

# INFO216: Advanced Modelling

Theme, spring 2017:  
Modelling and Programming  
the Web of Data

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



# Session S10-S11: Linked Open Data and Semantic Repositories

- Themes:
  - semantic vocabularies (*mostly S09*)
  - semantic repositories and LOD (*S10-S11*)
    - Linked Open Data (LOD)
    - the LOD cloud
    - Open semantic repositories:
      - DBpedia, Wikidata, GeoNames
      - some others (*WordNet, YAGO, SUMO, Facebook OGP, Graph API*)
    - *...some of them have their own vocabularies*



# Terms (→ S09)

- *Semantic vocabularies*
  - graphs/datasets (in RDFS, OWL...) that define:
  - standard IRIs for *types of resources*
  - standard IRIs for *properties*
  - standard types (identified by IRIs) for *literals*
- *Semantic repositories, open semantic datasets*
  - graphs/datasets (in RDF, RDFS, OWL...) that define:
  - standard IRIs for *individual resources*
  - facts (as triples) about those *individual resources*
  - *may* also define their own vocabularies



# Readings

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



# Open semantic datasets



# Places to start (→S02)

- Open and semantic:
  - open semantic data sets: <http://lod-cloud.net>
  - vocabularies: <http://lov.okfn.org/dataset/lov/>
  - statistics and overviews: <http://stats.lod2.eu/>
- Open data in general:
  - internationally: <http://datahub.io>
  - Norge: <http://data.norge.no>
  - EU: <http://publicdata.eu> (and others)
  - Storbritannia: <http://data.gov.uk>
  - USA: <http://data.gov>



# 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*)
    - requests be *SPARQL queries*
  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 Document 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 have 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)*



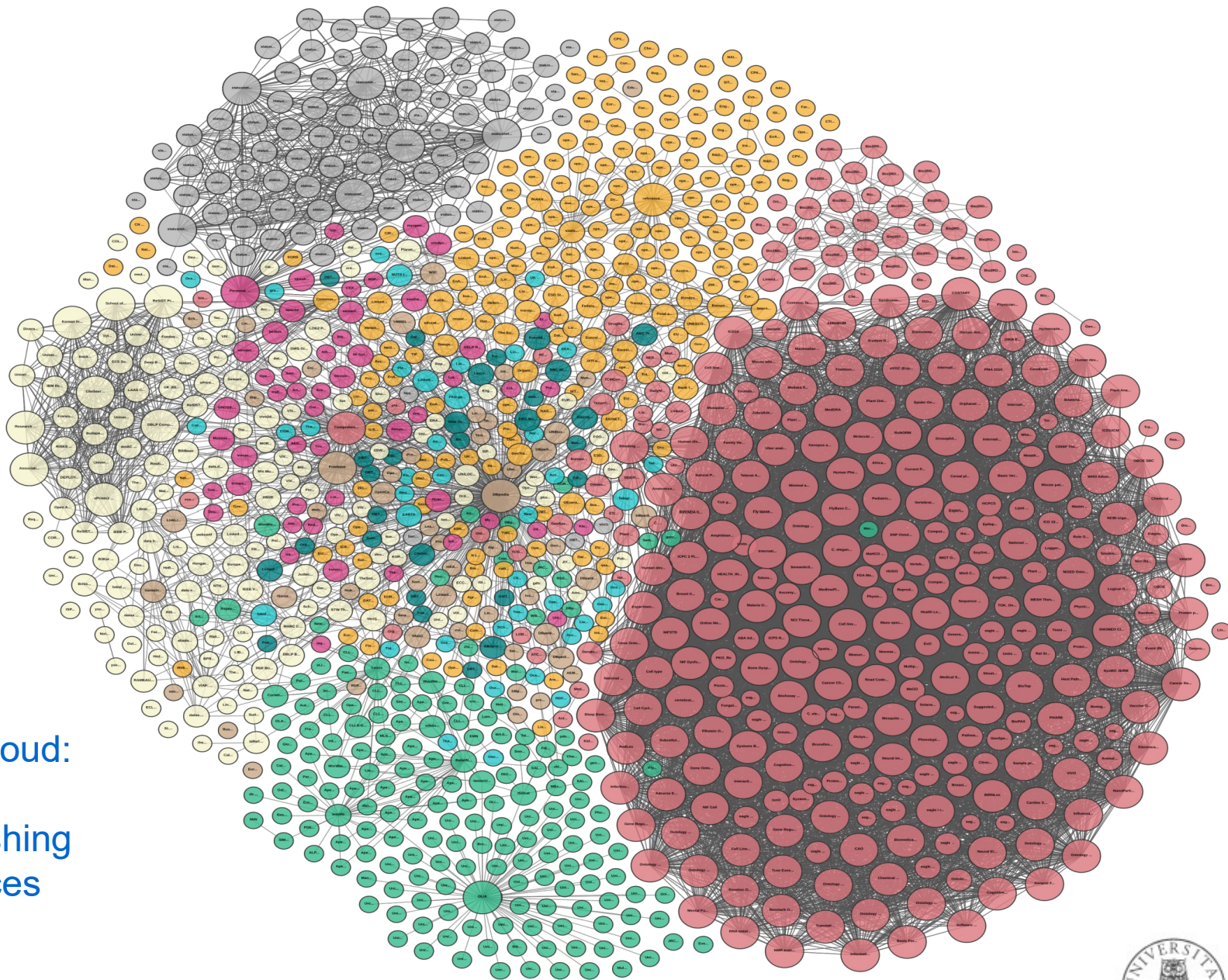
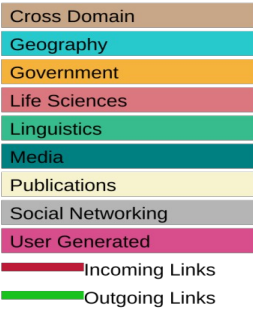


