

Welcome to INFO216:
Knowledge Graphs
Spring 2022

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

Session 5: Open KGs

- Themes:
 - Linked Open Data (LOD)
 - LOD cloud
 - *Open Knowledge Graphs:*
 - Wikidata, DBpedia, GDELT project, EventKG
GeoNames, WordNet, BabelNet
 - perhaps ConceptNet, DUL (DOLCE UltraLite)...
 - *...some of them have their own vocabularies*
 - Enterprise Knowledge Graphs (EKGs) (→ S06)
 - Ontologies and vocabularies (→ S08 and S09)



Readings

- Resources in the portal, including:
 - Wikidata
 - DBpedia
 - the GDELT project
 - EventKG
 - GeoNames
 - WordNet
 - BabelNet
 - and perhaps others

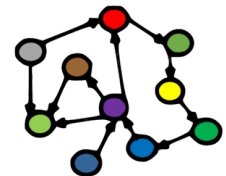


**THE KNOWLEDGE GRAPH
COOKBOOK**
RECIPES THAT WORK

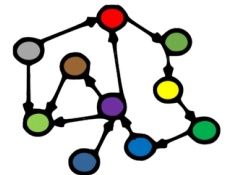


**ANDREAS BLUMAUER
AND HELMUT NAGY**

1st edition, 2020

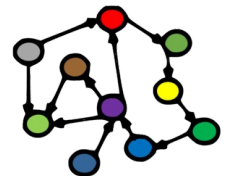


Linked Open Data (LOD)



Places to start

- Open and semantic:
 - open semantic data sets: <http://lod-cloud.net>
 - vocabularies: <https://lov.linkeddata.es/dataset/lov/>
- Open data in general:
 - internationally: <http://datahub.io> or <http://ckan.org>
 - Norge: <http://data.norge.no>
 - EU: <https://data.europa.eu/en>
 - UK: <http://data.gov.uk>
 - USA: <http://data.gov>



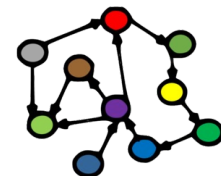
Linked Open Data (LOD)

- *Four basic principles (Berners-Lee 2006):*
 - 1 *URIs (Uniform Resource Identifier)*
 - *identify resources* <http://dbpedia.org/resource/Bergen>
 - 2 *URIs 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)
 - 4 *The information contains URI-s that identify related resources*



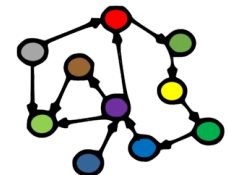
Best Practices for Data Provisioning

- Recommended directly by W3C
 - or emerged within the LOD community:
 1. *Provide dereferencable URIs*
 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)*



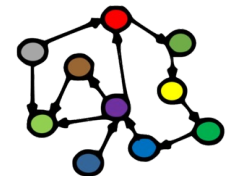
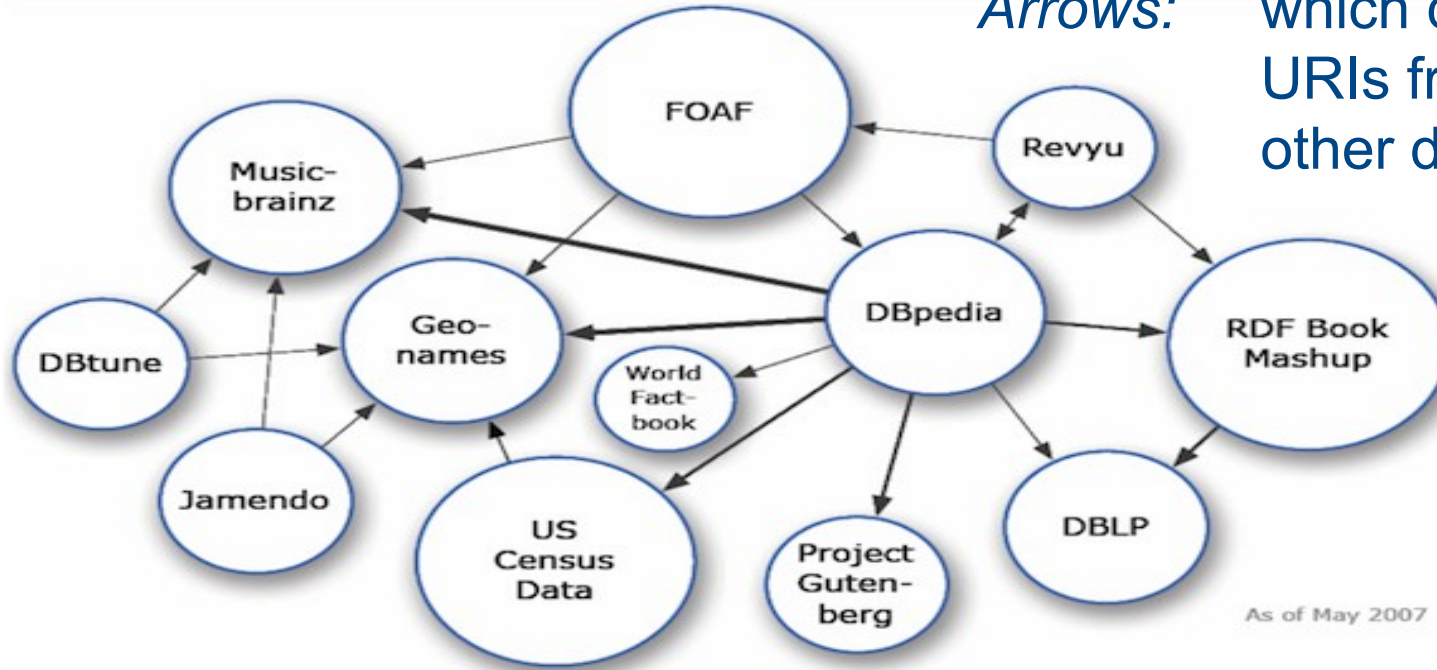
The LOD cloud

- <http://www.lod-cloud.net/>
 - started in 2007
 - exponential-like growth for a few years
 - still growing, but more slowly now
 - statistics at www.lod-cloud.net/state
 - >1300 data sources (February 2022)
 - based on data from DataHub (+ some crawling)
 - datahub.io or ckan.org
 - an open data portal
 - not necessarily semantic
 - ...also based on LOD crawling

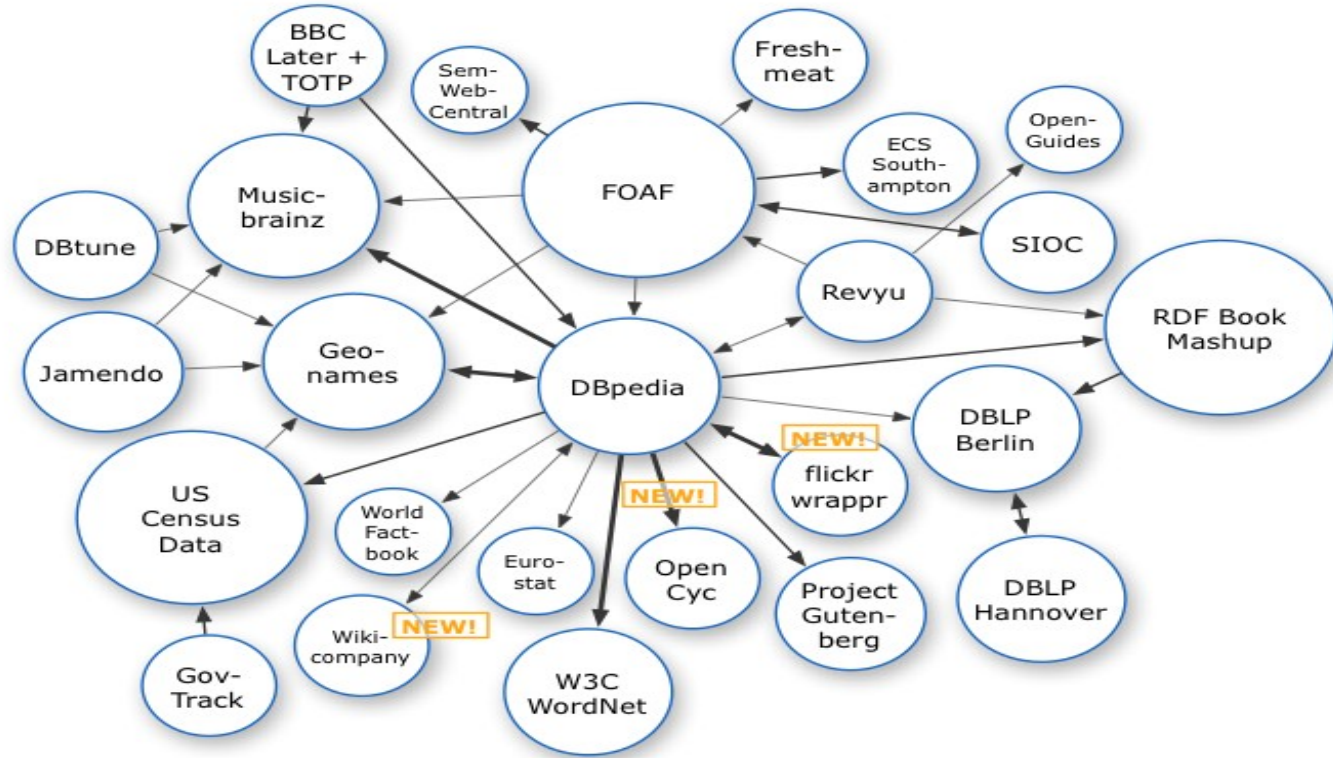


The LOD cloud (2007-05)

