

INFO216: Knowledge Graphs

Andreas L. Opdahl
<Andreas.Opdahl@uib.no>



Session S09-S10: Linked Open Datasets

- Themes:
 - semantic vocabularies (S07-S08)
 - linked open datasets (S09-S10)
 - Linked Open Data (LOD)
 - the LOD cloud
 - Open semantic repositories:
 - DBpedia, Wikidata, GeoNames
 - and others
 - a few proprietary ones?
 - *...some of them have their own vocabularies*



Readings

- Resources in the portal, including:
 - research papers
 - LOD cloud og LOD stats
 - DBpedia
 - Wikidata
 - GeoNames
 - WordNet



Linked open datasets



Places to start (→S02)

- Open and semantic:
 - open semantic data sets: <http://lod-cloud.net>
 - vocabularies: <https://lov.linkeddata.es/dataset/lov/>
 - overview and statistics: <http://lodstats.aksw.org/>
- Open data in general:
 - internationally: <http://datahub.io>
 - Norge: <http://data.norge.no>
 - EU: <https://open-data.europa.eu>
 - UK: <http://data.gov.uk>
 - USA: <http://data.gov>
 - ...many run on the <http://ckan.org> platform



Linked Open Data (LOD, →S02)

- 3-4 basic principles (Berners-Lee 2006):
 1. IRI-er (Uniform Resource Identifier) *identify resources*
 - <http://dbpedia.org/resource/Bergen>
 2. IRI-s *answer to HTTP requests (dereferencing)*
 - for example *SPARQL queries, Turtle files, ...*
 3. Returns *information about the resource* on standard format, e.g.,
 - *RDF-XML, Turtle, N3, JSON-LD (JSON, XML, CSV, TSV, HTML)*
 - *may use “303 redirection” to distinguish the Concept from the Information about it, e.g.,*
 - <http://sws.geonames.org/3161732/>
 - <http://sws.geonames.org/3161732/about.rdf>
 4. The information contains IRI-s that *identify related resources*



Best Practices for Data Provisioning

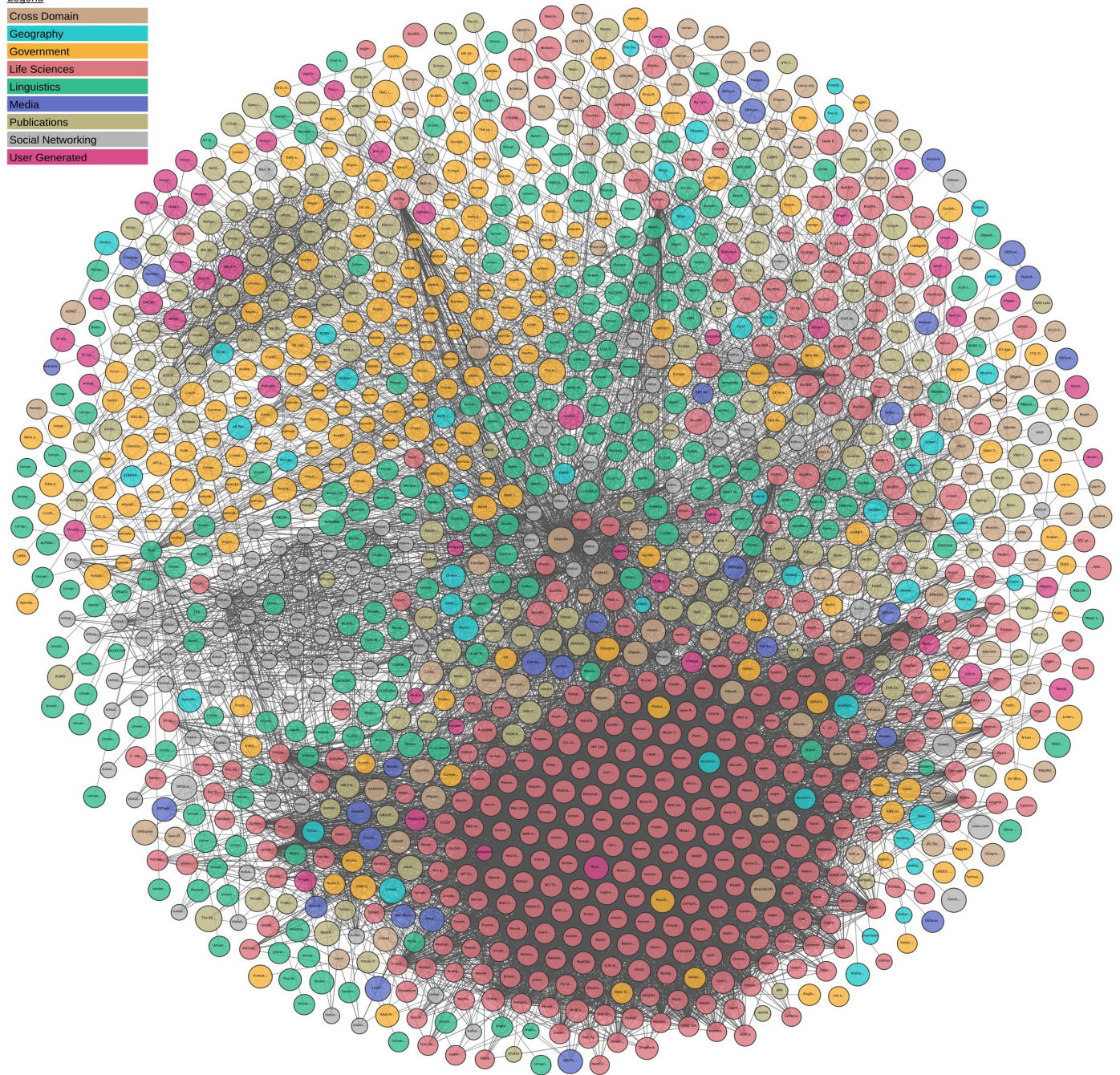
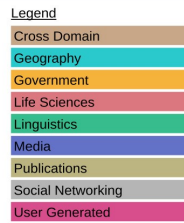
- Recommended directly by W3C
 - or emerged within the LOD community:
 1. *Provide dereferencable IRIs*
 2. *Set RDF links pointing at other data sources*
 3. *Use terms from widely deployed vocabularies*
 4. *Make proprietary vocabulary terms dereferencable*
 5. *Map proprietary vocabulary terms to other vocabularies*
 6. *Provide provenance metadata (e.g., PROV)*
 7. *Provide licensing metadata (e.g., CC)*
 8. *Provide dataset-level metadata (e.g., VANN, VS)*
 9. *Refer to additional access methods (e.g., SPARQL)*



The LOD cloud

- <http://lod-cloud.net/>
 - which datasets mention resources in other datasets?
 - 1239 datasets from datahub.io
 - ≥ 1000 triples, ≥ 50 links to other datasets
 - started in 2007
 - exponential-like growth for a few years
 - still growing, but more slowly now
 - statistics at http://lod-cloud.net/state/state_2014/
- <http://lodstats.aksw.org/> is less restrictive
 - ca 150G (150 000M) triples from ca 3000 data sets
 - most from SPARQL endpoints, some from file dumps





