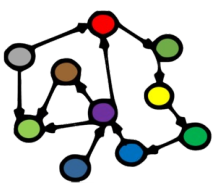
Welcome to INFO216:
Knowledge Graphs
Spring 2023

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

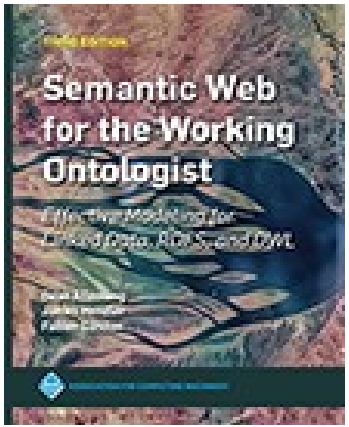
Session 4-5: Open KGs I and II

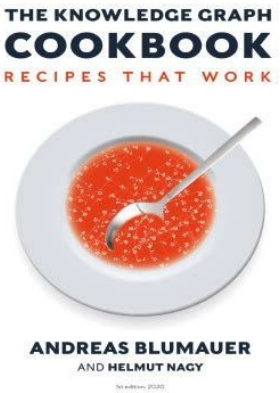
- Themes:
  - Linked Open Data (LOD)
  - LOD cloud
  - *Open Knowledge Graphs:*
    - Wikidata, DBpedia, GeoNames, GDELT project, WordNet, BabelNet, ConceptNet
    - *...some of them have their own vocabularies*
  - Enterprise Knowledge Graphs (EKGs) (→ *S06)*
  - Ontologies and vocabularies (→ *S07 and S08)*
  - *also: SPARQL programming (← S03 leftover)*



Readings

- Allemang, Hendler & Gandon (2020): Semantic Web for the Working Ontologist, 3rd edition chapter 5 (Linked Data)
- Materials in the wiki <http://wiki.uib.no/info216>, including:
  - Wikidata
  - DBpedia
  - GeoNames
  - GDELT
  - WordNet
  - BabelNet
  - ConceptNet





A brief history of KGs (←S02)

Tim Berners-Lee's call for a transition

- From around 1990: creation of a *Web of Documents*
  - the "plain old web" (PoW)
  - document-centric
  - document-to-document links
  - for humans
- From around 2000: transition to a *Web of Data*
  - document- *and data-centric*
  - doc-to-doc *and data-to-data links*
  - for humans *and machines*
  - also called the *Semantic Web*, *Web 3.0*, the *Web of Knowledge,* the *Giant Global Graph (GGG)*, the *Linked Open Data (LOD) cloud...*

Tim Berners-Lee
Inventor of the
World Wide Web
(WWW, 1989)

Tim Berners-Lee's call for a transition

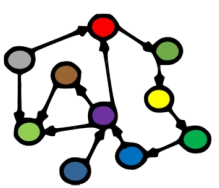
- There's an enormous amount of data on the web
  - ...but the data are mostly not linked
    (think of a world wide web without document links!)
  - availability, accessibility does not go all the way
  - *what if we had standard ways of representing data so that linkable data could always be automatically linked?*
  - *enormous potential to solve, simplify, speed up... many critical information handling problems*
- This is the purpose of *semantic technologies*
- This is the vision that led to today's *semantic knowledge graphs*

Tim Berners-Lee
Inventor of the
World Wide Web
(WWW, 1989)

