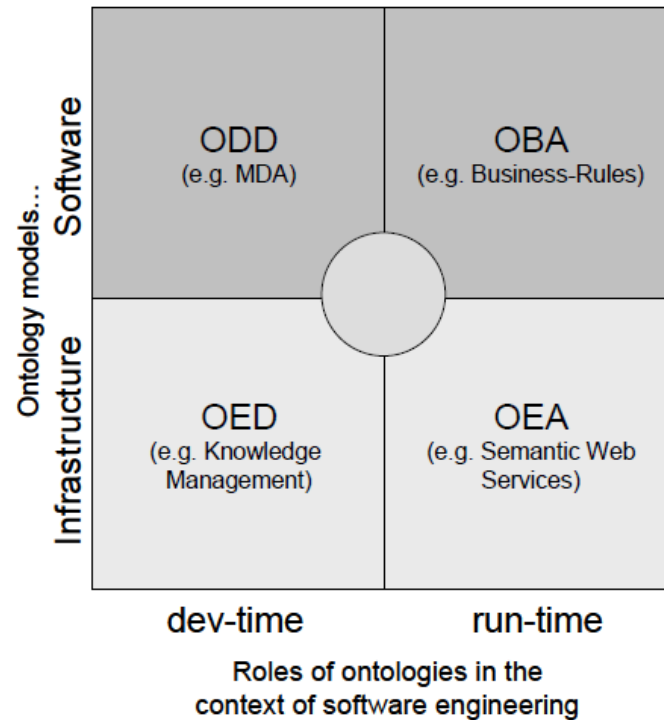
Applications of ontologies in Software Engineering

Ontology Driven Development

Use of ontologies through the software development life cycle.

Ontology Enabled Development

Knowledge management support - e.g., for the developer (component search, problem solving support...) or for project management.



Ontology Based Architecture

Declarative capture of business logic through business rules.

Ontology Enabled Architecture

Semantic web services.

(figure from Happel and Seedorf 2006)

Ontologies in Software Development ?

Ontology-driven
development.

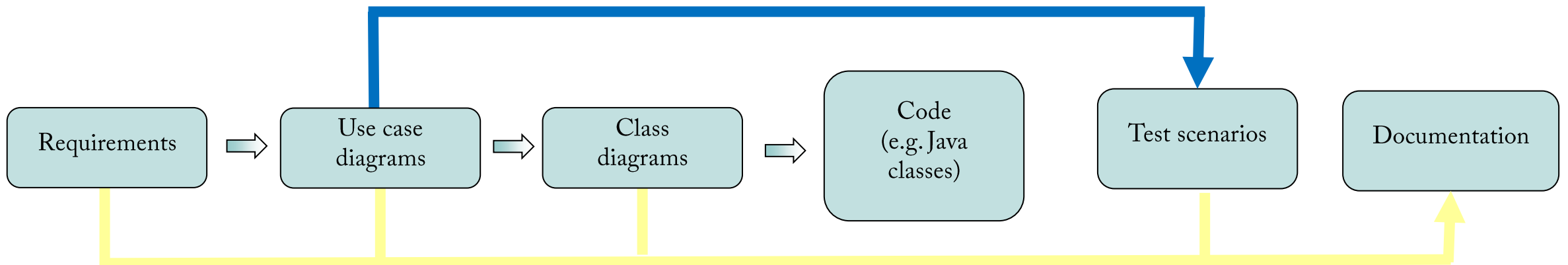
- Requirements.
 - Capturing and modeling domain knowledge.
 - Formally modeling and validating requirements (requirements engineering).
 - Evolutionary requirement management and traceability.
- Analysis, design, implementation.
 - Integration with software modeling languages (**Model Driven Software Engineering**):
 - Sharable terminologies, formal semantics and consistency, unique resource names.
 - Support for automatic transformation and validation of models.
 - Use an ontology as object model (including automatic generation of ontology APIs in an object-oriented programming language).

Ontology-enabled
development.

- Support
 - Component reuse (e.g., component retrieval based on semantic descriptions).
 - Knowledge-based coding support (e.g., autocompletion of API calls).
 - Documentation (software information systems with query facilities).