Domains:

- general
- geography
- government
- life sciences
- linguistics
- media
- publications
- social networki
- user generated

A “lumpy cloud”

Challenges

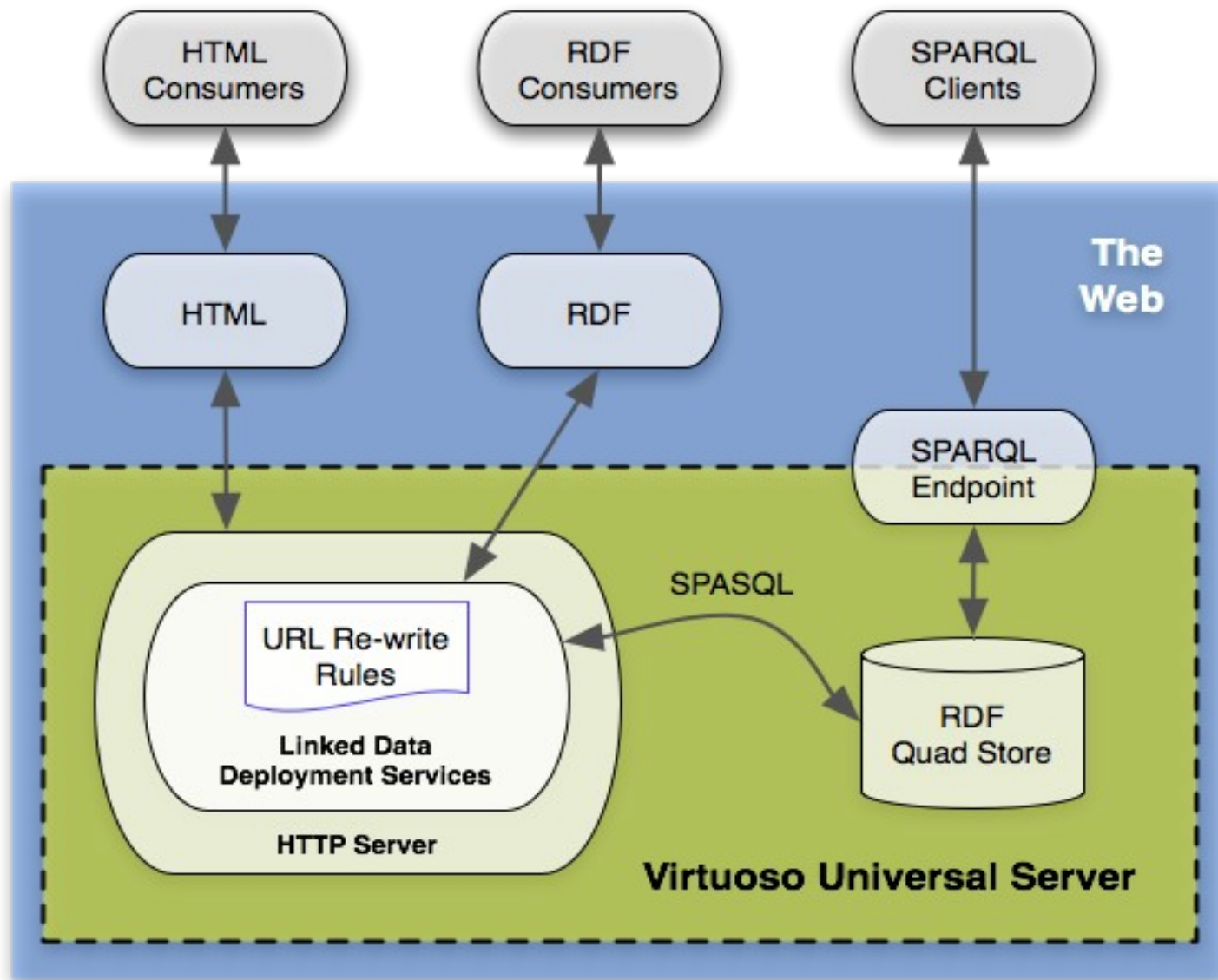
- Semantic technologies and the Web of Data and LOD has an enormous potential
 - ...but is not so much used *in its entirety* so far
 - parts of it are used
 - industry: biodata, publishing, music/media...
 - government: clean energy, libraries...
- Possible causes:
 - maintenance: individuals versus organisations
 - abstraction: general versus domain data
 - trust: open versus closed networks
- *Some of the “lumps” in the LOD cloud form such domain-specific and more tightly-knit subnetworks*



DBpedia

- Extracting structured information from Wikipedia
 - a crowd-sourced community effort
 - making this information available on the Web of Data
- Central source of individual IRIs:
 - <http://dbpedia.org/resource/<Res>>
- Available as:
 - RDF files, SPARQL endpoint (<http://dbpedia.org/sparql>)
 - HTML pages (<http://dbpedia.org/page/<Res>>)
 - faceted RDF browsing, powered by Virtuoso OpenLink
 - live SPARQL endpoint (<http://live.dbpedia.org/sparql>)
 - entity resolver service (<http://demo.dbpedia-spotlight.org/>)
 - lexicalizations dataset (maps names to DBpedia IRIs)





DBpedia: extraction

- Extracted approximately once a year
 - current version is 2019-08-30
 - but latest-core has more recent data
- Since January 2007:
 - first only in English
 - then 15 largest languages (since 3.7)
 - not Norwegian, but Swedish
 - today around 125 languages (since 3.8)
 - triple version + quad version with *provenance*
- Wikipedia's *infoboxes* are central
- ...also some full-text extraction and some NL parsing

