UNIVERSITÄT BONN

GEO-INFORMATION MANAGEMENT IN INTERDISCIPLINARY RESEARCH

LECTURE 7 – STANDARDISATION OF DATA

Prof. Dr. Juliane Fluck

Usage restriction:
Please use this presentation only in the context of this lecture.

SHORT REPETITION

- Challenges for re-use of data
- Research cycle
- Data management
- FAIR data principles
- OWL RDF
- Protege
- Graph databases

UNIVERSITÄT BONN

SHORT REPETITION

- Challenges for re-use of data
- Research cycle
- Data management
- FAIR data principles
- Protege and RDF

Today we adress the **I** of the FAIR Principles:

To be Interoperable:

- I1. (meta)data use a **formal, accessible, shared, and broadly applicable language for knowledge representation**.
- I2. **(meta)data use vocabularies that follow FAIR principles**

CONTENT TODAY

- What is Semantics?
- How can we make semantics explicit?
- How can we make computers to understand semantics?
- Where to search for relevant resources

WHAT IS SEMANTICS?

THE TREE
THE DOOR
THE HOUSE
SHIRT
PANTS
THE DOG
THE CAT
Larson

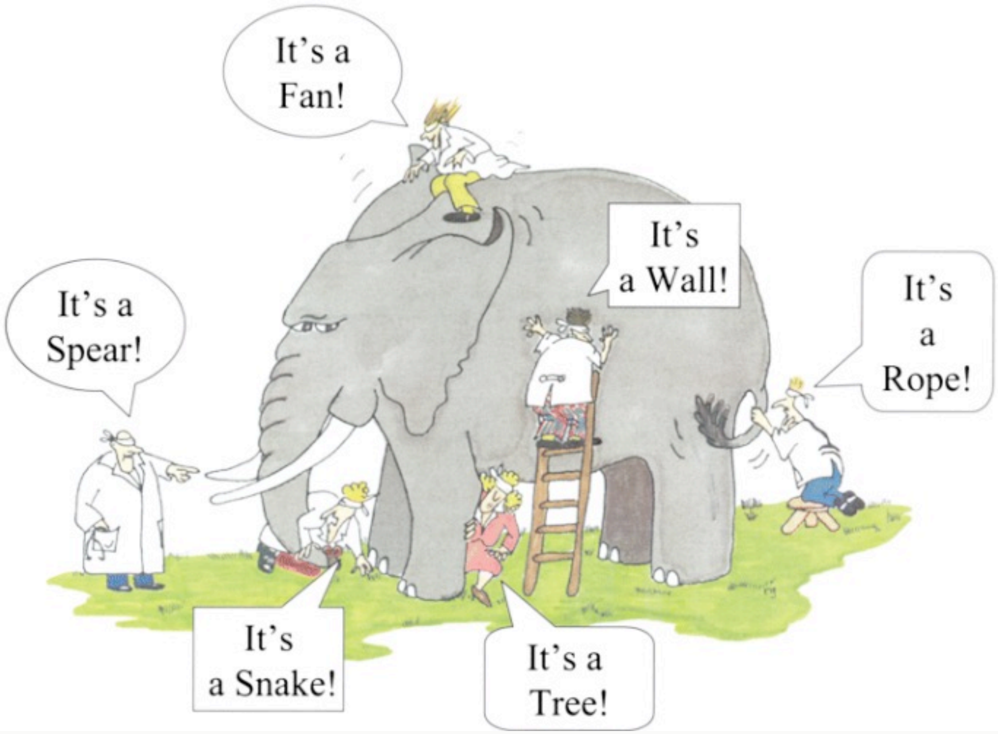
"Now! … *That* should clear up a few things around here!"

Semantics is (in linguistics) the study of meaning:

- ➢ how to give things a label,
- ➢ how to define them,
- ➢ attach attributes/features to them and
- ➢ how they are related to each other

Garry Larson The Far Side: Now! … That should clear up a few things around here!

WHAT IS SEMANTICS?



Well and ... how the labels depend or may depend on our perspective / view / understanding.

Pierre-Marc Daigneault The blind man and the elephant

DOI: 10.4256/mio.2013.015

WHAT IS THE DIFFERENCE?

- Word
- Term
- Synonym
- Acronym
- Concept

WHAT IS THE DIFFERENCE?

- Word – Usually regarded as the smallest isolable meaningful element of the language.
- Term - A word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject.

WHAT IS THE DIFFERENCE?

- Word – Usually regarded as the smallest isolable meaningful element of the language.
- Term - A word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject.

- Example:
- Crop, yield are words and terms
- Crop yield is a term

WHAT IS THE DIFFERENCE?

- Word – Usually regarded as the smallest isolable meaningful element of the language.
- Term - A word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject.
- Synonym
- Acronym
- Concept