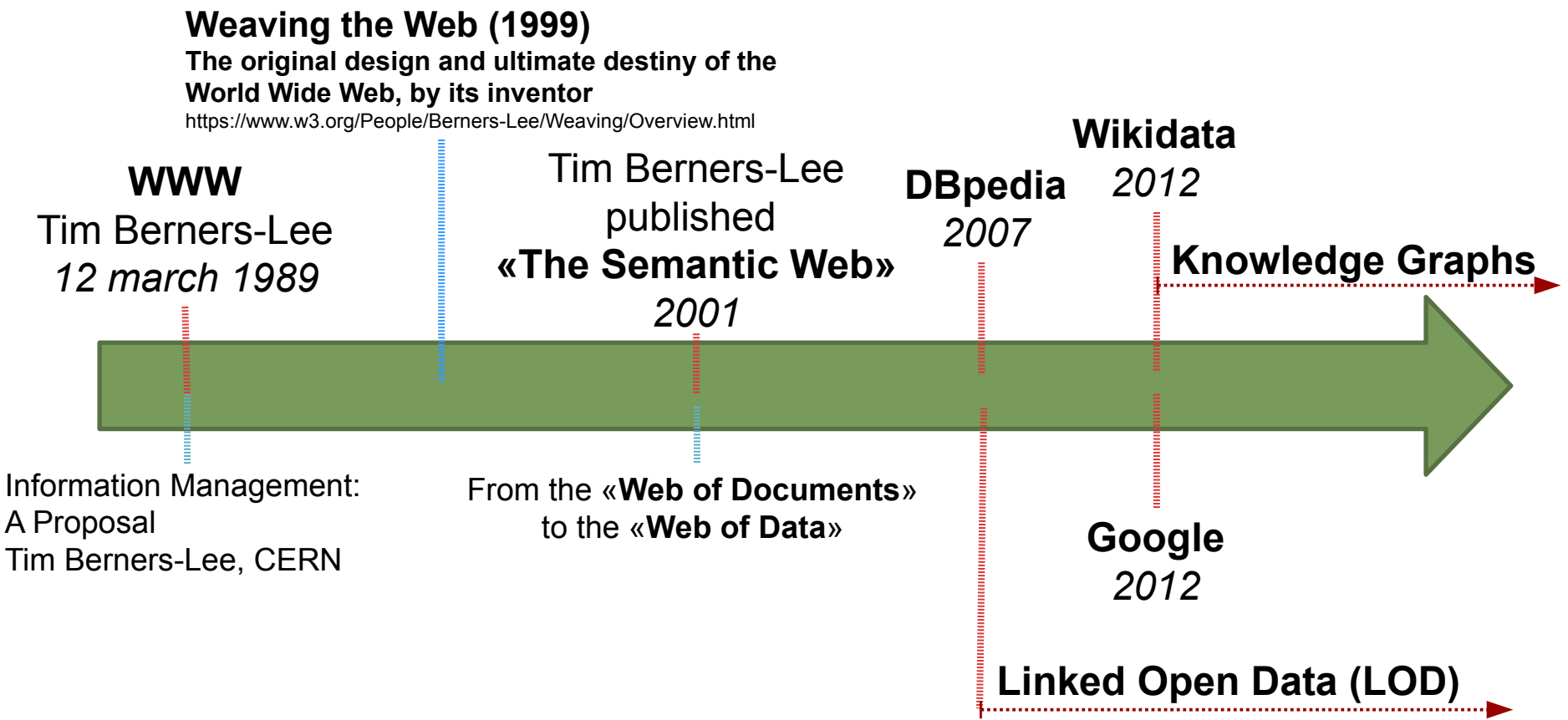
Tim Berners-Lee: <http://www.youtube.com/watch?v=HeUrEh-nqtU>

Many independent, but related developments

- The *Linked Open Data (LOD)* cloud:
  - interlinking semantic datasets, making them openly available: DBpedia (2007-), Wikidata (2012-), …
- *Knowledge graphs:*
  - currently popular term for semantic graph representations of (primarily) factual information (Google, 2012)
- *Enterprise knowledge graphs (→ S06):*
  - company-internal semantic data
  - linked open data and semantic-web technologies used inside an enterprise or cluster



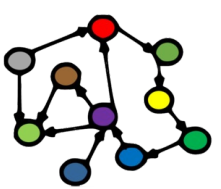
Semantic web and WWW history



Tim Berners-Lee: http://www.youtube.com/watch?v=HeUrEh-nqtU
Information Management: A Proposal: https://cds.cern.ch/record/369245/files/dd-89-001.pdf