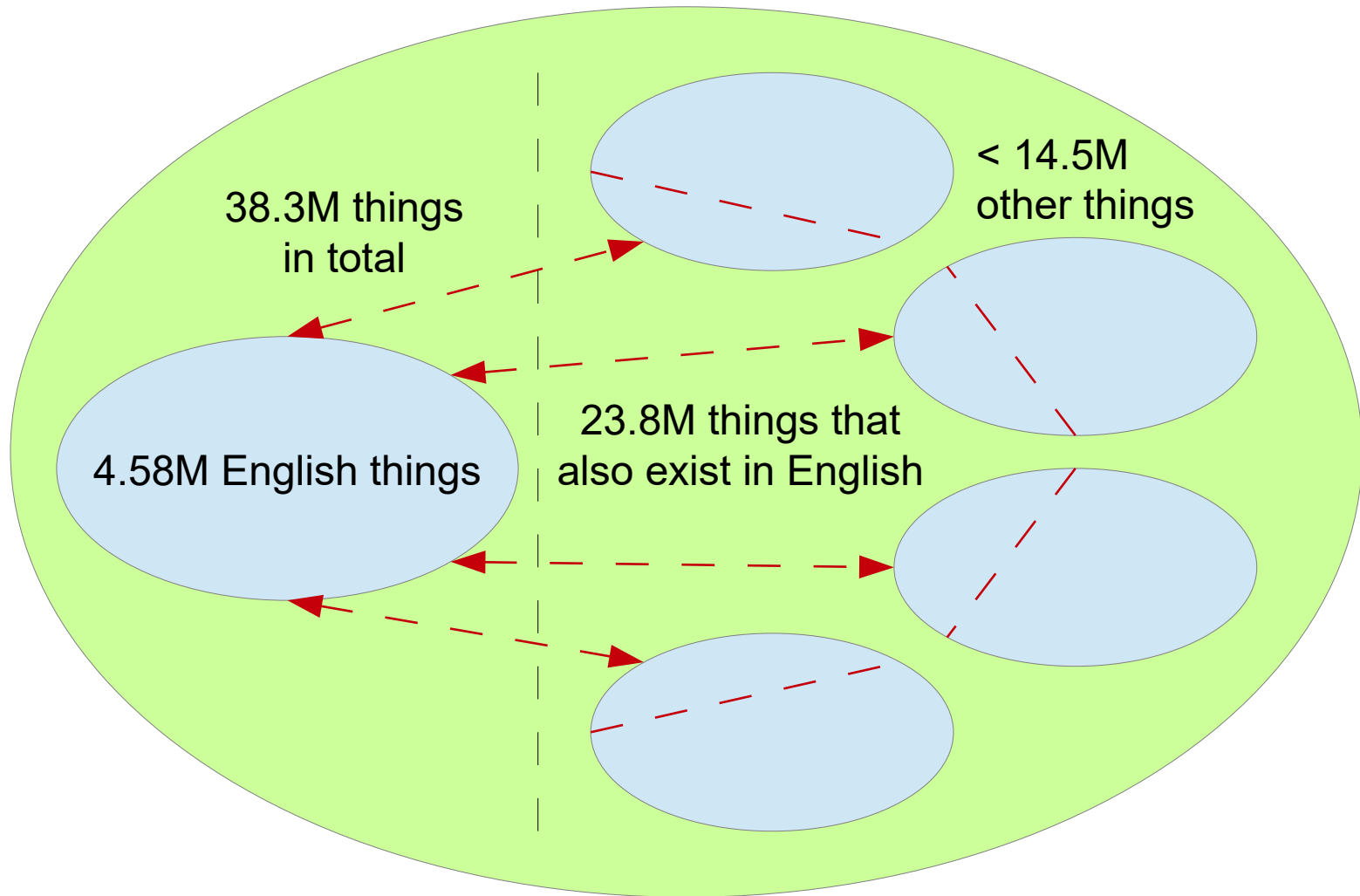


DBpedia: statistics

- Contents (around 2015):
 - 4.6M resources in the English DBpedia:
 - 1.5M persons, 810k places, 490k works, 275k organisations, 301k species...
 - also 1.7M SKOS concepts and other stuff
 - 38.3M resources in national DBpedia versions
 - around 125 languages
 - 23.8M resources are localised versions of resources that also exist in the English DBpedia
 - *may* describe more than 19M unique resources (things)



Canonical and normalised resources



(Example numbers from 2015-10.)

DBpedia: ontology and identities

- IRIs derived from Wikipedia, e.g.:
 - <http://en.wikipedia.org/wiki/Bergen> →
 - <http://dbpedia.org/resource/Bergen>
 - **English, canonical, dereferencable URIs**
- localised/national:
 - <http://no.dbpedia.org/resource/Bergen>
 - **not always dereferencable IRIs**
- Ontology (a stronger type of vocabulary):
 - 685 classes, 2795 properties
 - max depth: 5



DBpedia: some name spaces

- <http://dbpedia.org/> – language-independent base, URIs
- <http://nn.dbpedia.org/> – language-specific base, IRIs
 - approx. 125 languages, not all dereferencable
- <http://dbpedia.org/resource/> – resources (individuals)
- <http://dbpedia.org/property/> – raw infobox properties
- <http://dbpedia.org/ontology/> – mapped infobox properties and types
- <http://dbpedia.org/reference/> – external references
- [foaf:homepage](http://foaf.org/homepage) – external identifier reference
- [owl:sameAs](http://owl.w3.org/sameAs) – interlinking, e.g, across languages
- [rdf:type](http://rdf.type) – three classification schemes



DBpedia: raw and mapped extraction

- Wikipedia's *infoboxes* are central
 - raw transformation from *infoboxes* to triples:
 - generates national property names
 - infobox templates may be badly defined and used
 - inconsistent properties, no literal types
 - manual mapping (by scripts) from *infoboxes* to triples:
 - generates standardised properties
 - the DBpedia *ontology*
 - fixes many infobox problems
 - increasingly specific
 - own wiki for creating mappings



DBpedia: triples

- The full (international) data set:
 - 9.5G triples (*≈ 6-7% of <http://stats.lod2.eu/>*)
 - 1.3G from the English Wikipedia
 - 5.0G from other Wikipedias
 - the rest from Commons **and Wikidata**
 - 38M labels and abstracts (2015 here and below)
 - > 120M categorisation links
 - 67M links to Wikipedia categories
 - 24.6M links to images
 - 27.6M links to external web pages
 - 45M other external links: GeoNames, Freebase, Wikidata, Flickr wrapppr, YAGO, UMBEL...



DBpedia: classification schemes

- Wikipedia categories:
 - *81M links*
 - SKOS vocabulary and DCMI terms
- YAGO classification:
 - *41M links*
 - Yet Another Great Ontology
 - derived from Wikipedia using WordNet (also from GeoNames)
- WordNet synsets
- *...also 50M other links (30M to web pages)*



DBpedia: advantages

- Covers many domains
 - like Wikipedia, exploits *the long tail*
- Real community agreement
- Automatically evolves (as Wikipedia changes)
- Is truly multilingual



Freebase

- *A terminated free and open knowledge base that could be read and edited by both humans and machines*
 - from 2007
 - similar to DBpedia, but crowdsourced
 - acquired by Google in 2010
 - closed in 2014
 - data dumps still available
- Central information source for
 - Google's knowledge graph
 - Wikidata



Wikidata

- *A free and open knowledge base that can be read and edited by both humans and machines*
 - a Wikimedia project, crowdsourced, multi-lingual
 - *a Wikipedia for structured, secondary data*
 - verifiability, link to sources, perspectives
 - central storage for the structured data of its Wikimedia sister projects (Wikipedia etc.)
 - supports many other sites and services
 - free license (CC0), standard formats, interlinked
- Wikidata entities:
 - > 80M items (things)
 - 3000 properties