UNIVERSITÄT BONN

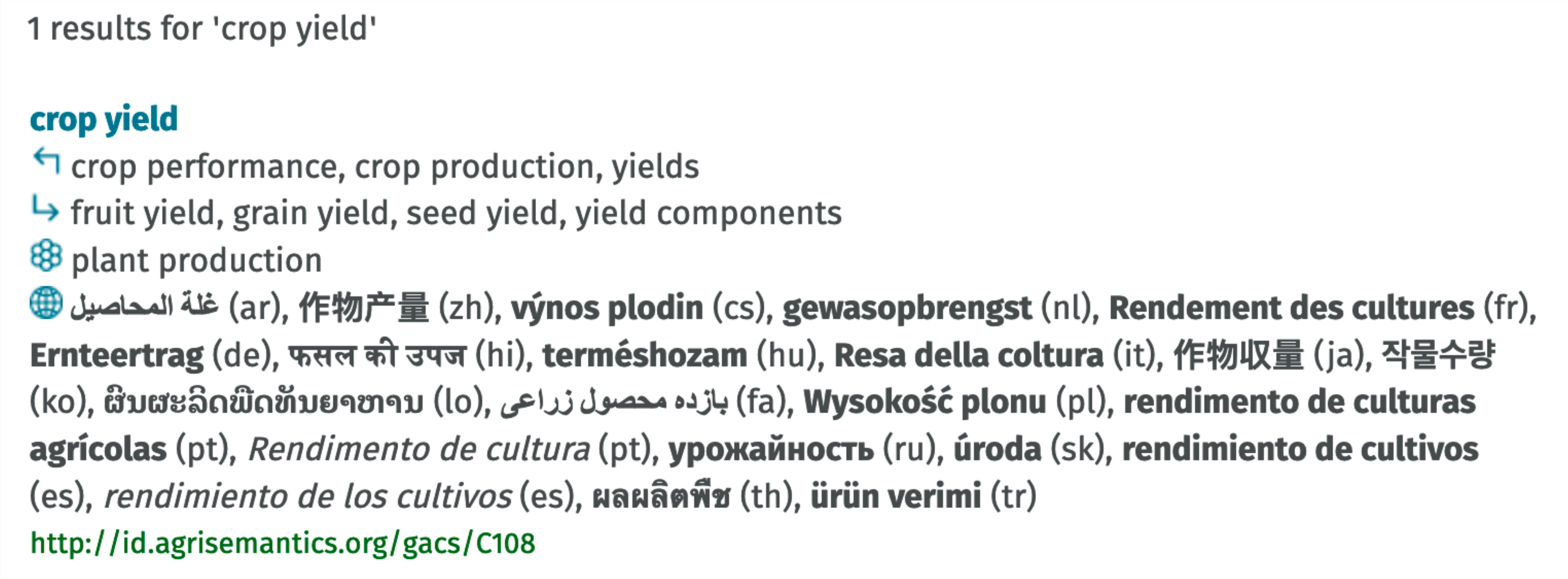
WHAT IS THE DIFFERENCE?

- Word – Usually regarded as the smallest isolable meaningful element of the language.
- Term – A word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject.
- Synonym – A word/term having the same or nearly the same meaning as another word or other words in a language.
- Acronym – Short form of a term often only understandable within the context or with the long form.
- Concept

WHAT IS THE DIFFERENCE?

- Word – Usually regarded as the smallest isolable meaningful element of the language.
- Term - A word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject.
- Synonym: Crop Performance
- Acronym: quantitative trait loci (QTL)
- Concept

GAGS TERMINOLOGY



1 results for 'crop yield'

crop yield

↰ crop performance, crop production, yields

↳ fruit yield, grain yield, seed yield, yield components

❁ plant production

🌐 **غلة المحاصيل** (ar), **作物产量** (zh), **výnos plodin** (cs), **gewasopbrengst** (nl), **Rendement des cultures** (fr), **Ernteertrag** (de), **फसल की उपज** (hi), **terméshozam** (hu), **Resa della coltura** (it), **作物収量** (ja), **작물수량** (ko), **ຜົນຜະລິດພືດທັນຍາຫານ** (lo), **بازده محصول زراعی** (fa), **Wysokość plonu** (pl), **rendimento de culturas agrícolas** (pt), *Rendimento de cultura* (pt), **урожайность** (ru), **úroda** (sk), **rendimiento de cultivos** (es), *rendimiento de los cultivos* (es), **ผลผลิตพืช** (th), **ürün verimi** (tr)