# The LOD cloud

- <http://lod-cloud.net/>
  - 1139 datasets from datahub.io
    - started in 2007, doubled since 2011
    - still growing, but also consolidating
  - graphs shows cross-dataset links:
    - triples with object or subject IRIs that belong to other datasets
  - statistics at [http://lod-cloud.net/state/state\\_2014/](http://lod-cloud.net/state/state_2014/)
- <http://stats.lod2.eu/> is less restrictive
  - 149G (149 423M) triples from 2973 data sets
  - mostly SPARQL endpoints, some from file dumps



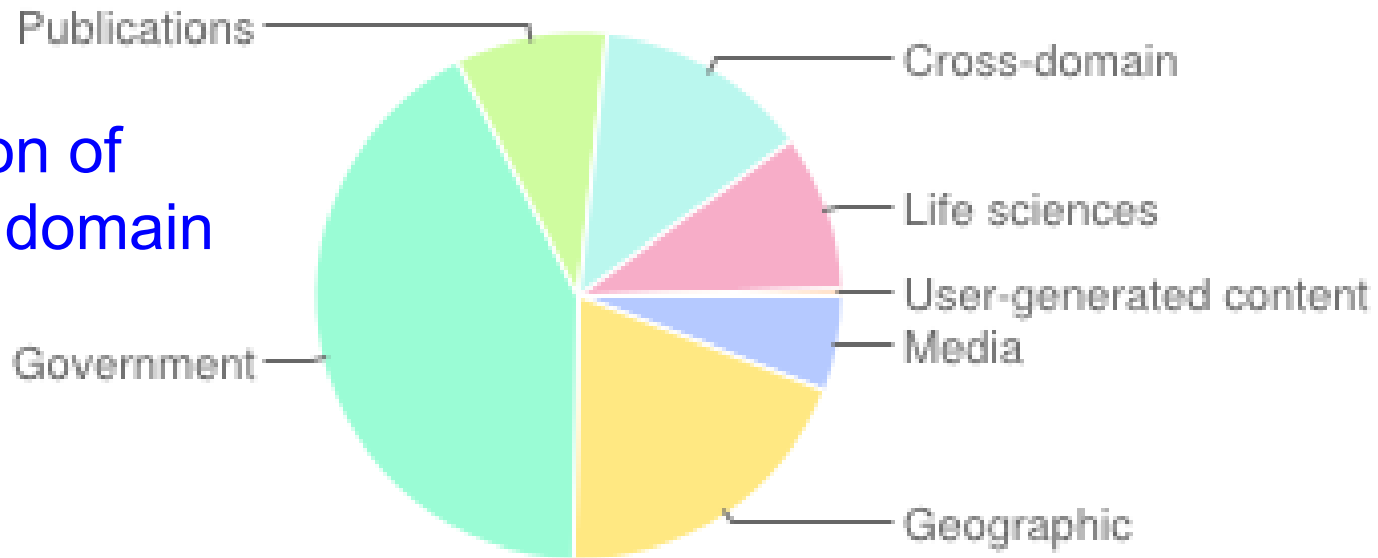
## Legend



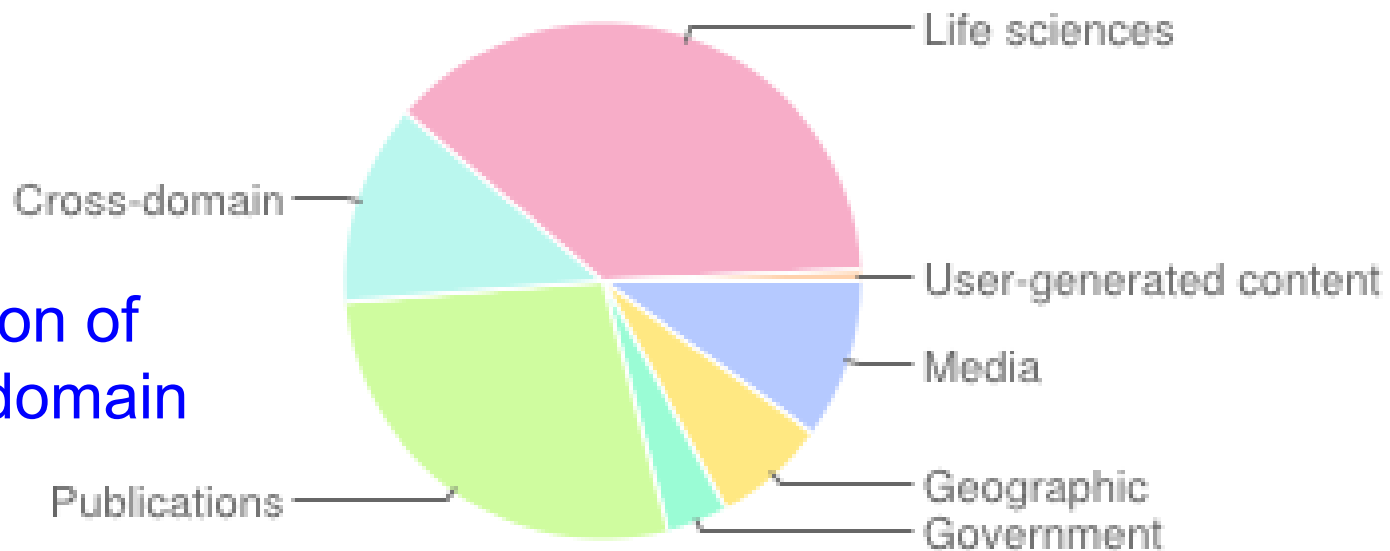
## A “lumpy” cloud:

- general
- sci. publishing
- life sciences
- social
- language

## Distribution of *triples* by domain (2014)



## Distribution of *links* by domain (2014)





# Challenges

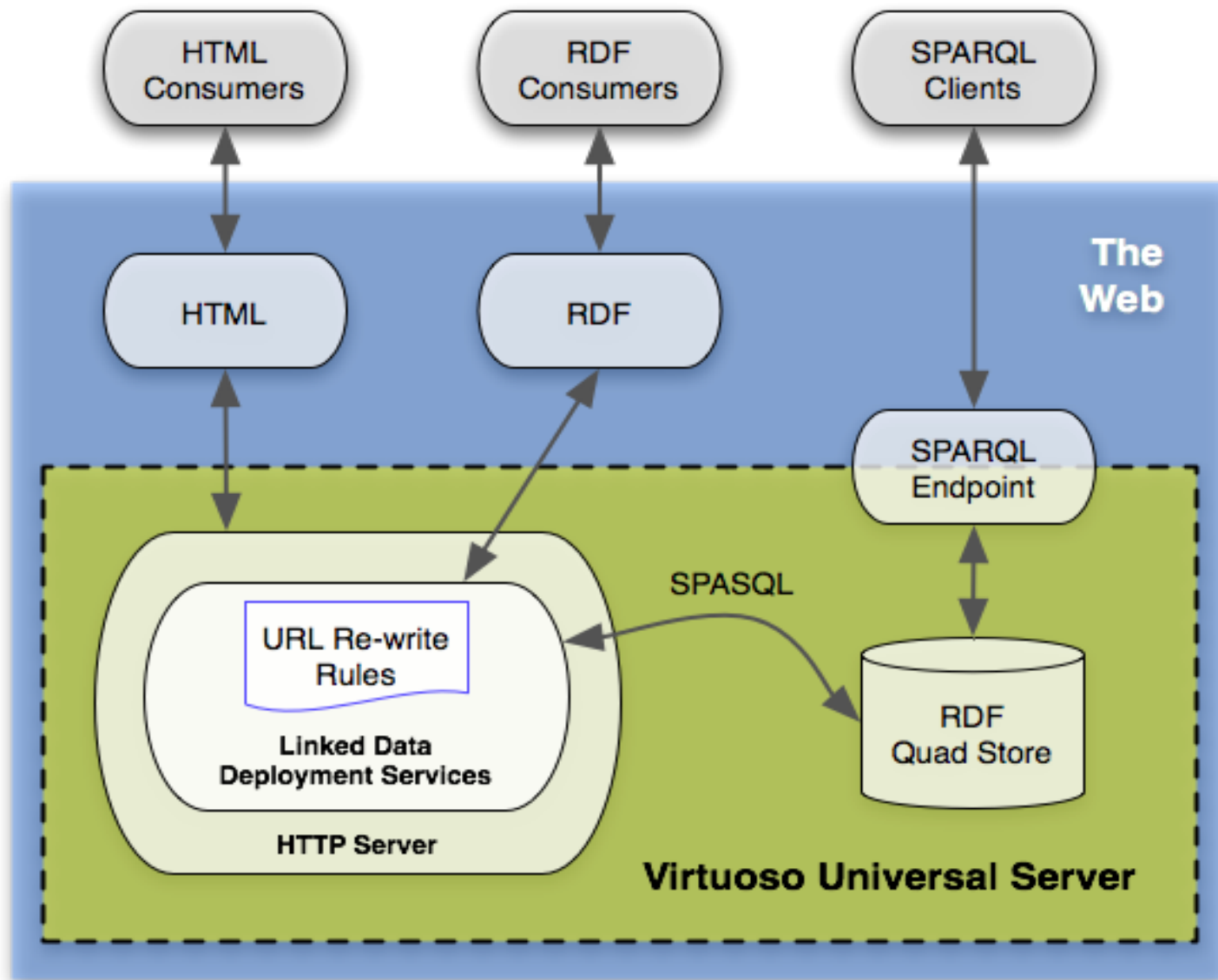
- Semantic technologies and the Web of Data and LOD has an enormous potential
  - ...but is not so much used *in its entirety* so far
  - parts of it are used
    - commercially: biodata, publishing, music/media...
    - publically: clean energy, libraries...
- Possible causes:
  - abstraction: general versus domain data
  - trust: open versus closed networks
  - maintenance: individuals versus organisations
- *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
- Important interlinking-hub for the Web of Data
  - <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 dataset (<http://demo.dbpedia-spotlight.org/>)
  - lexicalizations dataset (maps names to Dbpedia IRIs)





# DBpedia: extraction

- Extracted approximately once a year
  - current version is 2016-04
- Since January 2007:
  - first only in English
  - then 15 largest languages (since 3.7)
    - not Norwegian, but Swedish
  - today all 128 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: 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: 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
    - generic version
    - specific version



# DBpedia: name spaces

- <http://dbpedia.org/> – language-independent base, URIs
- <http://nn.dbpedia.org/> – language-specific base, IRIs
  - 128 languages, not always 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://www.w3.org/2002/07/owl#sameAs) – interlinking, e.g, across languages
- [rdf:type](http://www.w3.org/2000/01/rdf-schema#type) – three classification schemes