Wikidata access

- Available through
 - the Wikimedia API
 - HTTP: <http://www.wikidata.org/entity/Q42>
 - RDF: <http://www.wikidata.org/entity/Q42.ttl>
 - SPARQL endpoint: <http://query.wikidata.org>
 - Wikidata Query Service (WDQS)
 - for download (JSON, RDF, XML)
- Also as Linked Data Fragments:
 - <https://query.wikidata.org/bigdata/ldf>
- DBpedia also offers Wikidata compatible dumps



label — **Douglas Adams** (Q42) — unique identifier

description — English writer and humorist
Douglas Noël Adams | Douglas Noel Adams — aliases
▶ In more languages

Statements

property — **educated at** — value
St John's College
end time 1974
academic major English literature
academic degree Bachelor of Arts
start time 1971 — qualifiers

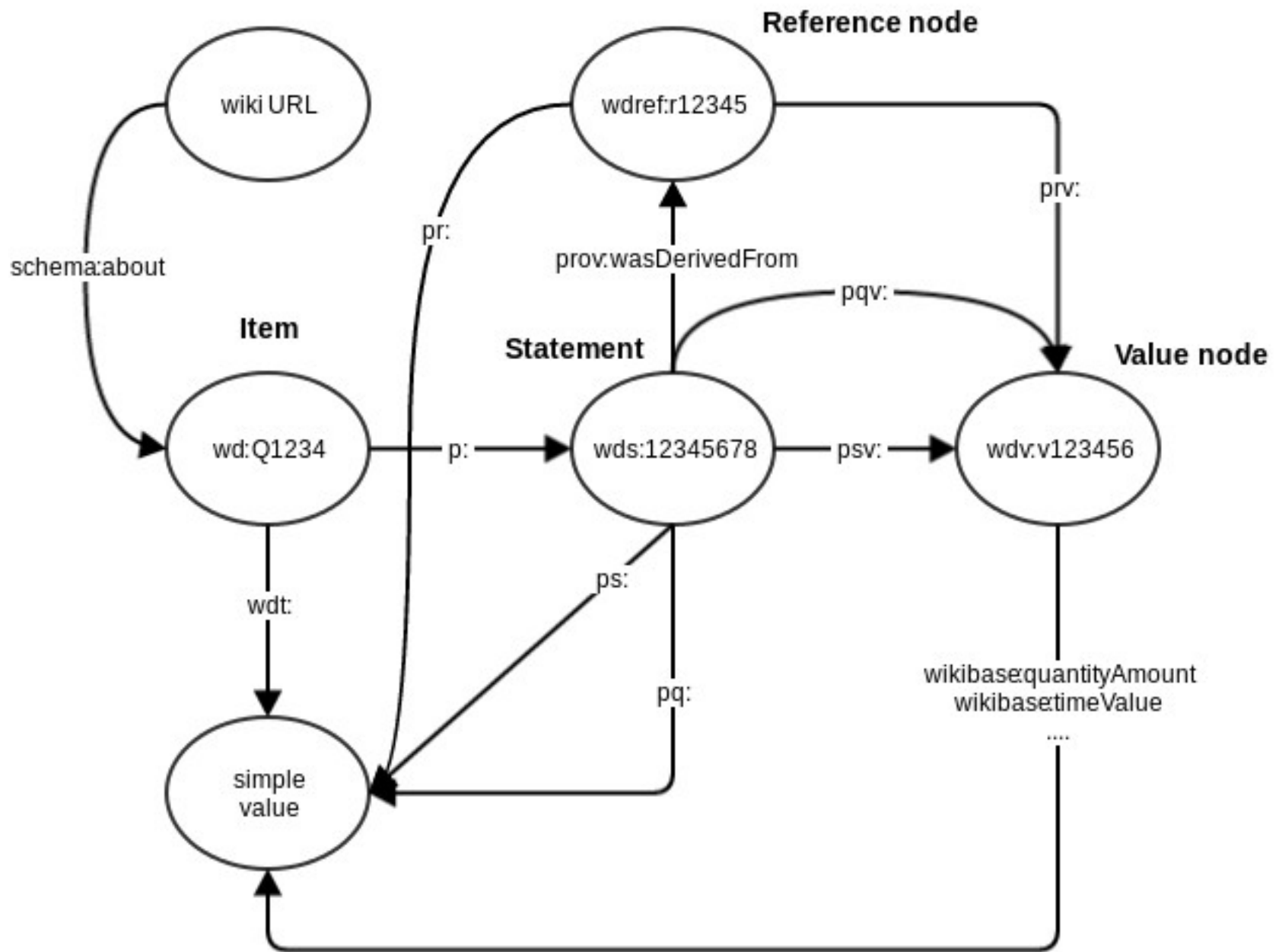
rank —
▼ 2 references
opened references
stated in Encyclopædia Britannica Online
reference URL <http://www.nndb.com/people/731/000023662/>
original language of work English
retrieved 7 December 2013
publisher NNDB
title Douglas Adams (English)

statement group —
+ add reference
Brentwood School
end time 1970
start time 1959
collapsed reference
▶ 0 references
+ add (statement)

Wikidata item structure

- Items:
 - item identifier (*Qnn*)
 - fingerprint:
 - multilingual label, description, aliases
 - statements, each:
 - claim: a property-value pair
 - qualifiers: additional property-value pairs *about the claim*
 - references (one or more property-value pairs)
 - rank
- Site links
- *Similar structure for properties!*





```
PREFIX wikibase: <http://wikiba.se/ontology#>
```

```
PREFIX wd: <http://www.wikidata.org/entity/>
```