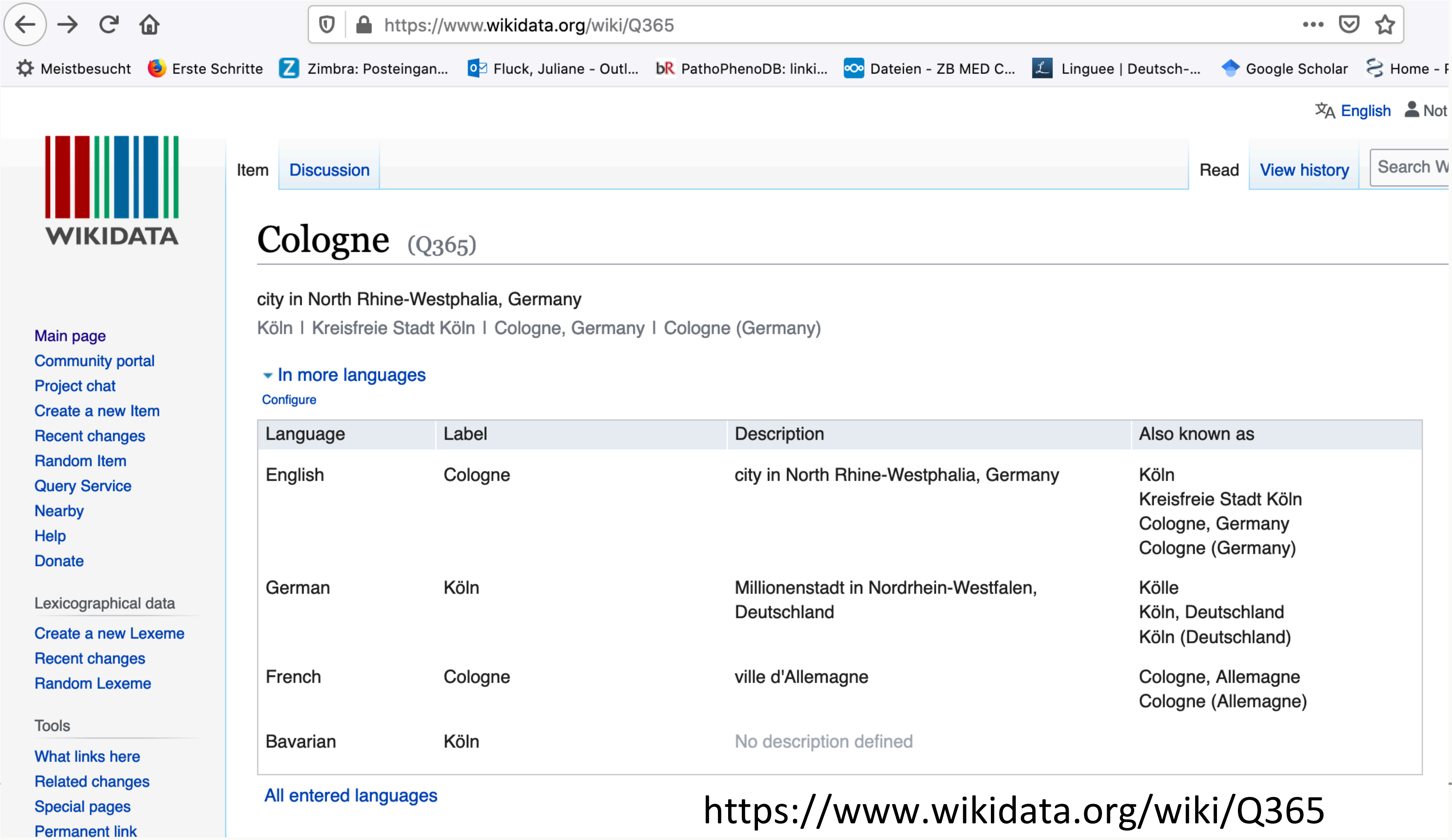
http://id.agrisemantics.org/gacs/C108

quantitative trait loci (QTL)

EXAMPLE CONCEPT

Wikidata:Cologne

Wikidata:Q365



Cologne (Q365)

city in North Rhine-Westphalia, Germany

Köln | Kreisfreie Stadt Köln | Cologne, Germany | Cologne (Germany)

In more languages

Configure

Language	Label	Description	Also known as
English	Cologne	city in North Rhine-Westphalia, Germany	Köln Kreisfreie Stadt Köln Cologne, Germany Cologne (Germany)
German	Köln	Millionenstadt in Nordrhein-Westfalen, Deutschland	Kölle Köln, Deutschland Köln (Deutschland)
French	Cologne	ville d'Allemagne	Cologne, Allemagne Cologne (Allemagne)
Bavarian	Köln	No description defined	

All entered languages

https://www.wikidata.org/wiki/Q365

WHAT IS THE DIFFERENCE?

- Word – Usually regarded as the smallest isolable meaningful element of the language.
- Term - A word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject.
- Synonym - A word/term having the same or nearly the same meaning as another word or other words in a language.
- Acronym – Short form of a term often only understandable within the context or with the long form.
- Concept - A concept is specified by its definitions, i.e. the semantics of the concept is defined in a textual description. For the computer: the concept is specified by a unique identifier.

WHAT IS THE DIFFERENCE?

- Terminology
- Catalogue
- Glossary
- Controlled vocabulary
- Taxonomy
- Thesaurus

TERM COLLECTIONS

- Terminology:

A system of terms belonging to specialised subject.

Examples:

Medical terminology

https://www.anerkennung-nrw.de/medizinische-terminologie-uebungen/

TERM COLLECTIONS

- Terminology:

A system of terms belonging to specialised subject.

- Catalogues:

A scattered lists of terms.

- Glossaries:

A scattered lists of terms plus glosses in natural language

TERM COLLECTIONS

- Terminology:

A system of terms belonging to specialised subject.

- Catalogues:

A scattered lists of terms.

- Glossaries:

A scattered lists of terms plus glosses in natural language

Example:

https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Essen_und_Trinken/Lebensmittel

TERM COLLECTIONS

- Controlled vocabulary (CV):

A set of terms authorized by a community established mandate. In theory, **the terms are defined excluding ambiguity** (possible use of preferred terms, synonyms).

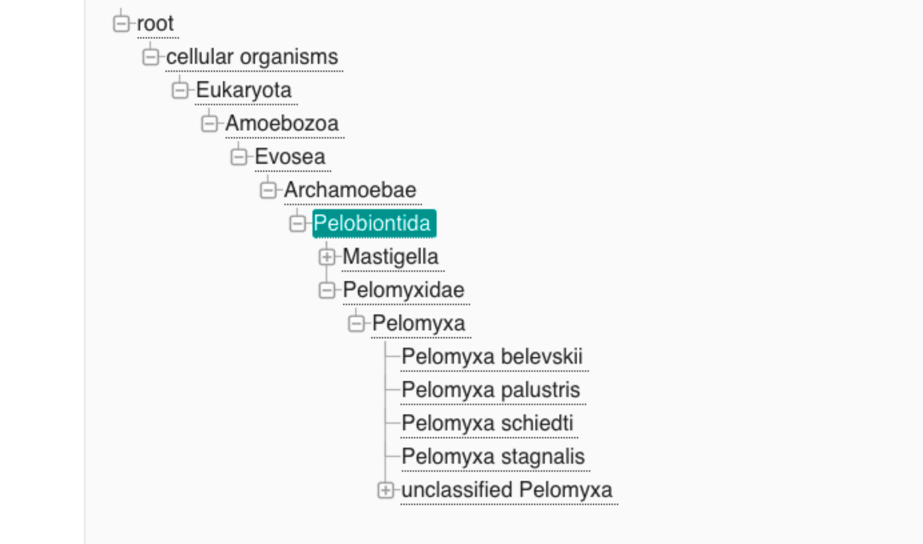
TERM COLLECTIONS

➢Taxonomies:

A CV organized into a hierarchical structure using the **basic parent-child relationship** (aka: whole-part, broader-narrower, genus-species, type-instance) and possibly others too.

Example: NCBI Taxonomy

https://www.ebi.ac.uk/ols/ontologies/ncbitaxon



```
⊟root
  ⊟cellular organisms
    ⊟Eukaryota
      ⊟Amoebozoa
        ⊟Evosea
          ⊟Archamoebae
            ⊟Pelobiontida
              ⊞Mastigella
              ⊟Pelomyxidae
                ⊟Pelomyxa
                  Pelomyxa belevskii
                  Pelomyxa palustris
                  Pelomyxa schiedti
                  Pelomyxa stagnalis
                  ⊞unclassified Pelomyxa
```

TERM COLLECTIONS

- Taxonomies:

A CV organized into a hierarchical structure using the basic parent-child relationship (aka: whole-part, broader-narrower, genus-species, type-instance) and possibly others too.