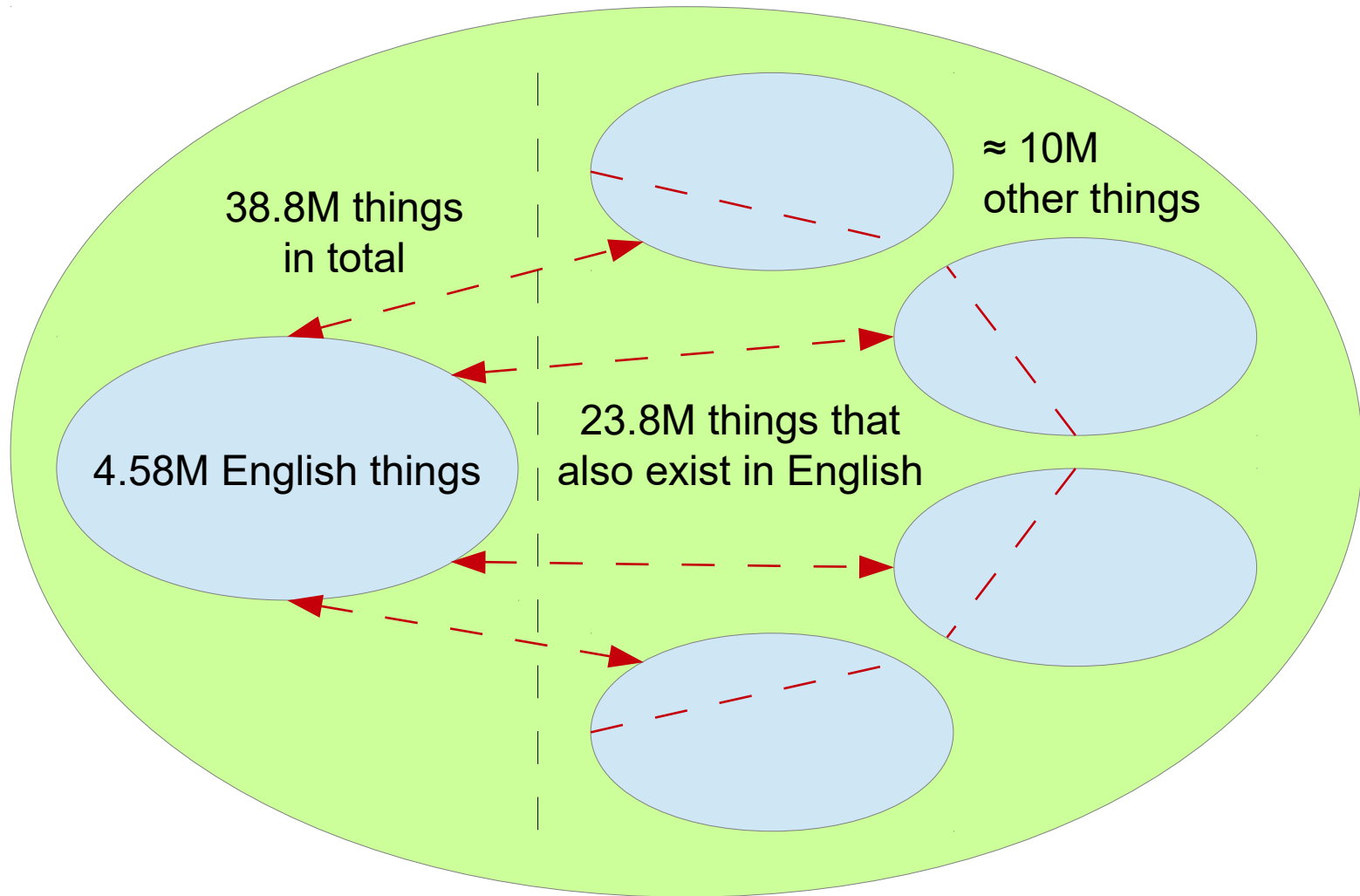


# DBpedia: statistics

- Contents:
  - 38.3M resources (“things”, in 2015)
    - 128 languages
    - 23.8M resources (2015) are localised versions of
    - 6.0M resources from the English Wikipedia
      - 1.5M persons, 810k places, 490k works, 275k organisations, 301k species...
      - also 1.7M SKOS concepts and other stuff



# Canonical and normalised resources



*(Example numbers from 2015-10.)*

# DBpedia: triples

- The full (international) data set:
  - 9.5G triples (*≈ 4% 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 wrappr, UMBEL...



# Dbpedia: advantages

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



# DBpedia: classification schemes

- Wikipedia Categories:
  - *81M links*
  - SKOS vocabulary and DCMI terms
- YAGO Classification:
  - *41M links*
  - Yet Another General/Great Ontology
  - derived from Wikipedia using WordNet (also from GeoNames)
- Word Net Synsets:
  - derived directly from the infoboxes
- *...also 50M other links (30M to web pages)*



# 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
- *...a central information source for Wikidata*





# Wikidata

- *A free and open knowledge base that can be read and edited by both humans and machines*
  - a Wikimedia project, crowdsourced
  - *a Wikipedia for structured data*
  - central storage for the structured data of its Wikimedia sister projects:
    - Wikipedia, Wikivoyage, Wikisource, etc.
  - supports many other sites and services
  - free license, standard formats, interlinked
- Wikidata entities:
  - 30M items (things)
  - 3000 properties