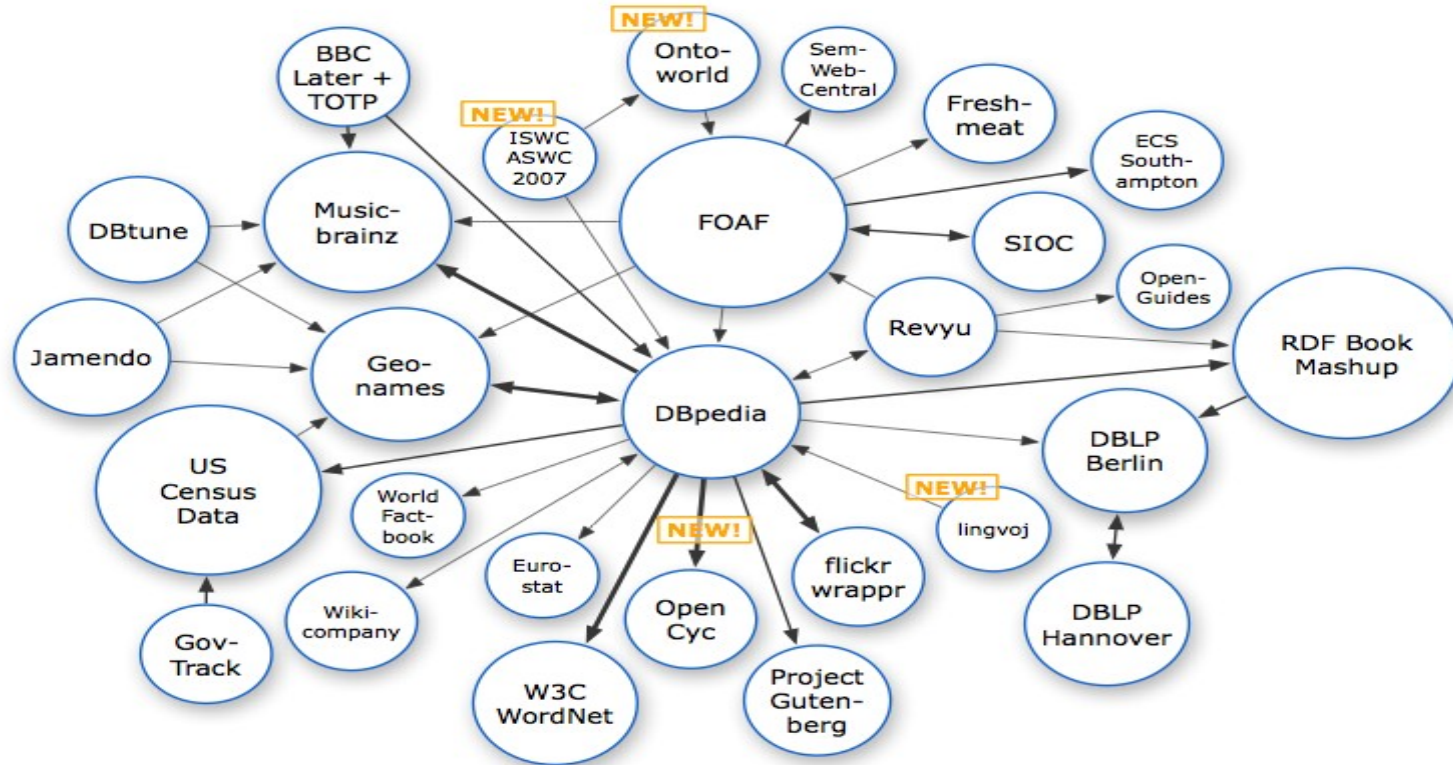
Nodes: semantic datasets
Arrows: which datasets use URIs from which other datasets



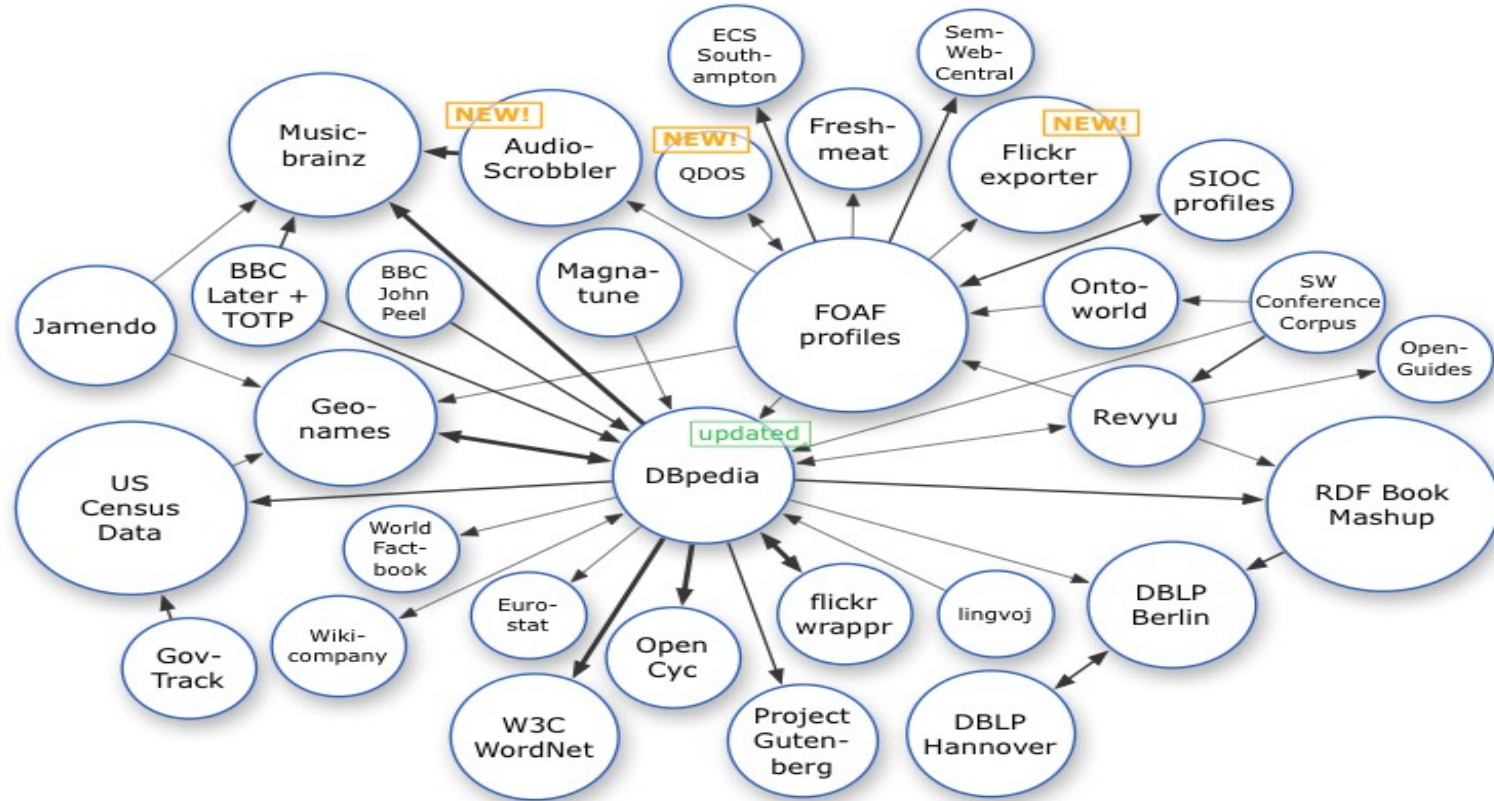
The LOD cloud (2007-10)



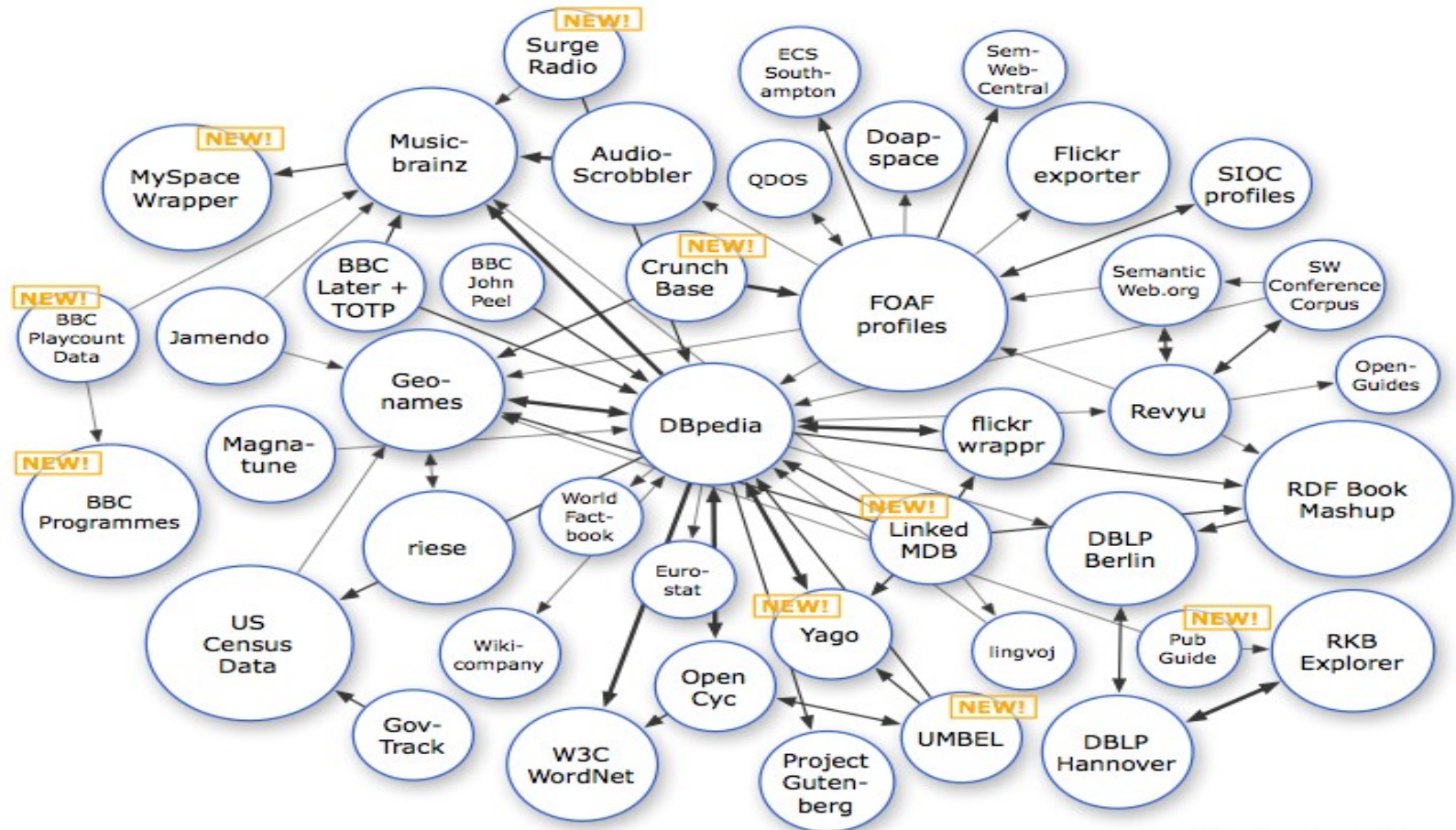
The LOD cloud (2007-11)



