V.2.4

Knowledge Graphs in the Life-Science Industry

Heiner Oberkampf, PhD

April 17th 2018, Knowledge Graph Lab, RWTH, Aachen

OSTHUS

WHO IS OSTHUS?

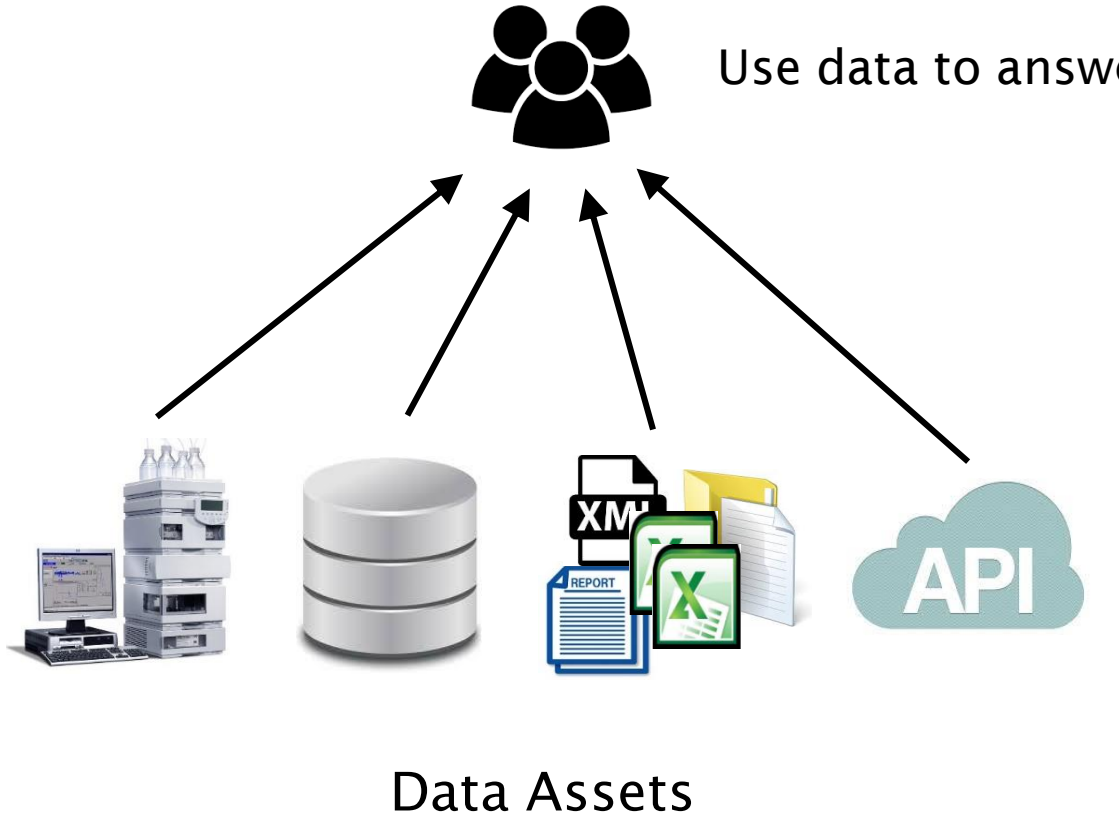
- Global organization
- 120+ employees and growing rapidly
- 16+ big pharma customers + many chemicals and lab-based companies.
- Our approach technology:

Connecting data, people and organizations

OSTHUS

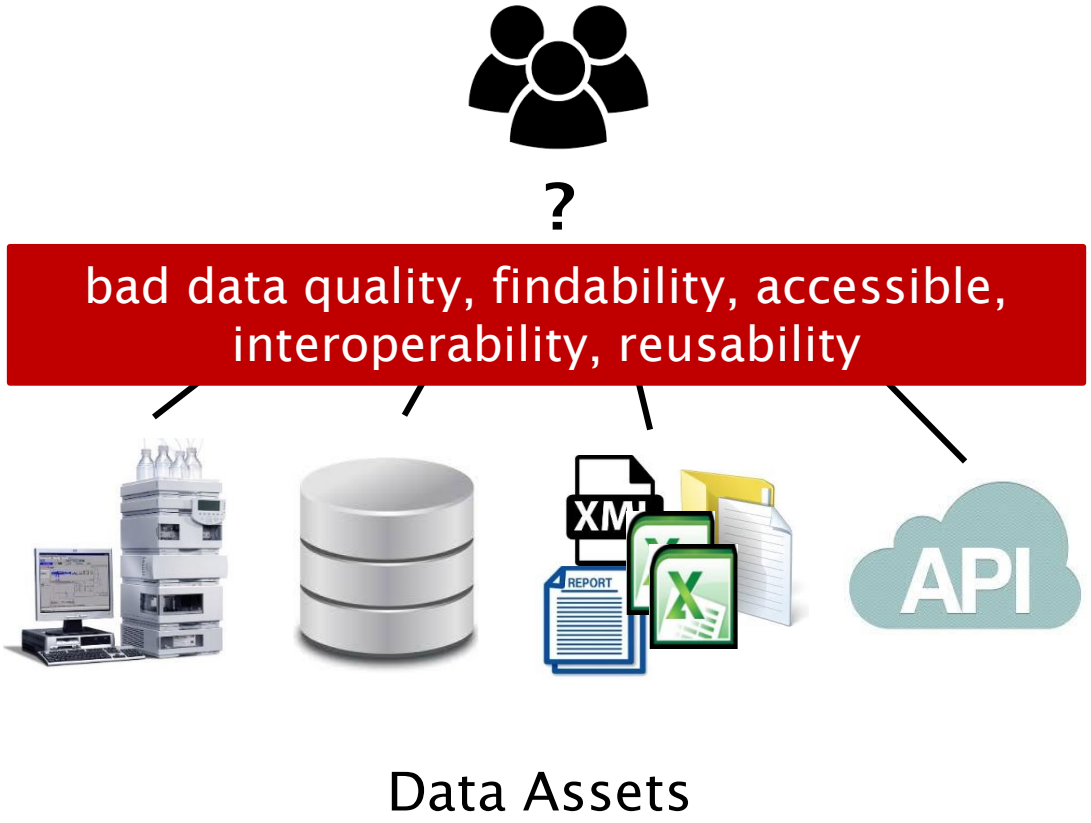
Industry Perspective

Value of Data is Realized by Usage in Decision Making and Insights



OSTHUS

Value of Data is Mostly Not Realized



Data Assets

“Only 3% of Companies’ Data Meets Basic Quality Standards”

Harvard Business Review: https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards

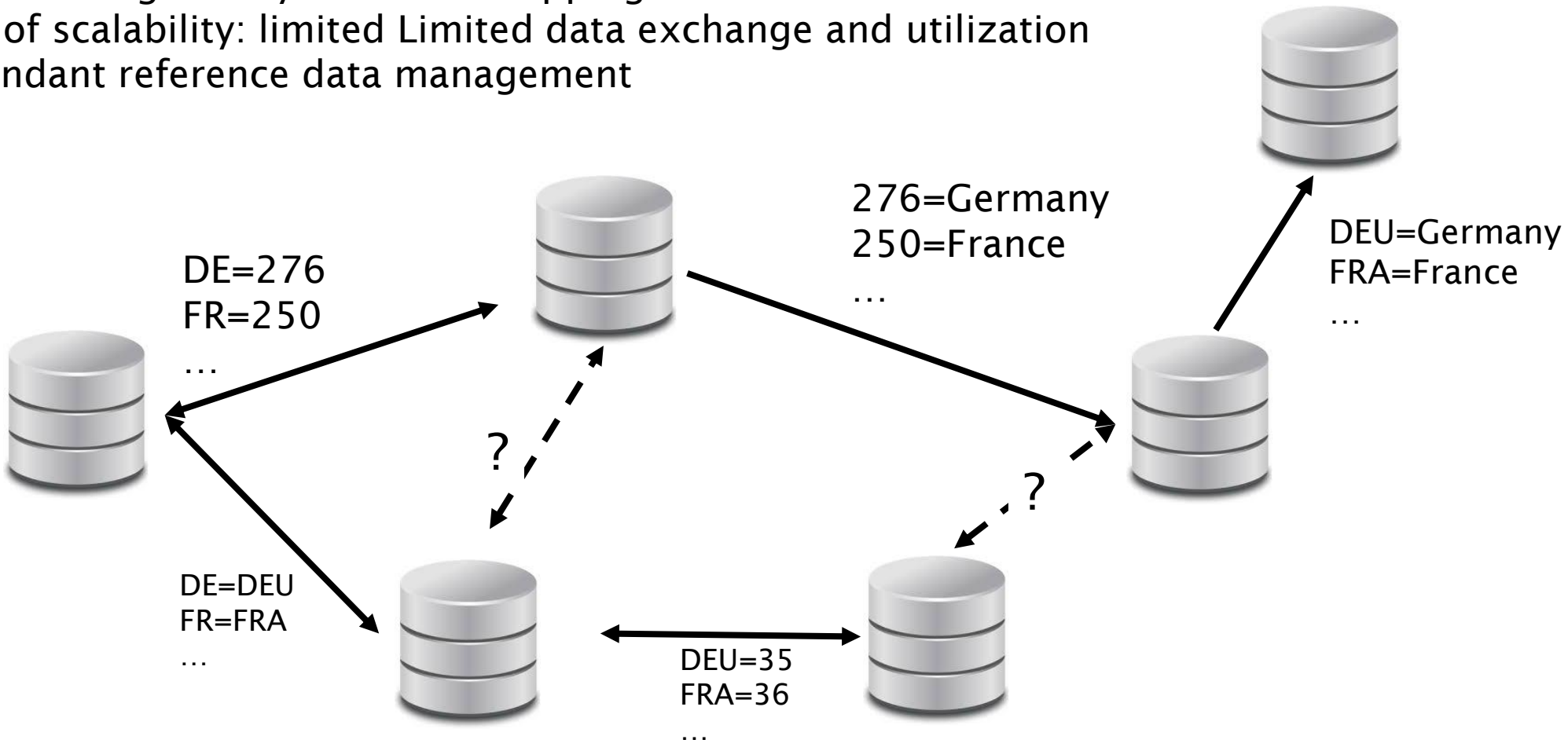
“For most large enterprises, the root of this problem lies in years of treating the data generated by their operational systems as a form of exhaust rather than as a fuel to deliver great services, build better products, and create competitive advantage.”