Thesauri:

Taxonomies enriched by relations for equivalence or association of a term (i.e. a term being “synonym of", ”related to", or “similar to" the preferred term), the most complex type of CV.

Example: MeSH Thesaurus:
https://meshb.nlm.nih.gov/record/ui?ui=D003922

TERM COLLECTIONS

- Ontology:

An ontology is a formal specification of domain knowledge and makes use of concepts, but also of formalisms to capture the semantics between concepts.

ONTOLOGY

An ontology

- is a knowledge model which defines a set of concepts and the relationship between those concepts within a specific domain
- supports automated reasoning and inference of data using logical rules
- provides knowledge sharing and reus among people or software agents

- **A simple tutorial on OWL Ontologies using Protege - Part 1**
- https://www.youtube.com/watch?v=t-Q0l4LwM2M

WHY DO WE NEED STANDARDISATION EFFORTS?

USAGE OF THESE RESOURCES

➢Lookup and Search

- Gives semantic information and inside about the term/concept
- Influence findability or resources when CV or concepts are used
- If source is enriched with synonyms it enhance search even when no CV or concept are used for data annoatation
- When hierarchies and relationships are included can be used to restricted or expand search

UNIVERSITÄT BONN

SEMANTIC SEARCH

Locate information by concept instead of using keyword or key phrase only.

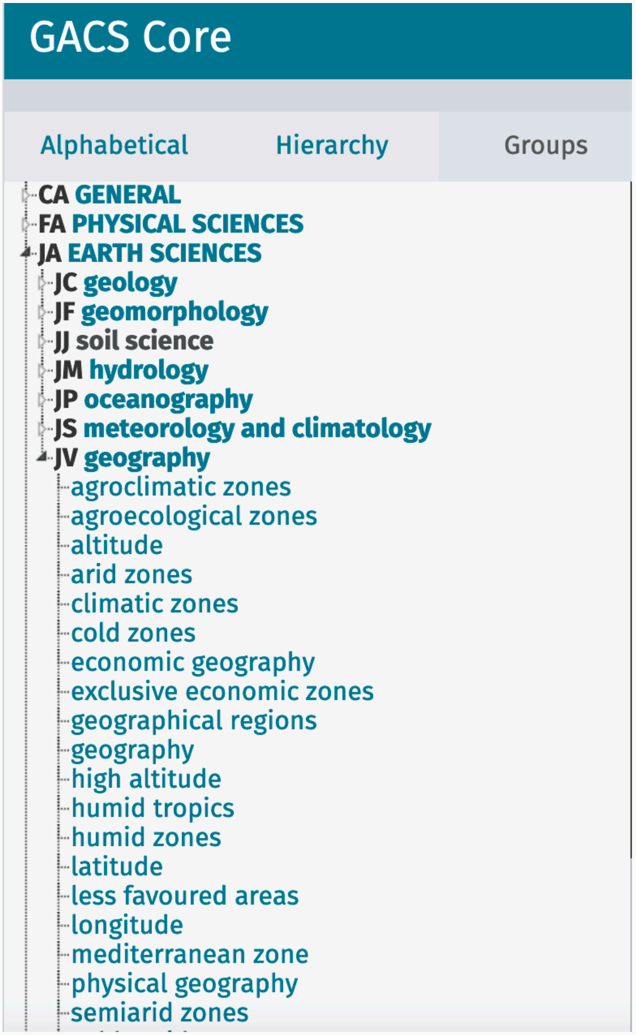
A simple example:

You are searching all data sets/literature about

Geography!

What do you need to get a rather complete set of data?

EXAMPLE OF THE USAGE OF A THESAURUS



GACS Core

Alphabetical | Hierarchy | Groups

- CA GENERAL
- FA PHYSICAL SCIENCES
- JA EARTH SCIENCES
  - JC geology
  - JF geomorphology
  - JJ soil science
  - JM hydrology
  - JP oceanography
  - JS meteorology and climatology
  - JV geography
    - agroclimatic zones
    - agroecological zones
    - altitude
    - arid zones
    - climatic zones
    - cold zones
    - economic geography
    - exclusive economic zones
    - geographical regions
    - geography
    - high altitude
    - humid tropics
    - humid zones
    - latitude
    - less favoured areas
    - longitude
    - mediterranean zone
    - physical geography
    - semiarid zones

USAGE OF THESE RESOURCES

➢(Meta-) Data Annotation

- ➢ For literature and research data
- ➢ Enhance Findability and Retrieval (can be used as index)
- ➢ Leads to interoperability
- ➢ Enable Semantics and lessens documentation needs
- ➢ Leads to the linkage of data to information and knowledge!

  (= linked data)

STANDARD IN LIFE SCIENCES

Example: Genes and Proteins

Databases are available for reference/normalisation/standardisation:

- ❖ Human Gene nomenclature commitee (HGNC) standardize human gene names

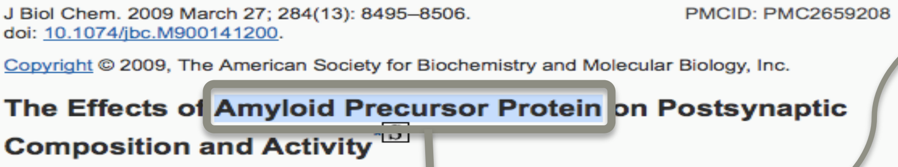
HGNC
HUGO Gene Nomenclature Committee

- ❖ Gene and Protein databases with standard identifiers

e!Ensembl NCBI UniProt

- ➢ All use HGNC name and HGNC IDs
- ➢ Linkout between different resources

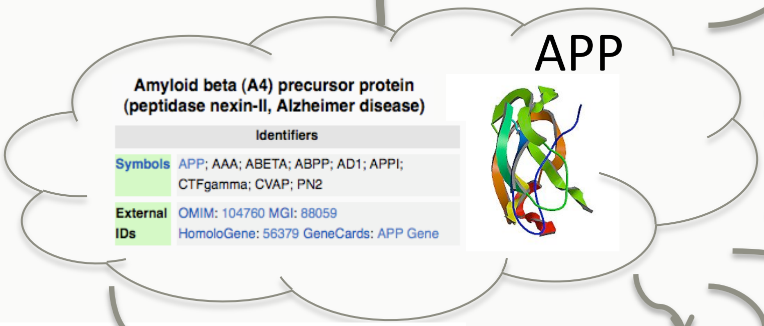
INTEROPERABILITY AND A LOT MORE CONNECTIONS