The LOD cloud (2008-02)

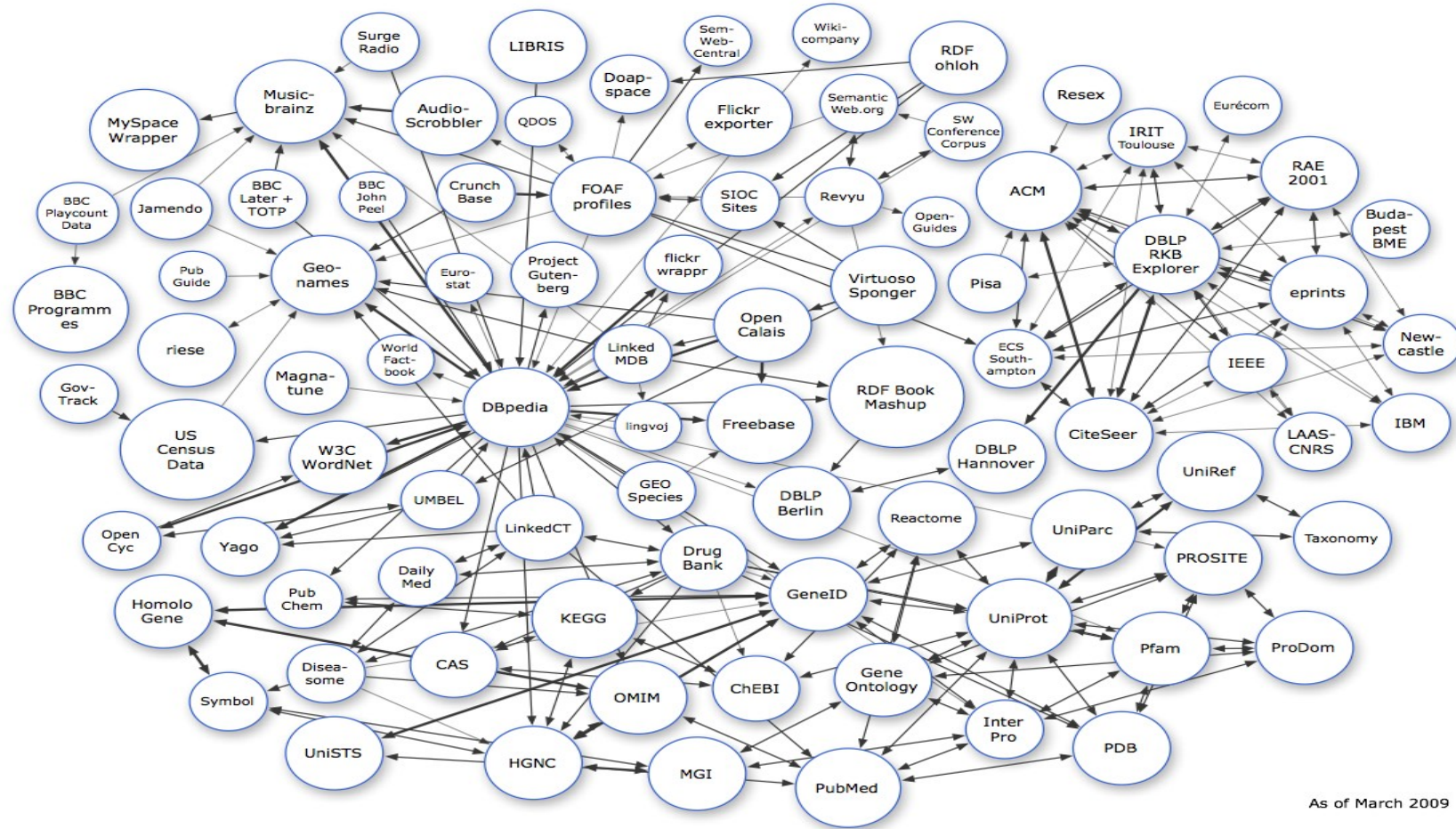


The LOD cloud (2008-09)



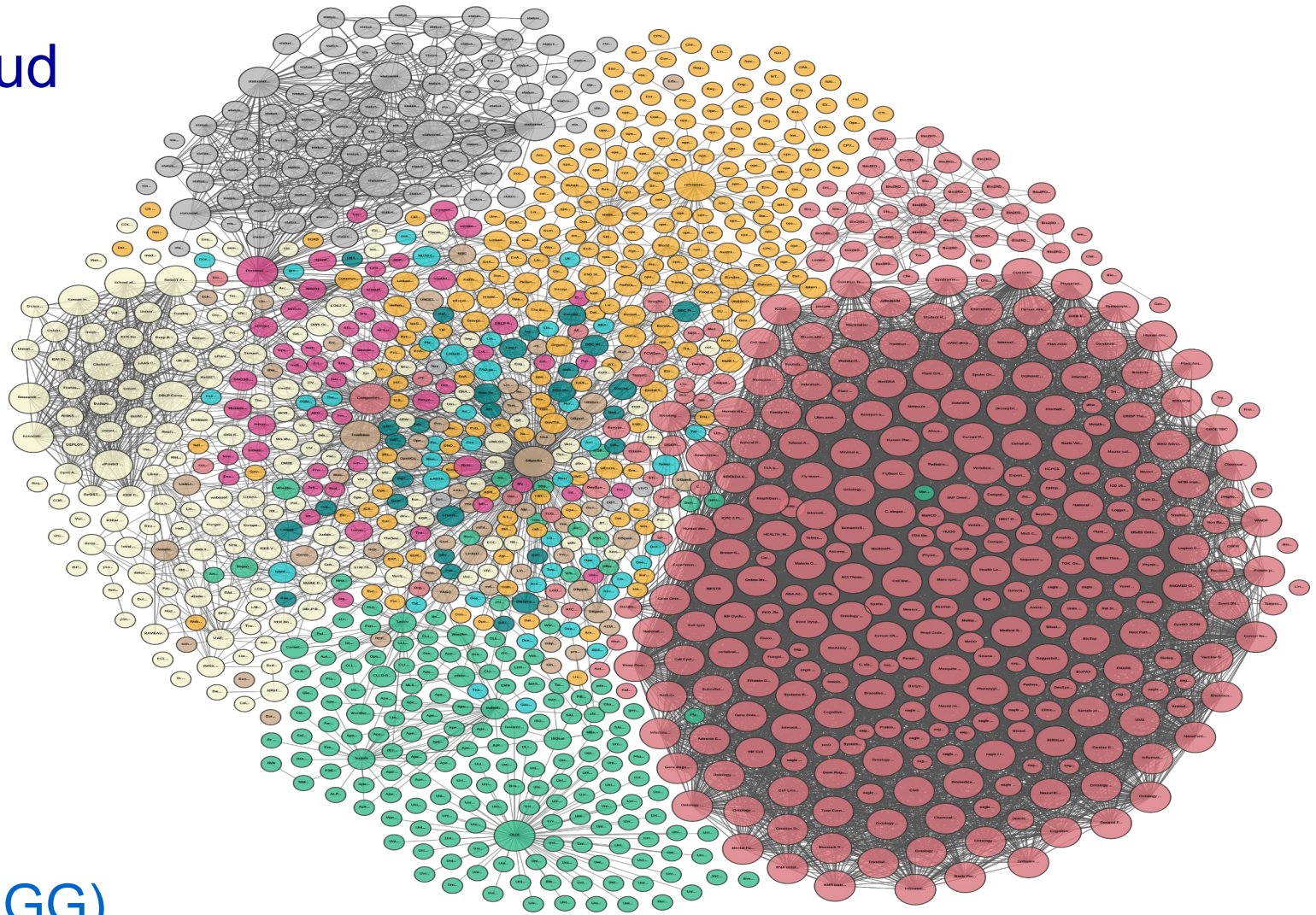
As of September 2008

The LOD cloud (2009-03)