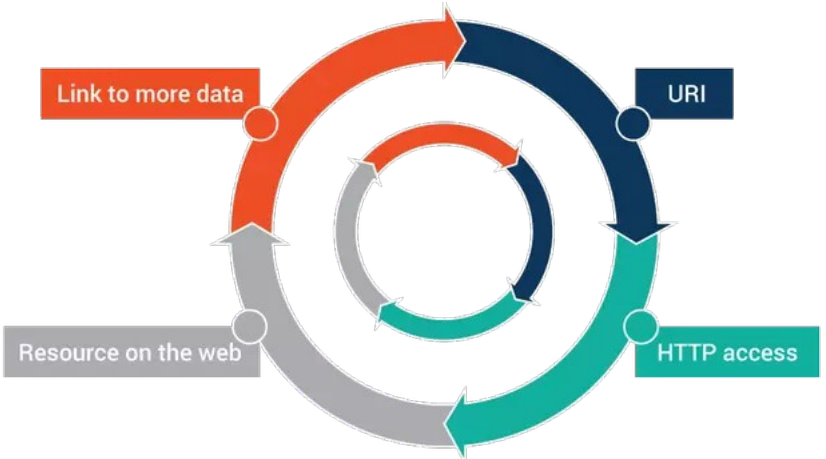
Common themes

- *Graph representations* of knowledge
  - RDF, RDFS, OWL, SPARQL
  - plus the more recent alternatives
- *Semantically tagged* data
  - well-defined tags (terms)
    - defined in standard vocabularies
    - formal ontologies, description logic
- *Global* and *interlinked*
  - standard formats, technologies, resource URIs, etc.
- From the start *open* and *community-based*



Linked Open Data (LOD)

Linked Open Data (LOD)