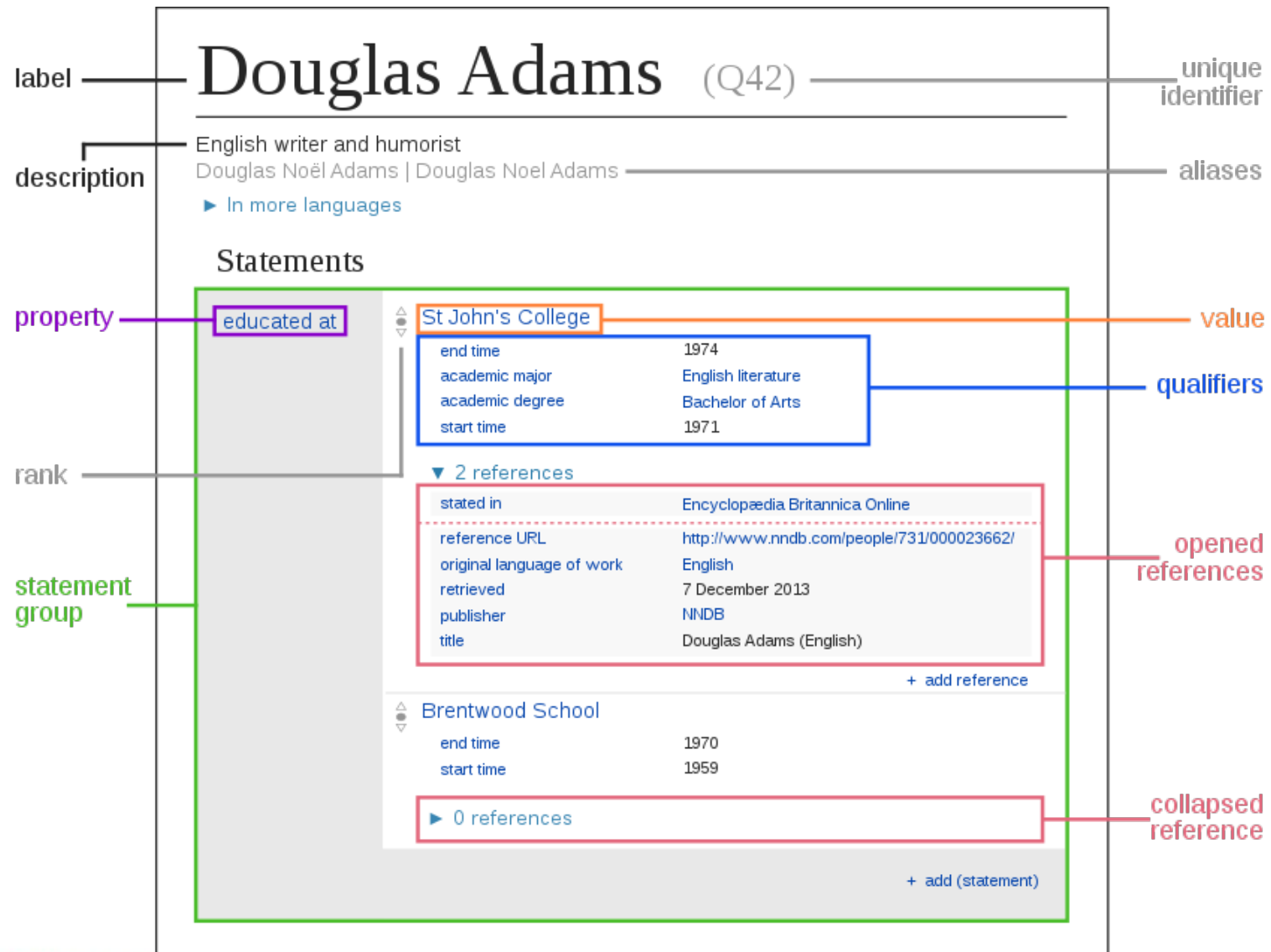


# Wikidata access

- 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)
    - based on BlazeGraph, OpenRDF/RDF4J
  - for download (JSON, RDF, XML)
- DBpedia also offers Wikidata compatible dumps