Database Trends and Applications: http://www.dbta.com/Editorial/Trends-and-Applications/The-Enterprise-Data-Debt-Crisis-123008.aspx

OSTHUS

Application-Centric World: Knowledge in Interfaces

- Silos are bridged with point to point interfaces or ETL processes
- Data exchange always involves mapping reference data
- Lack of scalability: limited Limited data exchange and utilization
- Redundant reference data management



OSTHUS

Two Attempts to Overcome Silos

Data Warehouse

- \+ Proven enterprise technology
- − Big DWHs require too great an effort
- − Not all data is suitable for rigid DWHs

Data Lake

- \+ Great flexibility and very little effort to store all sorts of data
- − Data lakes are too loose a construct
- − Tremendous efforts on retrieval

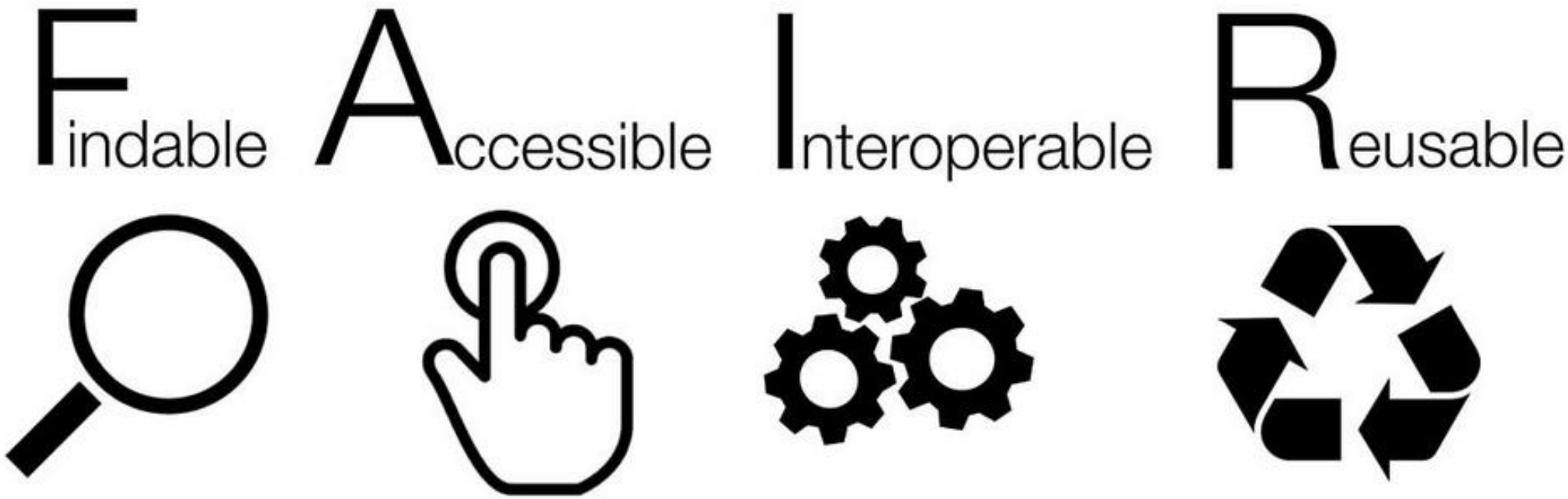
What is Problematic About Data Lakes?

This is not FAIR!!!

"Not if you have to clean up a data swamp!"

OSTHUS

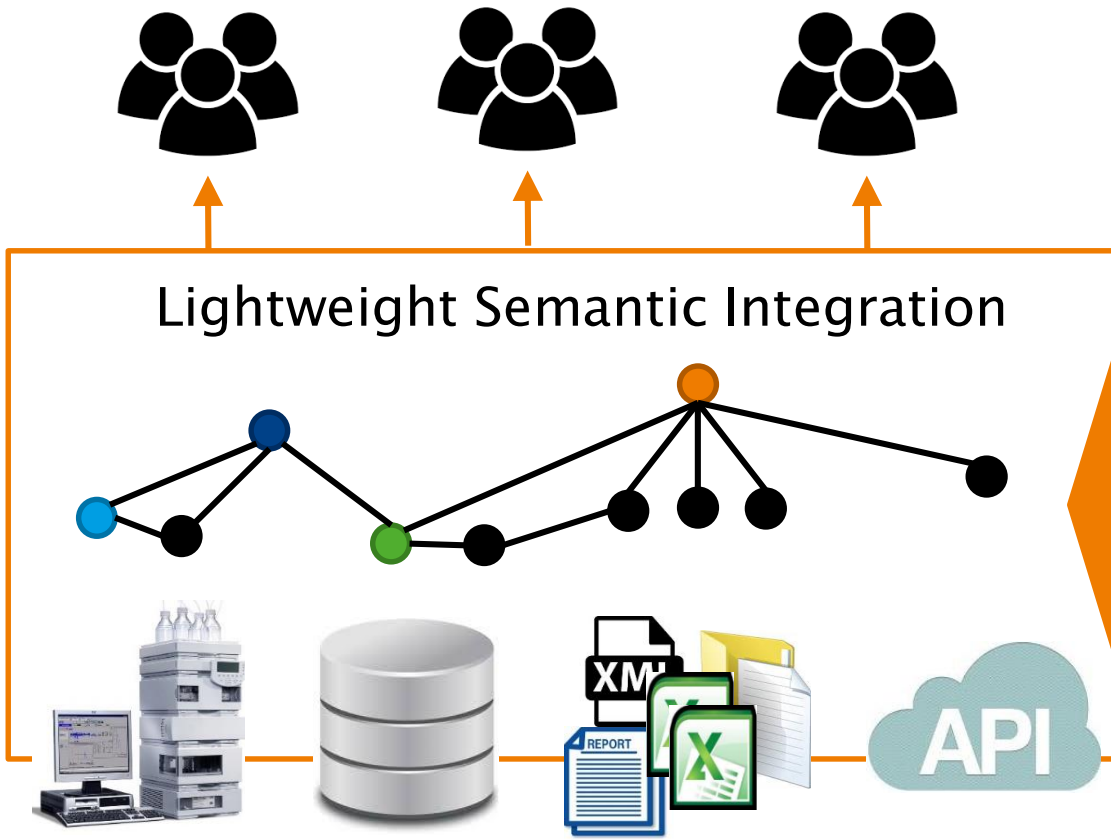
Guiding Principles for Scientific Data Management and Stewardship*



*Source: https://www.nature.com/articles/sdata201618
G20 endorse the FAIR principles: https://www.dtls.nl/2016/09/13/g20-endorse-fair-principles/

OSTHUS

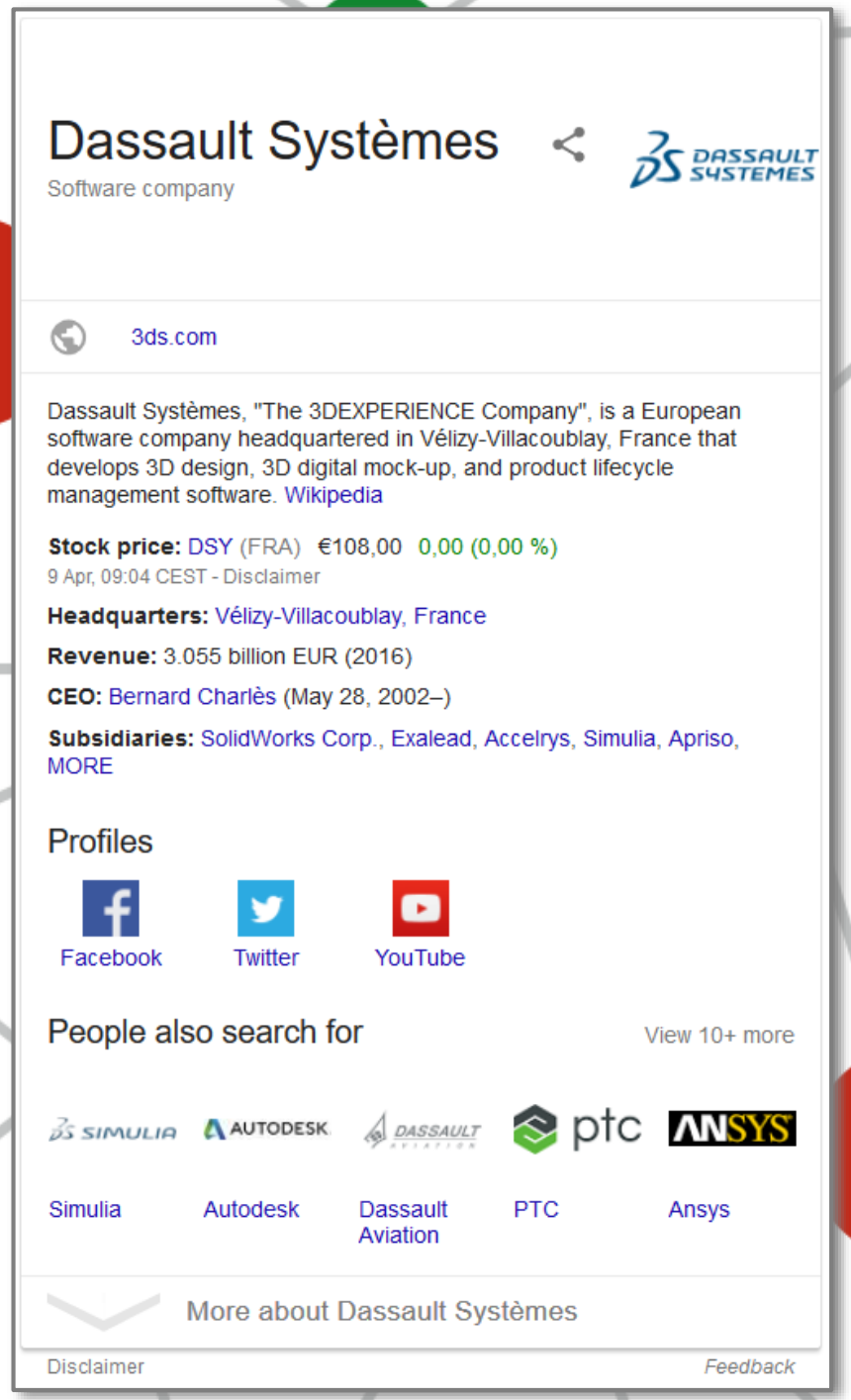
Make Data **F**indable, **A**ccessible, **I**nteroperable **R**eusable



1. reference master data mgmt.
2. data catalogs & metadata mgmt.
3. data governance & quality mgmt.
4. data integrity