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
```

```
#defaultView:BubbleChart
```

```
SELECT ?cLabel ?p WHERE {
```

```
  ?c wdt:P31 wd:Q6256 .
```

```
  ?c wdt:P30 wd:Q46 .
```

```
  ?c wdt:P1082 ?p .
```

```
  SERVICE wikibase:label {
```

```
    bd:serviceParam wikibase:language "en" .
```

```
  }
```

```
}
```



Wikidata Query Service (WDQS)

- SPARQL wrapper for Wikidata (<http://query.wikidata.org>)
 - based on BlazeGraph, OpenRDF/RDF4J
 - built-in prefixes
 - generate query IRIs
 - various entity/ontology explorers, e.g.,
 - SQID (<https://tools.wmflabs.org/sqid/#/>)
 - GraphBuilder
 - built-in visualisations
 - built-in SERVICES ([wikibase:label](#))
- Also:
 - Linked Data Fragments
(<https://query.wikidata.org/bigdata/ldf>)



WDQS visualisations

- Use a comment: `#defaultView:viewName`
- Supported viewNames:
 - **Table** - default view, displays the results as a table
 - **Map** - displays coordinate points if present
 - **ImageGrid** - displays result images as a grid
 - **BubbleChart** - displays numbers as bubble chart
 - **TreeMap** - displays hierarchical tree map for numbers
 - **Timeline** - displays timeline for results having dates
 - **Dimensions** - displays rows as lines between points
 - **Graph** - displays result as a connected graph
- (More limited) server-side alternative to Sgvizler



Wikidata versus DBpedia

- Similarities:
 - both publish **RDF data** about **entities/resources**
 - both use **standard IRIs**
 - both define **ontologies**
 - both are extensively **linked** to other semantic datasets
- Differences:
 - **source**: DBpedia is derived; Wikidata is crowdsourced
 - **direction**: DBpedia extracts data from Wikipedia;
Wikidata provides data to Wikipedia
 - **structure**: DBpedia adds structure to Wikipedia data;
Wikidata is natively structured
 - **maturity**: DBpedia is older; Wikidata is recent



GeoNames

- *Adding geospatial semantic information to the web*
 - a geographical database: <http://www.geonames.org>
 - collected from a large number of sources
 - > 10M geographical names (*toponyms*, Norway 68k),
> 9M unique features, ~ 2.8M populated places,
~ 5.5M alternate names
- Offers *dereferencable IRIs* for *toponyms / place names*
 - “*303 redirection*” for *Concept-Document distinction*
 - i.e., an entity and the information about it are different resources
 - <http://sws.geonames.org/3161732/>
 - <http://sws.geonames.org/3161732/about.rdf>



GeoNames

- Available as:
 - map-based HTML pages (POW – “Plain Old Web”)
 - web APIs (REST, XML, RDF)
 - SPARQL endpoints
 - dereferencable IRIs
 - downloadable (TSV)
 - Gazetteer lists
- Also as Linked Data Fragments:
 - <http://data.linkeddatafragments.org/geonames>



GeoNames ontology

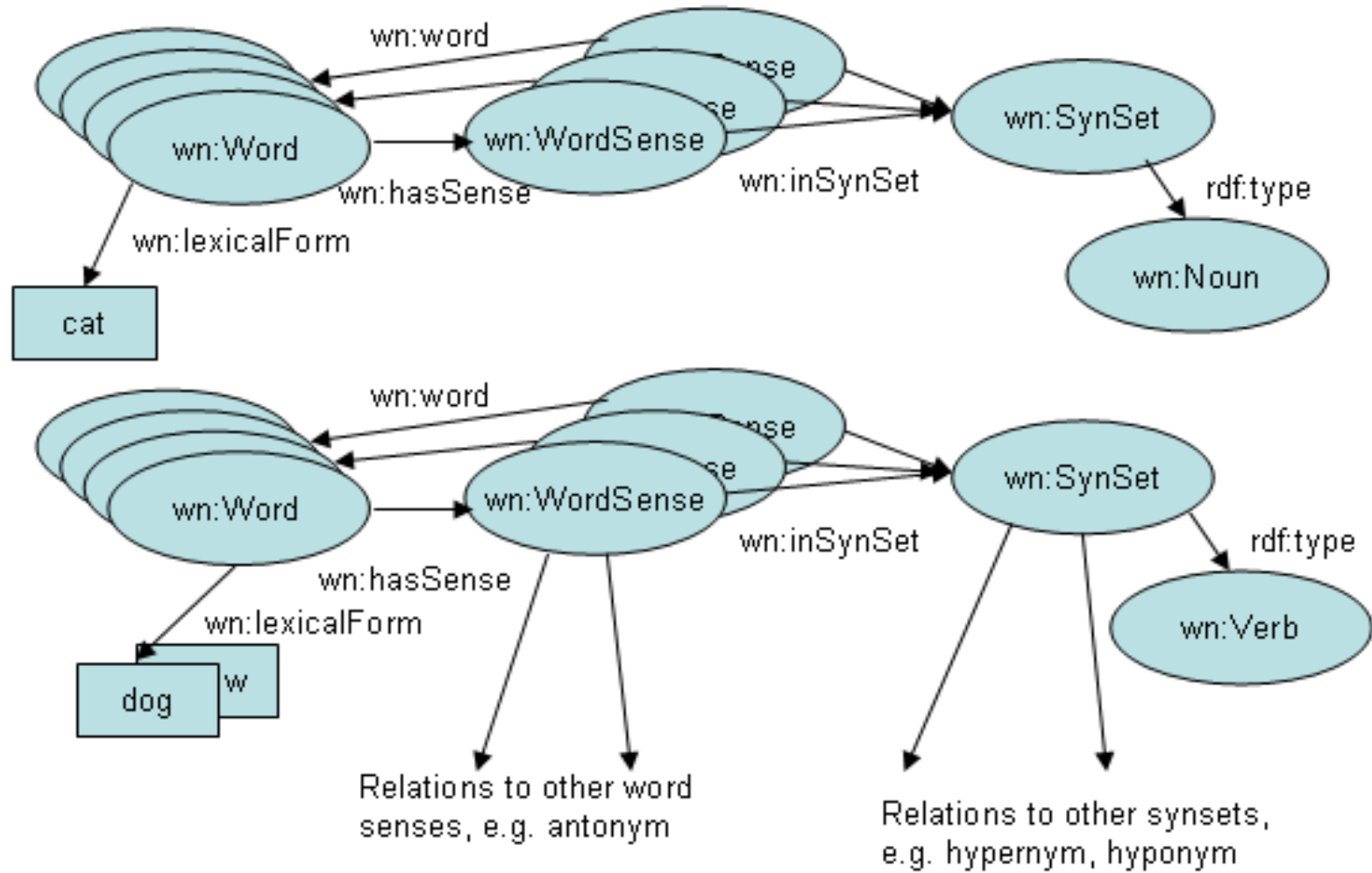
- Vocabulary in OWL:
 - @prefix gn: <<http://geonames.org/ontology#>> .
 - **gn:Feature** class
 - 9 top-level feature codes:
 - **A** country, state, region, ...; **H** stream, lake, ...;