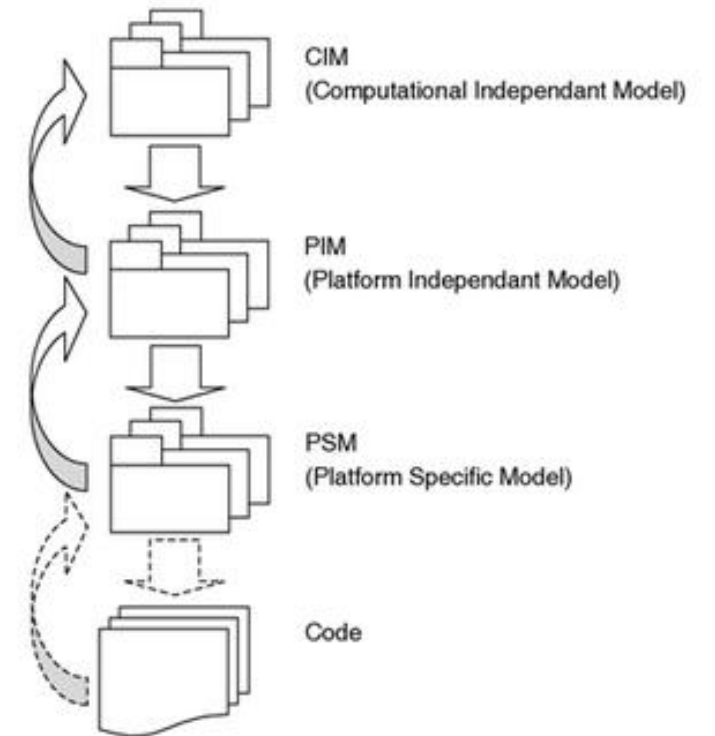
Model Driven Software Engineering

- ❑ Models support the development of running code which satisfies requirements.
- ❑ They are usually developed in successive steps.
 - In each phase, models are constructed.
 - Parts of the models of previous phases are used as input for other models.
- ❑ **Model driven software engineering** focuses on automatically generating code, or parts of models, from other models.



Model Driven Architecture (MDA)

- ❑ Model-driven approach to software design, development and implementation spearheaded by OMG.
- ❑ Separates business and application logic from underlying technology platform, using three layers of models :
 - **Computer Independent Model (CIM)**, also called Business Model or Domain Model.
 - **Platform Independent Model (PIM)**.
 - **Platform Specific Model (PSM)** or implementation models.
- ❑ Two ways to automate the path towards executable systems :
 - Apply transformation patterns to models or use a model interpretation engine.
- ❑ Ontologies can play several roles :
 1. Implement the CIM / Business model.
 2. Support the consistency checking and the transformation of models.