Data Assets

OSTHUS



KNOWLEDGE GRAPH

Linked Open Data

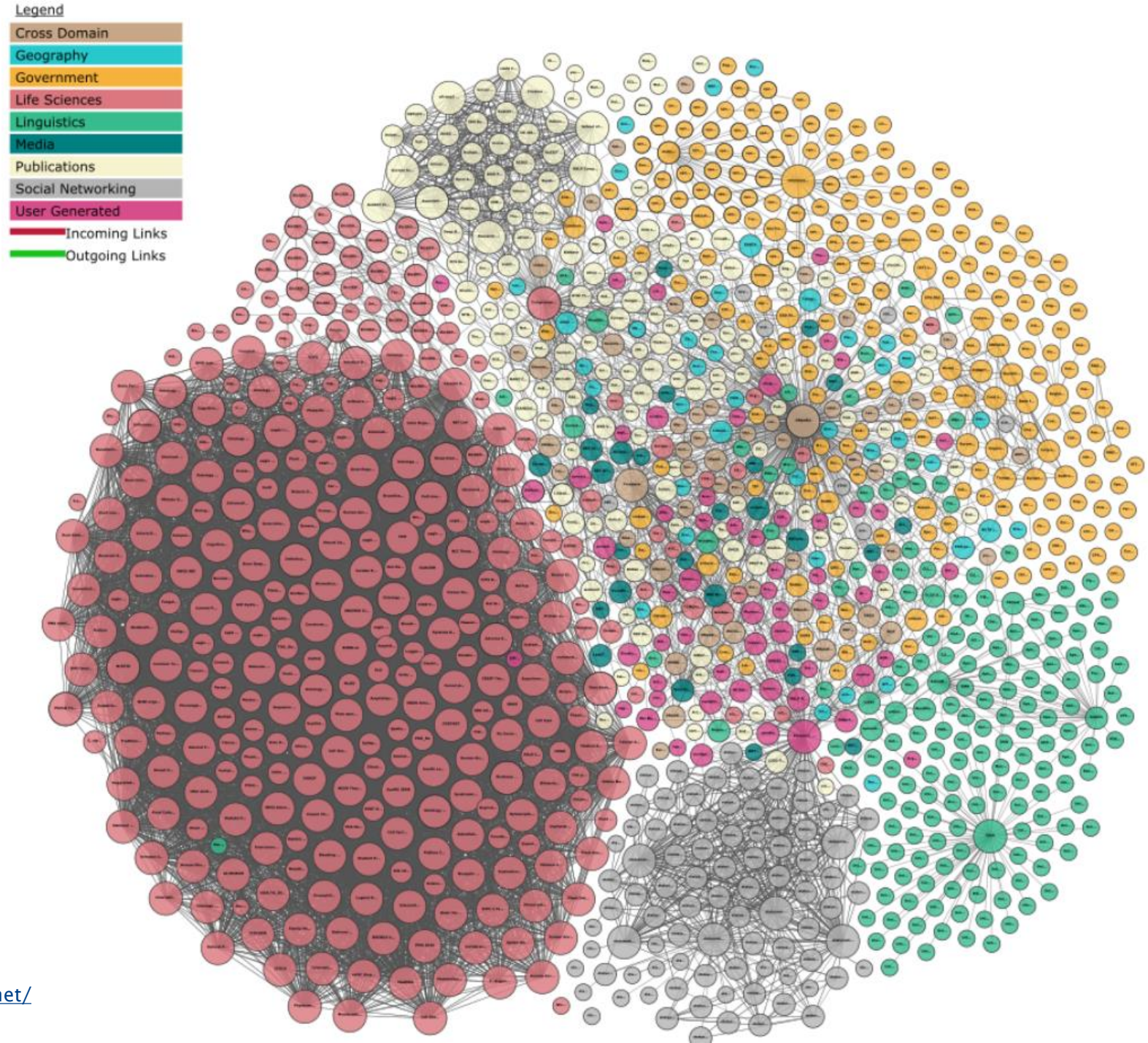


Image source: http://lod-cloud.net/

OSTHUS

Semantics

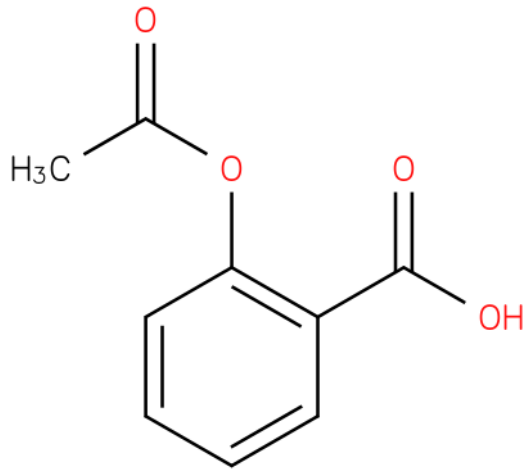
Words, Terms and Concepts

concept = ”An abstract entity signifying a general characterizing idea or universal which acts as a category for instances. The unit of semantics (meaning), the node in some mental or knowledge organization system.”
[Obrst2010]

Synonyms are …

different terms which represent the *same concept:*

- Colfarit
- Dispril
- Solupsan
- **Acetylsalicylic Acid**
- Acetysal
- 2-(Acetyloxy)benzoic Acid
- Micristin
- Polopiryna
- Benzoic acid, 2-(acetyloxy)-
- Ecotrin
- Magnecyl
- Zorprin
- Acylpyrin
- Solprin
- Easprin
- Acid, Acetylsalicylic
- Aloxiprimum
- Endosprin
- Polopirin
- Aspirin



Semantics

- Has its origins in philosophy - generally understood as the abstract study of *meaning*
- Distinguished from *syntax* – which is the rules-based grammar of a language

"Washington"

OSTHUS

How can we express meaning?

Textual Descriptions

Aspirin, also known as acetylsalicylic acid (ASA), is a medication, often used to treat pain, fever, and inflammation. Aspirin is also used long-term, at low doses, to help prevent heart attacks, strokes, and blood clot formation in people at high risk of developing blood clots. Low doses of aspirin may be given immediately after a heart attack to reduce the risk of another heart attack or the death of heart tissue. Aspirin may be effective at preventing certain types of cancer, particularly colorectal cancer. The main side effects of aspirin are gastric ulcers, stomach bleeding, and ringing in the ears, especially with higher doses. While daily aspirin can help prevent a clot-related stroke, it may increase risk of a bleeding stroke (hemorrhagic stroke). In children and adolescents, aspirin is not recommended for flu-like symptoms or viral illnesses, because of the risk of Reye's syndrome. Aspirin is part of a group of medications called nonsteroidal anti-inflammatory drugs (NSAIDs), but differs from most other NSAIDs in the mechanism of action. The salicylates have similar effects (antipyretic, anti-inflammatory, analgesic) to the other NSAIDs and inhibit the same enzyme cyclooxygenase (COX), but aspirin does so in an irreversible manner and, unlike others, affects the COX-1 variant more than the COX-2 variant of the enzyme. Aspirin also has an antiplatelet effect by stopping the binding together of platelets. The therapeutic properties of willow tree bark have been known for at least 2,400 years, with Hippocrates prescribing it for headaches. Salicylic acid, the active ingredient of aspirin, was first isolated from the bark of the willow tree in 1763 by Edward Stone of Wadham College, University of Oxford. Felix Hoffmann, a chemist at Bayer, is credited with the synthesis of aspirin in 1897, though whether this was of his own initiative or under the direction of Arthur Eichengrün is controversial. Aspirin is one of the most widely used medications in the world with an estimated 40,000 tonnes of it being consumed each year. In countries where "Aspirin" is a registered trademark owned by Bayer, the generic term is acetylsalicylic acid (ASA). It is on the WHO Model List of Essential Medicines, the most important medications needed in a basic health system.

[Wikipedia]

