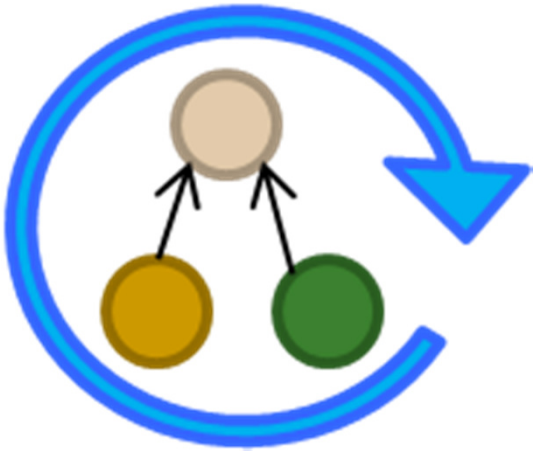
Ontologie-Management

Kapitel 3: Anwendungen

Dr. Michael Hartung

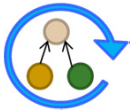
Wintersemester 2012/13

Universität Leipzig
Institut für Informatik
http://dbs.uni-leipzig.de

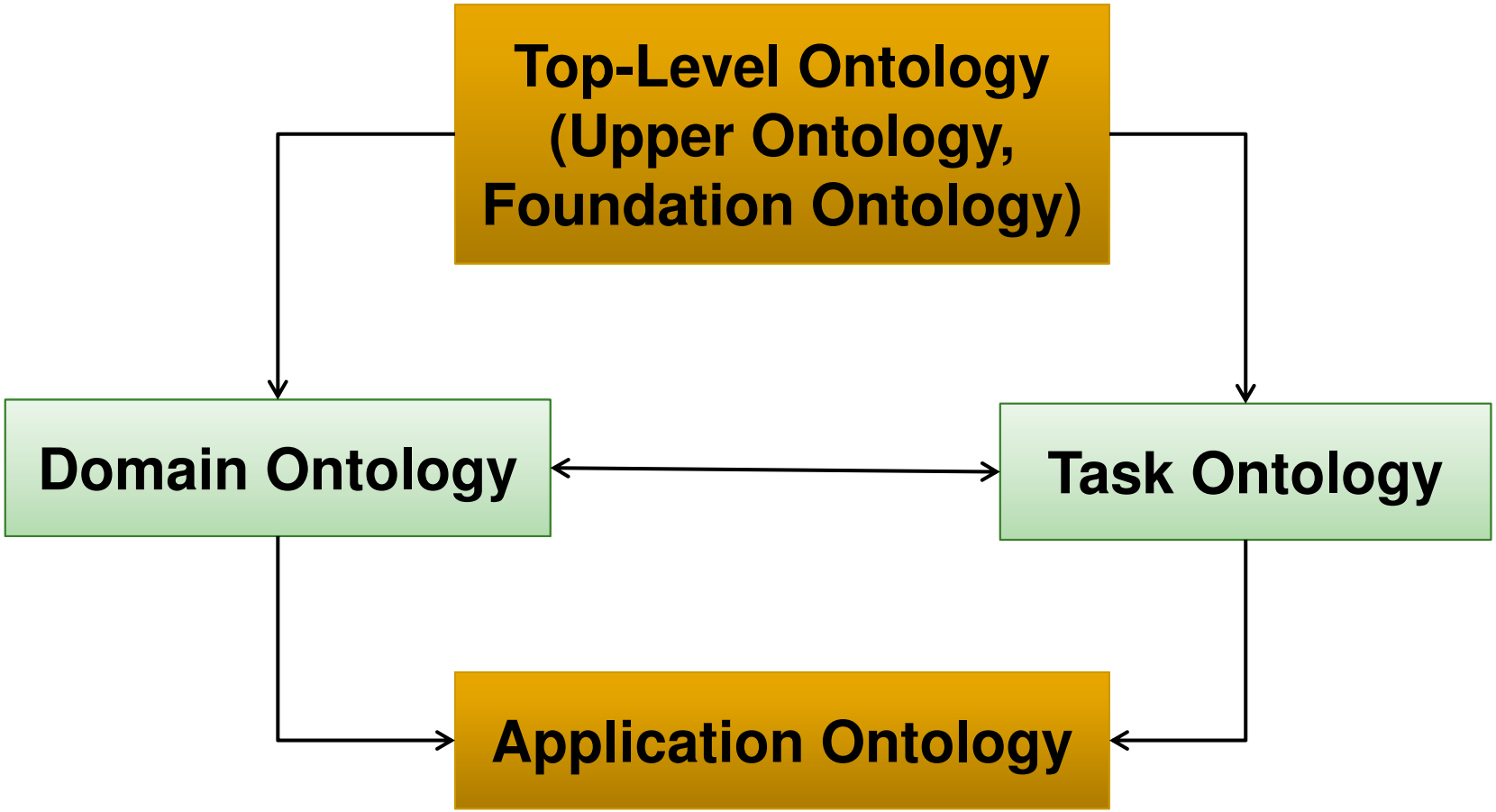


Inhalt

- Beispiele von Realwelt-Ontologien
- Annotation
- Semantische Suche und Navigation
- Linked Data



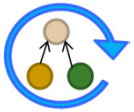
Klassifikation von Ontologien



Klassifikation nach Guarino, 1998

Ontologiebeispiele im Überblick

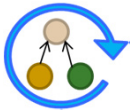
- **Top-Level Ontologien**
  - Basic Formal Ontology (BFO), General Formal Ontology (GFO)
  - Standard Upper Merged Ontology (SUMO)
  - OpenCyc
  - WordNet
  - ...
- **Domänenontologien (Lebenswissenschaften)**
  - Gene Ontology (GO)
  - National Cancer Institute Thesaurus (NCIT)
  - ...



BFO / GFO

- Basic Formal Ontology (BFO)
  - Unterstützung der Informationsintegration bei wissenschaftlichen Arbeiten
  - Hilfreich zur Entwicklung von Domänenontologien → Sicherung von Konsistenz sowie Vermeidung von Redundanzen
  - 39 Klassen: http://purl.bioontology.org/ontology/BFO
- General Formal Ontology (GFO)
  - Integration von Objekten und Prozessen
  - Informationen zu Zeit, Raum, Relationen, Rollen, ...
  - 45 Klassen: http://purl.bioontology.org/ontology/GFO

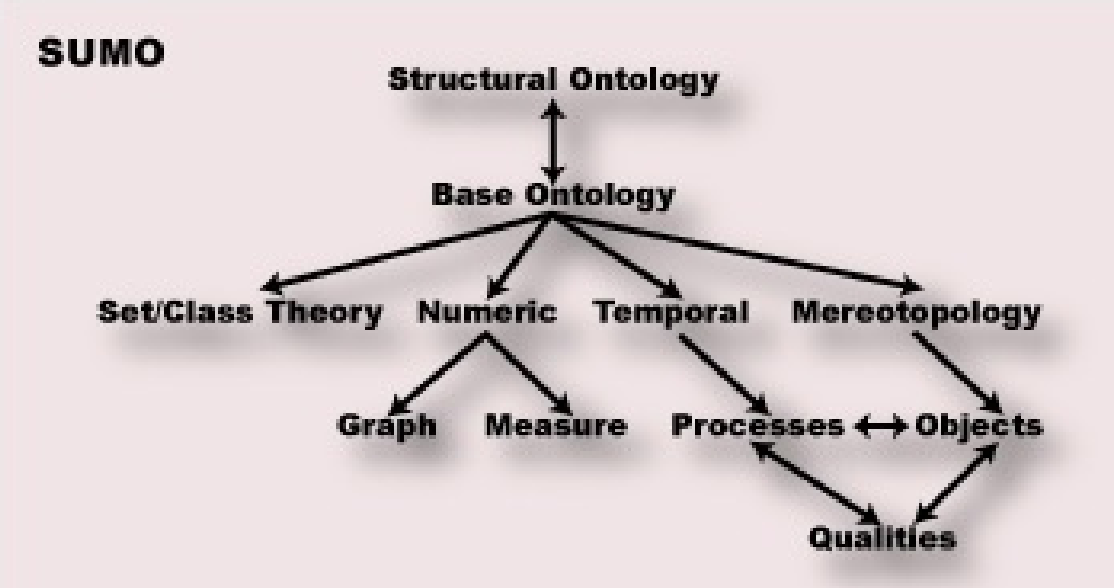
BFO: http://www.ifomis.org/bfo *GFO: http://www.onto-med.de/ontologies/gfo/*



SUMO

- Standard Upper Merged Ontology
  - Drei Teile
    - Upper Structural Ontology
    - Mid-level Base Ontology
    - Domain Ontologies: Wirtschaft, Transport, Geographie, ...
  - Ursprünglich nur generelle Konzepte → Erweiterung im Laufe der Zeit

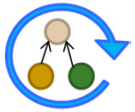
SUMO:
http://www.ontologyportal.org/



OpenCyc / Cyc

- Cyc
  - Von „encyclopedia“
  - Aufbau einer umfassenden Ontologie und Wissensbasis für Allgemeinwissen (seit 1984)
  - Teile frei verfügbar → OpenCyc, ResearchCyc
  - CycL sentences zur Wissensrepräsentation
    - Unterscheidung von Individuen und Kollektionen
    - (#$isa #$BillClinton #$UnitedStatesPresident)
    - (#$genls #$Tree-ThePlant #$Plant)
    - (#$capitalCity #$France #$Paris)

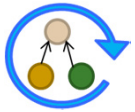
Cyc: http://www.cyc.com/cyc *OpenCyc: http://opencyc.org/*



WordNet

- Gruppierung englischsprachiger Wörter in Synonymmengen (synsets)
  - Kurze, allgemeine Definitionen
  - Semantische Relationen zwischen Synsets
  - Unterscheidung von Substantiven (N), Verben (V), Adjektiven (Adj) und Adverbien (Av)
- 117.659 Synsets in Version 3.0
- Relationen
  - Hypernym/hyponym (Ober- / Unterbegriffe)
  - Meronym/holonym (Teil-Ganzes)
  - Antonym (Gegensatzwort)
  - ...