 - L** parks, area, ...; **P** city, village, ...; **R** road, railroad;
 - S** spot, building, farm; **T** mountain, hill, rock, ...;
 - U** undersea; **V** forest, heath, ...
 - 645 detailed feature codes (in a hierarchy)
- **gn:name**, **gn:alternateName**, **gn:locationMap**,
gn:countryCode, **gn:featureClass**, **gn:featureCode**,
gn:nearbyFeatures, **gn:parentADM1**, **gn:parentADM2**,
gn:parentCountry, **gn:population**, **gn:wikipediaArticle**
- also uses properties from *geo*, *foaf*, *dcterms*, *cc*, *rdfs*...

WordNet

- An electronic open-source dictionary (Miller, 1985-):
 - 155k open-class words, 118k synonym sets (*synsets*), 207k Word-Sense pairs
 - hand-written definitions, common-use frequencies
 - version 3.1 available for download or online:
 - <http://wordnetweb.princeton.edu/perl/webwn>
 - APIs in many languages (Java, Python)
 - RDFS and OWL versions exist
 - WordNet in RDF:
 - <https://www.w3.org/TR/wordnet-rdf/>
 - <http://wordnet-rdf.princeton.edu/>
 - also versions for other languages



WordNet: Structure



WordNet: Standard synset IRIs

- *@prefix wn20schema:*
<<http://www.w3.org/2006/03/wn/wn20/schema/>> .
- *@prefix wn30:*
<<http://purl.org/vocabularies/princeton/wn30/>> .
- *@prefix wn31:*
<<http://wordnet-rdf.princeton.edu/wn31/>> .
- Example:
 - *wn31:synset-bank-noun-2*
- Other open semantic datasets – partly derived from WordNet – offer other IRI-schemas



WordNet: Synset structure

- Different *concept relations* for each *Part of Speech (PoS)*
- Nouns:
 - hyponyms/hypernyms
bat-n-1 is-kind-of placental_mammal-n-1
 - type / instance
Norway-n-1 instance-of Scandinavian_country-n-1
 - holonyms/meronyms
bat-n-1 has-part wing-n-1
 - *antonyms*
birth-n-1 has-antonym death-n-1
 - entailment, domains
bat-n-2 has-domain baseball-n-1



WordNet: Synset structure

- Verbs:
 - troponyms/hypernym
communicate-v-2 has-troponym talk-v-2
talk-v-2 has-troponym whisper-v-1
 - depending on semantic field:
run-v-1 has-troponym jog-v-3
like-v-2 has-troponym love-v-2
 - verb groups
 - antonyms
love-v-1 has-antonym hate-v-1
 - similarity, sister terms
bat-v-1 has-sister swat-v-1



WordNet: Synset structure

- Adjectives:
 - semantic, similarity, antonyms, indirect antonyms
- Adverbs:
 - similar to adjectives
- Also cross-PoS:
 - island – islander (derived from)
 - talk – speak for (phrasal)...
 - ...and others



WordNet: Norsk Ordvev

- Developed Kaldera språkteknologi
 - for Nasjonalbiblioteket (The national library)
 - both *bokmål* and *nynorsk*
 - \approx 50 000 words, 200 000 synsets each
- Available at
 - <http://www.nb.no/sprakbanken/show?serial=sbr-27&lang=nb>
 - <https://www.nb.no/sprakbanken/show?serial=sbr-7&lang=nb>
 - updated January / February 2016
- *So far not looking finished...*



International language resources

- Global Wordnet Grid (<http://globalwordnet.org/>)
 - building a *Global Multilingual Wordnet*
<http://compling.hss.ntu.edu.sg/omw/>
- DBpedia Wiktionary as Linked Data Fragments