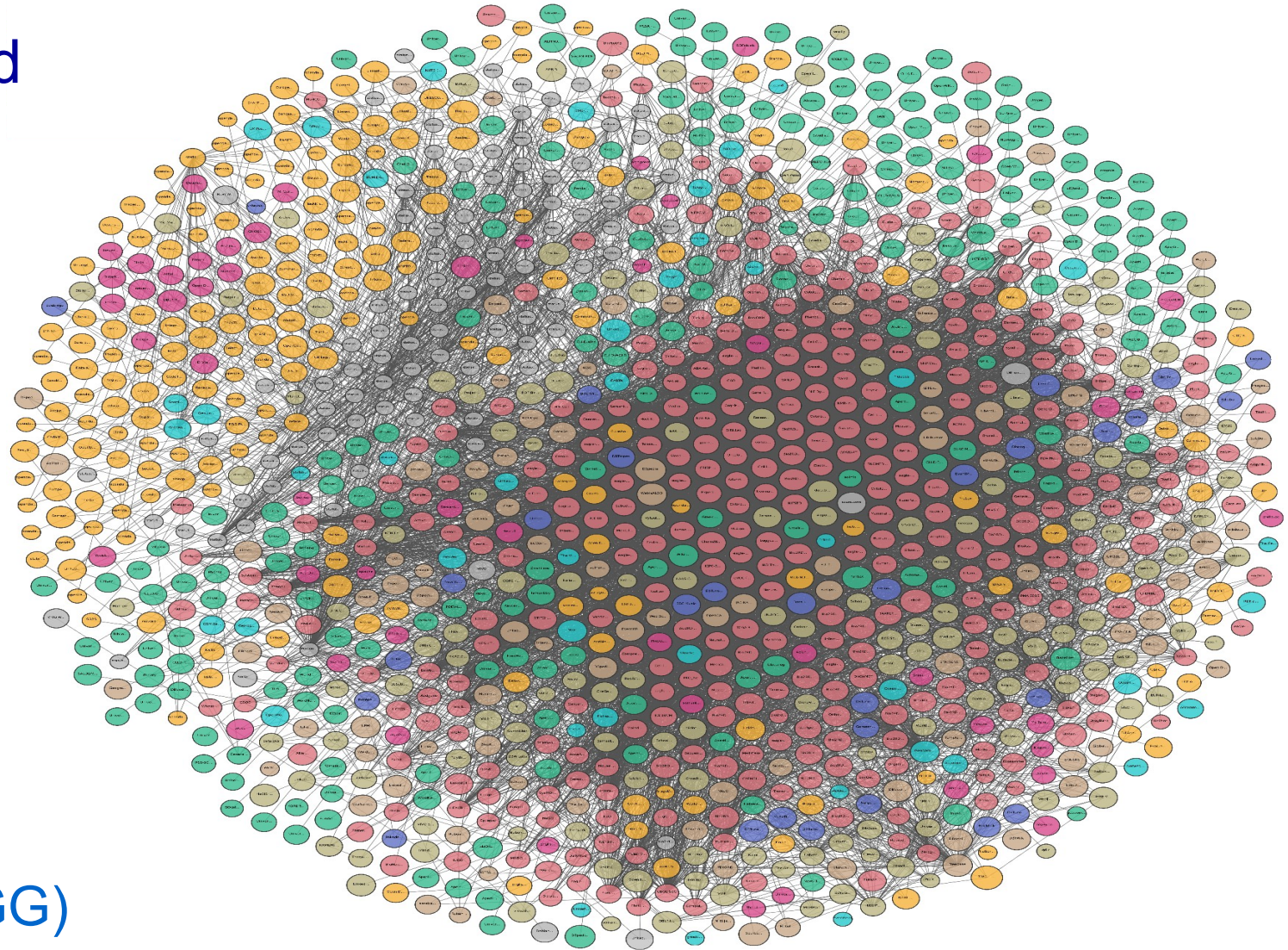
As of March 2009

The LOD cloud (2017-02)



The “Gigantic Global Graph” (GGG)

The LOD cloud (2018-07)



The “Gigantic Global Graph” (GGG)

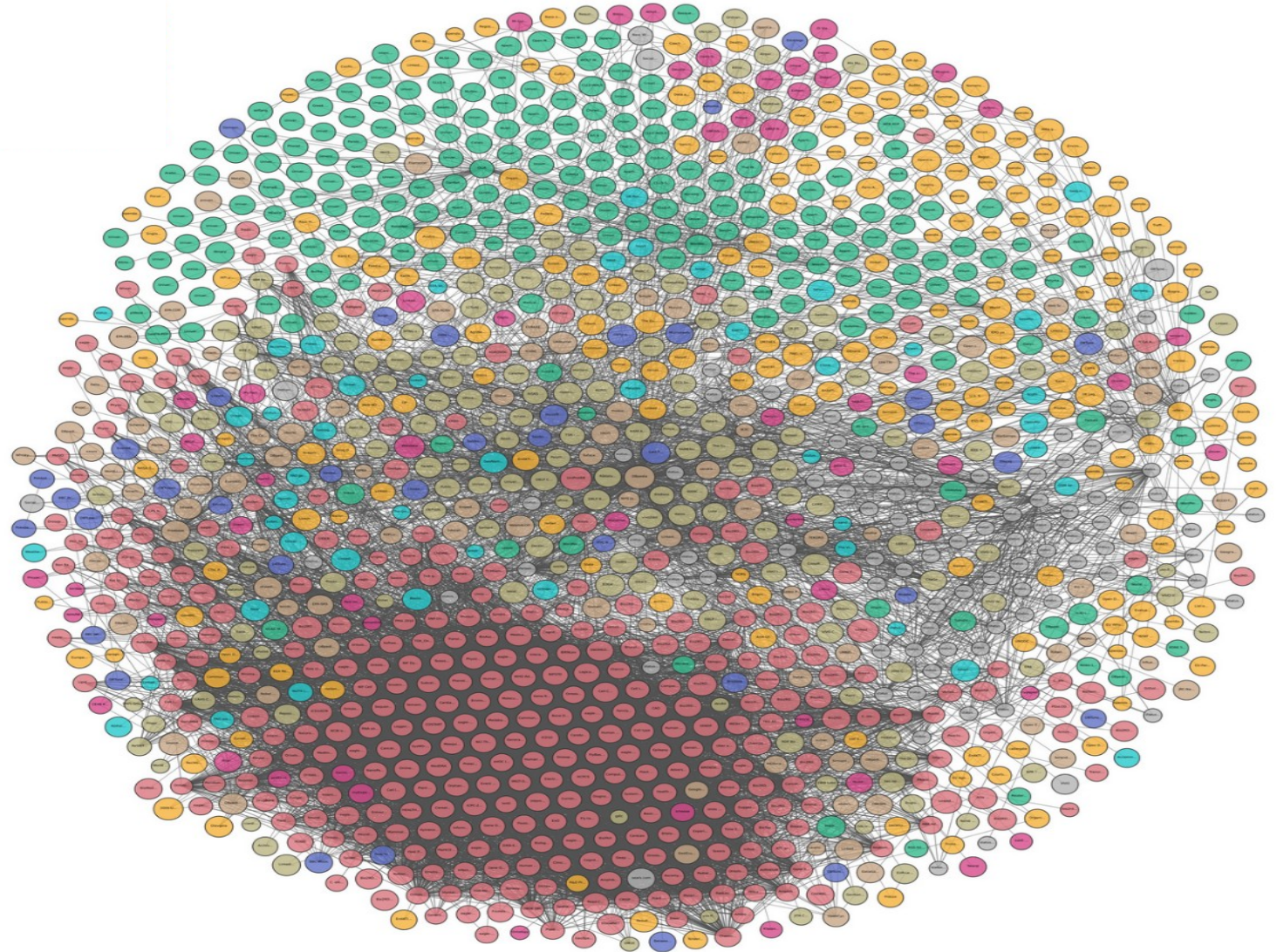
Legend

Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated

Domains:

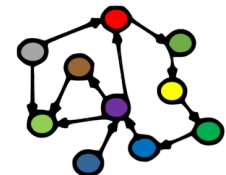
- general
- geography
- government
- life sciences
- linguistics
- media
- publications
- social networking
- user generated

A “lumpy cloud”



Challenges

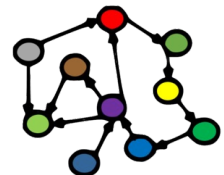
- Semantic technologies / Web of Data / LOD has enormous potential
 - Enterprise Knowledge Graphs are maturing
 - Google, Amazon, BBC, Reuters...
 - industry: biodata, publishing, music/media...
 - government: clean energy, libraries...
 - *“lumps” in the LOD cloud form domain-specific and more tightly-knit subnetworks around EKGs*
 - ...but Open Knowledge Graphs are not used to their fullest
 - *maintenance*: individuals versus organisations
 - *abstraction*: general versus domain data
 - *trust*: open versus closed networks



Open Knowledge Graphs

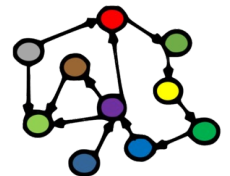
Wikidata (→ S01)

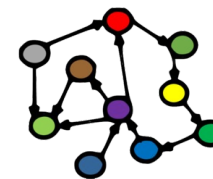
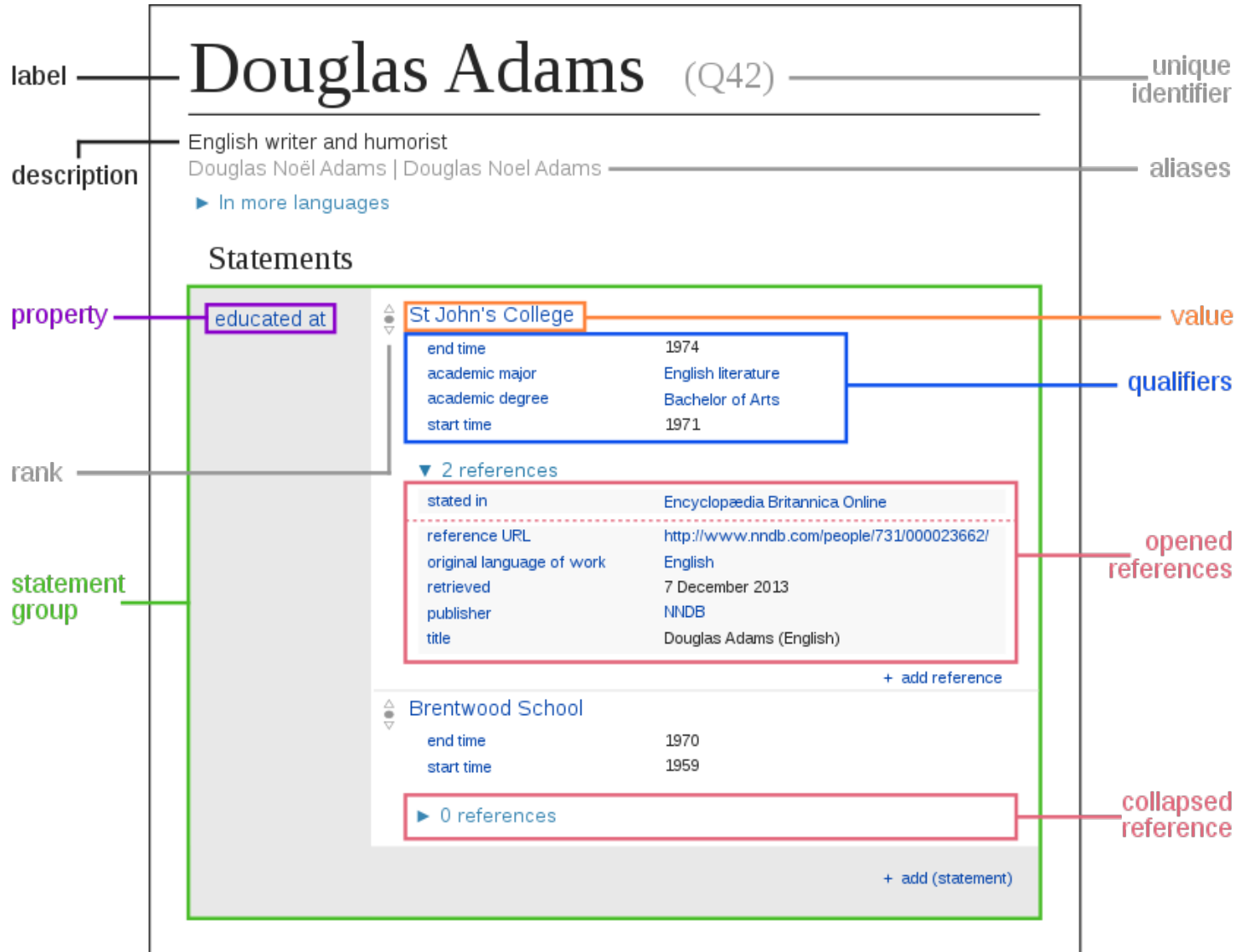
- *A free and open knowledge base that can be read and edited by both humans and machines*
 - a Wikimedia project, crowdsourced, multi-lingual
 - *a Wikipedia for structured, secondary data*
 - from managing Wikipedias *cross-language links*
 - to *central storage of structured data* for
 - Wikimedia sister projects (Wikipedia etc.)
 - many other projects, sites and services
 - verifiability, link to sources, perspectives
 - free license (CC0), standard formats, interlinked
- Wikidata entities:
 - >96M items (things), ≈1.4G statements (“primary” triples)