WordNet: http://wordnet.princeton.edu/



WordNet

WordNet Search - 3.1
- WordNet home page - Glossa

Word to search for: mouse

Display Options: (Select option to change

Key: "S:" = Show Synset (semantic) rel

Display options for sense: (gloss) "an e

Noun

- S: (n) **mouse** (any of numerous s having pointed snouts and small tails)
- S: (n) shiner, black eye, **mouse** (
- S: (n) **mouse** (person who is quie
- S: (n) **mouse**, computer mouse (coordinates of a cursor on your c the bottom of the device is a ball *much more room than a trackba*

Verb

- S: (v) sneak, **mouse**, creep, puss *around spying on the neighbor's*
- S: (v) **mouse** (manipulate the mo

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: mouse Search WordNet

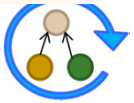
Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

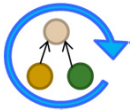
- S: (n) **mouse** (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
  - ***direct hyponym*** / *full hyponym*
    - S: (n) house mouse, Mus musculus (brownish-grey Old World mouse now a common household pest worldwide)
    - S: (n) harvest mouse, Micromyx minutus (small reddish-brown Eurasian mouse inhabiting e.g. cornfields)
    - S: (n) field mouse, fieldmouse (any nocturnal Old World mouse of the genus Apodemus inhabiting woods and fields and gardens)
    - S: (n) nude mouse (a mouse with a genetic defect that prevents them from growing hair and also prevents them from immunologically rejecting human cells and tissues; widely used in preclinical trials)
    - S: (n) wood mouse (any of various New World woodland mice)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*



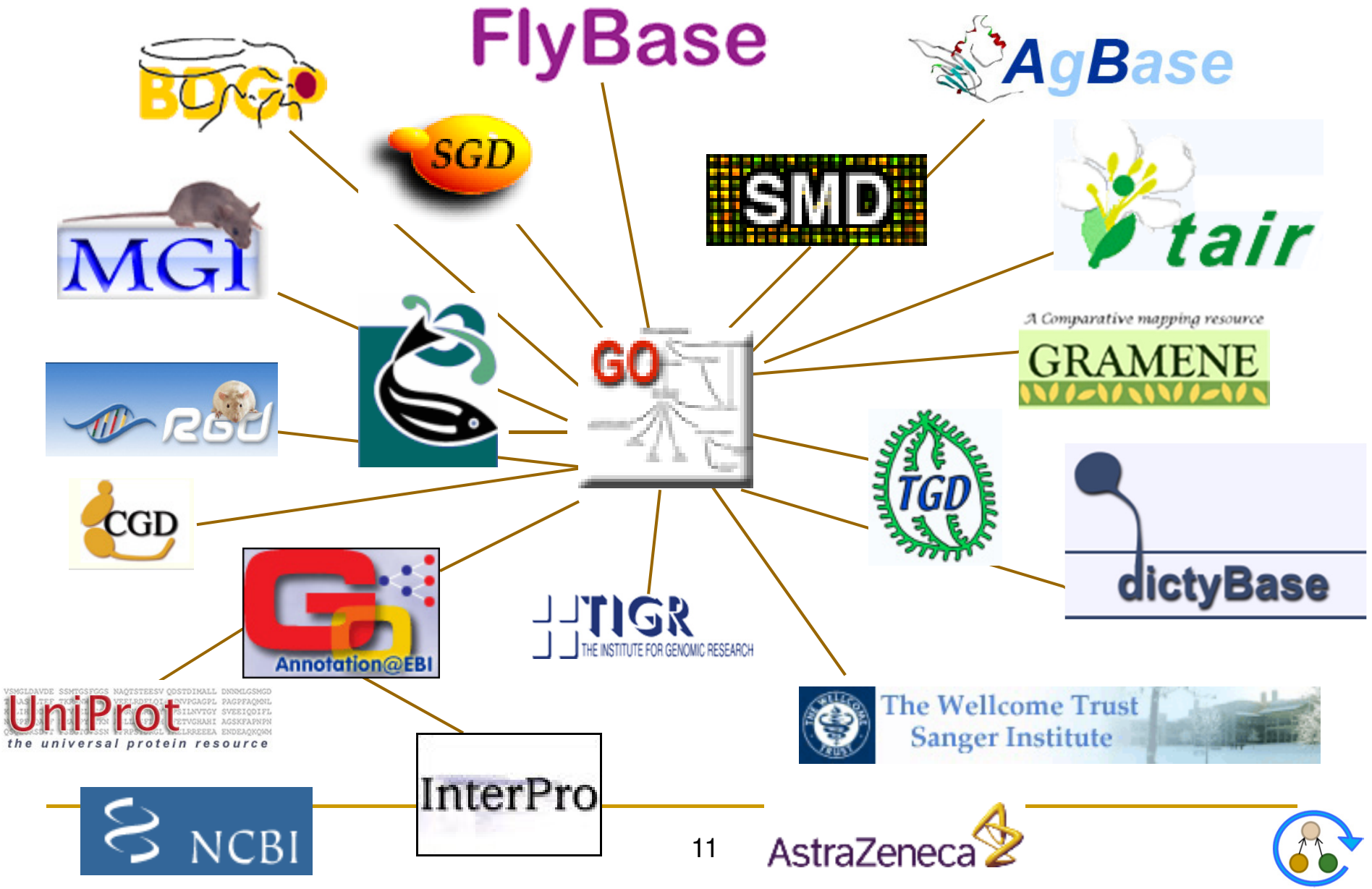
Gene Ontology (GO)

- Gene Ontology Consortium
  - Entwicklung einer gemeinsamen Sprache zur Annotation von molekularen Charakteristiken in Organismen → Wissensaustausch
  - Teilnehmer stellen GO-basierte Annotationen zur Verfügung → Möglichkeit neuartiger Datenanalysen und Zusammenarbeit
- Projektziele:
  - Management / Entwicklung der eigentlichen Ontologie
  - Annotation von Genen / Genprodukten
  - Tools / Werkzeuge für den Umgang mit GO

Gene Ontology: http://www.geneontology.org/

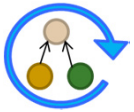


Verbreitung von GO

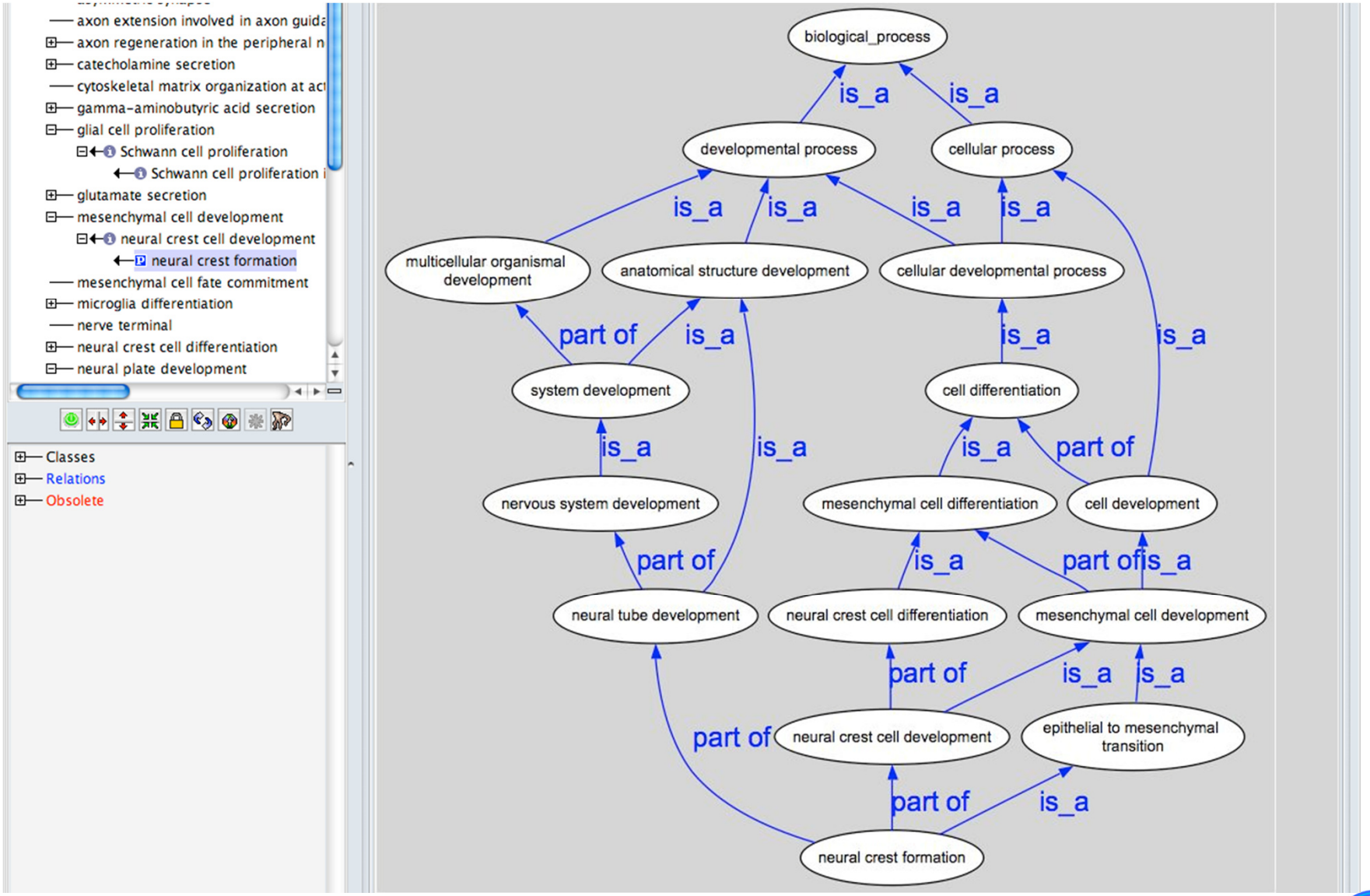


Gene Ontology (GO)

- Einteilung in drei Subdomänen
  - *Molecular Function* (ca. 10.000 Konzepte)
    - Elementare Aktivitäten / Aufgaben, durchgeführt von individuellen Genprodukten
    - Bsp.: carbohydrate binding, ATPase activity
  - *Biological Process* (ca. 20.000 Konzepte)
    - Biologisches Ziel realisiert durch eine Menge molekularer Funktionen
    - Bsp.: mitosis, purine metabolism
  - *Cellular Component* (ca. 3.000 Konzepte)
    - Ort oder Komplex (subzellulare Strukturen)
    - Bsp.: nucleus, telomere, RNA polymerase II holoenzyme



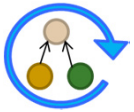
Konzepte in GO



National Cancer Institute Thesaurus (NCIT)

- NCI Thesaurus
  - Referenzterminologie des NCI
  - Krebsbezogene Themen für kollaborative Forschung
  - Ca. 90.000 Konzepte organisiert in 20 Hauptkategorien
    - Anatomie
    - Genprodukte
    - Krankheiten
    - Medikamente
    - ...
  - Konzeptinformationen
    - Bevorzugter Name, Synonyme, Definitionen, ...