J Biol Chem. 2009 March 27; 284(13): 8495–8506. PMCID: PMC2659208
doi: 10.1074/jbc.M900141200.

Copyright © 2009, The American Society for Biochemistry and Molecular Biology, Inc.

The Effects of Amyloid Precursor Protein on Postsynaptic Composition and Activity



Amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)

Identifiers	
Symbols	APP; AAA; ABETA; ABPP; AD1; APPI; CTFgamma; CVAP; PN2
External IDs	OMIM: 104760 MGI: 88059 HomoloGene: 56379 GeneCards: APP Gene

News and Hypotheses

ALZHEIMER RESEARCH FORUM — NETWORKING FOR A CURE

Search | Get Newsletter — Search Papers — Search Index

Home · Contact Us · How to Cite · Get Newsletter · Become a Member · Login

WHAT'S NEW
PAPERS OF THE WEEK
Current Papers
ARF Recommends
Milestone Papers
Search All Papers
Search Comments
NEWS
Research News
Drug News
Conference News
RESEARCH
AD Hypotheses
AlzSWAN
Current Hypotheses
Hypothesis Factory
Forums
Live Discussions
Virtual Conferences
Interviews
Enabling Technologies
Workshops
Research Tools
Compendia
AlzGene
AlzRisk
Antibodies
Biomarkers
Mutations
Protocols
Research Models
Video Gallery
Resources
Bulletin Boards
Conference Calendar
Grants
Jobs
EARLY-ONSET FAMILIAL

RESEARCH NEWS

- Mitochondria Stumble Their Way Along Axons in ALS Model
  6 January 2012. People with amyotrophic lateral sclerosis struggle to move, and it appears some of their organelles have the same problem...
- Muscle-Derived Stem Cells Slow Aging in Progeria Mouse Model
  6 January 2012. New research links defective stem cells to shortened lifespan and aging-related pathology in a mouse model of premature aging...
- Mutations in Microglia Protein Cause Rare Dementia
  6 January 2012. An international consortium has found the gene responsible for hereditary diffuse leukoencephalopathy with spheroids (HDLS)...
- Hippocampal Memory Markers: From Anatomy to Molecules
  5 January 2012. Shrinkage of the hippocampus, which plays a key role in learning and memory, is one potential marker of Alzheimer's disease...

GO TO ALL RESEARCH NEWS

IN THE SPOTLIGHT

- Webinar Recording and Slides Are Live

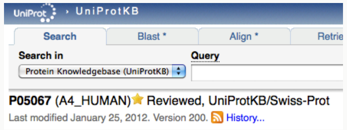
GM atrophy (ATRO) · WM atrophy (ATRO) · Hypometabolism (HYPO) · Amyloid deposition (AMYLO)

Different neuroimaging techniques reveal the spectrum of changes that occurs in the AD brain.

Print this page
Email this page
Alzforum News
Papers of the Week
Text size
Share & Bookmark

Most E-mailed | Most Commented
● Last 7 Days
○ Last 30 Days
- Book Review: *Memory Loss—A Practical Guide for Clinicians*
- Anti-Amyloid Treatment in Asymptomatic AD Trial
- Time Is Ripe for Clinical Trials in Frontotemporal Degeneration
- Top 10 Hits of 2011—Editors' Picks
- DC: Anti-Aβ Treatments Old and New

Forum Calendar
- See our recent Webinar: Adding Dynamics and Nuance to Alzheimer's Staging, held 5 January 2012.

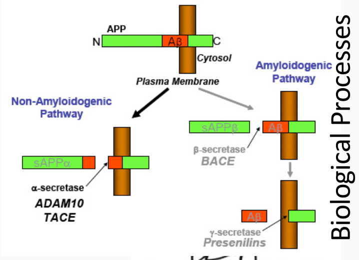


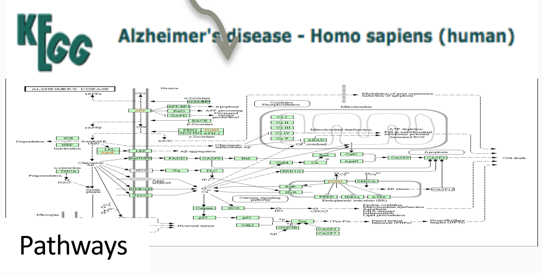
UniProt › UniProtKB

Search | Blast * | Align * | Retrie

Search in: Protein Knowledgebase (UniProtKB) — Query

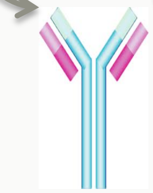
P05067 (A4_HUMAN) Reviewed, UniProtKB/Swiss-Prot
Last modified January 25, 2012. Version 200. History...



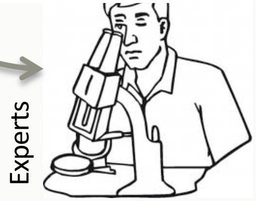


beta amyloid in alzheimer's

Videos/images



Antibodies



(Slide from Alan Rector)

INTRODUCTION BY BARRY SMITH

– Barry Smith is **the Ontology expert**

„ Since 2000 much of his research has been centered on the application of ontology in biomedical informatics, where he has worked on a variety of projects relating to biomedical terminologies and electronic health records. He is a founding Coordinating Editor of the OBO Foundry and has served as a member of the Scientific Advisory Board of the Gene Ontology Consortium, and of the Ontology for Biomedical Investigations (OBI). He contributes to the development of a number of biological and biomedical ontologies, including the Protein Ontology, the Plant Ontology, and others.