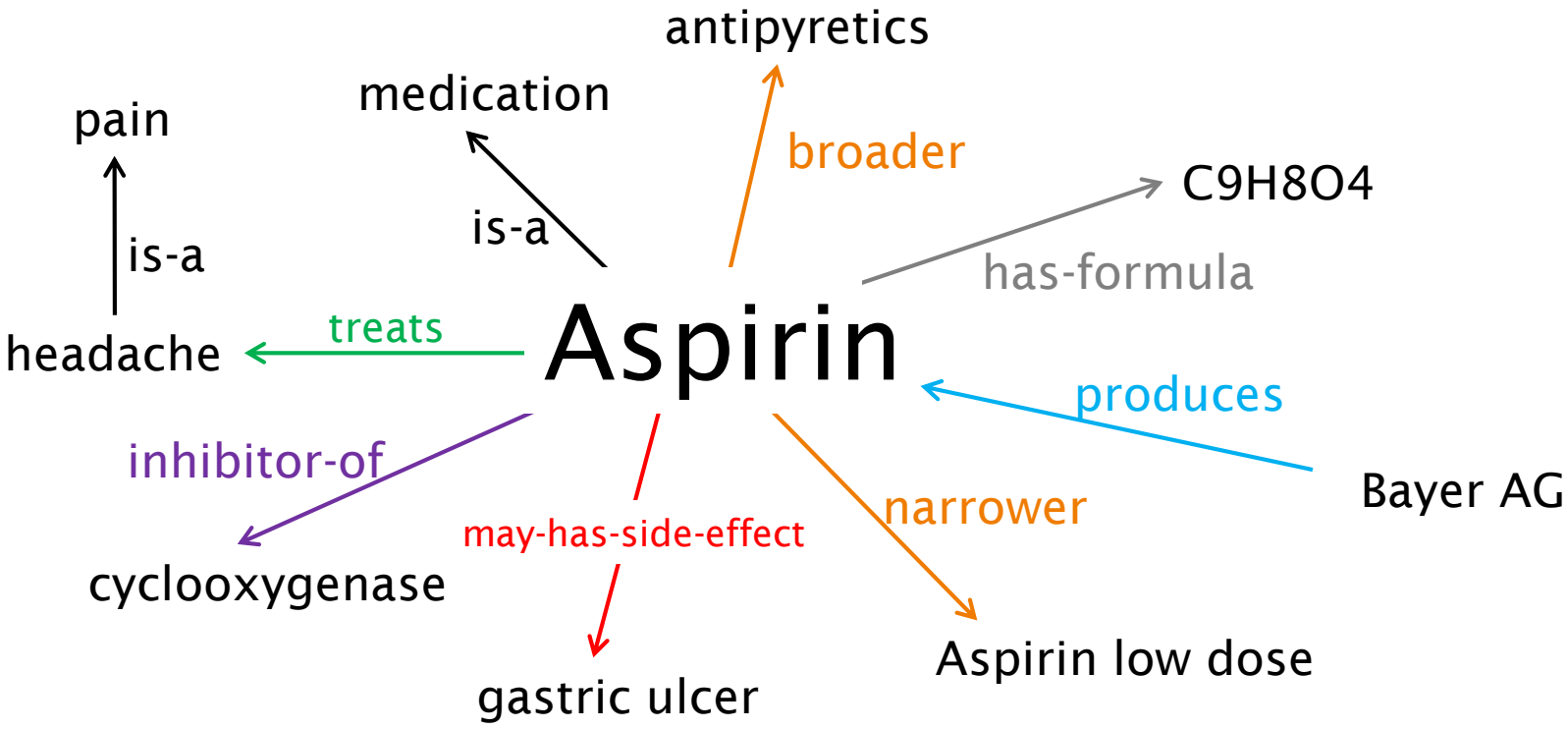
Textual Description + Links

Aspirin, also known as acetylsalicylic acid (ASA), is a medication, often used to treat pain, fever, and inflammation. Aspirin is also used long-term, at low doses, to help prevent heart attacks, strokes, and blood clot formation in people at high risk of developing blood clots. Low doses of aspirin may be given immediately after a heart attack to reduce the risk of another heart attack or the death of heart tissue. Aspirin may be effective at preventing certain types of cancer, particularly colorectal cancer. The main side effects of aspirin are gastric ulcers, stomach bleeding, and ringing in the ears, especially with higher doses. While daily aspirin can help prevent a clot-related stroke, it may increase risk of a bleeding stroke (hemorrhagic stroke). In children and adolescents, aspirin is not recommended for flu-like symptoms or viral illnesses, because of the risk of Reye's syndrome. Aspirin is part of a group of medications called nonsteroidal anti-inflammatory drugs (NSAIDs), but differs from most other NSAIDs in the mechanism of action. The salicylates have similar effects (antipyretic, anti-inflammatory, analgesic) to the other NSAIDs and inhibit the same enzyme cyclooxygenase (COX), but aspirin does so in an irreversible manner and, unlike others, affects the COX-1 variant more than the COX-2 variant of the enzyme. Aspirin also has an antiplatelet effect by stopping the binding together of platelets. The therapeutic properties of willow tree bark have been known for at least 2,400 years, with Hippocrates prescribing it for headaches. Salicylic acid, the active ingredient of aspirin, was first isolated from the bark of the willow tree in 1763 by Edward Stone of Wadham College, University of Oxford. Felix Hoffmann, a chemist at Bayer, is credited with the synthesis of aspirin in 1897, though whether this was of his own initiative or under the direction of Arthur Eichengrün is controversial. Aspirin is one of the most widely used medications in the world with an estimated 40,000 tonnes of it being consumed each year. In countries where "Aspirin" is a registered trademark owned by Bayer, the generic term is acetylsalicylic acid (ASA). It is on the WHO Model List of Essential Medicines, the most important medications needed in a basic health system.

[Wikipedia]

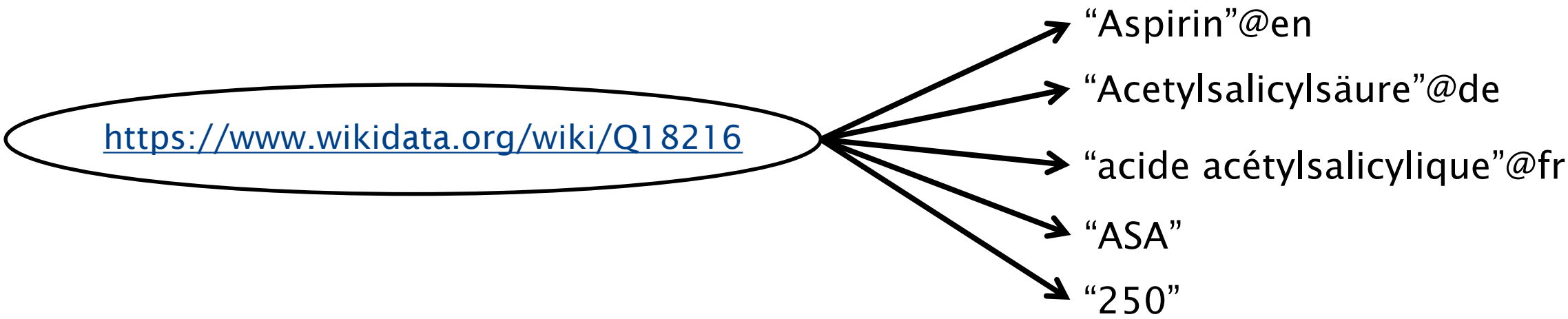
Semantic Networks

A simple, non-formal way to express the meaning of a concept through relations (links) to other concepts.



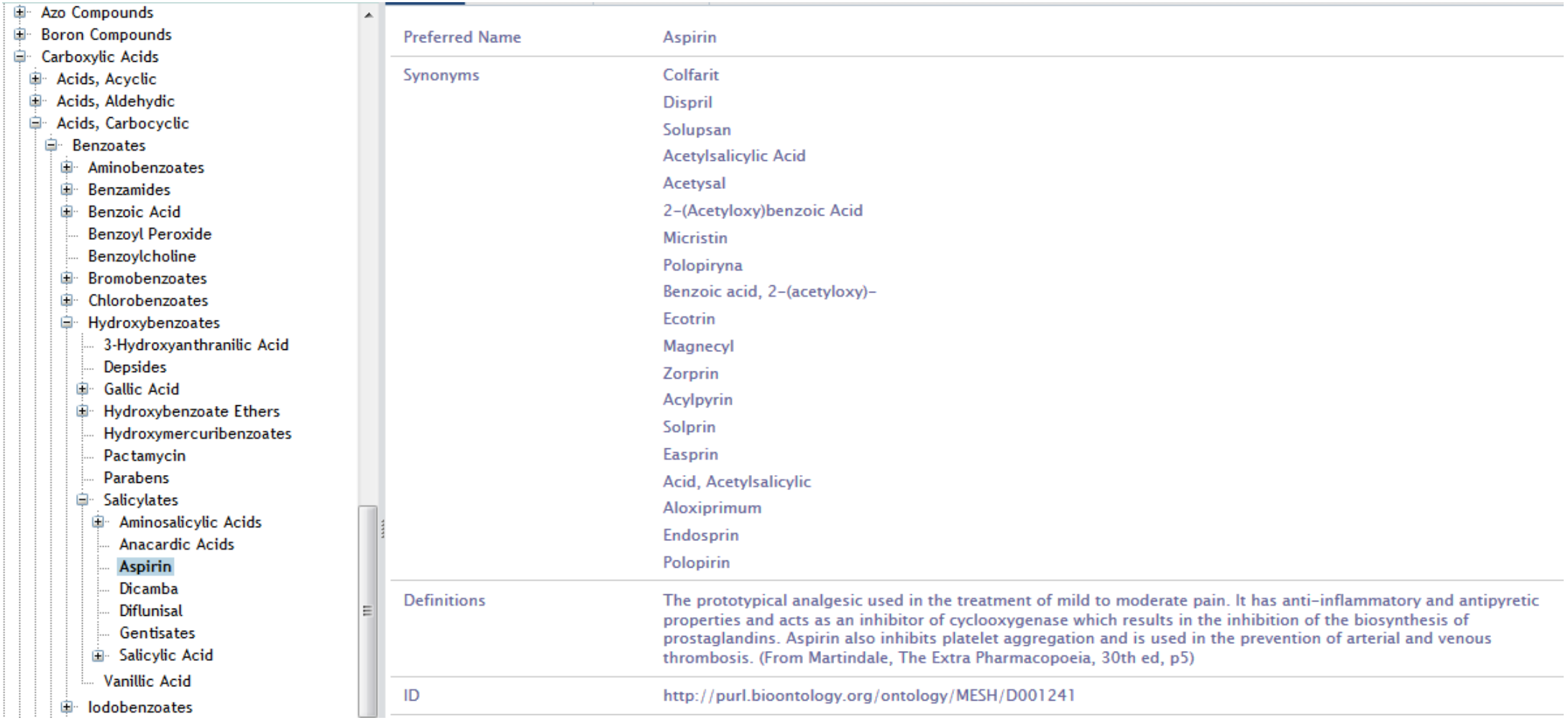
OSTHUS

Semantics: Bind Different Names & Identifiers to Unique Resources



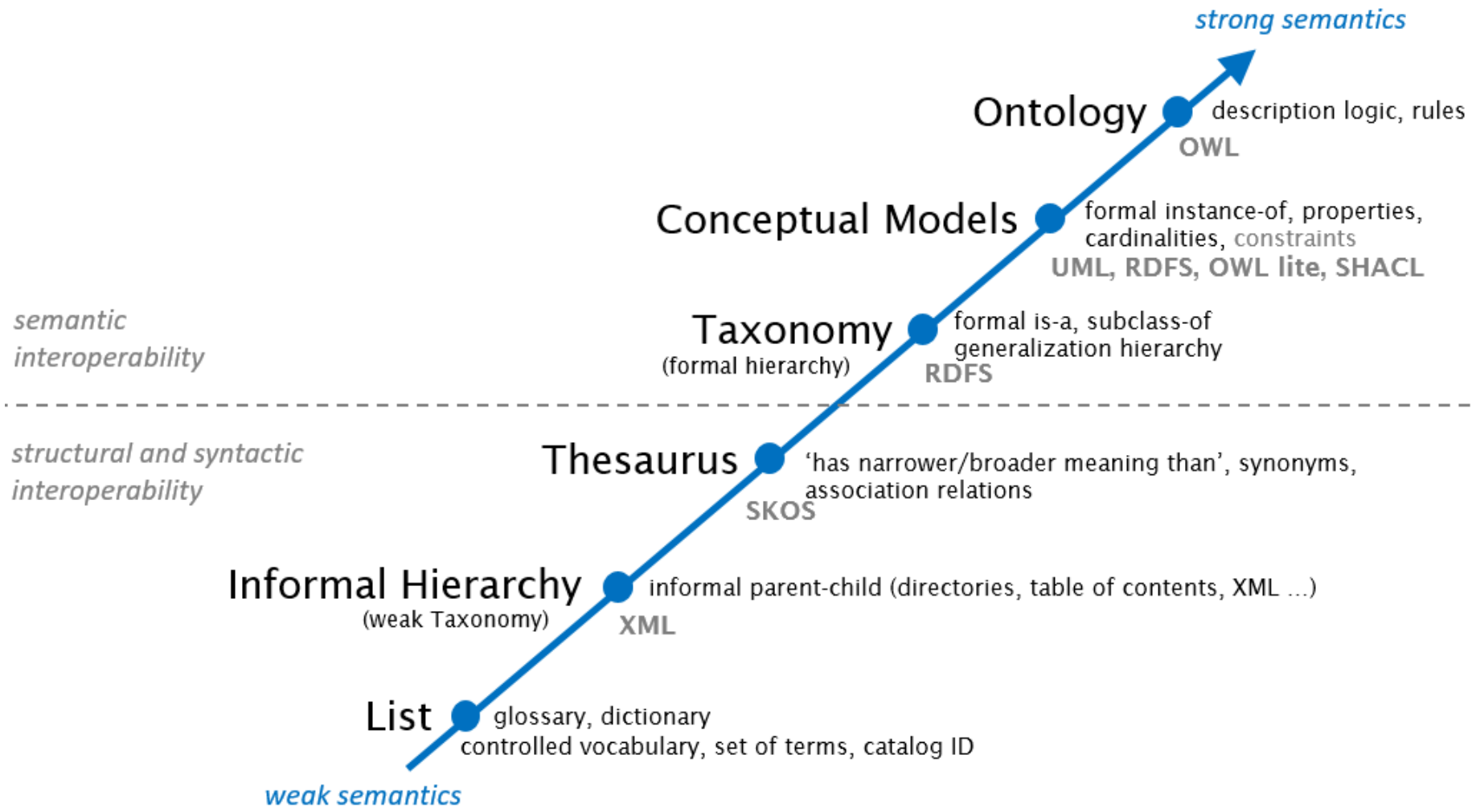
OSTHUS

Medical Subject Headings (MESH) Thesaurus



Source: http://bioportal.bioontology.org/ontologies/MESH

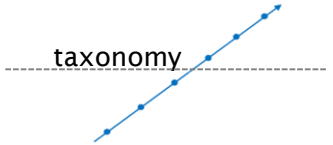
Spectrum of Semantic Knowledge Organization Systems