NCI Thesaurus: http://ncit.nci.nih.gov/ncitbrowser/



Annotation – Motivation

Informationsflut in den Lebenswissenschaften

- Zehntausende Gene und gen-regulatorische Elemente (in versch. Spezies)
- Zahlreiche Interaktionen in komplexen molekularen Netzwerken
- Millionen von Publikationen
- ...

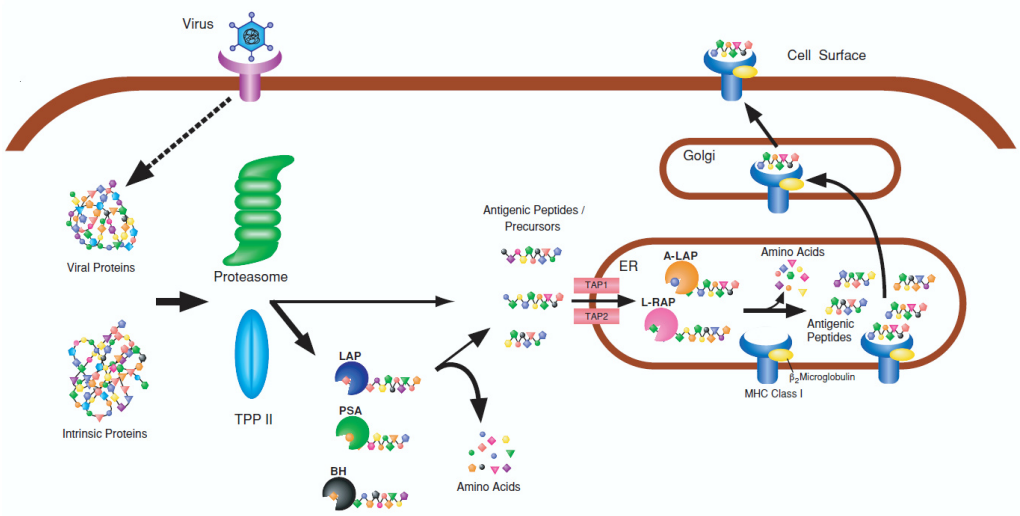
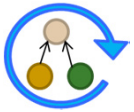


Fig. 1. Processing Pathway of Antigenic Peptides Presented to MHC Class I Molecules

Hattori A, Tsujimoto M.: Processing of antigenic peptides by aminopeptidases. Biol Pharm Bull., 2004

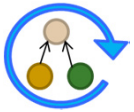
- Zuordnung
  - der Funktionen, Prozesse, Interaktionen zu biologischen Objekten
  - der enthaltenen Themen zu Publikationen
  - ...

→ Nutzen maschinenverstehbarer Annotationen

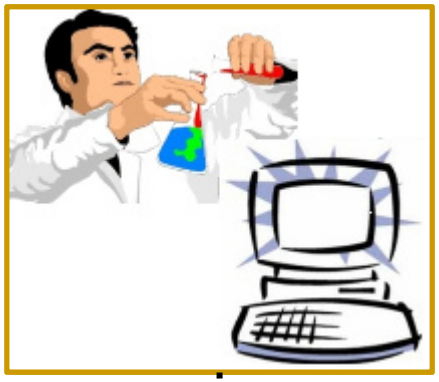


Begriff Annotation

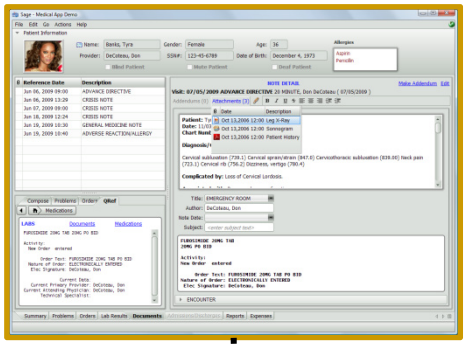
- Annotationen dienen der semantischen Beschreibung der Eigenschaften von Objekten
  - Erfassung von Metadaten
  - Möglichst einheitlicher Gebrauch der domänenspezifischen Begriffe
  - Assoziationen eines zu beschreibenden Objekts zu den Begriffen eines Vokabulars bzw. den Konzepten einer Ontologie
    - Genprodukte, Gene in Lebenswissenschaften mit zugeordneten Funktionen
    - Produkte in Online-Shops annotiert mit ihren Eigenschaften



Arten von Annotationen in den Lebenswissenschaften



- Biologische Objekte (Gene, Proteine, …)
- Pathways, Netzwerke
- …



Experiment-Annotation
MIAME, MGED…

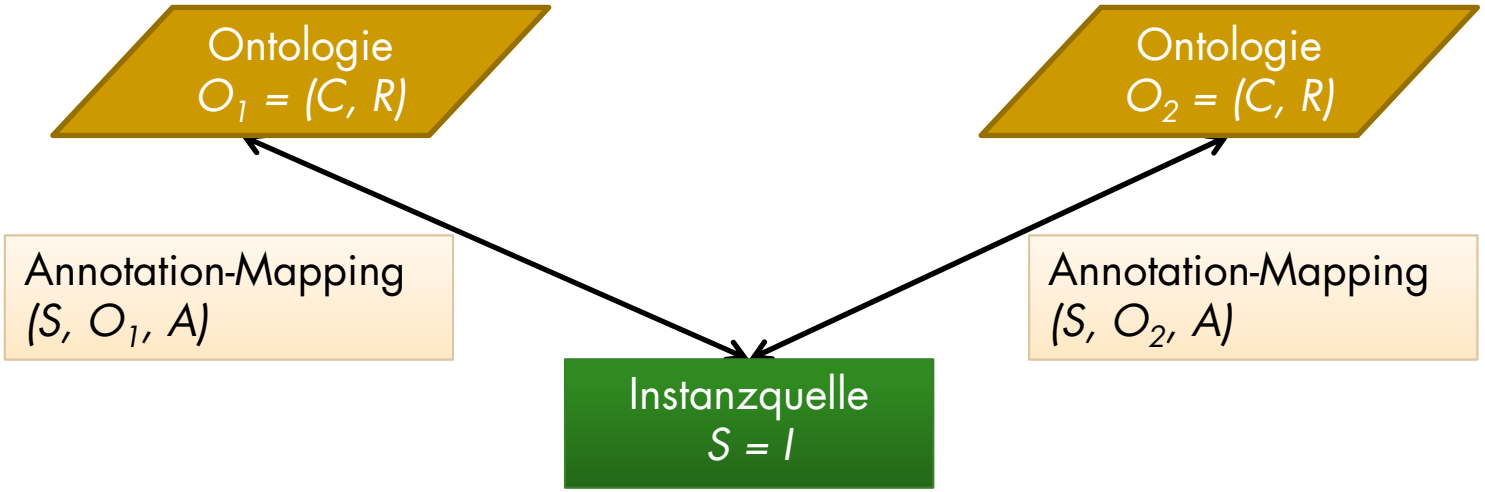
Annotation biologischer Objekte / Pathways
GO, KEGG, …

Annotation von Publikationen (indexing)
MeSH, GO,…

Annotation von klinischen Dokumenten
ICD, SNOMED, LOINC,…

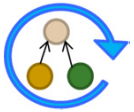
MIAME	*Minimum Information About a Microarray Experiment*
MGED	*Microarray Gene Expression Data*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
MeSH	*Medical Subject Headings*
ICD	*International Classification of Diseases*
SNOMED	*Systematized Nomenclature of Medicine*
LOINC	*Logical Observation Identifiers Names and Codes*
NCIT	*NCI (National Cancer Institute) Thesaurus*