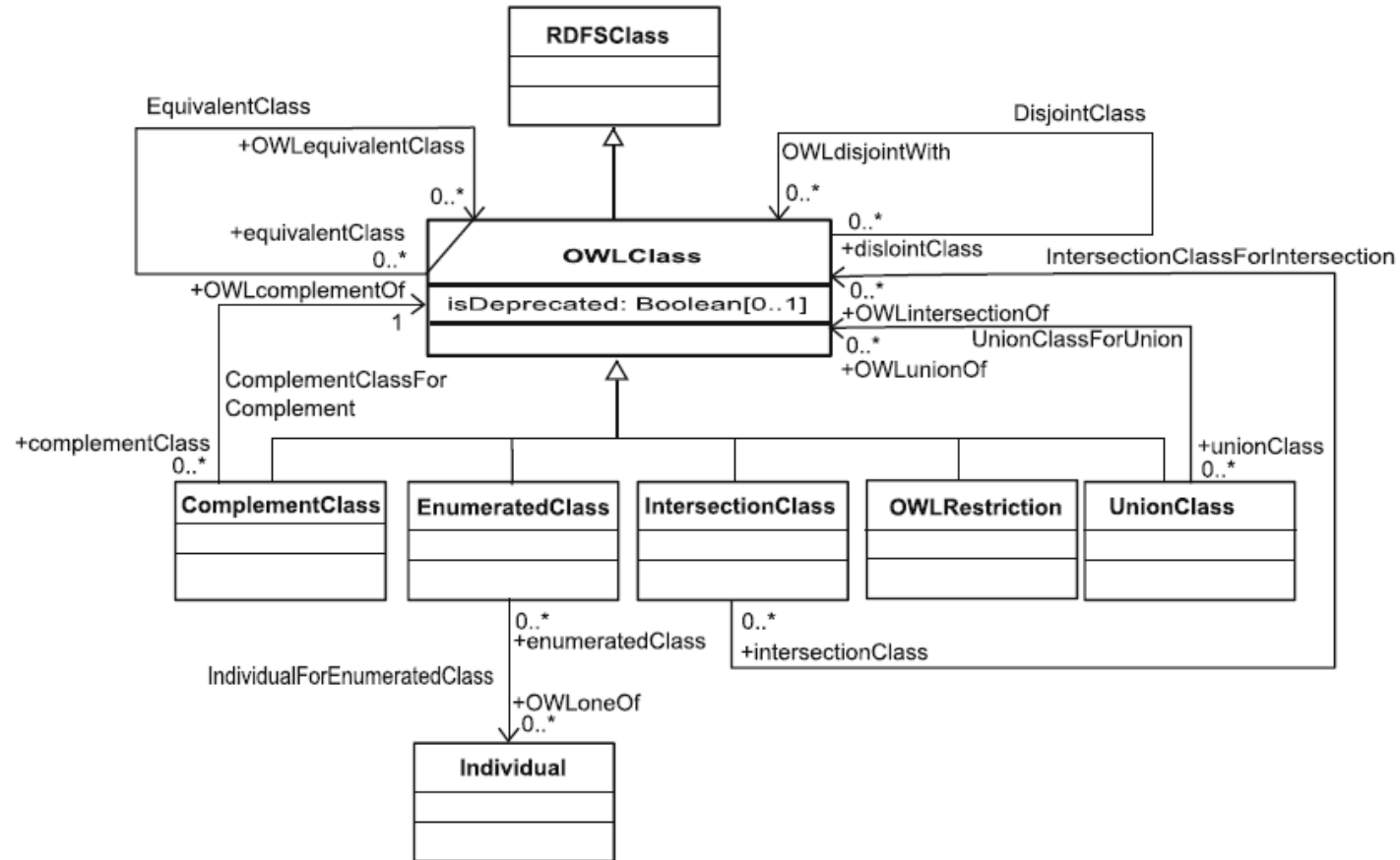
The Ontology Definition Metamodel (ODM)

- ❑ An OMG specification aiming to bring together SE and KR :
 - Mappings between UML, RDFS, OWL.
 - Last version (2014) supports OWL 2.

- ❑ Foundation for an important set of capabilities for Model Driven Architecture (MDA) :
 - Formal grounding for **representation, management, interoperability**, and **application** of **business semantics**.
 - Varying levels of expressivity, complexity, and form for conceptual models, ranging **from UML and ER methodologies to formal ontologies** represented in description logics or first order logic.
 - **Grounding in formal logic, through standards-based, model-theoretic semantics**, to enable reasoning engines to understand, validate, and apply ontologies developed using the ODM.
 - **Profiles and mappings** sufficient to support the exchange of models developed in various formalisms and to enable consistency checking and validation.
 - Basis for a family of specifications that marry **MDA and Semantic Web technologies to support semantic web services, ontology and policy-based communications and interoperability**.