# Wikidata item structure

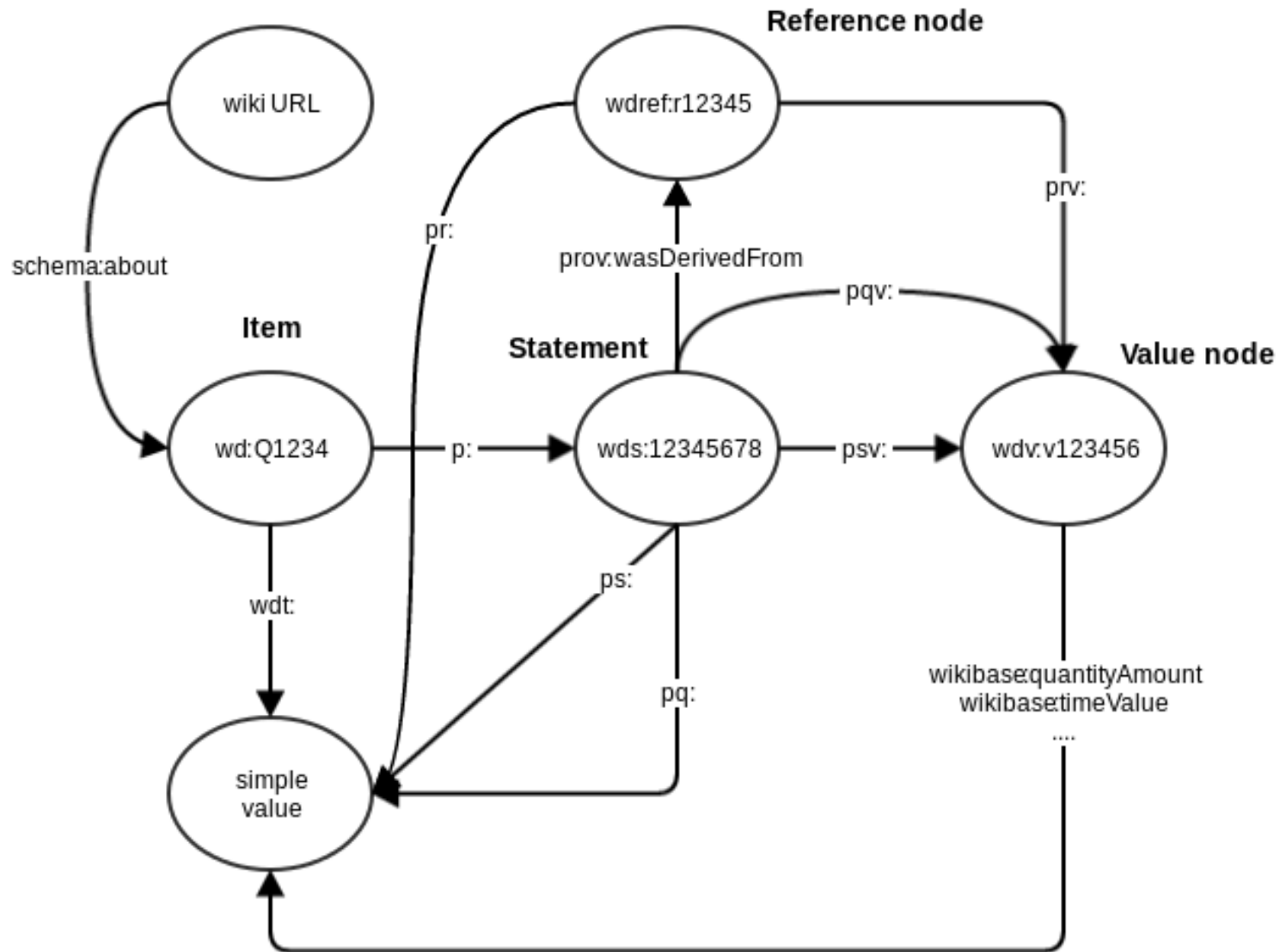


# Wikidata item structure

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



# Wikidata RDF mapping



# Wikidata versus DBpedia

- Similarities:
  - both publish RDF data about entities
  - both use standard IRIs derived from Wikipedia
  - both define ontologies
  - both are linked to other semantic repositories
- Differences:
  - source: DBpedia is derived; Wikidata is crowdsourced
  - direction: DBpedia extracts data from Wikipedia; Wikidata provides data to Wikipedia.
  - structure: DBpedia structures Wikipedia data; Wikidata is natively structured
  - maturity: DBpedia is older; Wikidata is getting started



# GeoNames ([www.geonames.org](http://www.geonames.org))

- Geographical database:
  - collected from a large number of sources
  - > 10M geographical names (Norway 68k)
  - > 8M unique features
  - ~ 2.8M populated places
  - ~ 5.5M alternate names
- Available as:
  - HTML pages (POW – “Plain Old Web”), map-based
  - web services (REST, RDF), SPARQL endpoints
  - downloadable (TSV)
  - dereferencable IRIs
  - Gazetteer lists



# GeoNames ontology

- Supports adding geospatial semantic information to the POW and WoD
- Provides dereferencable IRIs for all *toponyms*
  - “*303 redirection*” for *Concept-Document distinction*
- Vocabulary in OWL:
  - 9 top-level feature classes:
    - **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)
  - also uses: geo, foaf, dcterms, cc, rdfs