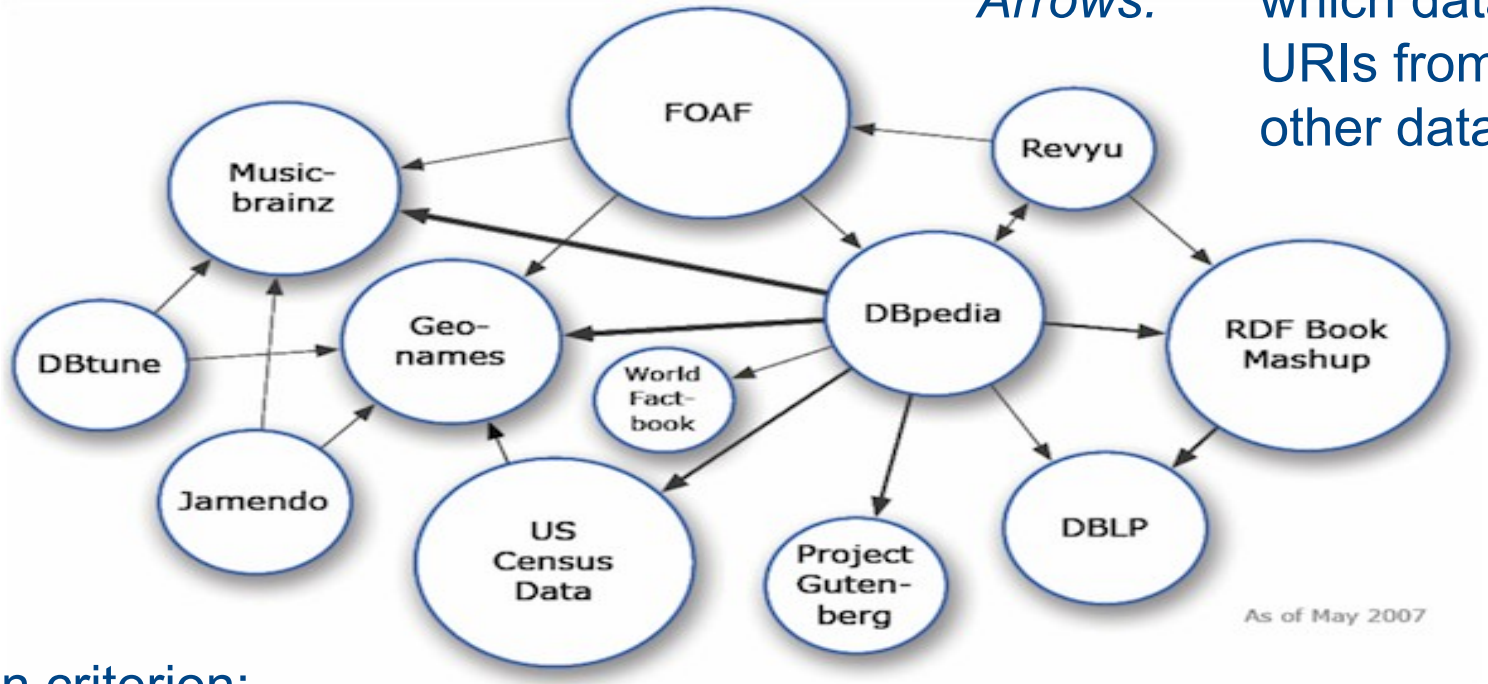
- *Four basic principles (Berners-Lee 2006):*
  1. *URIs (Uniform Resource Identifier)*
     - *identify resources, e.g.,* http://wikidata.org/entity/Q18692990
  2. *URIs answer to HTTP requests (dereferencing)*
     - for example *SPARQL queries, Turtle files, …*
  3. *Returns information about the resource on standard format,*
     - *semantic, e.g., Turtle, JSON-LD, RDF-XML, N-TRIPLES, N3*
     - *also non-semantic: HTML, JSON, XML, CSV, TSV...*
  4. *The information contains URI-s that identify related resources*

https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/

(c) Andreas L Opdahl, 2023

The LOD cloud (2007-05)

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



Inclusion criterion:

- minimum 1000 triples
- minimum 50 links to other datasets (in- or outbound)

The LOD cloud (2022-11)

Domains:

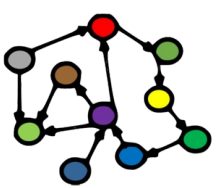
- general
- geography
- government
- life sciences
- linguistics
- media
- publications
- social networking
- user generated

A “lumpy cloud”

Diagram: John P. McCrae for the Insight Centre for Data Analytics

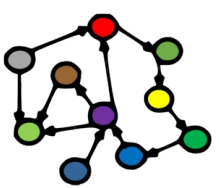
The LOD cloud

- The LOD cloud web page <http://lod-cloud.net/>
  - which datasets mention resources in other datasets?
  - >1250 datasets with >15000 links between them
    - started in 2007
    - very rapid growth for a few years
    - consolidating since ca 2017
- How big is the LOD cloud?
  - hard to measure exactly (old stats: http://*lodstats.aksw.org*)
  - approx. 150G (150 000M) triples from >3000 data sets (2020)
  - *Wikidata <http://wikidata.org>* is the largest general one:
    - >100M resources (items), >1,4G (1400M) triples



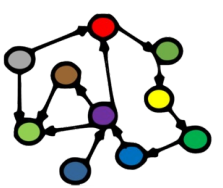
Challenges

- Enormous potential
  - ...but *Open KGs* are not used to their fullest
    - *maintenance:* individuals versus organisations
    - *abstraction:* general versus domain data
    - *trust:* open versus closed networks
  - *Enterprise Knowledge Graphs (EKGs)* have matured
    - *S01: 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*