Allgemeines Annotationsmodell



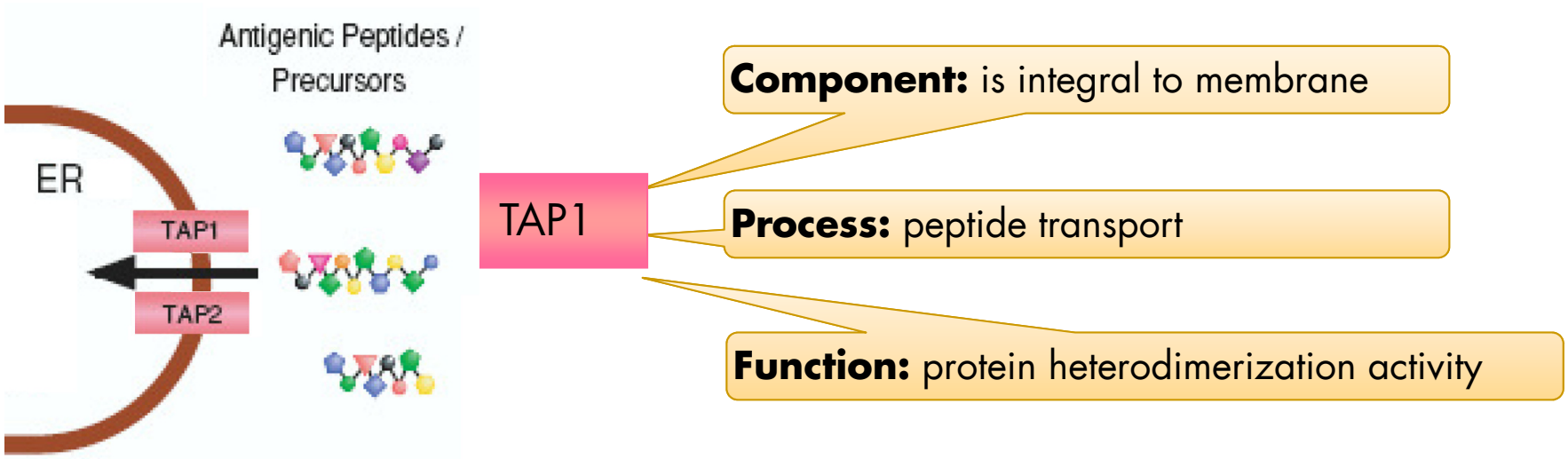
C Menge von Konzepten
R Menge von Relationen
I Menge von Instanzen/Objekten
A Menge von Annotationen / Korrepondenzen

Annotation $a=(i,c)$, $a \in A$, $i \in I$, $c \in C$
object_id - concept_id



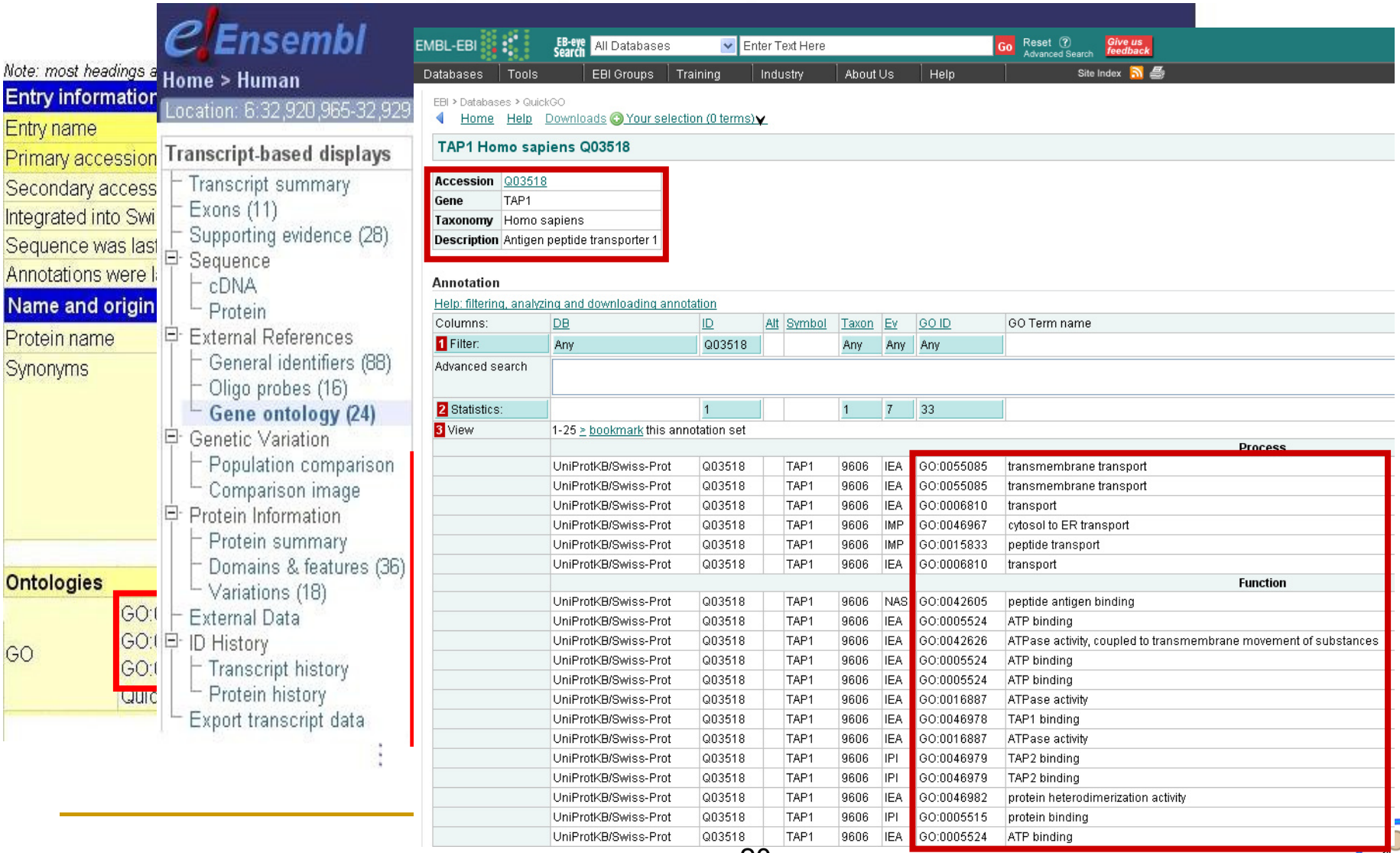
Beispiel: GO Annotationen

- Assoziation eines Proteins zu einem/r bestimmten biologischen Prozess / zellulären Komponente / molekularen Funktion