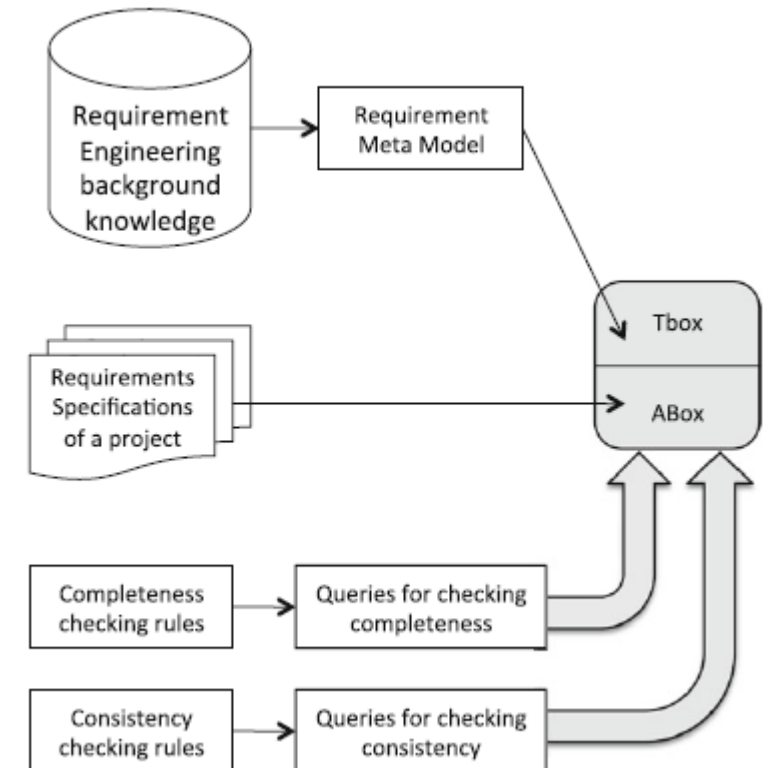
(source <https://www.omg.org/spec/ODM/>)

ODM OWL Class description diagram



Ontology driven requirement engineering

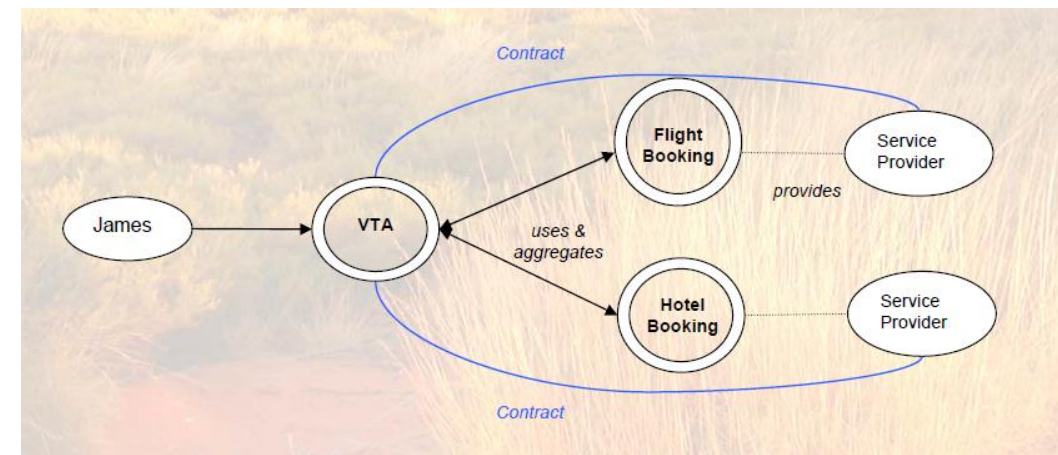
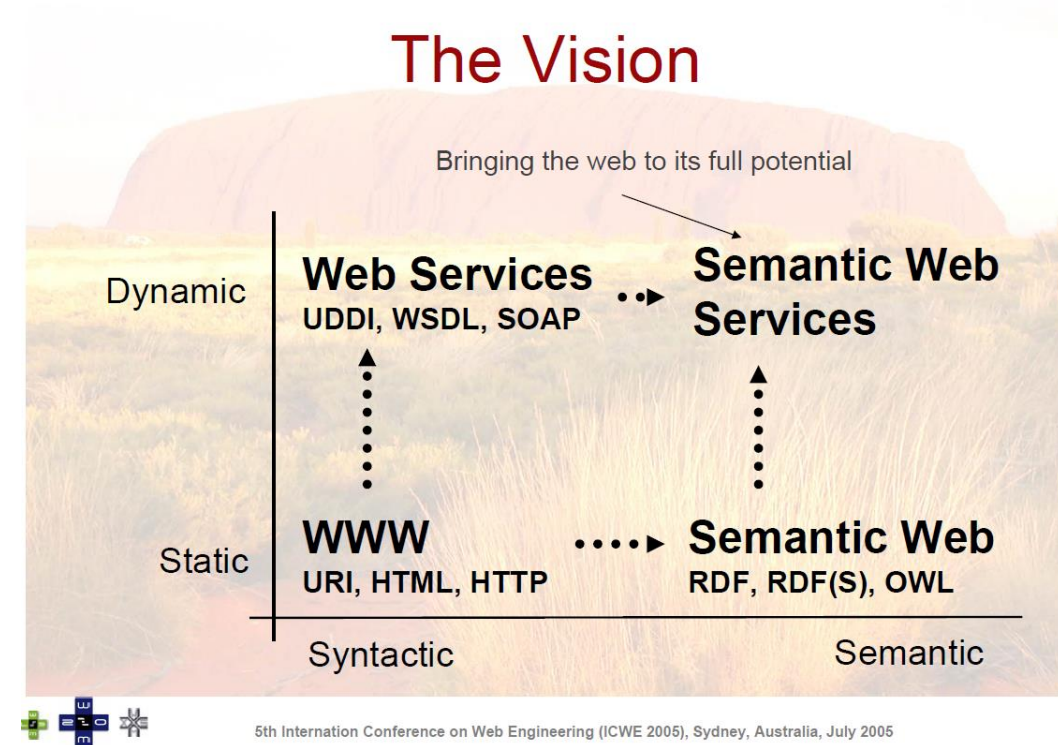
- ❑ Requirement engineering is the first task of a software development process.
- ❑ In a (partially) automated model-driven process, **consistency** and **completeness** of requirements are crucial.
- ❑ A DL ontology can be used for that :
 - A requirement metamodel is captured in the TBox. This model includes consistency and completeness axioms in DL.
 - Requirements of a specific project are captured in the ABox.
 - SPARQL based queries based on the consistency and completeness rules validate the requirements.