Places to start

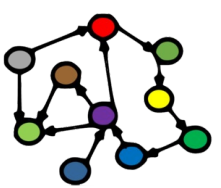
- Open and semantic:
  - open semantic data sets: *http://lod-cloud.net*
  - vocabularies: *https://lov.linkeddata.es/dataset/lov/*
- Open data in general:
  - internationally: *http://datahub.io* or *http://ckan.org*
  - Norge: *http://data.norge.no*
  - EU: *https://data.europa.eu/en*
  - UK: *http://data.gov.uk*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



(c) Andreas L Opdahl, 2023

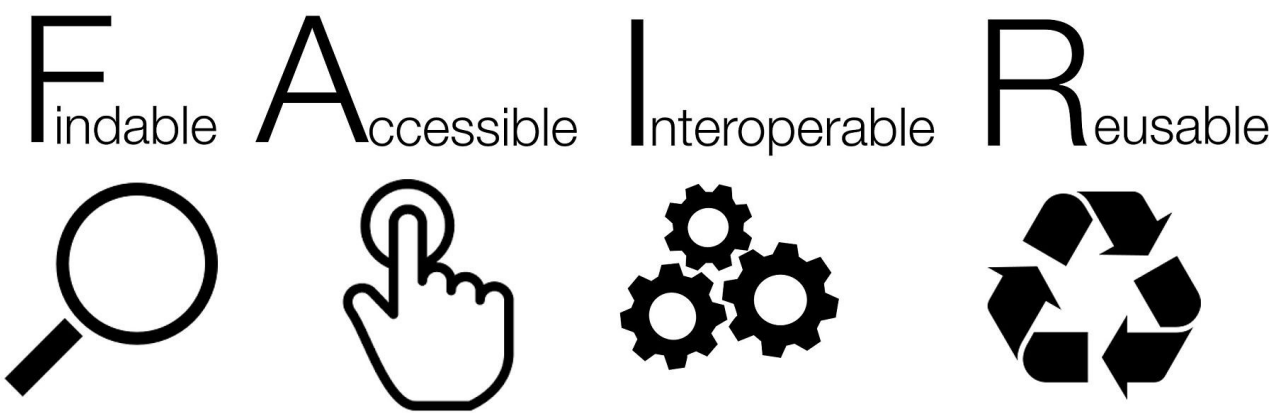
Best Practices for Data Provisioning

- 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)*



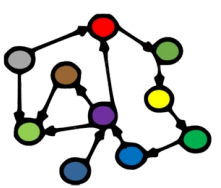
FAIR Principles

- (Scientific) data management and stewardship
- Three types of entities
  - data (or any digital object)
  - metadata (information about that digital object)
  - infrastructure
- Four principles for data *and metadata*:



FAIR Principles for data *and metadata*

- Findable
  - globally unique and persistent identifier
  - rich metadata
  - metadata include identifier
  - in a searchable resource
- Accessible
  - retrievable by identifier
  - standardised protocol
    - open, free, universally implementable
    - authentication and authorisation when needed
  - persistently accessible metadata
- Interoperable
  - formal, accessible, shared, broadly used language
  - FAIR vocabularies
  - qualified references
- Reusable
  - richly described, with many accurate and relevant attributes
  - clear and accessible license
  - detailed provenance
  - follow community standards



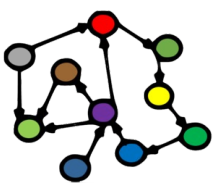
(c) Andreas L Opdahl, 2023

Open Knowledge Graphs

Wikidata

Wikidata (→ S01)

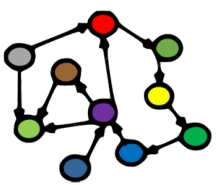
- *A free and open knowledge base that can be read and edited by both humans and machines*
  - a Wikimedia project, crowdsourced, multi-lingual
  - *a Wikipedia for structured, secondary data*
  - a central source of URIs: http://wikidata.org/entity/Q*nn* (“Q-codes”)
  - beginning: from managing Wikipedias *cross-language links*
  - today: to *central storage of structured data* for
    - Wikimedia sister projects (Wikipedia etc.), and many others
  - verifiability, link to sources, perspectives
  - free license (CC0 1.0), standard formats, interlinked
- Wikidata entities:
  - >100M items (things), >1.4G statements (“primary” triples)