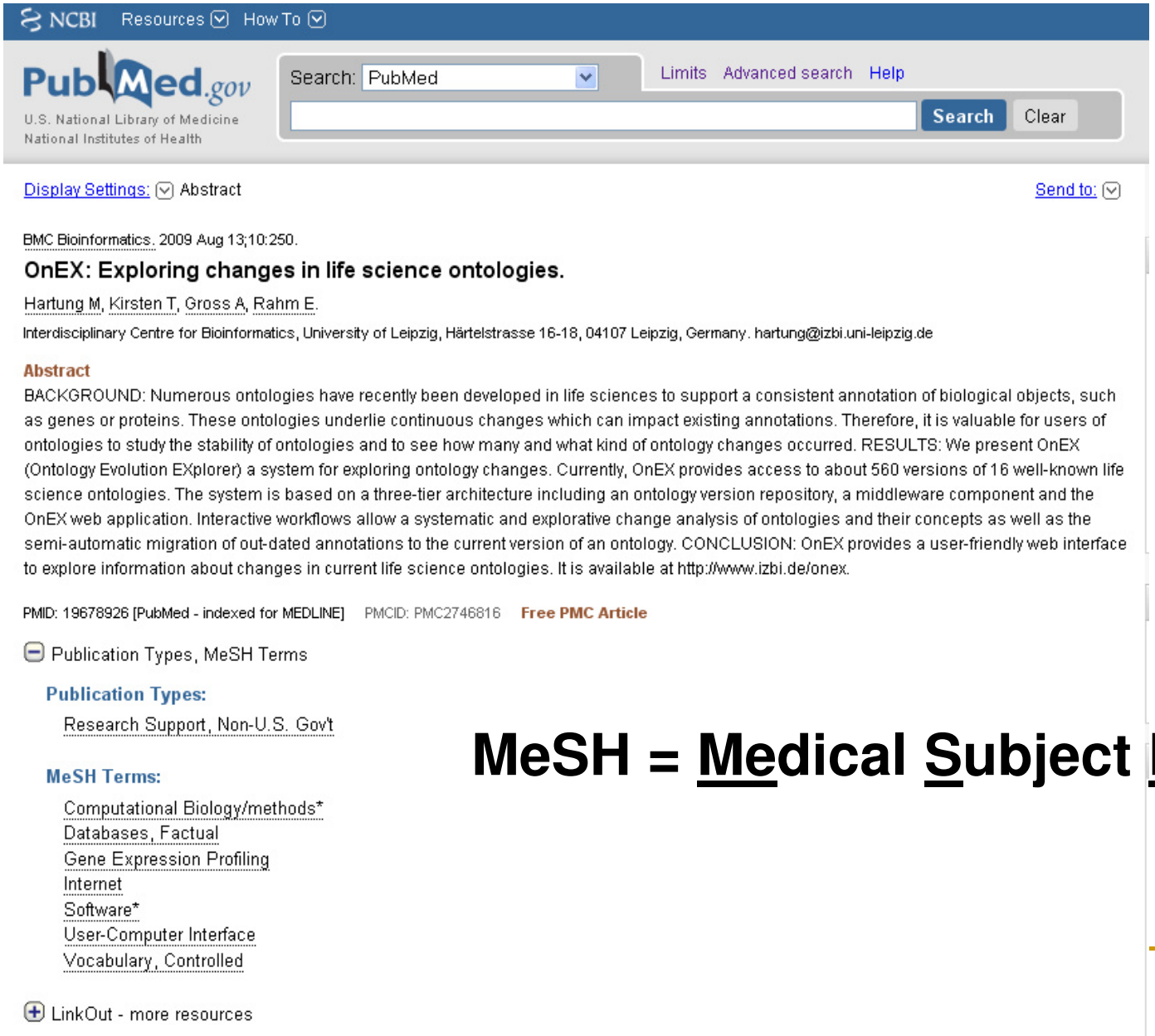


Beispiel: GO Annotationen

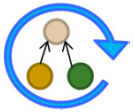
UniProtKB/Swiss-Prot entry Q03518



Beispiel: MeSH Annotationen

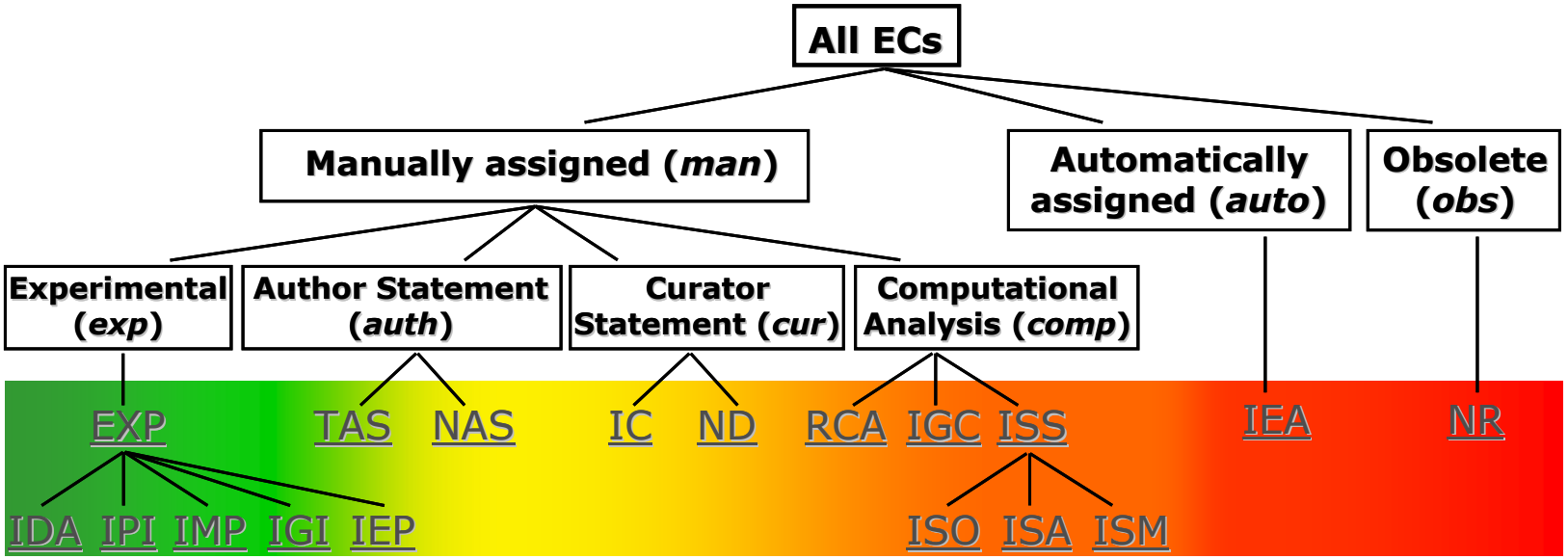


MeSH = Medical Subject Headings



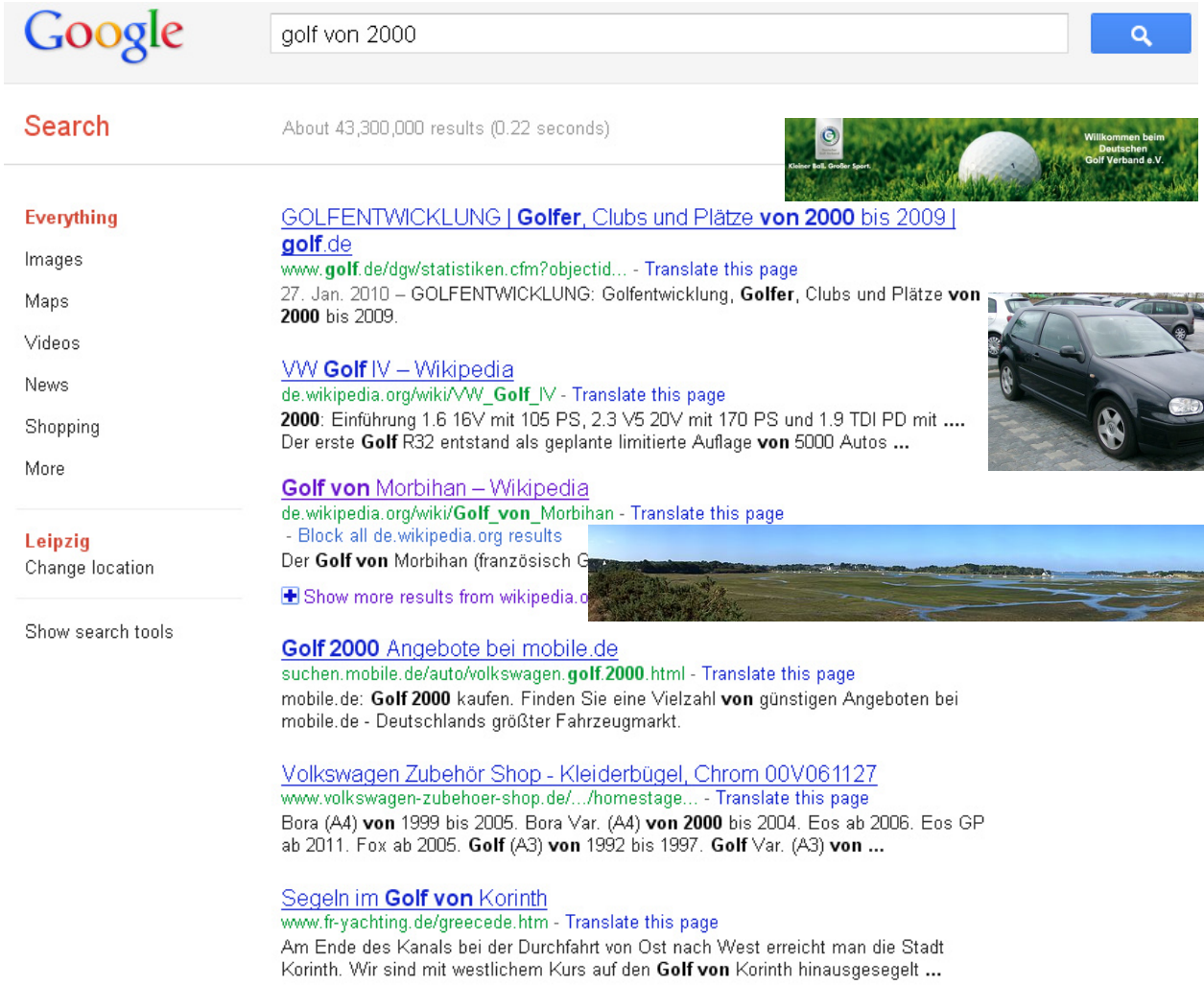
Qualität von Annotationen

- Annotationen können auf unterschiedlichste Art und Weise entstehen
  - Erfassung weiterer Metadaten zu Annotationen
    - Herkunft (provenance) → Evidence Codes
    - Kurator, Datum, Methode, Quelle, Referenzen, ...
  - GO Evidence Code (EC) Taxonomy



* http://www.geneontology.org/GO.evidence

Klassische Informationssuche



- Keyword-Suche oftmals mit **nicht relevanten** Ergebnissen
  - Bedeutungs-unterschiede
  - Homonyme
  - Unterschiedlicher Kontext
- Keyword-Suche findet **nicht alle relevanten** Ergebnisse
  - Synonyme
  - Fehlende Kontextpräzisierung

Semantische Suche

- Suchmaschinen – Document Retrieval
  - Eingabe: „Golf von 2000“

Nachfrage zu Golf:
(1) PKW, (2) Sportart, (3) Geographie, ...
→ Verfeinerung: PKW

Semantische Suche

- Suchmaschinen – Document Retrieval
  - Eingabe: „Golf von 2000“

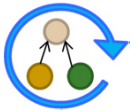
- Name: Volkswagen **Golf**
- Kategorie: **Auto**
  - ...
  - Baujahr <Integer>
  - Tech. Parameter
    - Hubraum <Integer>
    - Leistung <Integer>
    - ...
  - ...

Nachfrage zu 2000:
Baujahr oder technischer Parameter
→ Verfeinerung: Baujahr

Wissenrepräsentation
(Ontologie, Linked Data)

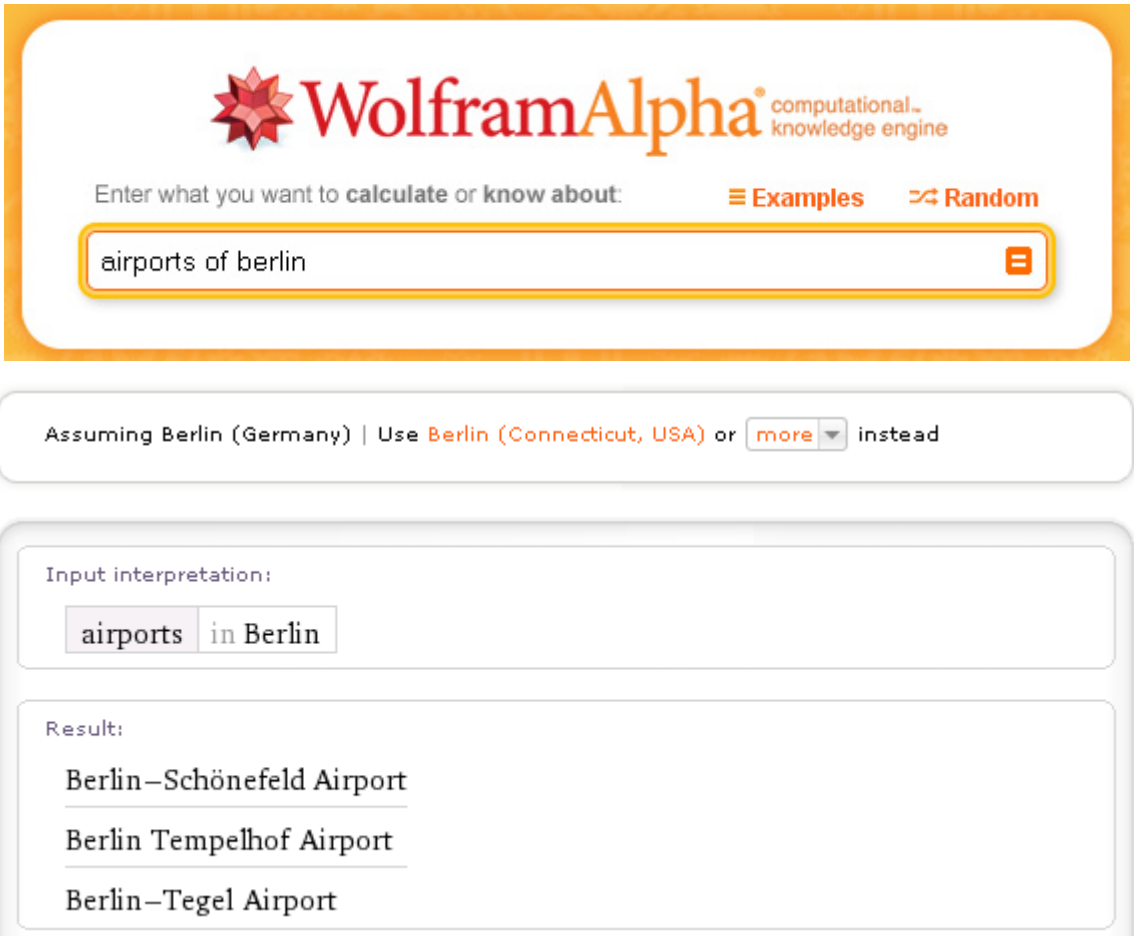
Semantische Suche

- Suchmaschinen – Document Retrieval
  - Generelle Probleme
    - Korrekte Interpretation der Suchphrase (Keywords)
    - Korrekte Identifikation von Entitäten
    - Automatische Disambiguierung
    - Usability
    - Personalisierung