Wikidata access

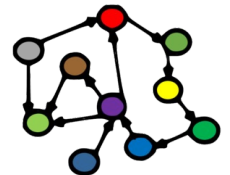
- Available through
 - the Wikimedia API
 - HTTP: <http://www.wikidata.org/entity/Q42>
 - RDF: <http://www.wikidata.org/entity/Q42.ttl>
 - SPARQL endpoint: <http://query.wikidata.org>
 - Wikidata Query Service (WDQS)
 - for download (JSON, RDF, XML)
- Also as Triple Pattern Fragments:
 - <https://query.wikidata.org/bigdata/ldf>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

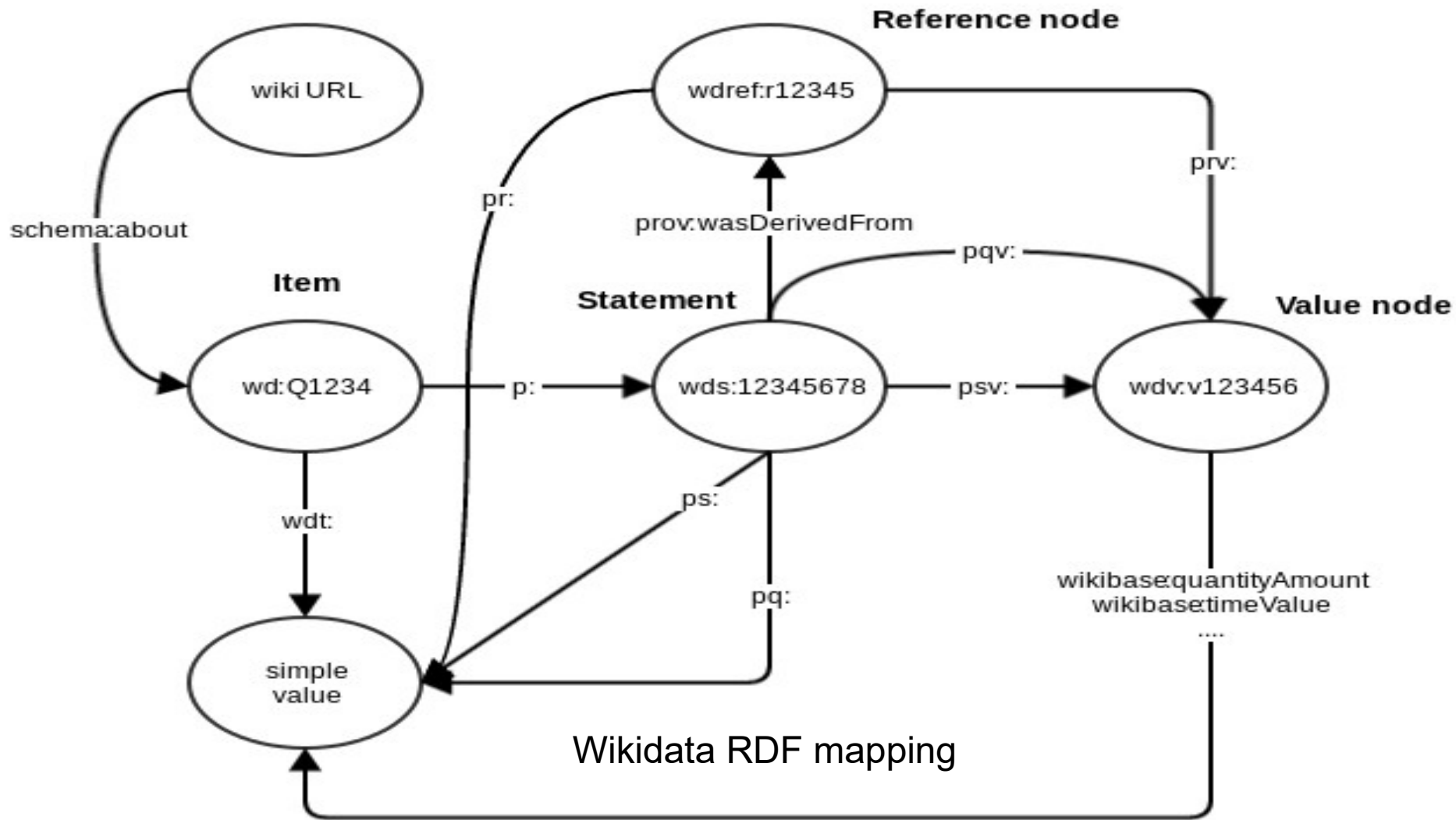




Wikidata item structure

- 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!*





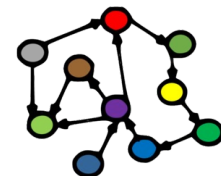
```
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
#defaultView:BubbleChart
```

```
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" .
  }
}
```



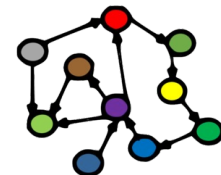
Wikidata Query Service (WDQS)

- SPARQL wrapper for Wikidata (<http://query.wikidata.org>)
 - based on Blazegraph, OpenRDF/RDF4J
 - built-in prefixes
 - generate query URIs
 - 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>)



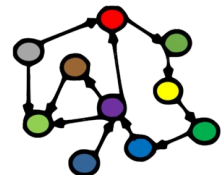
WDQS visualisations

- 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



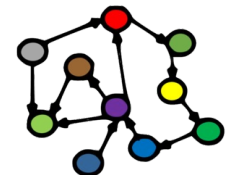
DBpedia

- Extracting structured information from Wikipedia
 - a crowd-sourced community effort
 - making this information available on the Web of Data
- Central source of individual URIs:
 - <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 URIs)



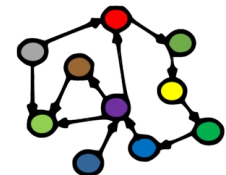
DBpedia: old extraction

- Since January 2007:
 - first only in English
 - the 15 largest languages (since 3.7)
 - around 125 languages (since 3.8)
 - Wikipedia's *infoboxes* are central, but also
 - inter-language links, redirects, disambiguation pages, categories, links to external pages
 - ...also full-text extraction and some NL parsing
 - triple version + quad version with *provenance*



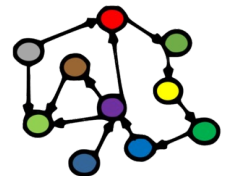
DBpedia: raw and mapped extraction

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

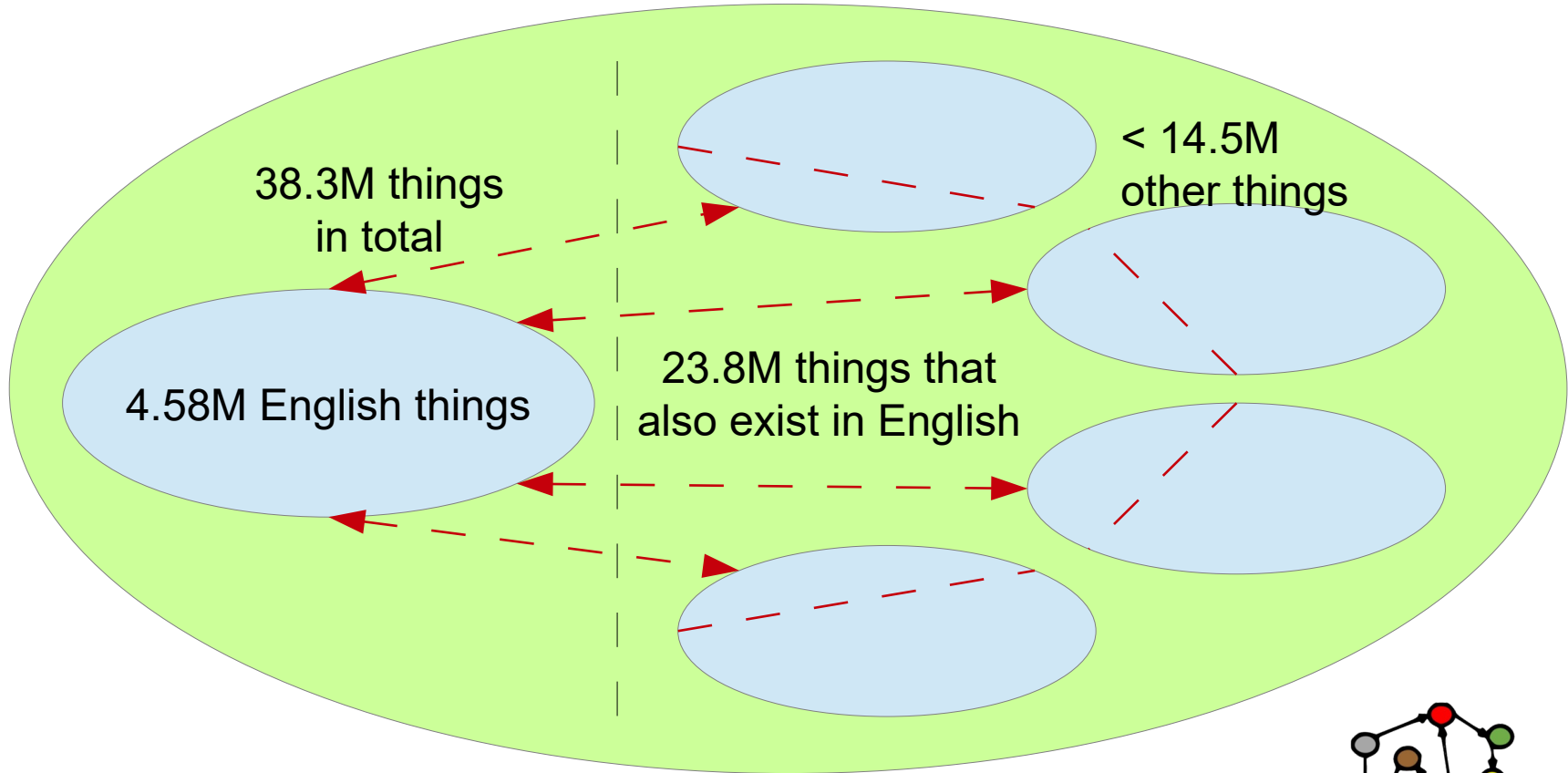


DBpedia: ontology and identities

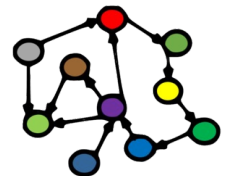
- URIs derived from Wikipedia, e.g.:
 - <http://en.wikipedia.org/wiki/Bergen> →
 - <http://dbpedia.org/resource/Bergen>
 - **English, canonical, always dereferencable URIs**
- localised/national:
 - <http://no.dbpedia.org/resource/Bergen>
 - **not always dereferencable**
 - ...they are *URNs*, but not always URIs



Canonical and localised resources



(Example numbers from 2015-10.)