 - *extracting a DBpedia from Wiktionary*
 - <http://data.linkeddatafragments.org/wiktionary>
- Dbnary (<http://kaiko.getalp.org/about-dbnary/>)
 - *extracting a DBpedia from Wiktionary*
 - automatic extraction of RDF graphs from Wiktionary
- BabelNet (<http://babelnet.org>)
 - multilingual text analysis and translation
 - *BabelNet is very active at the moment!*



BabelNet

- A multilingual encyclopedic dictionary and a semantic network of concepts and named entities
 - both lexicographic and encyclopedic coverage
 - 15 million Babel synsets
 - integrates data from *WordNet*, *Open Multilingual Wordnet*, *Wiktionary*, *Wikidata*, *Wikipedia*, *Wikiquotes*, *GeoNames* and several others



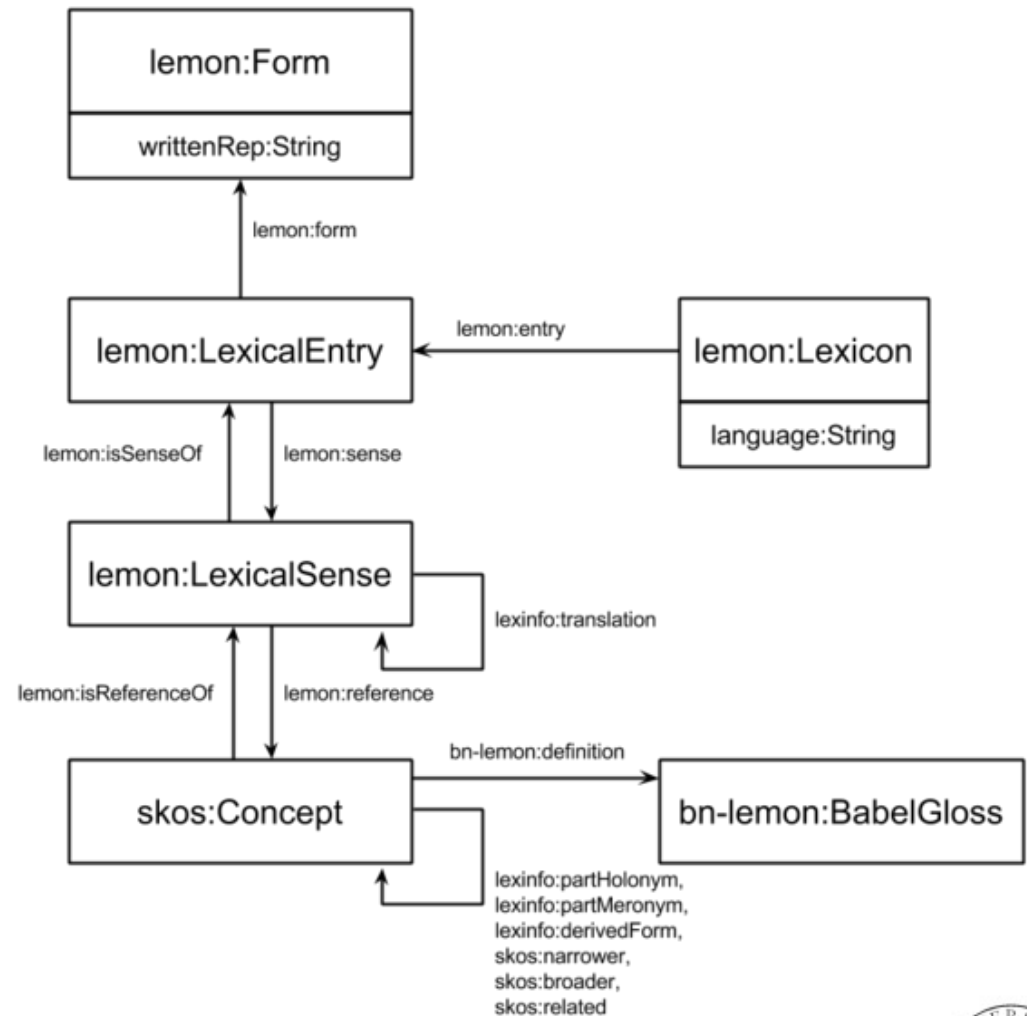
BabelNet availability

- Available as:
 - web lookup service
 - web translation service
 - web API (JSON) with Java library
 - SPARQL endpoint
 - linked data interface
 - <http://babelnet.org/rdf/page/>
 - the Linguistic LOD (LLOD) cloud



BabelNet conceptual model

- Making BabelNet part of the LLOD cloud
- Vocabularies:
 - Lemon
 - BabelNet-lemon
 - LexInfo
 - SKOS
 - RDFS
 - DC elements
 - DC terms
- Lemon is the backbone



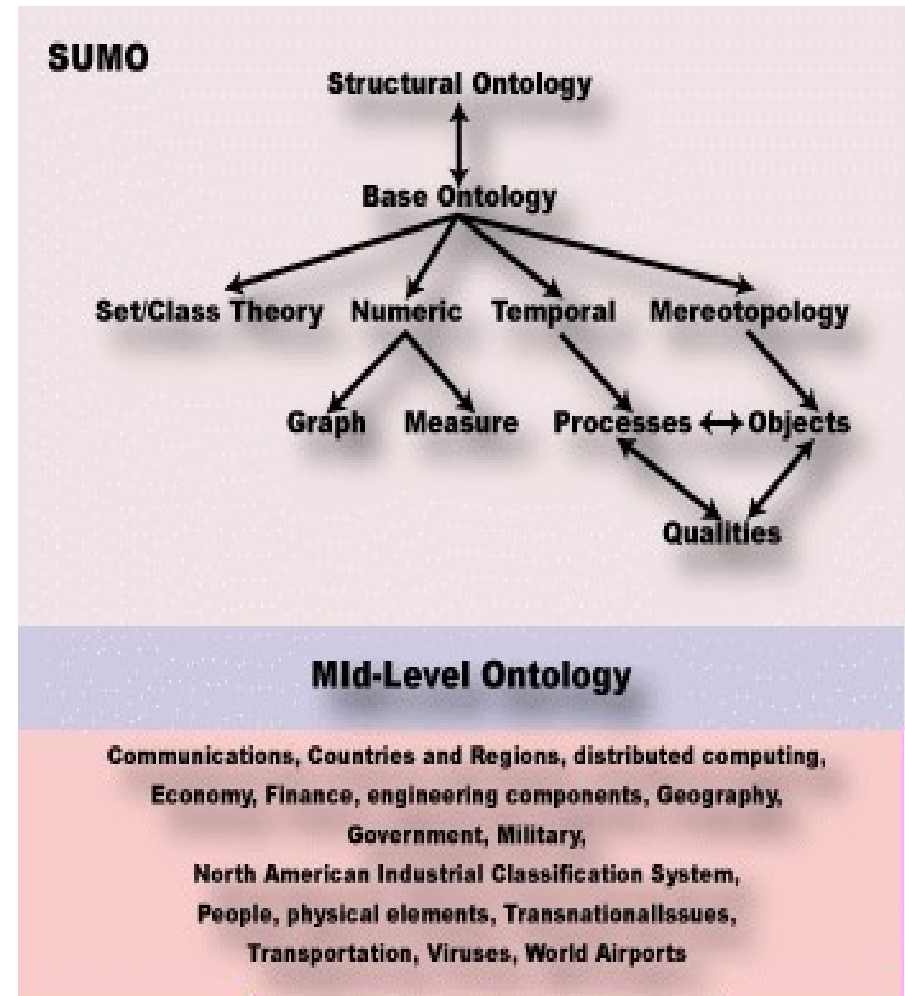
YAGO2

- *Yet Another General/Great Ontology:*
 - facts extracted from Wikipedia (the category structure), WordNet and GeoNames...
 - 10M entities, 120M triples/facts
 - places facts and entities in time and space
 - ...and in WordNet domains
 - integrated with DBpedia and SUMO
 - used to categorise DBpedia's resources
 - a commonsense knowledge base
 - used in IBM's Watson system
 - downloadable as RDF
 - querying and browsing at
<http://www.mpi-inf.mpg.de/yago-naga/yago/>

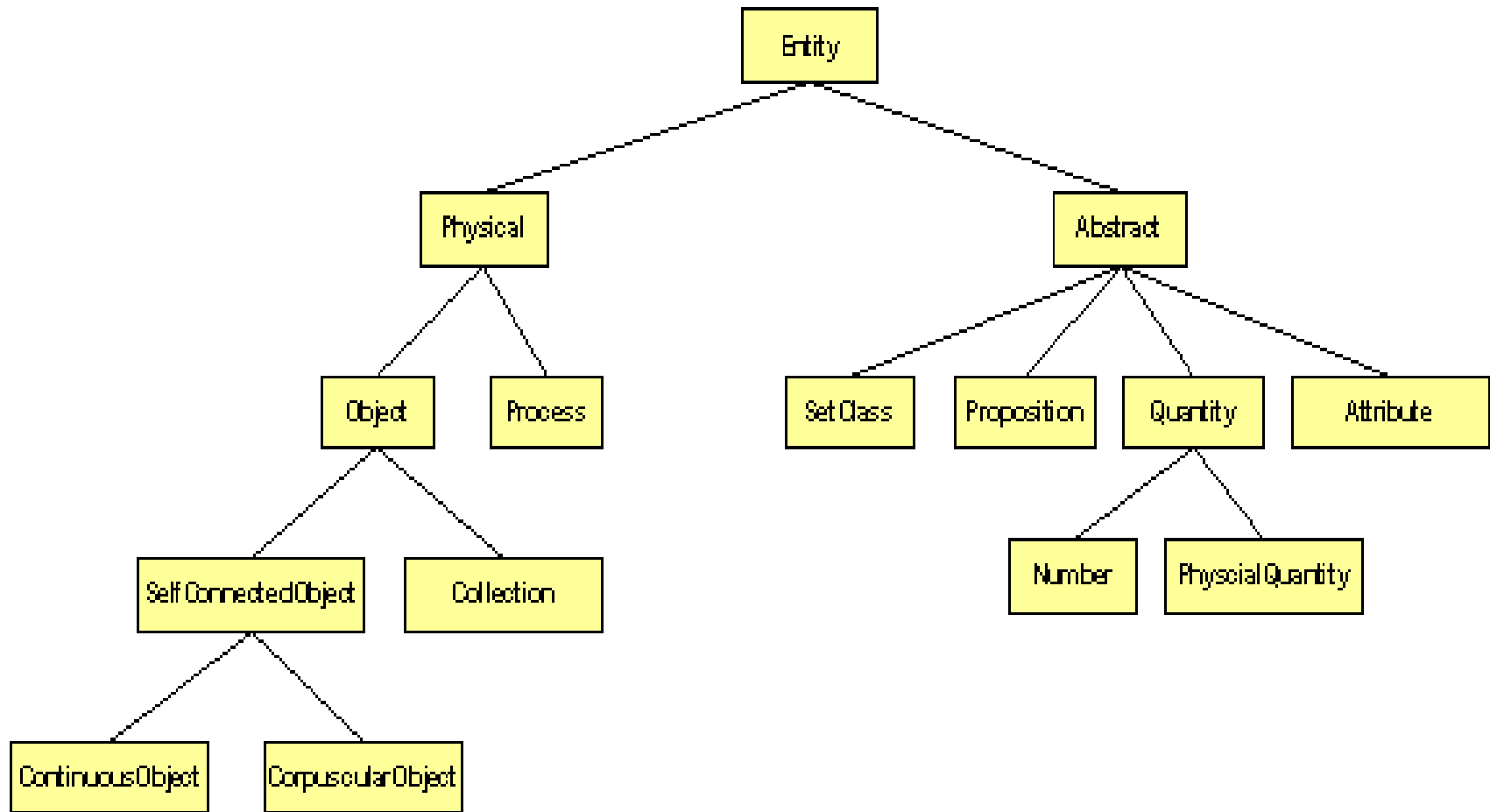


Suggested Upper Merged Ontology (SUMO)

- Defines and organises concepts *formally and philosophically*:
 - 22.6k terms
 - 307k axioms
 - 5k rules
- Mapped to WordNet, DBpedia, YAGO...
- Available online, KIF and OWL dumps
- IEEE working group



High-level concepts



Subclass Hierarchy Tree

- entity
 - physical
 - object
 - process
 - dual object process
 - intentional process
 - intentional psychological process
 - recreation or exercise
 - organizational process
 - guiding
 - keeping
 - maintaining
 - repairing
 - poking
 - content development
 - making
 - constructing
 - manufacture
 - publication
 - cooking
 - searching
 - social interaction
 - maneuver
 - motion
 - internal change
 - shape change
 - abstract

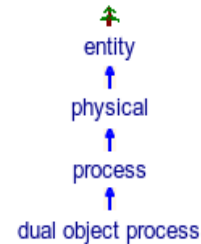
dual object process (DualObjectProcess)

Any **Process** that requires two, nonidentical patients.

Ontology

SUMO / BASE-ONTOLOGY

Superclass(es)



Subclass(es)

substituting transaction comparing attaching detaching combining separating

Coordinate term(s)

intentional process internal change motion shape change

Axioms (1)

If **process** is an instance of dual object process, then there exist **obj1,obj2** so that **obj1** is a patient