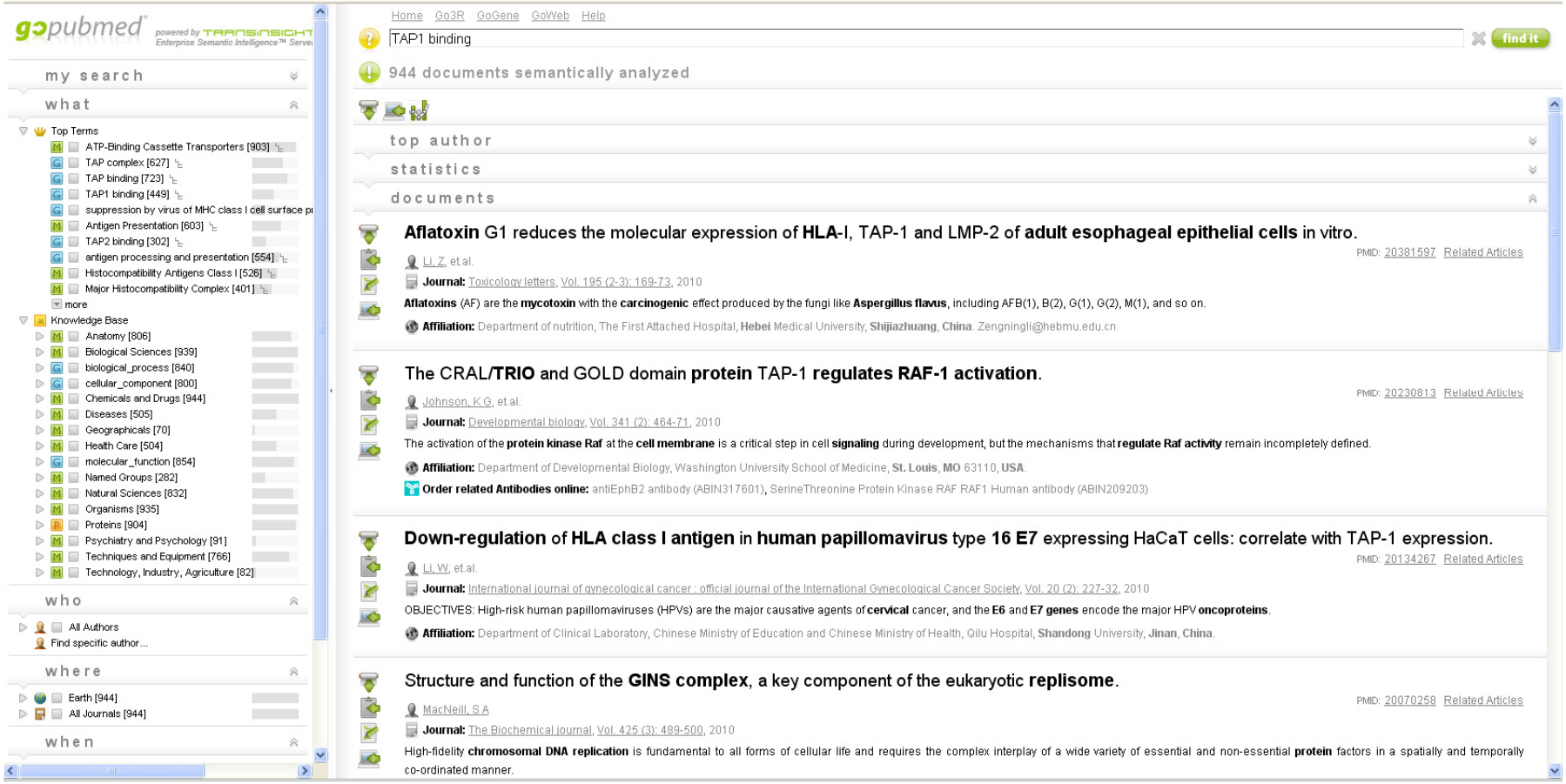
Semantische Suche

- Suchmaschinen – Fact Retrieval
  - Eingabe in Suchmaschine: „Welche Flughäfen gibt es in Berlin?“ (airports of Berlin)