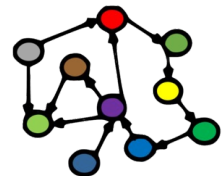
DBpedia: some name spaces

- <http://dbpedia.org/> – language-independent base, URIs
- <http://nn.dbpedia.org/> – language-specific base, URIs
 - 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.org/sameAs) – interlinking, e.g, across languages
- [rdf:type](http://rdf.org/type) – three classification schemes



DBpedia: new extraction

- Since 2020:
 - monthly extraction in four groups
 - generic
 - generic parsers, language-specific RDF properties
 - mappings
 - editable ontology mappings: mappings.dbpedia.org
 - text
 - abstract and article full-text extraction
 - Wikidata
 - mapped and cleaned Wikidata data
using the DBpedia Ontology

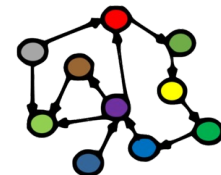


DBpedia: resource statistics (old extraction)

- Contents:
 - 4.6M resources in the English DBpedia
 - 4.22M in a consistent ontology
 - 1.445M persons, 735k places, 411k works, 241k organisations, 251k species, 6k diseases...
 - 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)

DBpedia: triple statistics (old extraction)

- The full (international) data set:
 - 9.5G triples (*perhaps 6-7% of the LOD cloud*)
 - 1.3G from the English DBpedia
 - 5.0G from other DBpedias
 - 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, YAGO, UMBEL...



DBpedia: concept schemes (old extraction)

- 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)*



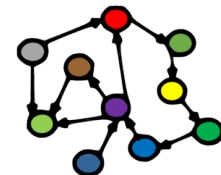
DBpedia: advantages

- Covers many domains
 - like Wikipedia, exploits *the long tail*
- Community
- Automatically evolves (as Wikipedia changes)
- Is truly multilingual
- Includes Wikidata
 - claim: in a better managed ontology



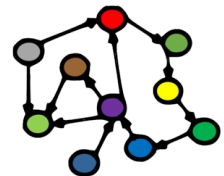
DBpedia ↔ Wikidata

- Similarities:
 - both publish **RDF data** about **entities/resources**
 - both offer **standard URIs** and 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
- Recently: DBpedia also extracts data from Wikidata



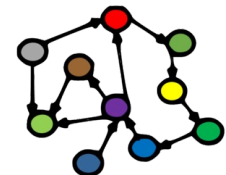
DBpedia and Wikidata ↔ Freebase

- *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



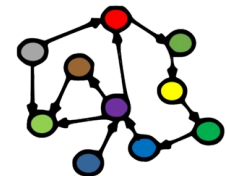
The GDELT project

- Global Database of Events, Language, and Tone (GDELT)
 - free open platform
 - monitors the world’s broadcast, print, and web news
 - focus on crises, but much broader in practice
 - globally in over 100 languages
 - identifies people, locations, organizations, themes, sources, emotions, counts, quotes, images, events
 - *“can we map happiness and conflict, provide insight to vulnerable populations and even potentially forecast global conflict in ways that allow us as a society to come together to deescalate tensions, counter extremism, and break down cultural barriers?”*



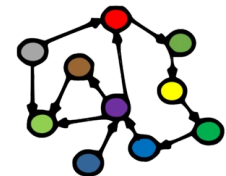
The GDELT project

- Archives back to 1979 (expanding back to 1800)
- Increasingly integrating social media
- Translations from 65 languages into English
- Supported by Google
 - runs in the Google Cloud
- Almost a knowledge graph, but
 - not native RDF
 - not fully linked
 - no ontology



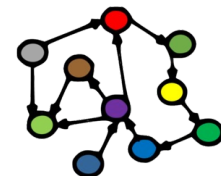
The GDELT project: data streams

- Downloadable CSV files (every 15 minutes)
 - <http://data.gdeltproject.org/gdeltv2/lastupdate.txt>
 - *Events* (...export.CSV, ~400k)
 - low-level actor - event type – actor triples
 - *Mentions* (...mentions.CSV, ~600k)
 - where in and which source is each event mentioned?
 - *Global Knowledge Graph* (...gkg.CSV, ~50M)
 - which people, locations, organizations, themes, sources, emotions, counts, quotes, images, events are mentioned where and in which source?
 - Also available as Google BigQuery tables
- Lots of other datasets and streams, raw and analysed, native language or translated to English



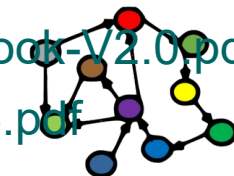
The GDELT project: data streams

- Other data streams:
 - *Visual GKG*
 - codifying the world's news images in real time
 - random sampling, Google's Vision API
 - *Global Entity Graph*
 - experimental, random sampling of news articles
 - deep learning, Google's Natural Language API
 - provides Wikidata links for entities
 - *Global Relationship Graph*
 - experimental, related to the global entity graph
 - extracts verbs and the words in their context
 - groups new articles with similar verbs-in-context



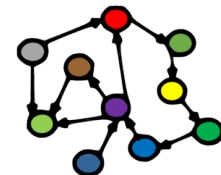
The GDELT project: Events 2.0

- For each event:
 - global event id and datetime
 - actor 1 and 2:
 - name (person, organisation, location, ethnicity, religion, type) and CAMEO code
 - event:
 - CAMEO code and importance of event type
 - numbers of mentions and sources, tone
 - geography
- Codebooks
 - http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf
 - <http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf>



The GDELT project: Mentions 2.0

- For each event
 - global event id and datetime
 - mention type and datetime
 - source name and identifier (e.g., a URL)
 - sentence number
 - actor 1 and 2 mentions (character indices)
 - confidence
 - source length and tone
- Codebook
 - http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf



The GDELT project: GKG 2.0

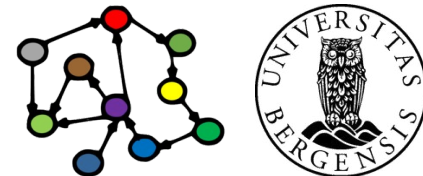
- For each document:
 - record id and datetime
 - source and document identifier (e.g., a URL)
 - keywords/themes (taxonomies of 50k keywords)
 - person and organisation names and types
 - locations, their types, names, geo-coordinates
 - counts, their types and counted objects
 - average tone, positive/negative score, polarity
 - ...and lots of other stuff
- Codebook
 - http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf



EventKG



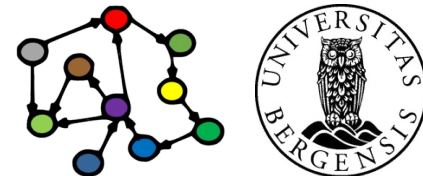
- The LOD cloud is *entity* centric
 - people, organisation, places, works...
- For many purposes *events* are more important
 - a type of entity, but less focussed on
 - ...and harder to extract/describe
- EventKG is a *Multilingual Event-Centric Temporal Knowledge Graph*
- Extracted from:
 - Wikidata, DBpedia and YAGO
 - Wikipedia Current Events Portal (WCEP)
 - Wikipedia event lists in 15 languages



EventKG



- EventKG 3.0:
 - > 1.3 million events
 - 56.25% from the existing KGs
 - > 676 million triples
 - > 4.5 million temporal relations
- Example uses:
 - “interactive spatio-temporal biographies”
 - “cross-lingual event timelines”
 - “semantic event-centric queries and their verbalisations in three languages”
 - “event-centric user interaction traces”



Example

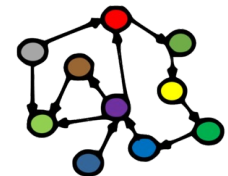
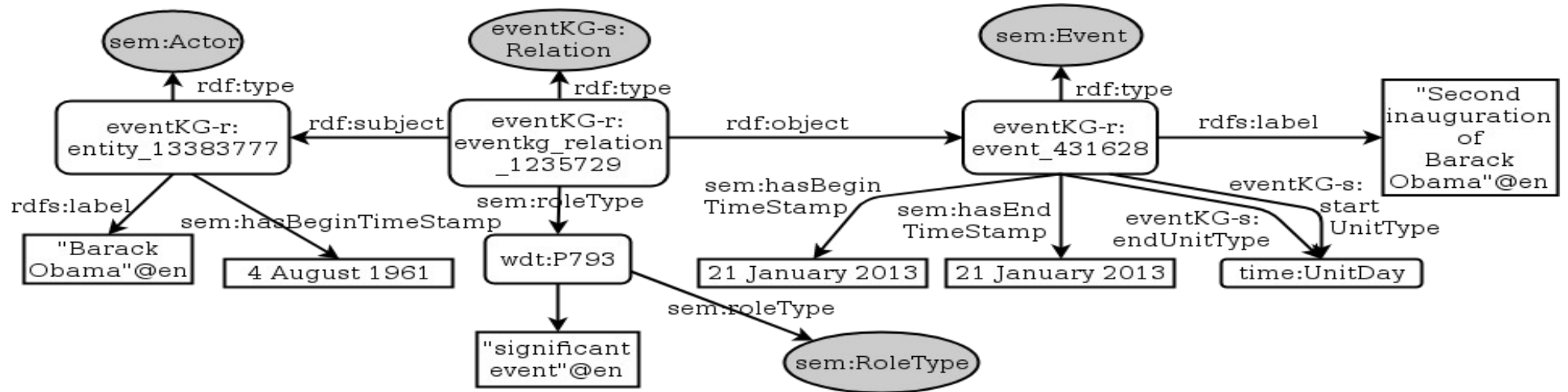
- The Brexit referendum
 - <http://www.wikidata.org/entity/Q21812812>
- Local EventKG
 - <https://imdb.uib.no/bg-eventkg>
- PREFIX owl: <<http://www.w3.org/2002/07/owl#>>
SELECT * WHERE {
 GRAPH ?g
 { ?s ?p <<http://www.wikidata.org/entity/Q21812812>> . }
}
- DESCRIBE
 <http://eventKG.l3s.uni-hannover.de/resource/event_61609>
 - Total results: 2930, displaying...
 - ~ 2800 are link relations
<http://eventkg.l3s.uni-hannover.de/>