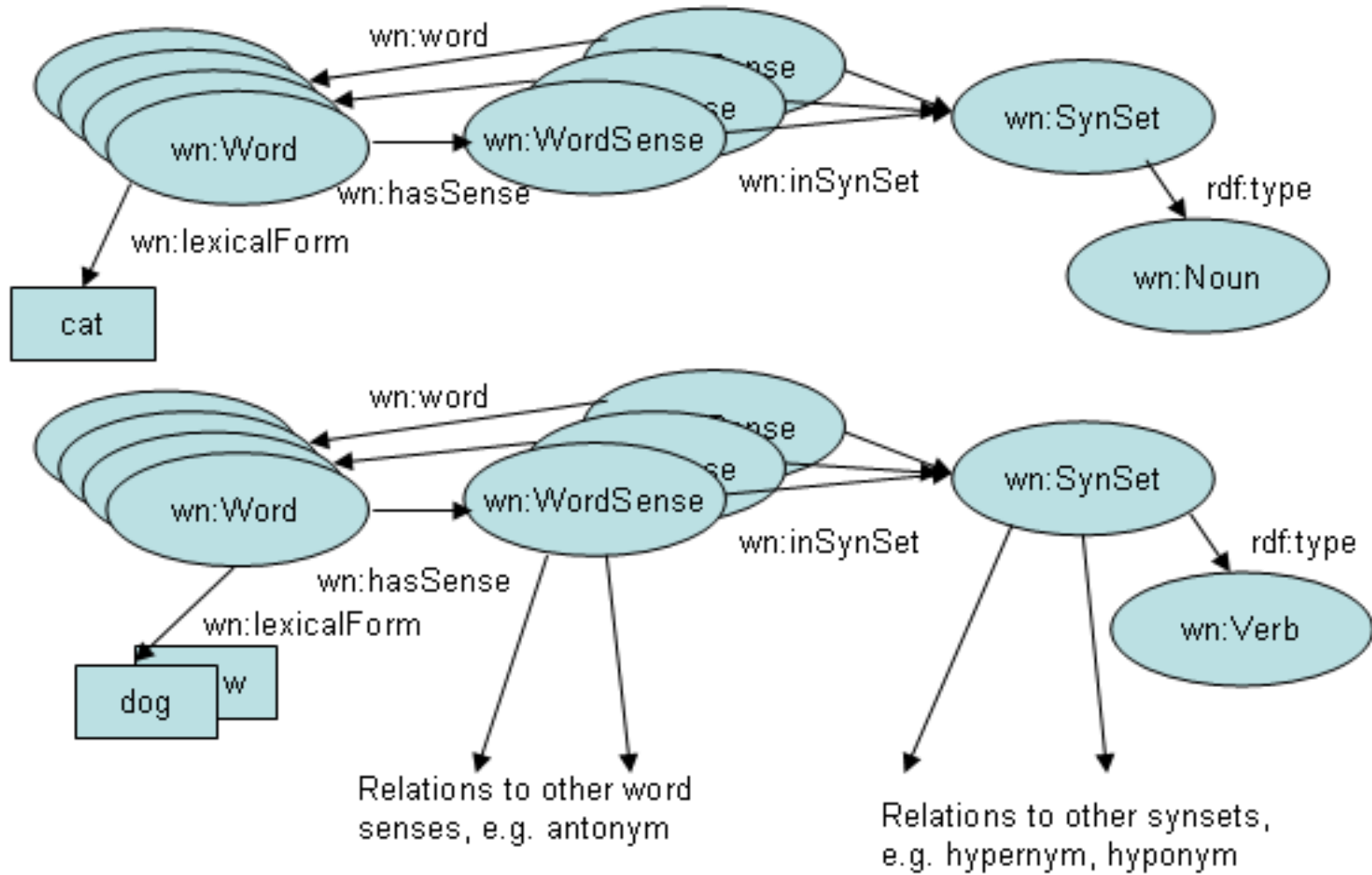


# WordNet

- An electronic open-source dictionary:
  - George A Miller (from 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
  - RDFS and OWL versions exist
    - WordNet in RDF:  
<https://www.w3.org/TR/wordnet-rdf/>
  - also versions for other languages



# WordNet: Structure



# WordNet: Organisation

- Organisation of synsets:
  - hypernyms/hyponyms (direct/full), troponyms:
    - *bat-n-1 is-kind-of placental\_mammal-n-1*
  - holonyms/meronyms
    - *bat-n-1 has-part wing-n-1*
  - antonyms
    - *birth-n-1 has-antonym death-n-1*
  - similarity, sister terms
    - *bat-v-1 has-sister SWAT-v-1*
  - entailment, domains
    - *bat-n-2 has-domain baseball-n-1*
- ...separate organisation per Part of Speech (PoS)

# WordNet: Synset IRIs

- @prefix wn20schema:  
<<http://www.w3.org/2006/03/wn/wn20/schema/>> .
- @prefix wn30:  
<<http://purl.org/vocabularies/princeton/wn30/>> .
- Example:
  - wn30:synset-bank-noun-2



# 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



# Other language resources

- Global Wordnet Grid:
  - <http://globalwordnet.org/>
  - building a Global Multilingual Wordnet  
<http://compling.hss.ntu.edu.sg/omw/>
- Dbnary:
  - <http://kaiko.getalp.org/about-dbnary/>
  - “DBpedia for Wiktionary”
  - automatic extraction of RDF graphs from Wiktionary
- BabelNet:
  - <http://babelnet.org>
  - multilingual text analysis and translation
- *...and many other :-)*