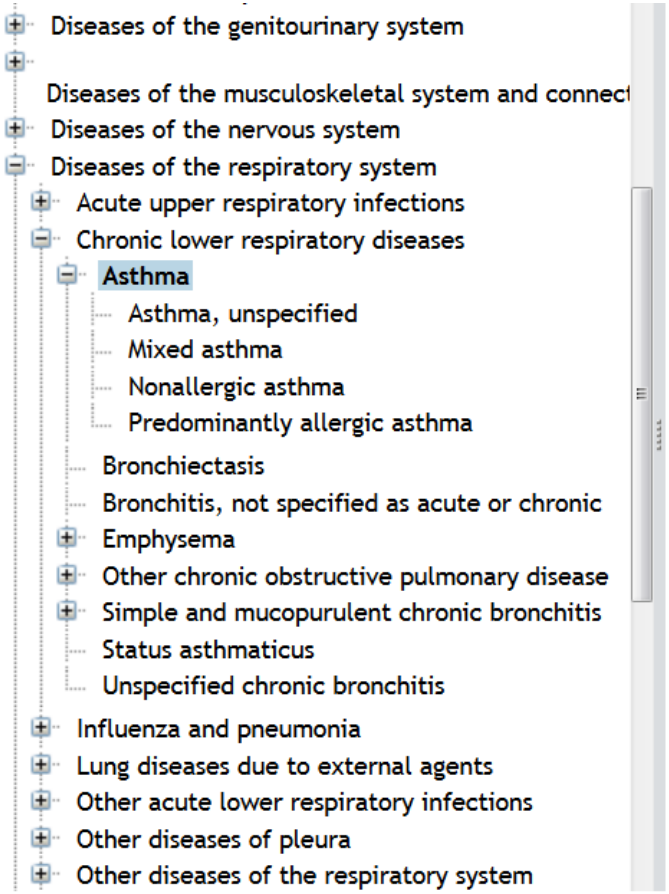
Sources

- Deborah L. McGuinness. "Ontologies Come of Age". In Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2003.
- Michael Uschold and Michael Gruninger "Ontologies and semantics for seamless connectivity" *SIGMOD Rec.* 33, 4 (December 2004), 58-64. DOI=http://dx.doi.org/10.1145/1041410.1041420
- Leo Obrst "The Ontology Spectrum". Book section in of Roberto Poli, Michael Healy, Achilles Kameas "Theory and Applications of Ontology: Computer Applications". Springer Netherlands, 17 Sep 2010.
- Leo Obrst and Mills Davis "Semantic Wave 2008 Report: Industry Roadmap to Web 3.0 & Multibillion Dollar Market Opportunities". 2008.

Strong Taxonomy: Classification



A ***taxonomy*** is a formal generalization-specialization (subclass or is-a) hierarchy. It allows inference along the class hierarchy.



Preferred Name	Asthma
ID	http://purl.bioontology.org/ontology/ICD10/J45
cui	C0004096
notation	J45
prefLabel	Asthma
tui	T047
subClassOf	Chronic lower respiratory diseases

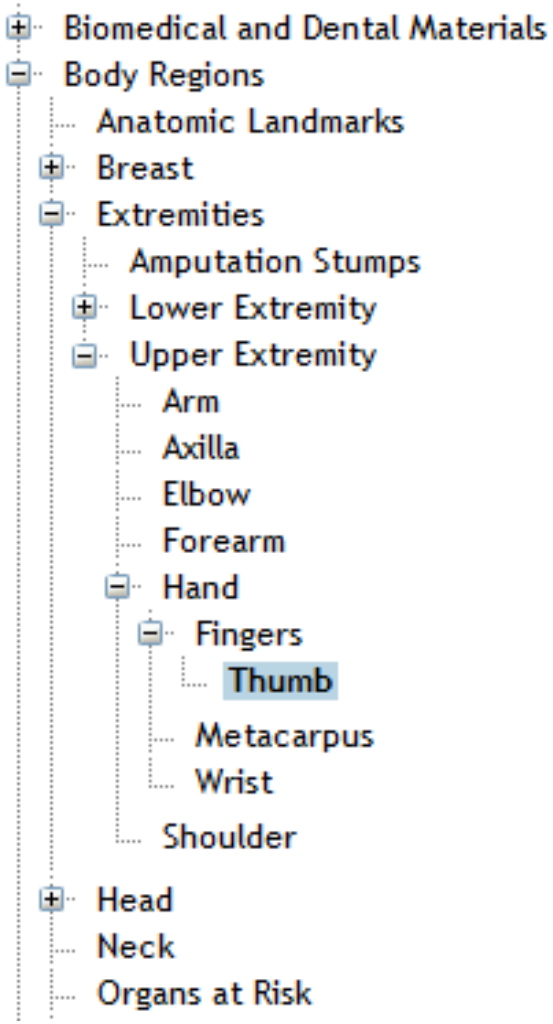
Commonly different classification systems are used to express different perspectives typically in a single inheritance hierarchy.

Application: Statistics, Analytics

Source: http://bioportal.bioontology.org/ontologies/ICD10/

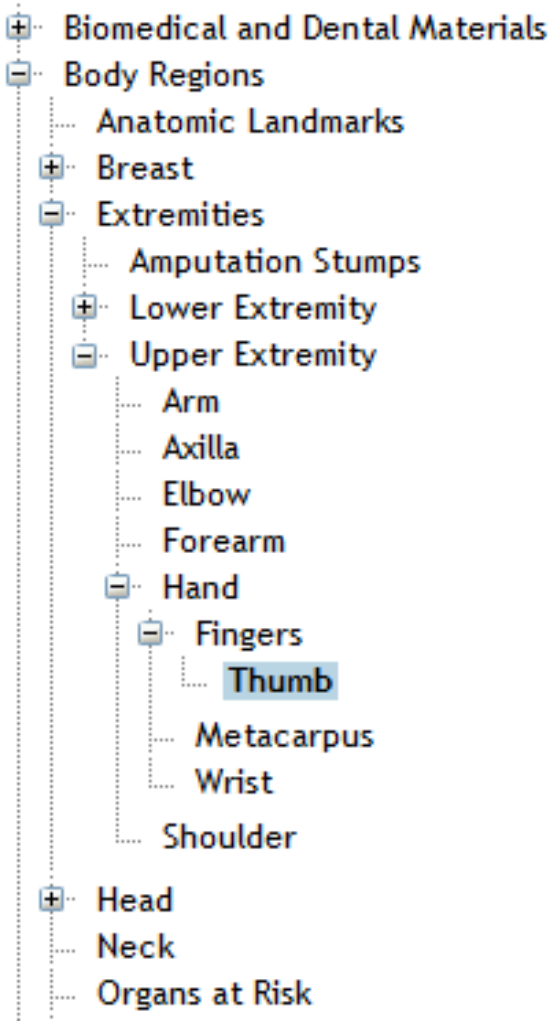
OSTHUS

Question: Thesaurus or Taxonomy?



- How are terms “Thumb” and “Finger” categorized here?
- Examine the relationships

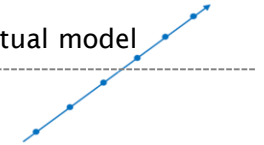
Answer: Thesaurus (not Taxonomy)



MeSH Thesaurus

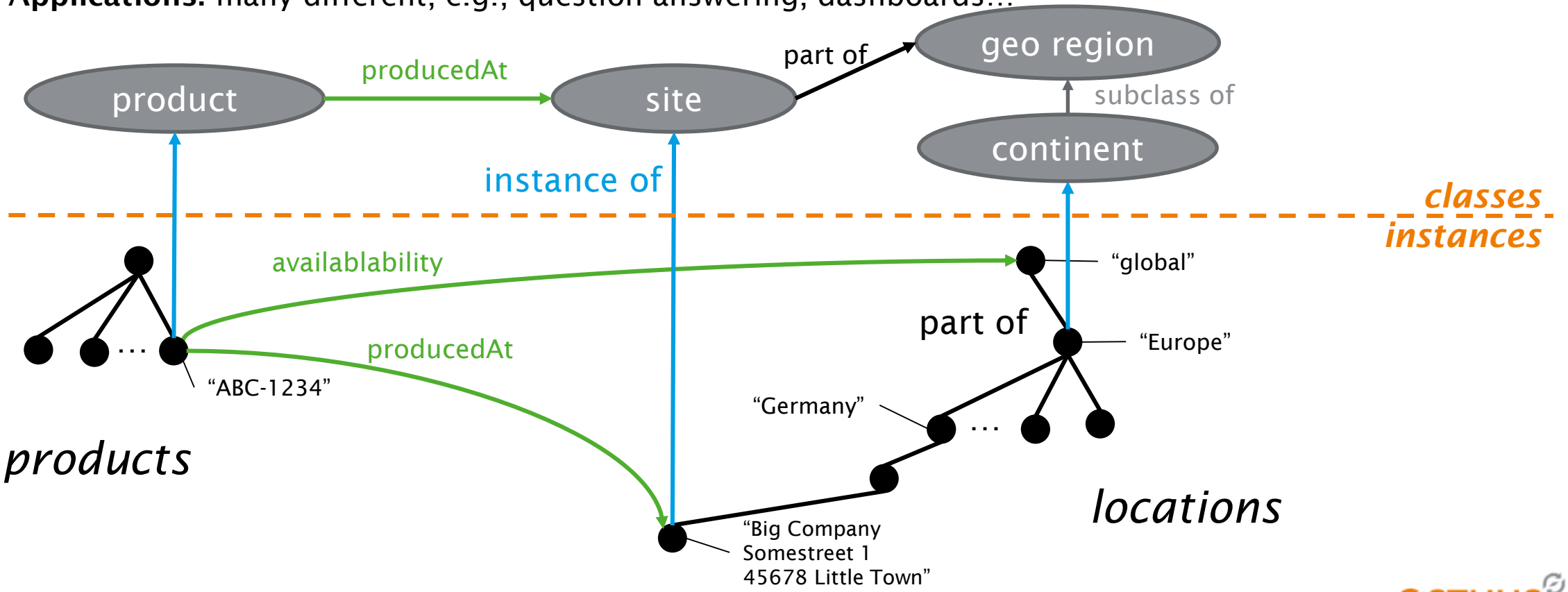
"**MeSH hierarchical links are not subclass relations.** If you interpret them as such you get strange inferences such as 'Every thumb is a hand'. This would do injustice to MeSH , which is a great resource, which fulfils it goals without subscribing to OWL semantics."