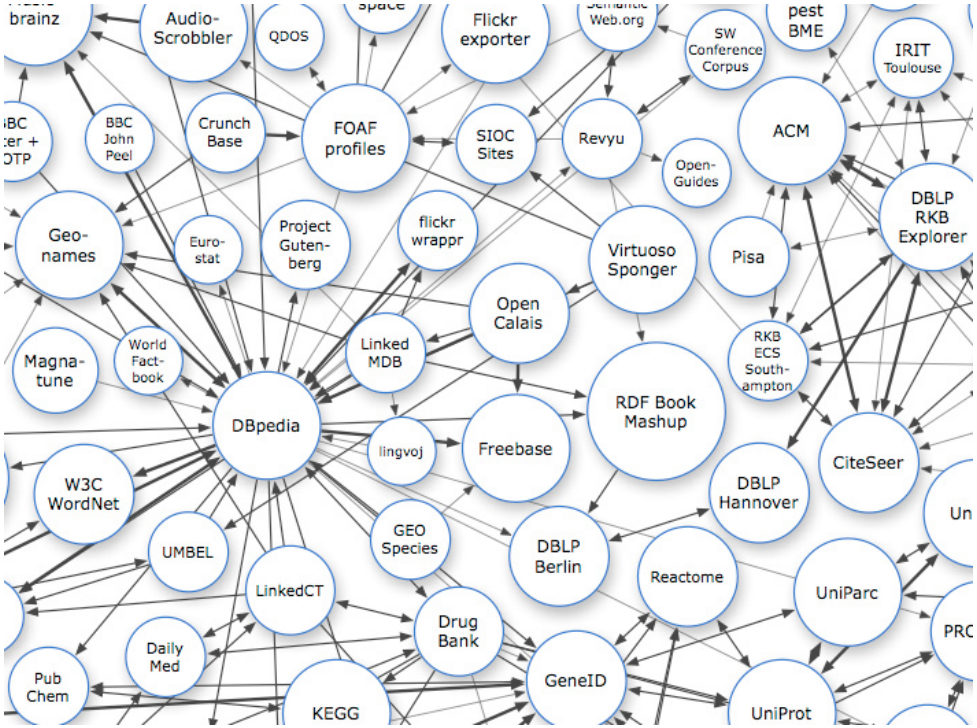
GOPubMed

- Klassifikation von PubMed-Artikeln unter Verwendung von GO und MeSH

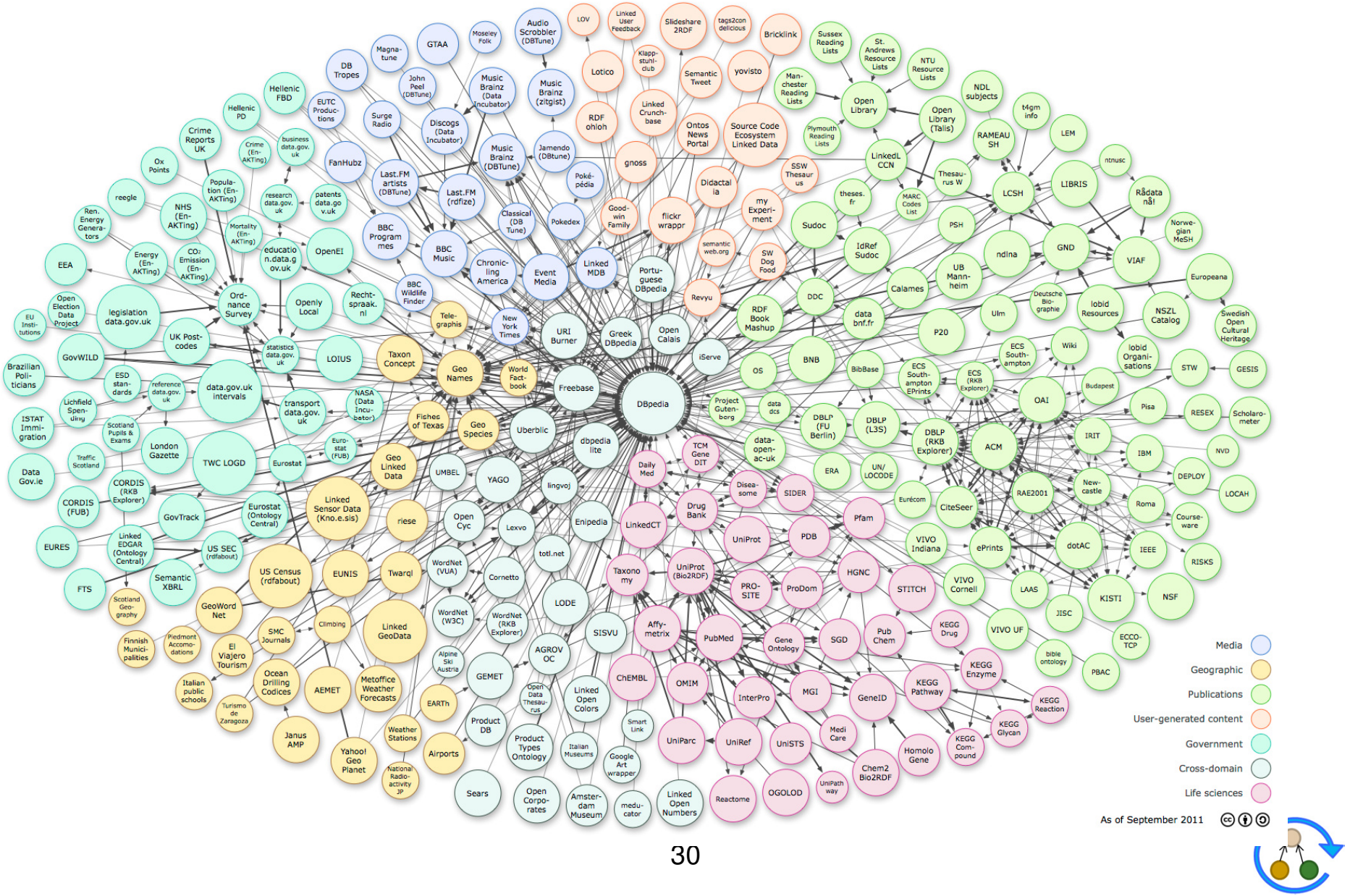


Linked Data

- Linked Open Data (LOD) bezeichnet im WWW frei verfügbare (RDF)-Daten
  - Per URI identifiziert → direkt via HTTP abrufbar
  - Verweis auf andere Daten via URI möglich



LOD (September 2011)



Beispiel BBC Music

a-ha

Formed 1983. Disbanded 2010.

PLAYED MOST ON BBC RADIO 2

Biography

A-ha (spelled "a-ha" in lower-case on their releases) were a Norwegian pop band formed in Oslo in 1982. The band was founded by Morten Harket (vocals), Magne Furuholmen (keyboard), and Pål Waaktaar (guitars). The group initially rose to fame during the mid 1980s after being discovered by musician and producer John Ratcliff and had continued global success in the 1990s and 2000s.

Read more at Wikipedia...

WIKIPEDIA This entry is from Wikipedia, the user-contributed encyclopedia. It may not have been reviewed by professional editors and is licensed under the GNU Free Documentation License. If you find the biography content factually incorrect, defamatory or highly offensive you can edit this article at Wikipedia. Find out more about our use of this data.

Share This Page 3 so far

Share facebook twitter

BBC Music Showcase

BBC MUSIC SHOWCASE
Watch and listen to exclusive music clips

Latest Tracks Played On The BBC

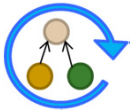
The Living Daylights
BBC RADIO 2 | KEN BRUCE 31/10/2011

Take On Me
BBC RADIO 2 | PICK OF THE POPS 1978 AND 1985

Foot Of The Mountain
BBC RADIO 2 | KEN BRUCE 17/10/2011

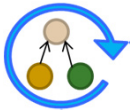
The Sun Always Shines On TV
BBC RADIO 2 | KEN BRUCE 29/09/2011

Hunting High & Low
BBC RADIO 2 | JEREMY VINE 08/09/2011



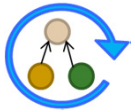
BBC Music

- Was ist das Besondere?
  - Informationen werden aus externen, öffentlich verfügbaren Daten automatisch aktuell zusammengestellt (Wikipedia, MusicBrainz, ...)
  - Kein Screen Scraping der Website
  - Kein spezielles API
  - Daten liegen im (Semantic) Web als Linked Open Data vor
  - Zugriff über gewöhnlichen HTTP Request
  - Daten immer hoch aktuell, kein manueller Eingriff nötig



BBC Music als RDF

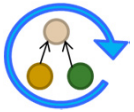
```
– <rdf:RDF>
  – <rdf:Description rdf:about="/music/artists/7364dea6-ca9a-48e3-be01-b44ad0d19897.rdf">
      <rdfs:label>Description of the artist a-ha</rdfs:label>
      <foaf:primaryTopic rdf:resource="/music/artists/7364dea6-ca9a-48e3-be01-b44ad0d19897#artist"/>
    </rdf:Description>
  – <mo:MusicArtist rdf:about="/music/artists/7364dea6-ca9a-48e3-be01-b44ad0d19897#artist">
      <rdf:type rdf:resource="http://purl.org/ontology/mo/MusicGroup"/>
      <foaf:name>a-ha</foaf:name>
      <ov:sortLabel>a-ha</ov:sortLabel>
    – <bio:event>
      – <bio:Birth>
          <bio:date rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1983</bio:date>
        </bio:Birth>
      </bio:event>
      <foaf:page rdf:resource="/music/artists/7364dea6-ca9a-48e3-be01-b44ad0d19897.html"/>
      <owl:sameAs rdf:resource="http://dbpedia.org/resource/A-ha"/>
      <mo:musicbrainz rdf:resource="http://musicbrainz.org/artist/7364dea6-ca9a-48e3-be01-b44ad0d19897.html"/>
      <mo:image rdf:resource="http://static.bbc.co.uk/music/images/artists/234x132/7364dea6-ca9a-48e3-be01-b44ad0d19897.jpg"/>
      <foaf:homepage rdf:resource="http://www.a-ha.com"/>
      <mo:wikipedia rdf:resource="http://en.wikipedia.org/wiki/A-ha"/>
      <mo:imdb rdf:resource="http://www.imdb.com/name/nm2165250/"/>
      <mo:myspace rdf:resource="http://www.myspace.com/aha"/>
      <mo:member rdf:resource="/music/artists/1ac73985-2cfa-433a-a6f8-379480e87179#artist"/>
      <mo:member rdf:resource="/music/artists/2093207d-cb69-4789-866b-5ac772c9ebd3#artist"/>
      <mo:member rdf:resource="/music/artists/0ef9b4c4-4a41-4ad8-9c64-98c0419c0150#artist"/>
    – <foaf:made>
      – <mo:Record>
          <dc:title>25 (disc 1)</dc:title>
          <mo:musicbrainz rdf:resource="http://musicbrainz.org/release/27ae8b3f-a4f8-408c-bbfe-1b4e3d56b3c7.html"/>
          <rev:hasReview rdf:resource="/music/reviews/qrd3#review"/>
        </mo:Record>
      </foaf:made>
    – <foaf:made>
      – <mo:Record>
          <dc:title>Foot of the Mountain</dc:title>
```

Linked Data Grundkonzept

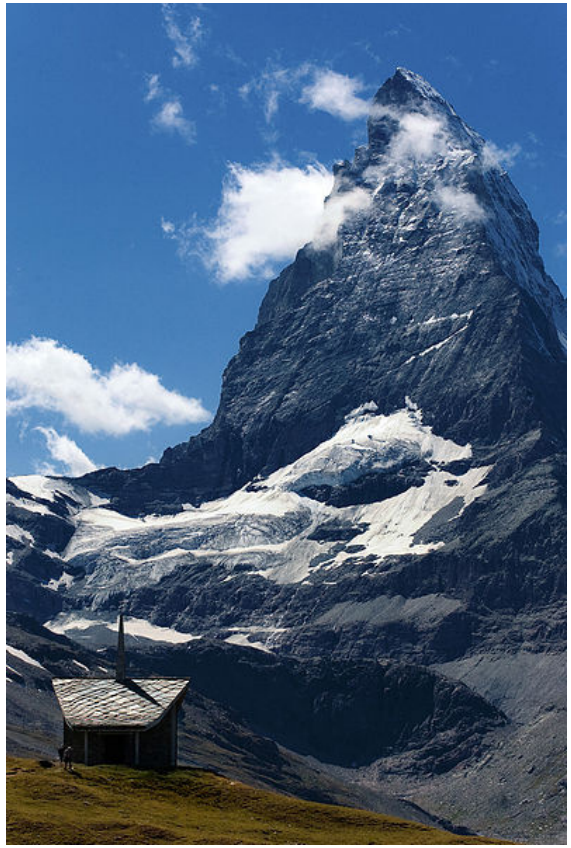
- Aus dem Netz von Webseiten soll ein Netz aus Daten entstehen
- „4 Regeln“ von Berners Lee
  1. Use URIs as names for things
  2. Use HTTP URIs so that people can look up those names
  3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
  4. Include links to other URIs, so that they can discover more things.