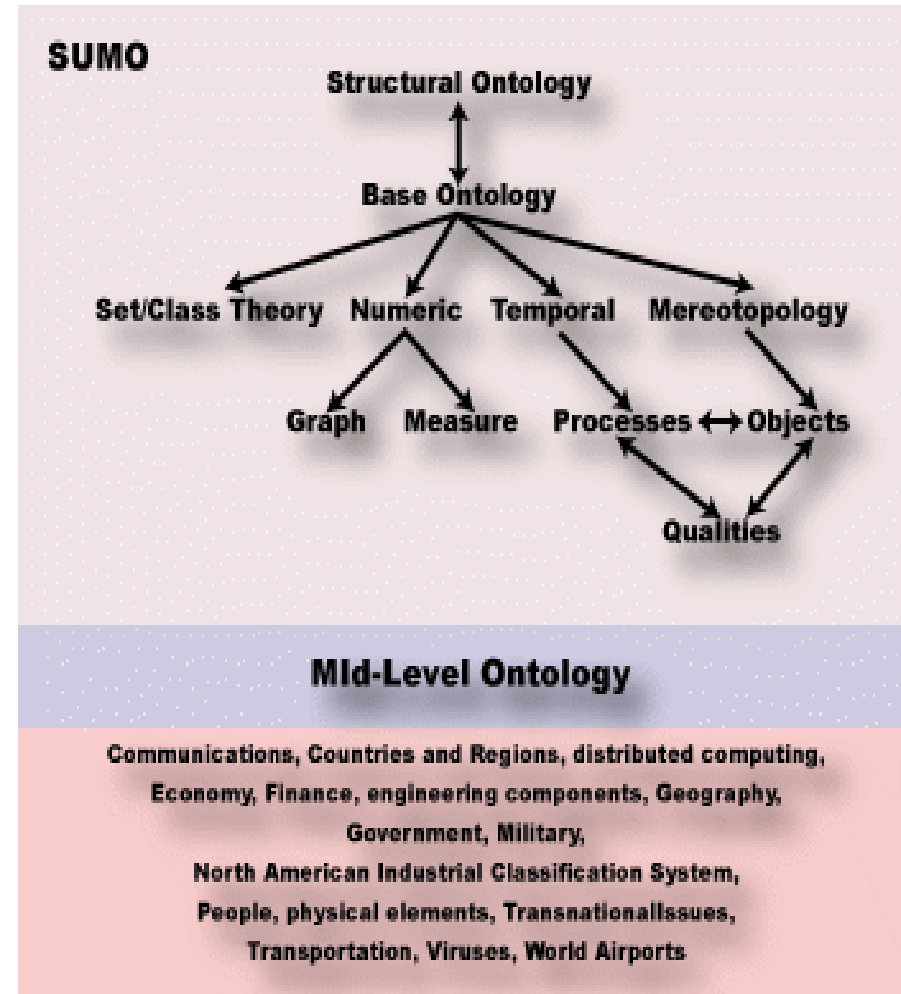
# YAGO2

- *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/>



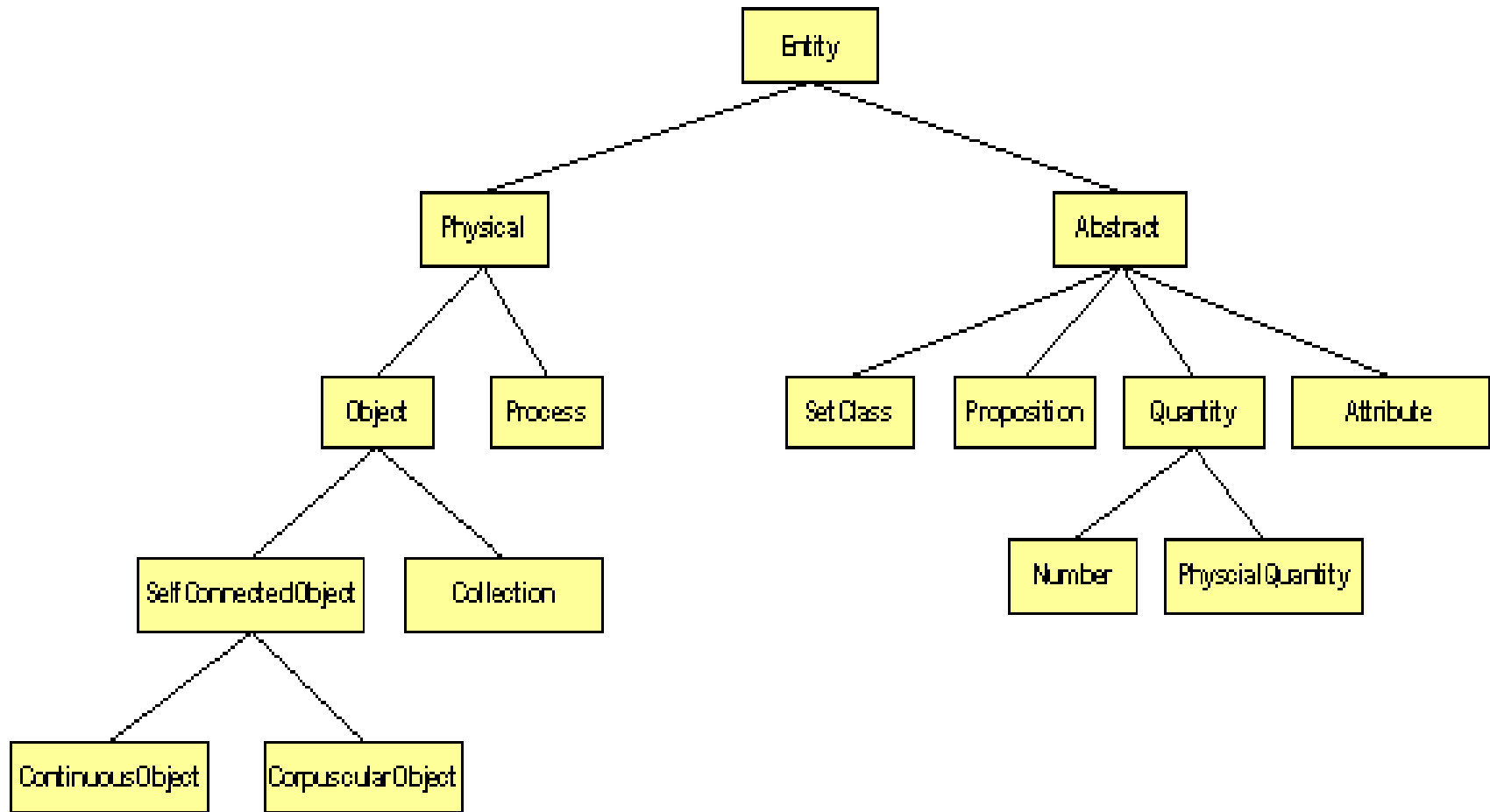
# Suggested Upper Merged Ontology (SUMO)

- Defines and organises concepts formally and philosophically:
  - 25k terms
  - 80k axioms
- WordNet mapping
- Links to DBpedia, YAGO
- Available in OWL
- 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

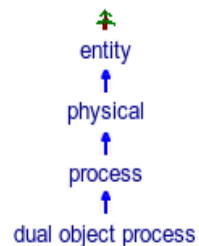
## dual object process (DualObjectProcess)

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



# Facebook “Open” Graph Protocol (OGP)

- Including IRIs (in particular web pages) in social graphs
  - also letting webmasters describe their resources as part of social graphs
- @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
- *Uses:*
  - *originally developed by Facebook to extend the “Likes” mechanism to resources outside Facebook*

# Facebook “Open” Graph Protocol (OGP)

- Including IRIs (in particular web pages) in social graphs
  - also letting webmasters describe their resources as part of social graphs
- @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
- *Uses:*
  - *originally developed by Facebook to extend the “Likes” mechanism to resources outside Facebook*

# 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



# OGP data types

- Boolean
- Date/Time (ISO 8601)
- Enum
- Float
- Integer
- String
- URL



# Facebook 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":
  - a Page's Photos, a Photo's Comments...
- *Fields* contain information about the "things":
  - the birthday of a User, the name of a Page...
- *Seriously limited since version 2.0... (Privacy!)*
  - *the idea remains important*
  - *open, user-owned alternatives are emerging*
    - *StatusNet, Diaspora and others*





# 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](http://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}](http://graph.facebook.com/{node-id})
- Edge paths:
  - **GET** [graph.facebook.com/{node-id}/{edge-name}](http://graph.facebook.com/{node-id}/{edge-name})
- With access token:
  - **GET** [graph.facebook.com/me](http://graph.facebook.com/me)
- **POST** and **DELETE** are also used

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