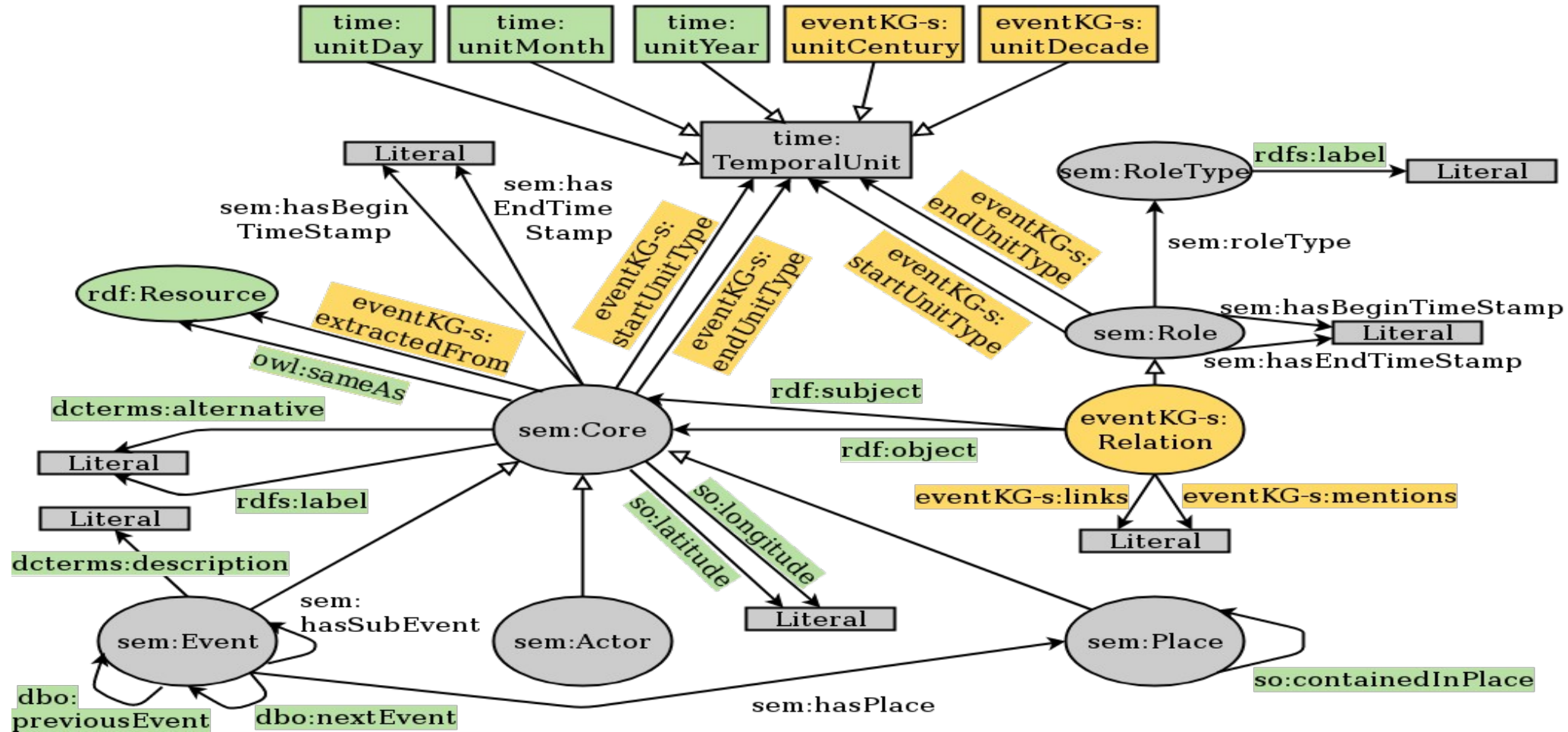
Example

- The Brexit referendum
 - <http://www.wikidata.org/entity/Q21812812>
- PREFIX eventKG-r: <<http://eventkg.l3s.uni-hannover.de/resource/>>
DESCRIBE eventKG-r:eventkg_relation_1705321
- eventKG-r:eventkg_relation_1705321
 - rdf:type eventKG-s:Relation;
 - sem:roleType wdp:P92 ; # “main regulatory text”
 - rdf:subject eventKG-r:event_61609 ;
 - rdf:object eventKG-r:entity_8402669 .
- eventKG-r:entity_8402669
 - ... ;
 - owl:sameAs dbr:European_Union_Referendum_Act_2015 ;
 -

A visual relation

- Relation from Barack Obama's second inauguration

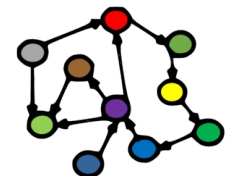




SEM = Simple Event Model (S08-09)

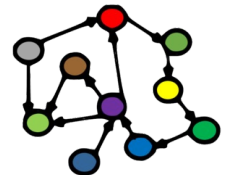
EventKG ↔ GDELT

- EventKG
 - native RDF, historic / static (but updated)
 - linked resources (DBpedia, Wikidata, YAGO)
 - ontology (DBpedia ontology, Wikidata, YAGO)
 - limited to existing KGs + Wikipedia
 - quality-controlled by crowdsourcing
- GDELT
 - real time, 15 minute updates, does its own lifting
 - thousands of live sources: text, audio, video
 - sentiment, topic, keyword and other analyses
 - metadata from codebooks
 - much noise, imprecisions, errors



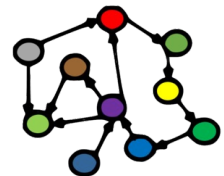
GeoNames

- *Adding geospatial semantic information to the web*
 - a geographical database: <http://www.geonames.org>
 - collected from a large number of sources
 - > 25M geographical names (*toponyms*, Norway 68k),
> 11M unique features, ~ 4.8M populated places,
~ 13M alternate names
- Offers *dereferencable URIs* 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 URIs
 - downloadable (TSV)
 - Gazetteer lists
- Also as Linked Data Fragments:
 - <http://data.linkeddatafragments.org/geonames>

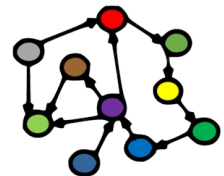


GeoNames ontology

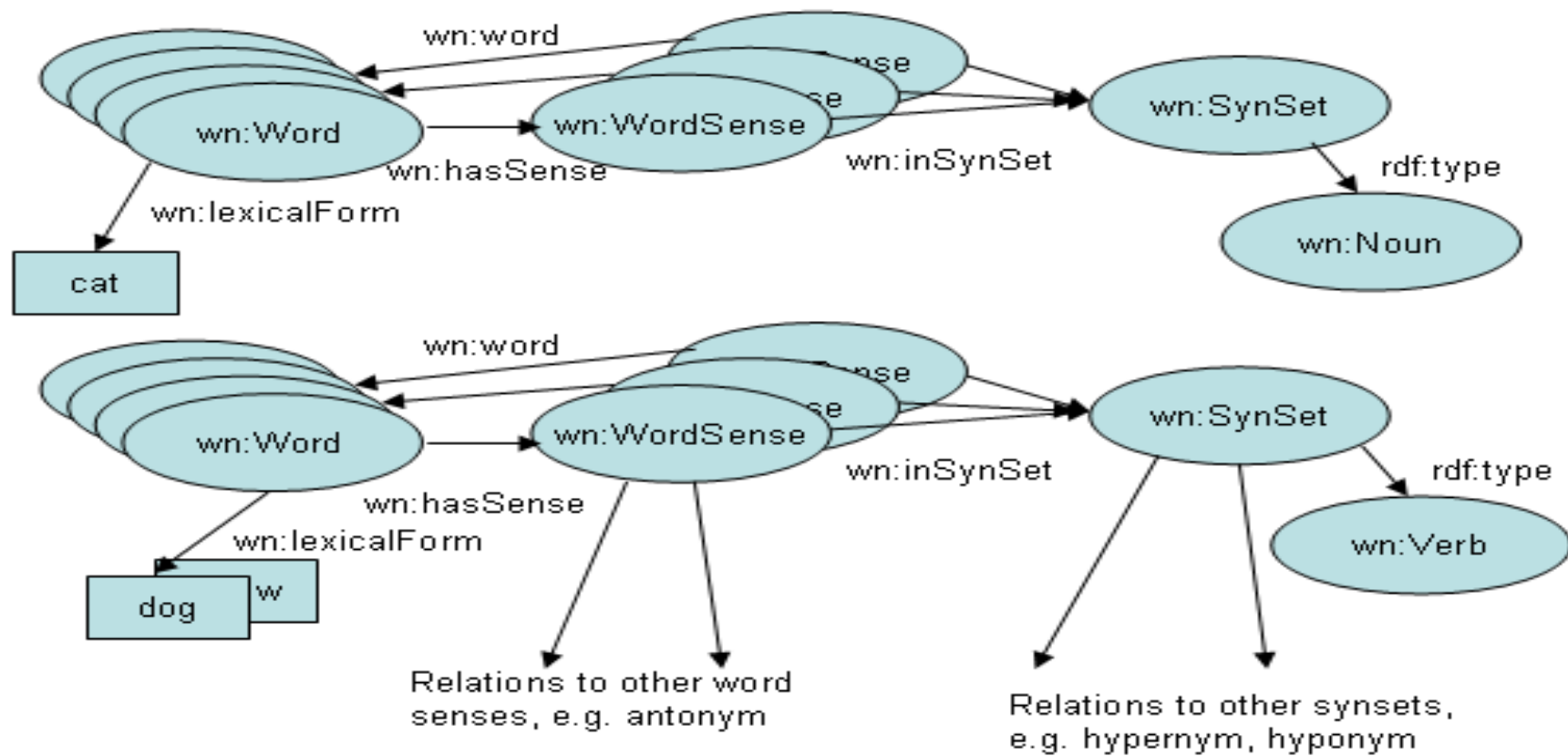
- Vocabulary in OWL:
 - @prefix gn: <<http://geonames.org/ontology#>> .
 - gn:Feature class
 - 9 top-level feature codes:
 - **A** country, state, region, ...; **H** stream, lake, ...;
 - L** parks, area, ...; **P** city, village, ...; **R** road, railroad;
 - S** spot, building, farm; **T** mountain, 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