Source: https://en.wikipedia.org/wiki/Barry_Smith_(ontologist), accessed 2020-11-18

https://www.youtube.com/watch?v=bj8mSbHh-qA

Start at 4:20 -20.52

INTRODUCTION TO ONTOLOGY

- **Ontology =** a representation of types of entities an a given domain and of the relations between them

- **What is an ontology for?**

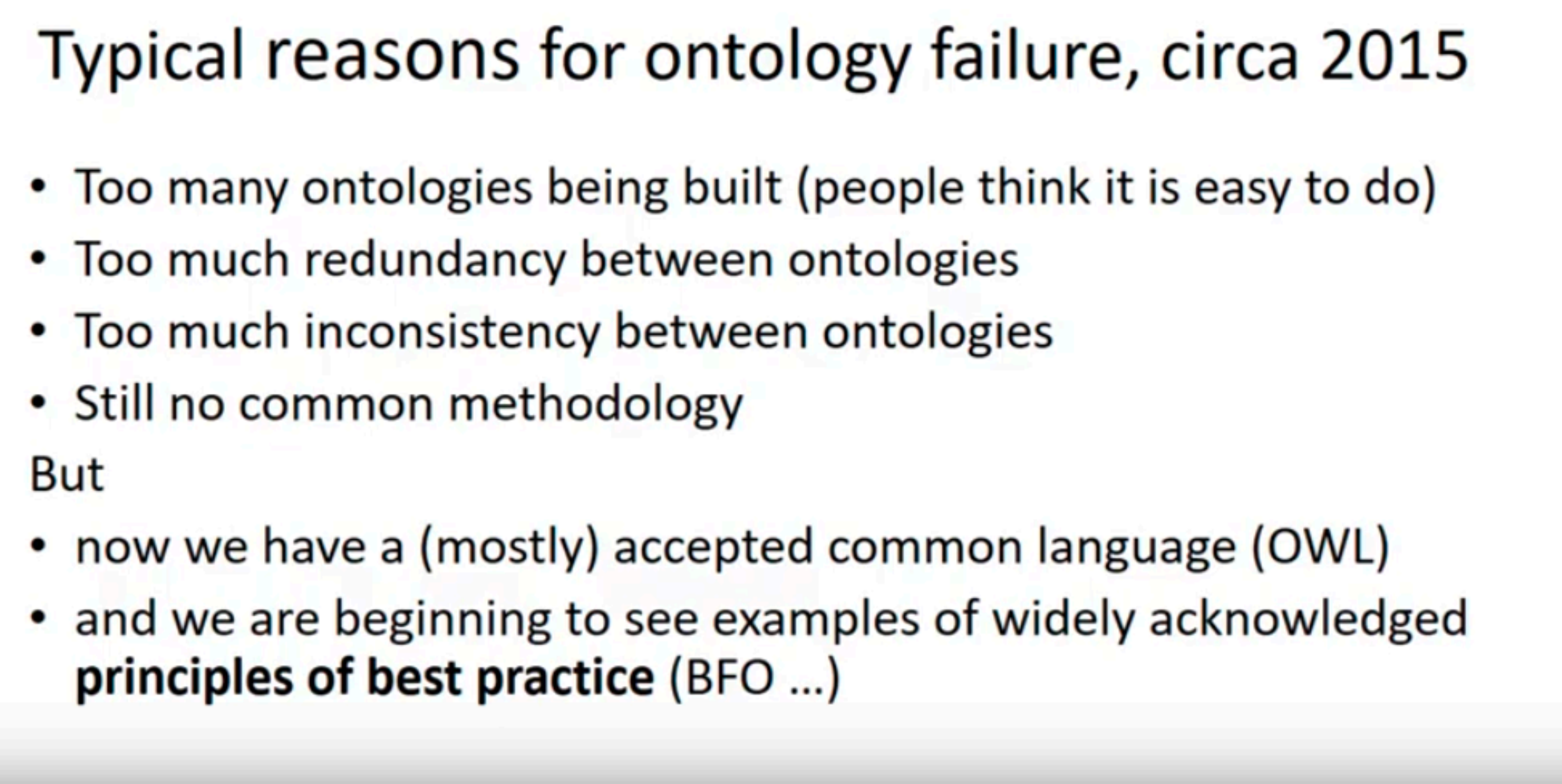
  To promote interoperability across heterogeneous data systems

- **How?**

  By exploiting **relative stability and ubiquity of natural language** vs changeability of computer hardware and *ad hockery* of data engineering software