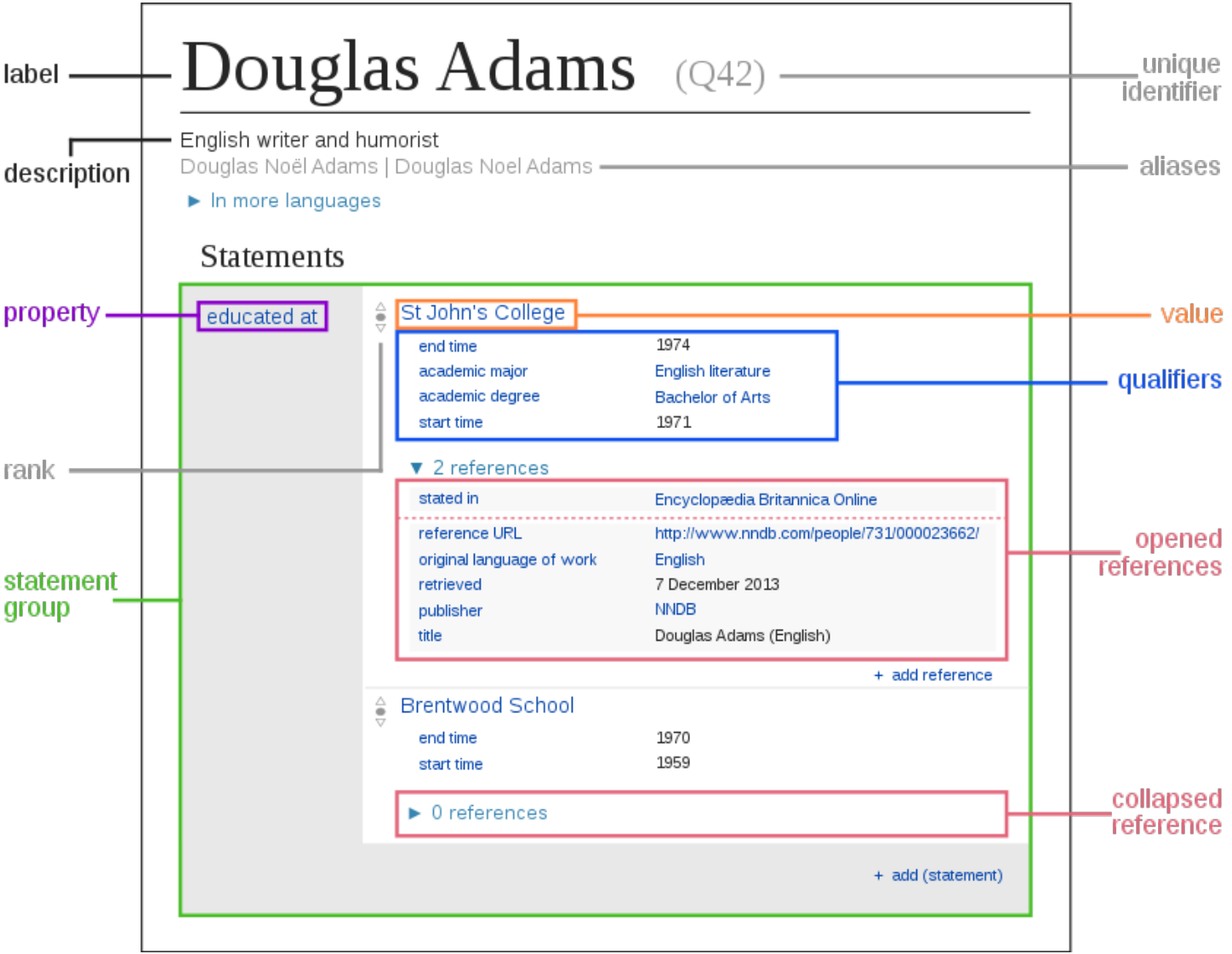


https://grafana.wikimedia.org/d/000000175/wikidata-datamodel-statements?orgId=1&refresh=30m