Conceptual Models (Knowledge Graphs)



A ***conceptual model*** formally distinguishes between classes and instances and allows to define properties for classes and instances and corresponding inheritance. It also allows to specify qualified relationships between instances.

Applications: many different, e.g., question answering, dashboards…



Triples

subject	predicate	object
RWTH	type	university
RWTH	has location	Aachen
Aachen	part of	Germany

OSTHUS

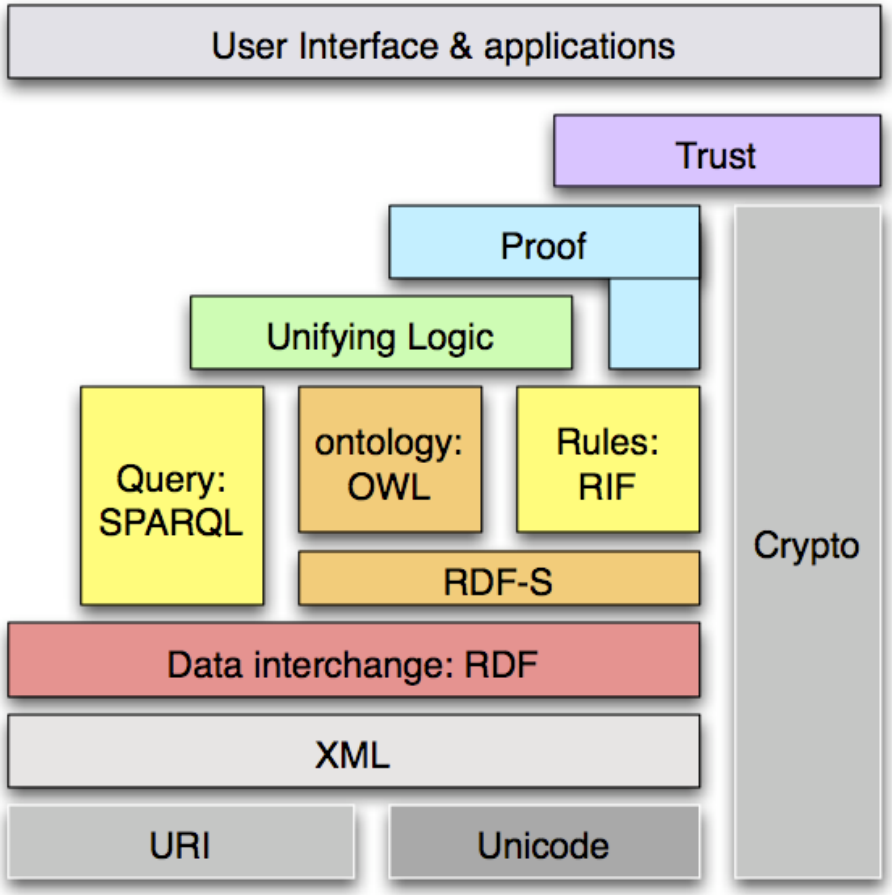
Uniform **Resource** Identifiers (URIs)

https://www.wikidata.org/wiki/Q1017

namespace identifier

OSTHUS

W3C Technology Stack

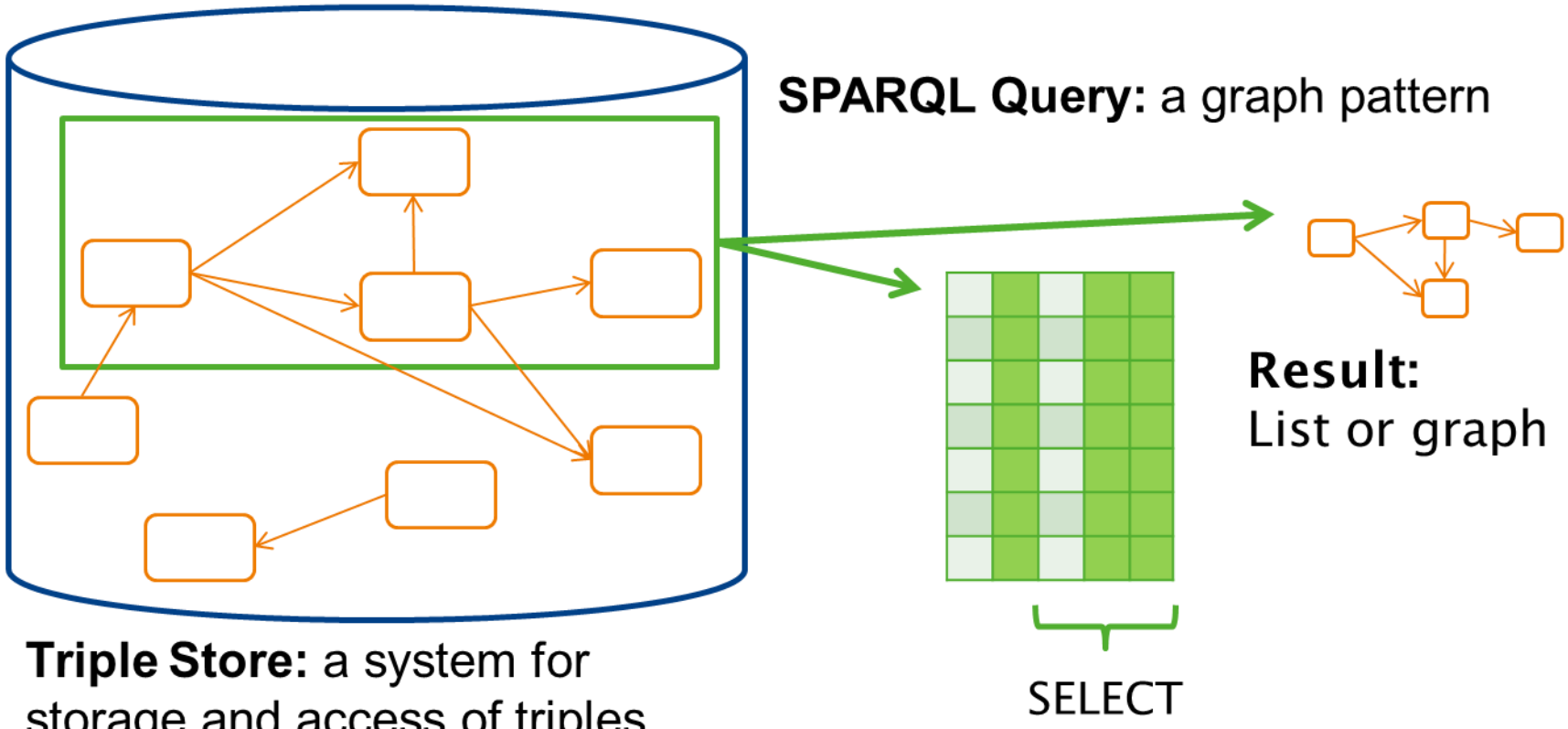


Standards-driven notational semantic stack to implement vendor-agnostic solutions.

Source: Artificial Intelligence and the Semantic Web: AAAI2006 Keynote. 2006. http://www.w3.org/2006/Talks/0718-aaai-tbl/Overview.html

Storage and Access: RDF-based Triple Stores

RDF graph: a set of triples



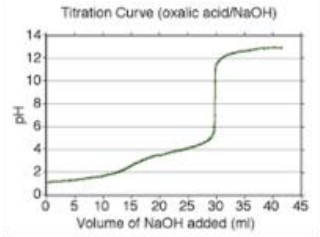
Triple Store: a system for storage and access of triples