- **Ontologists work only** when aggressively used by influential constituencies

Source: Presentation Barry Smith: https://www.youtube.com/watch?v=bj8mSbHh-qA



Typical reasons for ontology failure, circa 2015

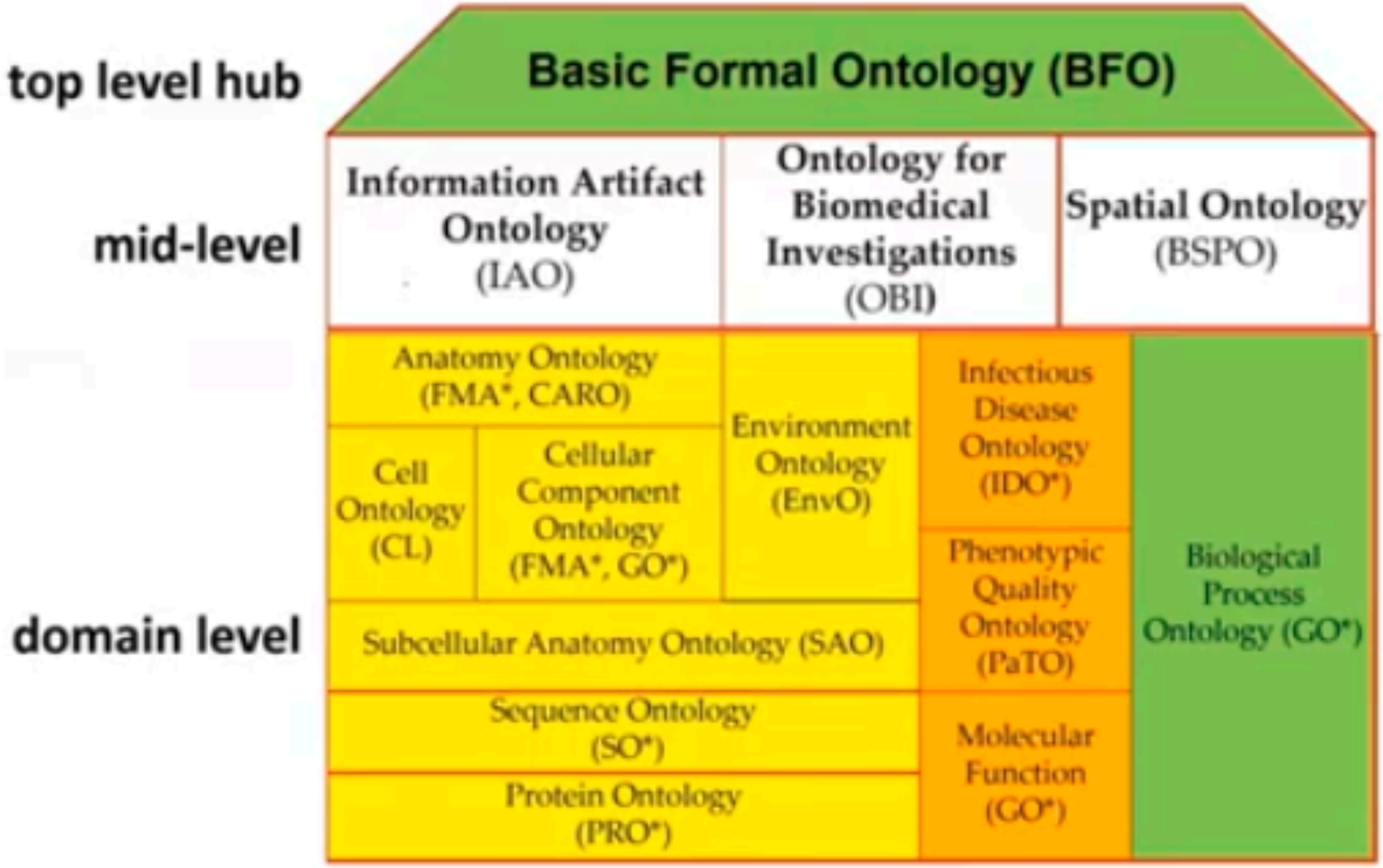
- Too many ontologies being built (people think it is easy to do)
- Too much redundancy between ontologies
- Too much inconsistency between ontologies
- Still no common methodology

But

- now we have a (mostly) accepted common language (OWL)
- and we are beginning to see examples of widely acknowledged **principles of best practice** (BFO ...)

Source: Presentation Barry Smith: https://www.youtube.com/watch?v=bj8mSbHh-qA

THE OPEN BIOMEDICAL ONTOLOGIES (OBO) FOUNDRY



top level hub

mid-level

domain level

Basic Formal Ontology (BFO)

Information Artifact Ontology (IAO)

Ontology for Biomedical Investigations (OBI)

Spatial Ontology (BSPO)

Anatomy Ontology (FMA*, CARO)

Cell Ontology (CL)

Cellular Component Ontology (FMA*, GO*)

Environment Ontology (EnvO)

Infectious Disease Ontology (IDO*)

Phenotypic Quality Ontology (PaTO)

Biological Process Ontology (GO*)

Subcellular Anatomy Ontology (SAO)

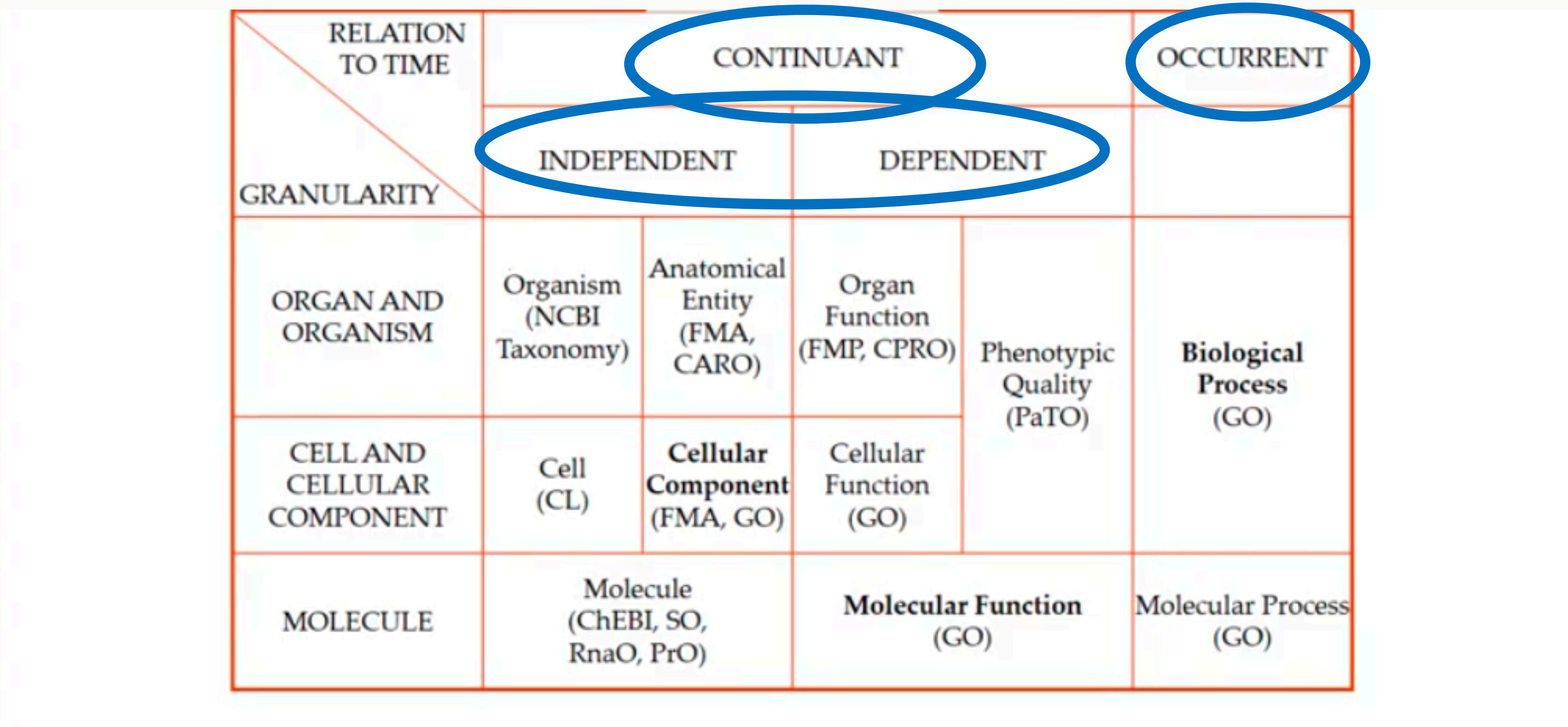
Sequence Ontology (SO*)