```
(=>
  (instance ?PROCESS DualObjectProcess)
  (exists
    (?OBJ1 ?OBJ2)
    (and
      (patient ?PROCESS ?OBJ1)
      (patient ?PROCESS ?OBJ2)
      (not
        (equal ?OBJ1 ?OBJ2))))))
```

UMBEL

- CYC:
 - base of common sense knowledge
 - Douglas Lenat (since 1984)
 - > 1M assertions, rules or common sense ideas
- OpenCYC:
 - 47k concepts and 306k facts
 - available in OWL
- UMBEL:
 - a lightweight reference structure of 20,000 subject concept classes and their relationships derived from OpenCyc
 - can act as binding classes to external data
 - linked to 1.5M named entities from DBpedia and YAGO



Google's Knowledge Graph

- Google Hummingbird (2013)
 - one of many updates of Google's search engine
 - attempts to leverage user context and intention
 - a move towards semantic search
- Google Knowledge Graph
 - seeded from Freebase
 - facts from Wikipedia, Wikidata, CIA World Factbook
 - explicit entities
 - also enriched by natural-language parsing (NLP)
 - implicit entities

Caution: *The public documentation is limited, so this is compiled based on presentations, technical notes, forums etc.*



Google's Knowledge Vault

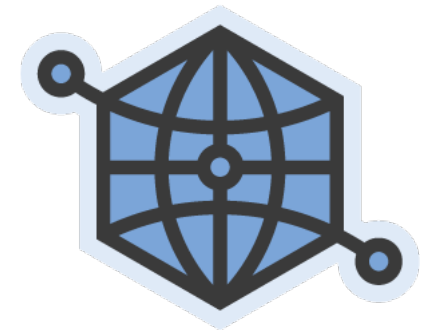
- Google Knowledge Vault
 - extends the Knowledge Graph
 - covers resources not from open semantic datasets
 - facts extracted from the whole web
 - NLP of text documents
 - HTML trees and tables
 - human annotated pages (e.g., schema.org)
 - probabilistic reasoning
 - graph-based priors
 - knowledge fusion

Caution: *The public documentation is limited, so this is compiled based on presentations, technical notes, forums etc.*



Facebook's “Open” Graph Protocol (OGP)

- Including resources (in particular web pages), through their IRIs, in social graphs
 - targetting webmasters and content-management system (CMS) developers
- @prefix og: <<http://ogp.me/ns#>>
- Main properties:
 - required: og:title, og:type, og:image, og:url
 - optional: og:audio, og:description, og:determiner, og:locale, og:locale:alternate, og:site_name, og:video
 - ...some of them combines with more specific ones
 - ...markup with *RDFa* <meta>-tags



OGP uses

- Uses:
 - originally developed by Facebook to extend the “Likes” mechanism to resources outside Facebook
 - also taken up by some other graph maintainers (claim: used by Google)
 - publishing side:
 - IMDb, Microsoft, Rotten Tomatoes, Yelp

Caution: *The public documentation is limited, so this is compiled based on presentations, technical notes, forums etc.*



OGP resource types

- `<meta property="og:type" content="ResType" />`
- Some predefined resource types for:
 - music: music.song, music.album, music.playlist...
 - video: video.movie, video.episode, video.tv_show...
 - others: article, book, profile, website
- Each predefined resource type has further type-specific properties, e.g.,
 - music:duration, music:album:track, music:musician
- Data types:
 - boolean, date/time (ISO 8601), enum, float, integer, string, URL



Facebook's Graph API

- Letting external applications access the information in Facebook's social graph
 - inspired by *social networks*
- *Nodes* represent “things”: *User, Photo, Page, Comment*
- *Edges* represent connections between the "things":
 - Users' *friends*, Pages' *photos*, Photos' *comments*...
- *Fields* contain information about the "things":
 - the *birthday* of a User, the *name* of a Page...
- *Seriously restricted since version 2.0... (Privacy!)*
 - *the idea remains important*
 - *open, user-owned alternatives are emerging*
 - *GNU social (StatusNet), Diaspora...*



Facebook Graph API

- *REST*-based (REpresentational State Transfer)
 - an example of a *web service*
 - all nodes have IRIs
 - GET, POST, DELETE over HTTP
- GET `graph.facebook.com/facebook/picture?redirect=false`
 - this is sent over HTTP (at least):
GET `/facebook/picture?redirect=false` HTTP/1.1
Host: `graph.facebook.com`
- Many API operations are based on *access tokens*
 - returned by *Facebook login*
 - mandatory for POST and DELETE
 - *friends' information must be explicitly granted*



Facebook Graph API

- Most HTTP-requests go to:
 - <http://graph.facebook.com/...>
 - <http://graph-images.facebook.com/...>
- Node paths:
 - **GET** graph.facebook.com/{node-id}
- Edge paths:
 - **GET** graph.facebook.com/{node-id}/{edge-name}
- With access token:
 - **GET** graph.facebook.com/me
- **POST** and **DELETE** are also used

Try it out: <https://developers.facebook.com/tools/explorer>