Adapted from: http://www.learningsparql.com/

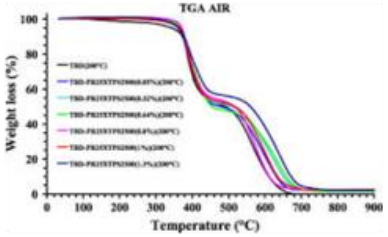
OSTHUS

Use Cases

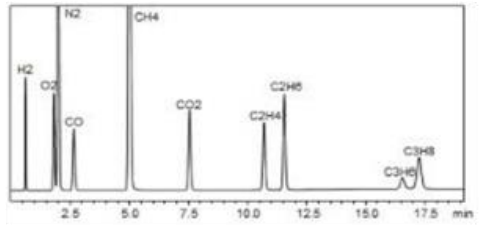
Laboratory Data



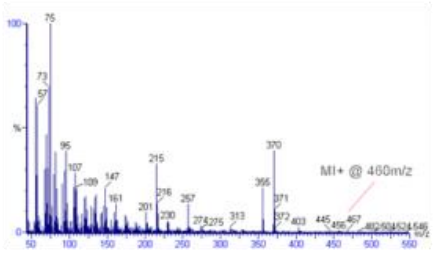
pH



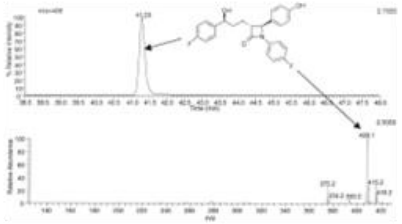
thermogravimetry



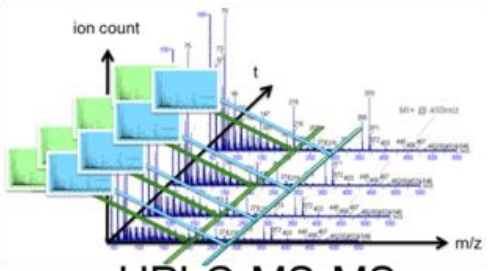
chromatography



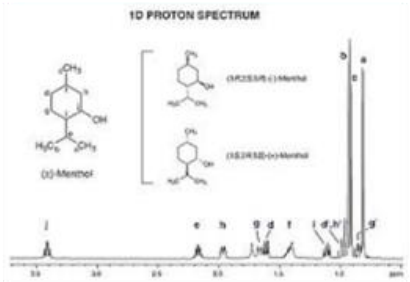
mass spectroscopy



HPLC-MS



HPLC-MS-MS



NMR

cell counter

■ ■ ■

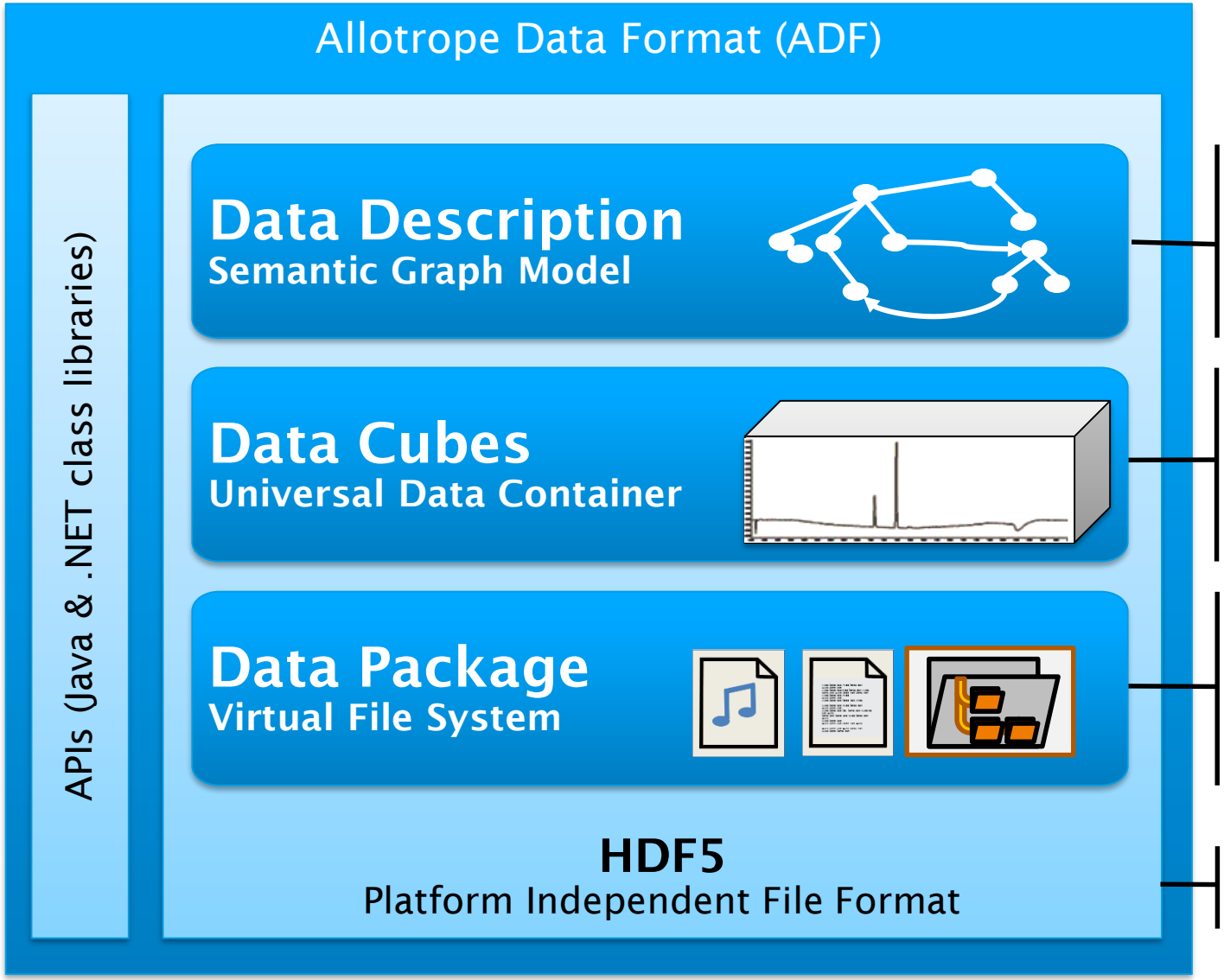
OSTHUS

Allotrope

Astrix Technology Group
BSSN Software
Elemental Machines
Erasmus MC
Fraunhofer IPA
LabAnswer
Mettler Toledo
NIST
SciBite
Stanford University
University of Illinois at Chicago
University of Southampton

OSTHUS

Allotrope Data Format (ADF)



Descriptive metadata about
- Method, instrument, sample, process, result, etc.
- Provenance, audit trail
- Data Cube, Data Package

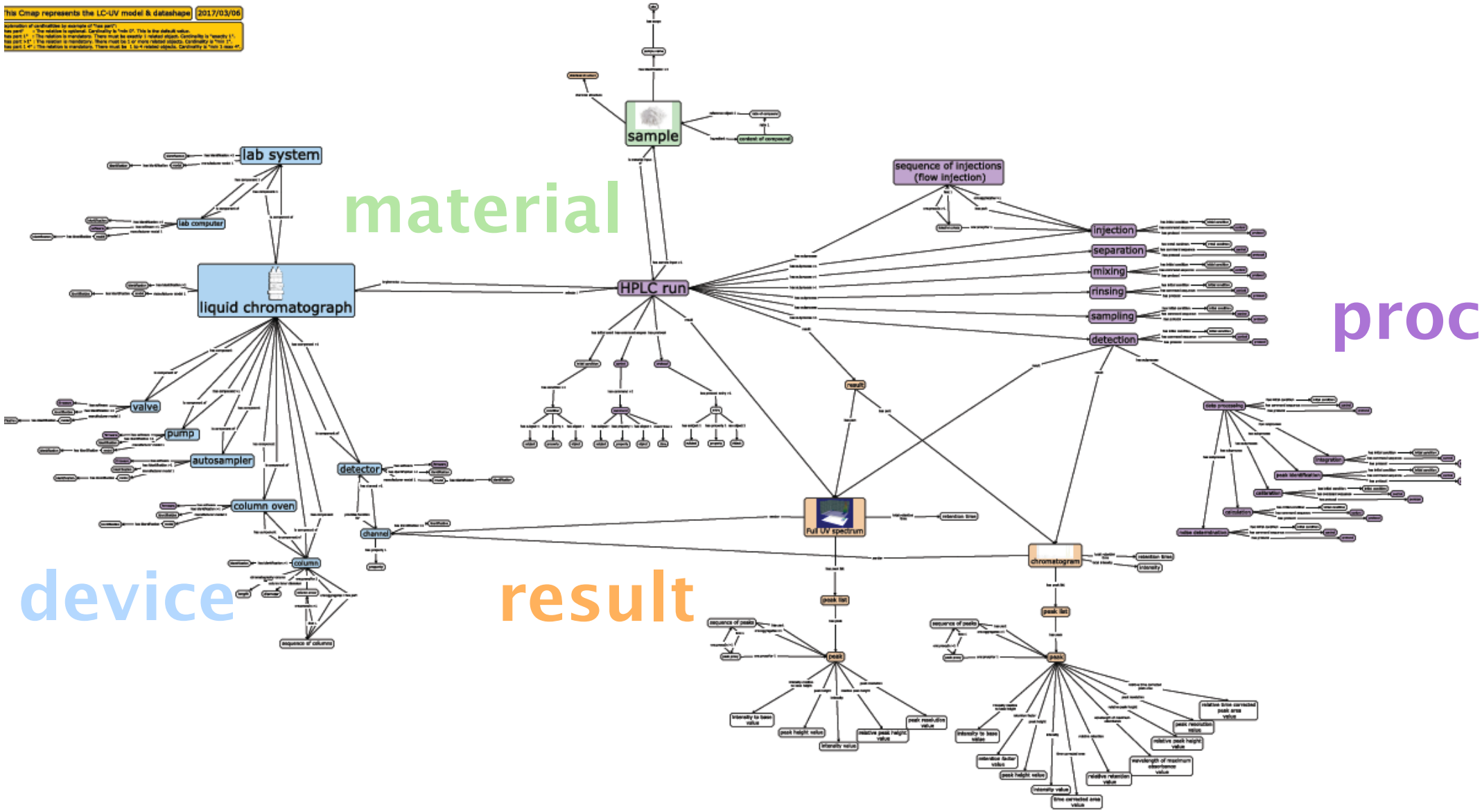
Analytical data represented by one- or multidimensional arrays of homogeneous data structures.

Analytical data represented by arbitrary formats, incl. native instrument formats, images, pdf, video, etc.

Specifically designed to store and organize large amounts of scientific data.

OSTHUS

Ontology for High Performance Liquid Chromatography (HPLC)



WHO Meeting Minutes: flu strains, regions & seasons

Influenza Activity in the WHO European Region, October 2013 - February 2014

Weekly reporting on influenza activity started in week 40/2013. The 2013-2014 season began later compared to the 2012-2013 season with levels of transmission increasing slowly following week 50/2013. The numbers of weekly influenza detections has been much lower than the numbers of detections reported in the previous year and remained so in week 05/2014 [...]

Samples, viruses or clinical specimens, with collection dates after 2013-08-31 have been received from 38 countries in Europe, Africa, the Middle East and the Far East. The large majority (90%) were type A viruses, with A(H3N2) viruses predominating over A(H1N1)pdm09 viruses by a ratio of 1.3:1 (Table 2). Of the type B viruses (just under 10% of all specimens received) B/Yamagata lineage viruses predominated over those of the Victoria lineage by a ratio of 4:1.

The vast majority (99%) of the H1N1 viruses collected after 2013-08-31 were antigenically similar to the vaccine virus (Table 4); only two test viruses (1% of the total tested) showed 4- fold reductions and none showed =8-fold reductions in HI titre compared with the titre of the vaccine virus (A/California/7/2009) with the homologous post-infection ferret antiserum. The HI results for all viruses tested since the September 2013 Vaccine Composition Meeting (VCM) are shown in Tables 6-1 to 6-8. Viruses for which gene sequences are included in phylogenetic trees are highlighted and, where known, the HA genetic group is indicated. The test viruses A/Denmark/106/2013 and A/Lyon/2694/2013, an egg-propagated cultivar of A/Kazakhstan/3314/2014 ...