Protein Ontology (PRO*)

Molecular Function (GO*)

Source: Presentation Barry Smith: https://www.youtube.com/watch?v=bj8mSbHh-qA

THE OPEN BIOMEDICAL ONTOLOGIES (OBO) FOUNDRY



RELATION TO TIME / GRANULARITY	CONTINUANT				OCCURRENT
	INDEPENDENT		DEPENDENT		
ORGAN AND ORGANISM	Organism (NCBI Taxonomy)	Anatomical Entity (FMA, CARO)	Organ Function (FMP, CPRO)	Phenotypic Quality (PaTO)	**Biological Process** (GO)
CELL AND CELLULAR COMPONENT	Cell (CL)	**Cellular Component** (FMA, GO)	Cellular Function (GO)		
MOLECULE	Molecule (ChEBI, SO, RnaO, PrO)		**Molecular Function** (GO)		Molecular Process (GO)

Source: Presentation Barry Smith: https://www.youtube.com/watch?v=bj8mSbHh-qA

EXAMPLE ONTOLOGY

The FoodOn Food Ontology (FOODON)

FoodOn is an ontology built to represent entities which bear a “food role” and is initially focused on categorizing and processing of food for humans. We aim to develop semantics for food safety, food security, the agricultural and animal husbandry practices linked to food production, culinary, nutritional and chemical ingredients and processes. FoodOn belongs to the OBOFoundry.org family of ontologies.

https://bioportal.bioontology.org/ontologies/FOODON

EXPLANATION CONTINUANTS AND OCCURENCES

- Entities are either **continuants** or **occurrents**.
- A **continuant** is something *existing at an instant in time*, such as
  - a *person*,
  - a *country*,
  - a *smile*,
  - the *smell of a flower*, or
  - an *email*.

Continuants maintain their identity though time.

EXPLANATION CONTINUANTS AND OCCURENCES

- An **occurrent** is something that has *temporal parts* such as
  - a *life*,
  - *smiling*,
  - the *opening of a flower*, and
  - *sending an email*.

Occurrents can be

- *processes that last through time and*
- events that occur at an instant in time

EXPLANATION CONTINUANTS AND OCCURENCES

Continuants participate in occurrents.

One way to think about the difference is to consider the entity's parts:

a finger is part of a person, but is not part of a life;

infancy is part of a life, but is not part of a person.

EXPLANATION CONTINUANTS

A continuant is

- an **independent continuant**,
- a **dependent continuant** or
- a **spatial region**.

EXPLANATION CONTINUANTS

- An **independent continuant** is an entity that can exist by itself or is part of another entity. For example, a person, a face, a pen, the surface of an apple, the equator, a country, and the atmosphere are independent continuants.

- A **dependent continuant** only exists by virtue of another entity and is not a part of that entity. For example, a smile, the smell of a flower, or the ability to laugh can only exist in relation to another object.

- A **spatial region** is a region in space, for example, the space occupied by a doughnut now, the boundary of a county, or the point in a landscape that has the best view.

WHERE TO SEARCH FOR STANDARDS AND DATABASES?

Overall place to search for FAIR standards and data:

https://fairsharing.org/

Registry of Research Data Repositories

https://www.re3data.org/

WHERE TO SEARCH FOR STANDARDS AND DATABASES?

Overall place to search for FAIR standards and data:

https://fairsharing.org/

Registry of Research Data Repositories

https://www.re3data.org/

Bio related (but with broad coveradge)

http://www.obofoundry.org/

http://bioportal.bioontology.org/ontologies

https://www.ebi.ac.uk/ols/ontologies

WHERE TO SEARCH FOR STANDARDS AND DATABASES?

Geo and Agriculture related

https://inspire.ec.europa.eu/

http://agroportal.lirmm.fr/

https://agrisemantics.org/

https://www.eionet.europa.eu/gemet/en/themes/

https://www.bonares.de/

https://gardian.bigdata.cgiar.org/exploration.php#!/

WHERE TO SEARCH FOR STANDARDS AND DATABASES?

Please use the last thirty minutes to look up the relevant standards for

(1) Long term field experiments

(2) Sensor data

(3) Spatiotemporal Data

Furthermore, as recommended, here again the link to the youtube recording of a presentation of Barry Smith

https://www.youtube.com/watch?v=bj8mSbHh-qA

Start at 4:20 -20.52