(source Pan & al. 2013)

Semantic web services

- ❑ **Service Oriented Architecture (SOA)** : approach to developing software by combining reusable and possibly distributed components called *services*.
- ❑ The systematic use of SOA remains limited.
 - Difficulty: the need for software developers to handle **manually** the discovery, usability, usage and integration of services.
- ❑ **Semantic web services** refer to the approach where :
 - Automatic discovery, selection, composition and execution of services across heterogeneous users and domains is possible,
 - Based on semantic ontologies and data descriptions.
- ❑ Typical example :
 - A customer books a trip from a virtual travel agency.
 - Flight booking and hotel booking services are identified and aggregated automatically.



(Stollberg et al. 2005)

Agenda

- 1 Bioinformatics: data deluge and ontologies
- 2 Medical informatics: a semantic data lake
- 3 Systems engineering: systems & models
- 4 Ontologies in software engineering
- 5 Thoughts on semantic data

Semantics 2019 conference keynote speak

Try not to move the data.

- ❑ **Hybrid architectures** : on-premise + cloud. De-centralized business logic. Structured + unstructured. Multiple data silos, technologies, vendors... the typical enterprise today...
- ❑ **RDF** -- particularly cloud-based RDF (and linked data) – **a viable, non-disruptive, universal solution as global metadata manager** for the enterprise.
- ❑ Wrapping a virtual model around multiple domain models + RDF views of relational data + materialized entailments ... seems to be a game changer.
- ❑ Among **CIO digital transformation issues** :
 - Widespread lack of familiarity with knowledge/semantics in general — usually because it is outside the scope of typical computer science.
 - Data migration, storage, and **ETL** are taking more and more bandwidth.



Michael J. Sullivan, Principal Cloud
Solutions Architect at Oracle
(<https://2019.semantics.cc/try-not-move-data>)

Semantics 2019 conference keynote speak

FAIR data and services

- ❑ Semantic technologies have proven crucial to ... represent and reason about complex biomedical knowledge, ... make knowledge available.
- ❑ **FAIR** - **F**indable, **A**ccessible, **I**nteroperable, **R**eusable : principles for improving how we discover and reuse high-value research data.
 - Endorsed by the G20, the European Commission, Horizon 2020, ... NIH ...
 - **Findable** : Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.
 - **Accessible** : Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.
 - **Interoperable** : Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
 - **Reusable** : Data and collections have a clear usage licenses and provide accurate information on provenance.