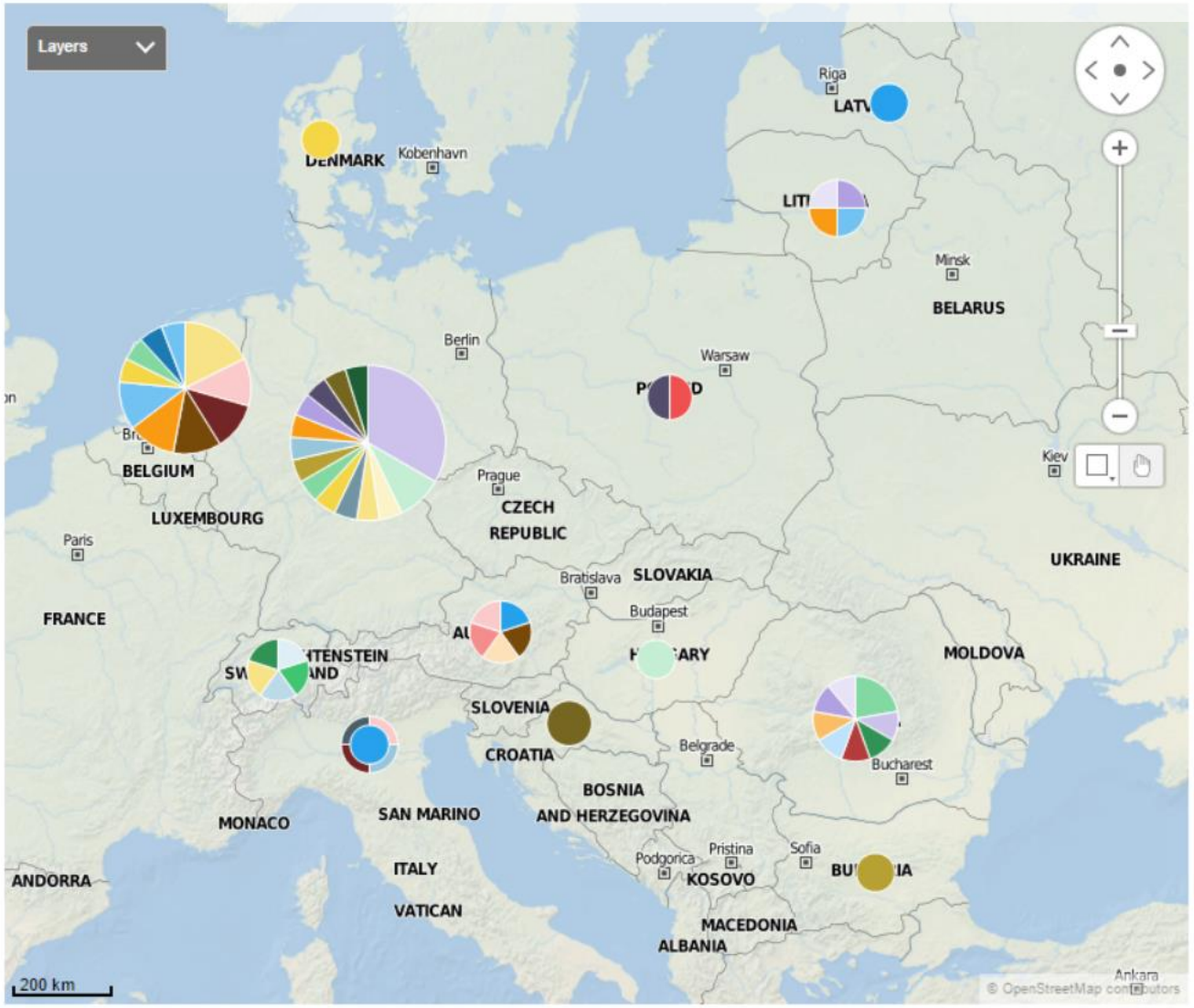
DrugNameSynonyms_Aspirin_WithQuery2 - TIBCO Spotfire

File Edit View Insert Tools Help

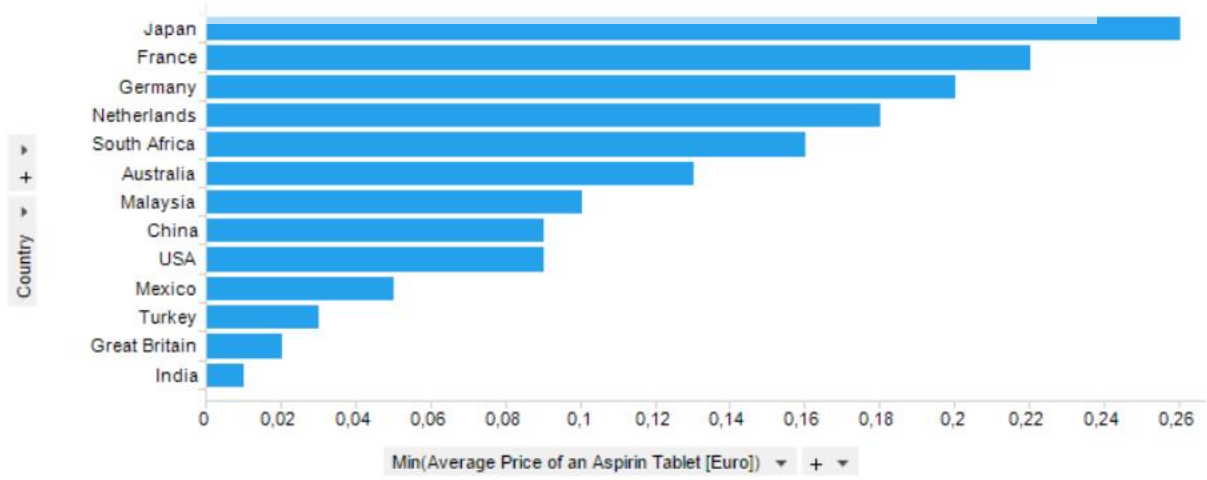
Overview | Marketing

Brand names per Country and Company

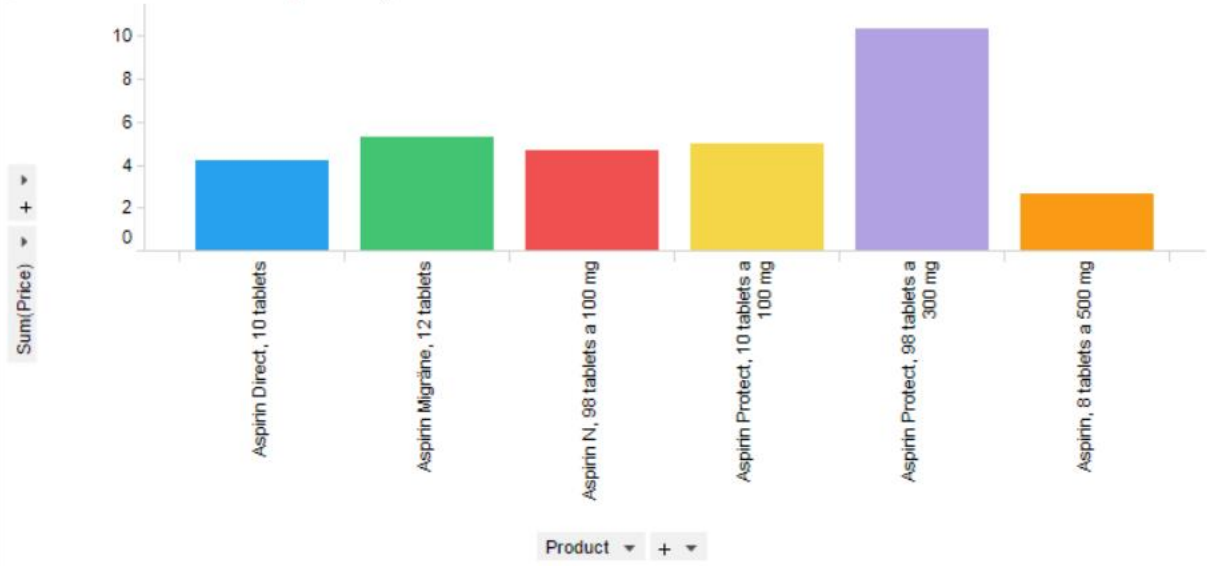
Number of brand names per country as size and producers as pie charts



Average Price of an Aspirin Tablet [Euro] per Country



Price Per Product at ABC pharmacy

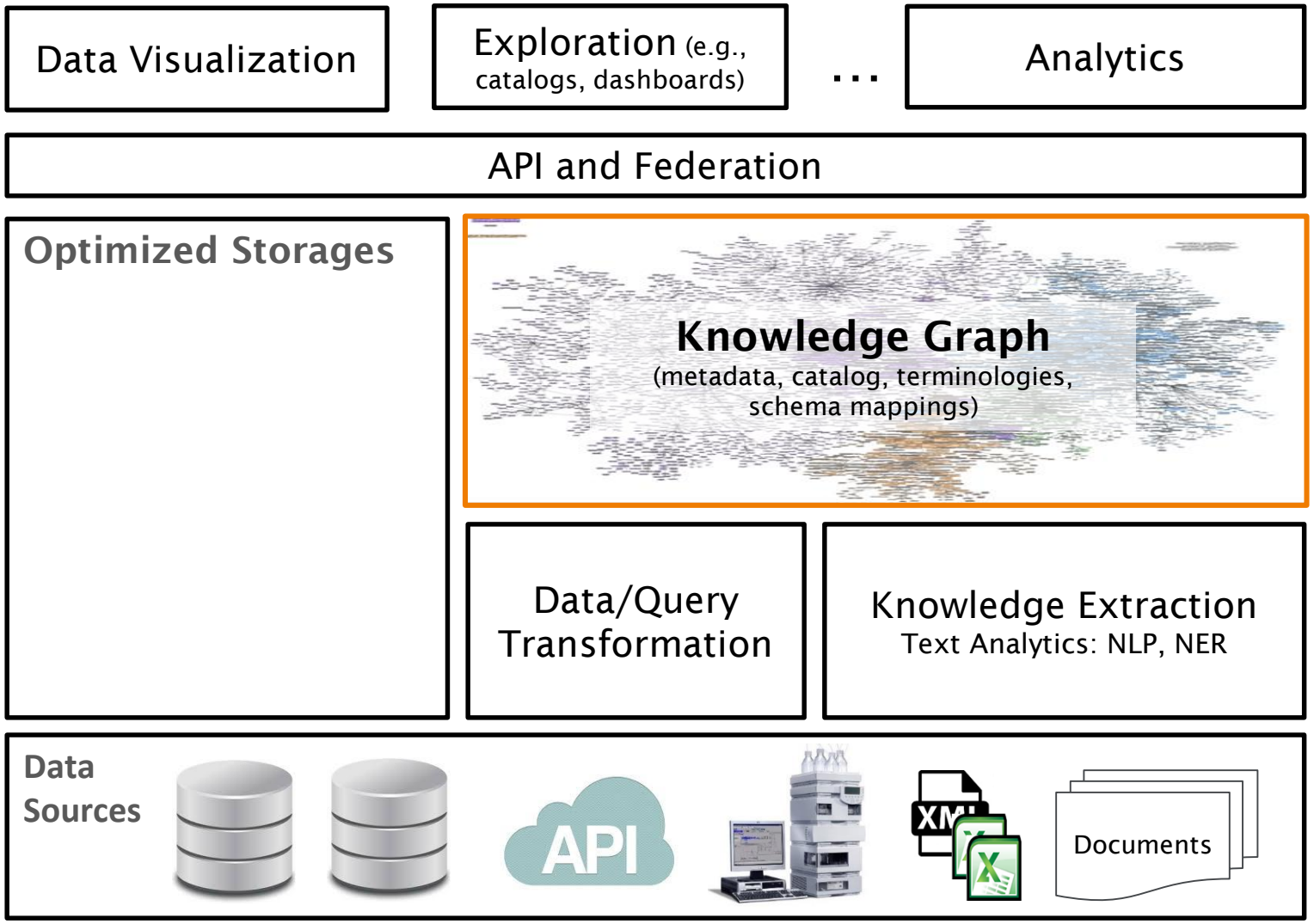


Online

6 of 6 rows | 0 marked | 5 columns | Aspirin_Preise - Sheet1

15:50
13.04.2018

Reference Architecture



OSTHUS

OSTHUS

Connecting data, people and organizations

Heiner Oberkampf, PhD
heiner.oberkampf@osthus.com