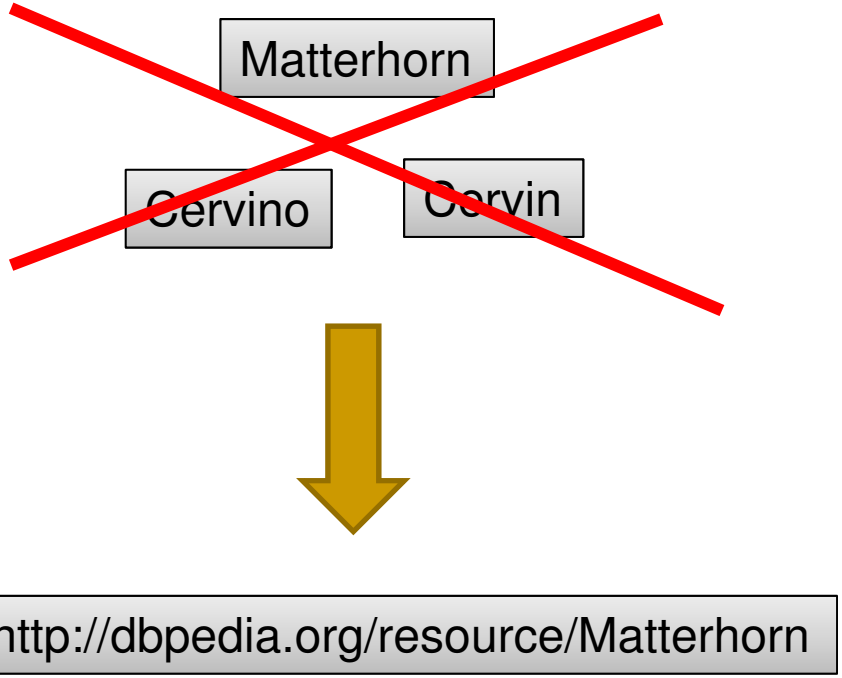
Linked Data Design Issues: http://www.w3.org/DesignIssues/LinkedData.html



Regel 1

- Eindeutige URIs als Bezeichnung für Objekte

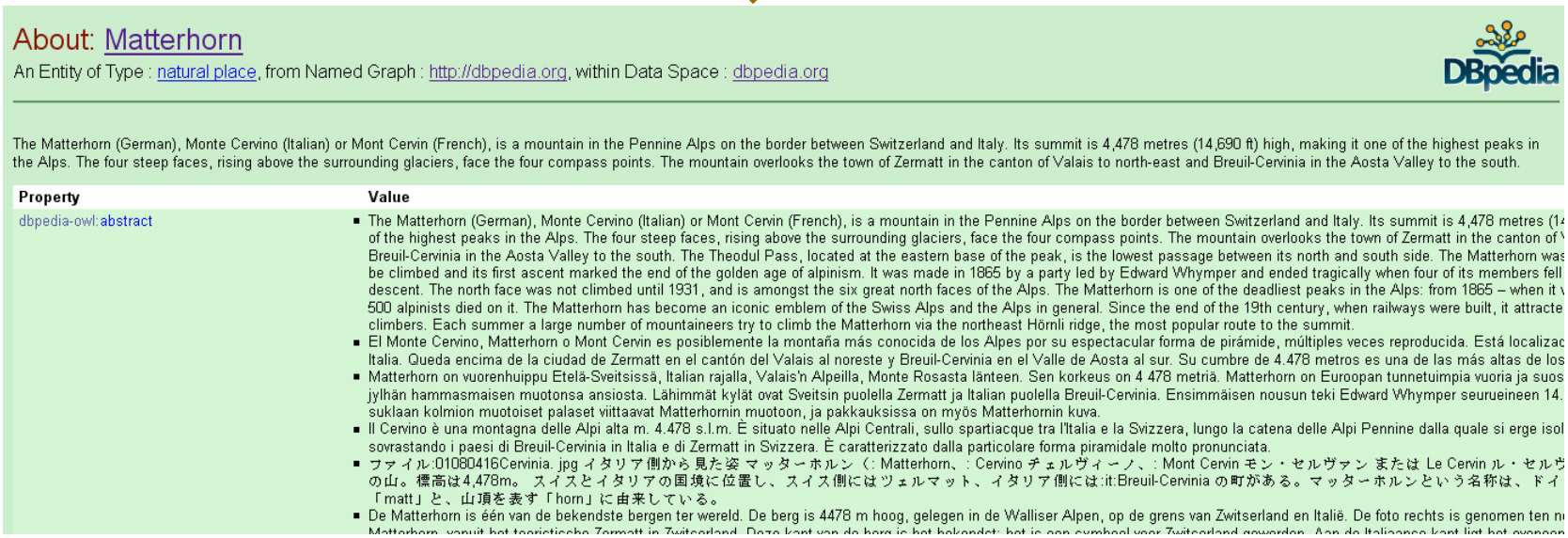




Regel 2

- HTTP URIs für Lookup

http://dbpedia.org/resource/Matterhorn



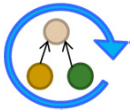
About: Matterhorn

An Entity of Type : natural place, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

DBpedia

The Matterhorn (German), Monte Cervino (Italian) or Mont Cervin (French), is a mountain in the Pennine Alps on the border between Switzerland and Italy. Its summit is 4,478 metres (14,690 ft) high, making it one of the highest peaks in the Alps. The four steep faces, rising above the surrounding glaciers, face the four compass points. The mountain overlooks the town of Zermatt in the canton of Valais to north-east and Breuil-Cervinia in the Aosta Valley to the south.

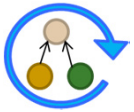
Property	Value
dbpedia-owl:abstract	▪ The Matterhorn (German), Monte Cervino (Italian) or Mont Cervin (French), is a mountain in the Pennine Alps on the border between Switzerland and Italy. Its summit is 4,478 metres (14 of the highest peaks in the Alps. The four steep faces, rising above the surrounding glaciers, face the four compass points. The mountain overlooks the town of Zermatt in the canton of V Breuil-Cervinia in the Aosta Valley to the south. The Theodul Pass, located at the eastern base of the peak, is the lowest passage between its north and south side. The Matterhorn was be climbed and its first ascent marked the end of the golden age of alpinism. It was made in 1865 by a party led by Edward Whymper and ended tragically when four of its members fell descent. The north face was not climbed until 1931, and is amongst the six great north faces of the Alps. The Matterhorn is one of the deadliest peaks in the Alps: from 1865 – when it w 500 alpinists died on it. The Matterhorn has become an iconic emblem of the Swiss Alps and the Alps in general. Since the end of the 19th century, when railways were built, it attracte climbers. Each summer a large number of mountaineers try to climb the Matterhorn via the northeast Hörnli ridge, the most popular route to the summit. ▪ El Monte Cervino, Matterhorn o Mont Cervin es posiblemente la montaña más conocida de los Alpes por su espectacular forma de pirámide, múltiples veces reproducida. Está localizad Italia. Queda encima de la ciudad de Zermatt en el cantón del Valais al noreste y Breuil-Cervinia en el Valle de Aosta al sur. Su cumbre de 4.478 metros es una de las más altas de los ▪ Matterhorn on vuorenhuippu Etelä-Sveitsissä, Italian rajalla, Valais'n Alpeilla, Monte Rosasta länteen. Sen korkeus on 4 478 metriä. Matterhorn on Euroopan tunnetuimpia vuoria ja suos jylhän hammasmaisen muotonsa ansiosta. Lähimmät kylät ovat Sveitsin puolella Zermatt ja Italian puolella Breuil-Cervinia. Ensimmäisen nousun teki Edward Whymper seurueineen 14. suklaan kolmion muotoiset palaset viittaavat Matterhornin muotoon, ja pakkauksissa on myös Matterhornin kuva. ▪ Il Cervino è una montagna delle Alpi alta m. 4.478 s.l.m. È situato nelle Alpi Centrali, sullo spartiacque tra l'Italia e la Svizzera, lungo la catena delle Alpi Pennine dalla quale si erge isol sovrastando i paesi di Breuil-Cervinia in Italia e di Zermatt in Svizzera. È caratterizzato dalla particolare forma piramidale molto pronunciata. ▪ ファイル:01080416Cervinia. jpg イタリア側から見た姿 マッターホルン（: Matterhorn、: Cervino チェルヴィーノ、: Mont Cervin モン・セルヴァン または Le Cervin ル・セルヴ の山。標高は4,478m。 スイスとイタリアの国境に位置し、スイス側にはツェルマット、イタリア側には:it:Breuil-Cervinia の町がある。マッターホルンという名称は、ドイ 「matt」と、山頂を表す「horn」に由来している。 ▪ De Matterhorn is één van de bekendste bergen ter wereld. De berg is 4478 m hoog, gelegen in de Walliser Alpen, op de grens van Zwitserland en Italië. De foto rechts is genomen ten n



Regel 3

- Nutzvolle Informationen bei Lookup

rdf:type	owl:Thing dbpedia-owl:Mountain http://schema.org/Mountain yago:MountainsOfItaly yago:MountainsOfSwitzerland yago:Mountain109359803 dbpedia-owl:NaturalPlace dbpedia-owl:Place http://schema.org/Place yago:MountainsOfTheAlps gml:_Feature yago:MountainsOfAostaValley yago:MountainsOfValais http://umbel.org/umbel/rc/Mountain http://umbel.org/umbel/rc/Location_Underspecified yago:InternationalMountainsOfEurope yago:PennineAlps
geo:geometry	POINT(7.65833 45.9764)
geo:lat	45.976387 (xsd:float)
geo:long	7.658333 (xsd:float)
rdfs:label	Matterhorn Cervino Matterhorn Cervin Cervino マッターホルン Matterhorn Matterhorn Matterhorn Matterhorn Matterhorn Маттерхорн 馬特洪峰



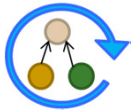
Regel 4

- Links zu weiteren Ressourcen („semantische Umgebung“)
  - In Datenquelle selbst

dbpedia-owl:firstAscentPerson	▪ dbpedia:Edward_Whymper ▪ dbpedia:Douglas_Robert_Hadow ▪ dbpedia:Lord_Francis_Douglas ▪ dbpedia:Charles_Hudson_(climber) ▪ dbpedia:Michel_Croz
dbpedia-owl:firstAscentYear	▪ 1865-01-01 00:00:00 (xsd:date)
dbpedia-owl:locatedInArea	▪ dbpedia:Italy
dbpedia-owl:mountainRange	▪ dbpedia:Pennine_Alps
dbpedia-owl:nationalTopographicSystemMapNumber	▪ Swisstopo 1347 Matterhorn
dbpedia-owl:parentMountainPeak	▪ dbpedia:Weisshorn

  - Zu anderen LOD Datenquellen

owl:sameAs	▪ http://linkedgeodata.org/triplify/node/26863664#id ▪ fbase:Matterhorn ▪ http://sws.geonames.org/2659729/



Zusammenfassung

- Beispielontologien
  - Top-Level vs. Domänenontologien
- Ontologien zur Annotation
  - Semantisch eindeutige Beschreibung von Objekten
  - Zahlreiche Vorteile (z.B. Suche, Wissensaustausch, ...)
- Semantische Suche
- Linked Data
  - Ontologien zur Gestaltung eines Web of Data
  - Designprinzipien