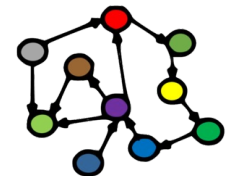


WordNet: Structure



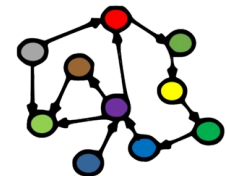
WordNet: Synset structure

- 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



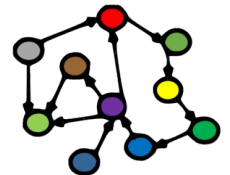
WordNet: Synset structure

- 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



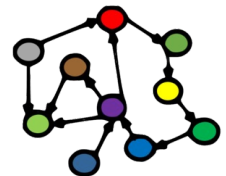
WordNet: Synset structure

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



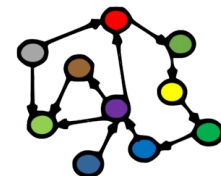
WordNet: Norsk Ordvev

- Developed Kaldera språkteknologi
 - for Nasjonalbiblioteket (The national library)
 - both *bokmål* and *nynorsk*
 - \approx 50 000 words, 200 000 synsets each
- Available at
 - <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-27/>
 - <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-7/>
 - updated 2013-2014



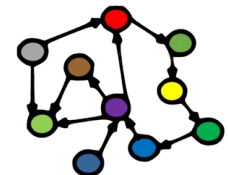
International language resources

- 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
 - *not fully open...*



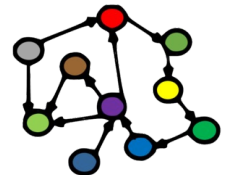
BabelNet

- A multilingual encyclopedic dictionary and a semantic network of concepts and named entities
 - both lexicographic and encyclopedic coverage
 - 16 million Babel synsets
 - > 800 million word senses
 - > 280 languages
 - integrates data from *WordNet*, *Open Multilingual Wordnet*, *Wiktionary*, *Wikidata*, *Wikipedia*, *Wikiquotes*, *GeoNames* and several others



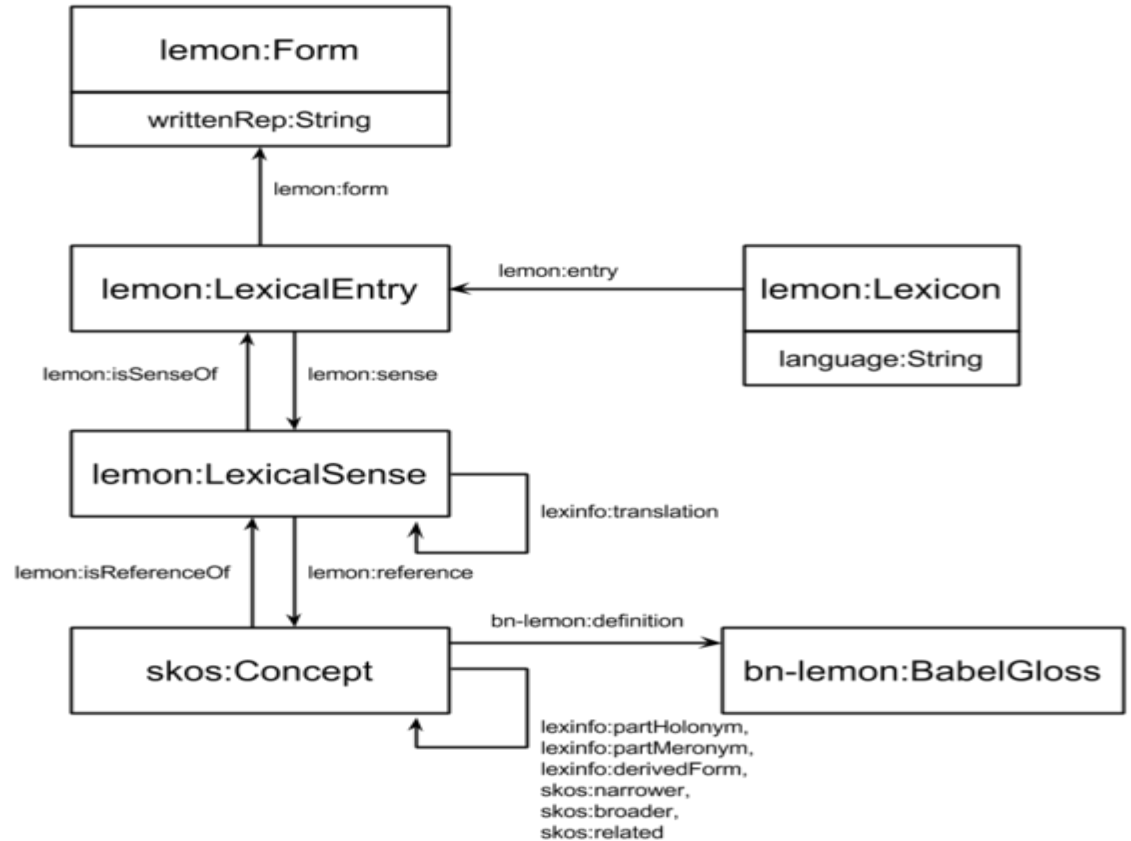
BabelNet availability

- 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
 - Attribution-NonCommercial-ShareAlike 3.0



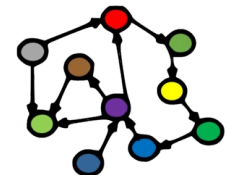
BabelNet conceptual model

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



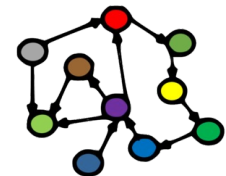
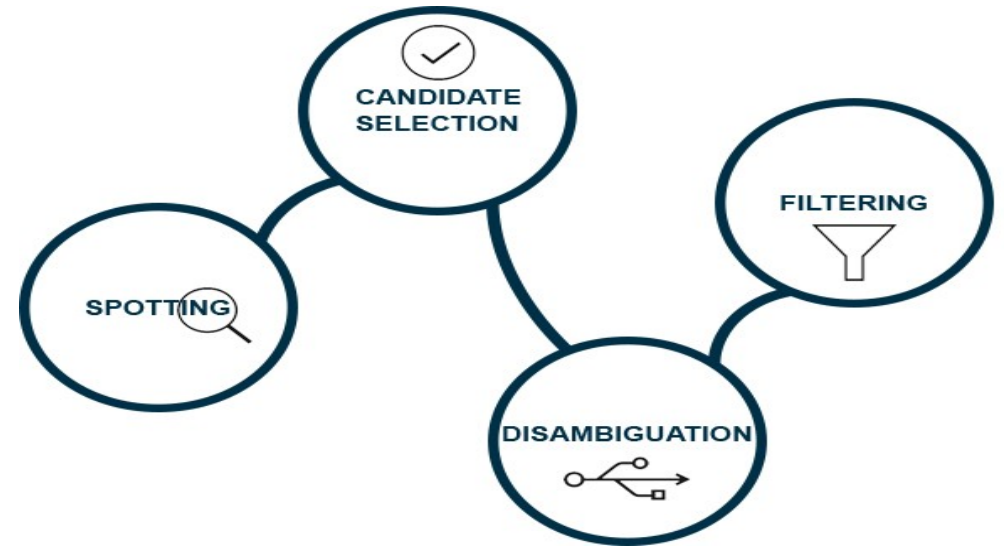
DBpedia Spotlight

- A tool and web API for lifting text:
 - automatically annotating mentions of DBpedia resources in text
 - linking unstructured information sources to the LOD cloud through DBpedia
 - available as:
 - online demo <https://demo.dbpedia-spotlight.org/>
 - web API: api.dbpedia-spotlight.org/en
[https://api.dbpedia-spotlight.org/en/candidates?text=“...”](https://api.dbpedia-spotlight.org/en/candidates?text=...)
 - download:
 - open source code
 - Docker



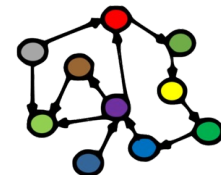
DBpedia Spotlight

- **Spotting:** identify potential *mentions* (substrings) of *named entities* in texts
- **Candidate selection:** find DBpedia *resources* that may match the *surface form* of a mention
- **Disambiguation:** select the more likely candidate resource for each surface form
- **Filtering:** adjust to user-specific requirements (e.g., confidence)
- **Limitation:** only DBpedia entries
 - focus on named entities
 - fewer *concepts, events, relations, times...*



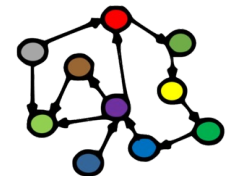
Text lifting tasks

- DBpedia Spotlight covers *entity extraction*:
 - entity recognition (detection)
 - entity disambiguation (name resolution)
 - linking
- Does not focus on:
 - word-sense disambiguation (WSD)
 - topic extraction
 - text classification
 - relation extraction
 - sentiment analysis, attitudes, negation



YAGO4

- *Yet Another General/Great Ontology:*
 - top-level classes from schema.org and bioschemas.org
 - facts extracted from Wikidata
 - 50M entities, 2B triples/facts
 - places facts and entities in time and space
 - YAGO1-3:
 - based on Wikipedia and WordNet, GeoNames, DBpedia, SUMO...
 - used to categorise DBpedia's resources
 - used in IBM's Watson system
 - downloadable as RDF



Next week:
Enterprise KGs