Michel Dumontier, Distinguished
Professor of Data Science at
Maastricht University
(<https://2019.semantics.cc/beyond-hype-and-pessimism-fair-data-and-services>)

International Semantic Web Conference 2019 – keynote speak

For knowledge

- ❑ Written expression and communication ... allow learning directly from elaborated knowledge instead of by experience.
- ❑ The worldwide web facilitating cultural exchange is a culminating point in this story.
- ❑ The idea of a semantic web allowing machines to have a grasp on this knowledge is a tremendous idea.
- ❑ *Alas, after twenty years, the semantic web field is mostly focused on data, even when it is made of so-called knowledge graphs.*
- ❑ *The grand goal of formally expressing knowledge on the web must be rehabilitated.*
- ❑ We do not need consistent knowledge at the web scale, but local theories that can be combined...



Jérôme Euzenat, senior research
scientist at INRIA, France
(<https://iswc2019.semanticweb.org/keynote-euzenat/>)

Any thought you want to add ?

Summary

- ❑ The role of ontologies in software engineering is multiform; contributions have been identified and investigated in all stages of the software life cycle.
- ❑ From a global software engineering point of view, the Ontology Definition Metamodel from OMG offers a basis to bring together knowledge engineering and software engineering through a common metalanguage and mappings.
- ❑ Implementation of business rules, as in expert systems, can rely i.a. on the Semantic Web Rule Language SWRL, which is merging OWL DL with Horn logic rules. A restriction, DL-Safe rules, must be respected to keep the result decidable.
- ❑ As a general conclusion, **data integration**, with the support of semantic metadata and technologies, **and** - ideally web-wide - **knowledge representation**, are the two main contributions from the field of semantic data.

References

- ❑ [Blake and Bult 2006]: Blake J. and Bult C., Beyond the data deluge: Data integration and bio-ontologies, *Journal of Biomedical Informatics* 39, 314–320, Elsevier, 2006.
- ❑ [Dessimoz and Škunca 2017]: Dessimoz, C. and Škunca, N., The Gene Ontology Handbook, Springer, 2017.
- ❑ [Fensel et al. 2011]: J., Fensel, D., Hendler, J. (Eds.), Handbook of Semantic Web Technologies, Springer, 2011.
- ❑ [Gene Ontology Consortium 2019]: The Gene Ontology Resource: 20 years and still GOing strong, *Gene Ontology Consortium, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019*.
- ❑ [Happel and Seedorf 2006]: H.-J. Happel and S. Seedorf, Applications of Ontologies in Software Engineering, *2nd International Workshop on Semantic Web-Enabled Software Engineering (SWESE 2006), 5th International Semantic Web Conference (ISWC 2006), Athens, USA, 2006*.
- ❑ [Merelli et al. 2014]: Merelli I., Pérez-Sánchez H., Gesing S. and D'Agostino D., Managing, Analysing, and Integrating Big Data in Medical Bioinformatics : Open Problems and Future Perspectives, *BioMed Research International, volume 2014*.
- ❑ [Pan & al. 2013]: Pan J.Z., Staab S., Aßmann U., Ebert J., Zhao Y., (Eds.), Ontology driven software development, Springer, 2013.
- ❑ [Pollack 2011]: Pollack, A., DNA Sequencing Caught in Deluge of Data, *The New York Times, November 2011*.
- ❑ [Schatz & Langmead 2013]: Schatz M. and Langmead B., The DNA data deluge, *IEEE Spectrum* July 2013; 50(7): 26–33.
- ❑ [Smith et al. 2007]: Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L.J., Eilbeck K., Ireland A., ... Lewis S., The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration, *Nature biotechnology* 25.11, 2007.
- ❑ [Stephens et al. 2015]: Stephens Z., Lee S., Faghri F., Campbell R., Zhai C., Efron M, Lyer R., Schatz M., Sinha S., Robinson G., *Big Data: Astronomical or Genomical?*, *PLoS Biol* 13(7):e1002195. doi:10.1371/journal.pbio.1002195, July 7, 2015.
- ❑ [Stollberg et al. 2005]: M. Stollberg, M. Moran, J. Domingue and L. Cabral, Semantic Web Services Tutorial, *5th International Conference on Web Engineering (ICWE 2005) Sydney, Australia, 2005*.

THANK YOU