Wikidata access

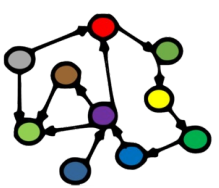
- Available through
  - the WikiMedia API
  - HTTP: http://www.wikidata.org/entity/Q42
  - RDF: http://www.wikidata.org/entity/Q42.ttl
  - SPARQL endpoint: http://query.wikidata.org
    - Wikidata Query Service (WDQS)
  - for download (JSON, RDF, XML)
    - https://www.wikidata.org/wiki/Wikidata:Database_download
  - Triple Pattern Fragments:
    - https://query.wikidata.org/bigdata/ldf
- Lots of other/third party tools
- 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
St John's College
value
end time 1974
academic major English literature
academic degree Bachelor of Arts
start time 1971
qualifiers
rank
▼ 2 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)
opened references
+ add reference
statement group
Brentwood School
end time 1970
start time 1959
► 0 references
collapsed reference
+ add (statement)

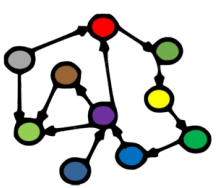
Wikidata item structure

- Items:
  - item identifier (Q*nn*)
  - 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!*



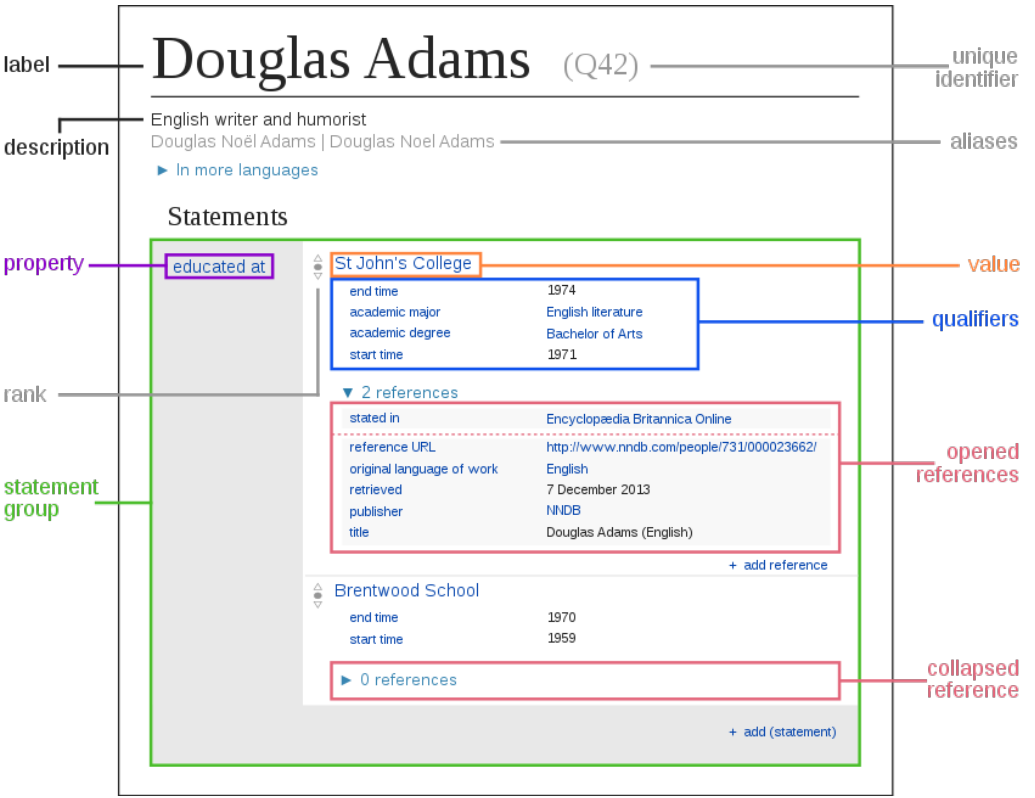
Wikidata contents

- *Wikidata does not represent facts, but statements!*
- Fact: taken-for-granted truth about the world
  - (perhaps) qualified
- Statement: claims about truth made by a social actor
  - (perhaps) qualified, (should have) backing in references
- *Contradictory claims are allowed!*
  - can be explained in free-text qualifications
  - statement rankings: preferred, normal, deprecated
- Truthy (versus full) statements: the best statement for an item+property
  - *this is why you should stay with the (truthy) wdt: properties*



(c) Andreas L Opdahl, 2023

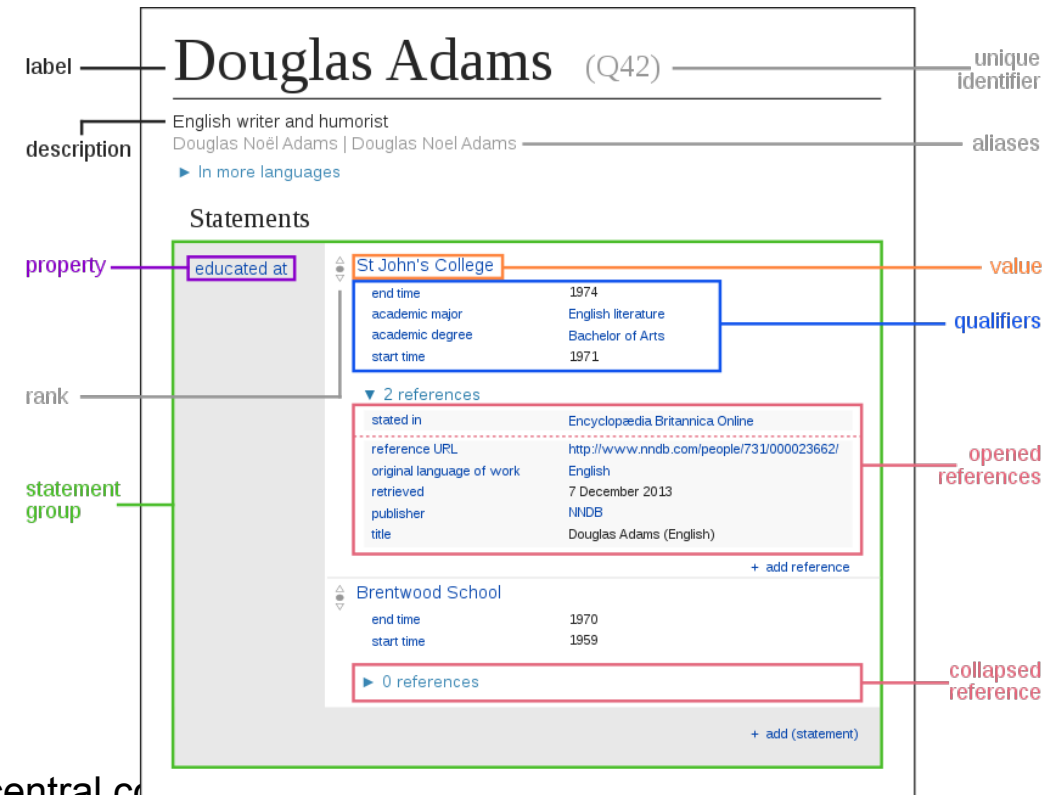
Wikidata item structure



(c) Andreas L Opdahl, 2023

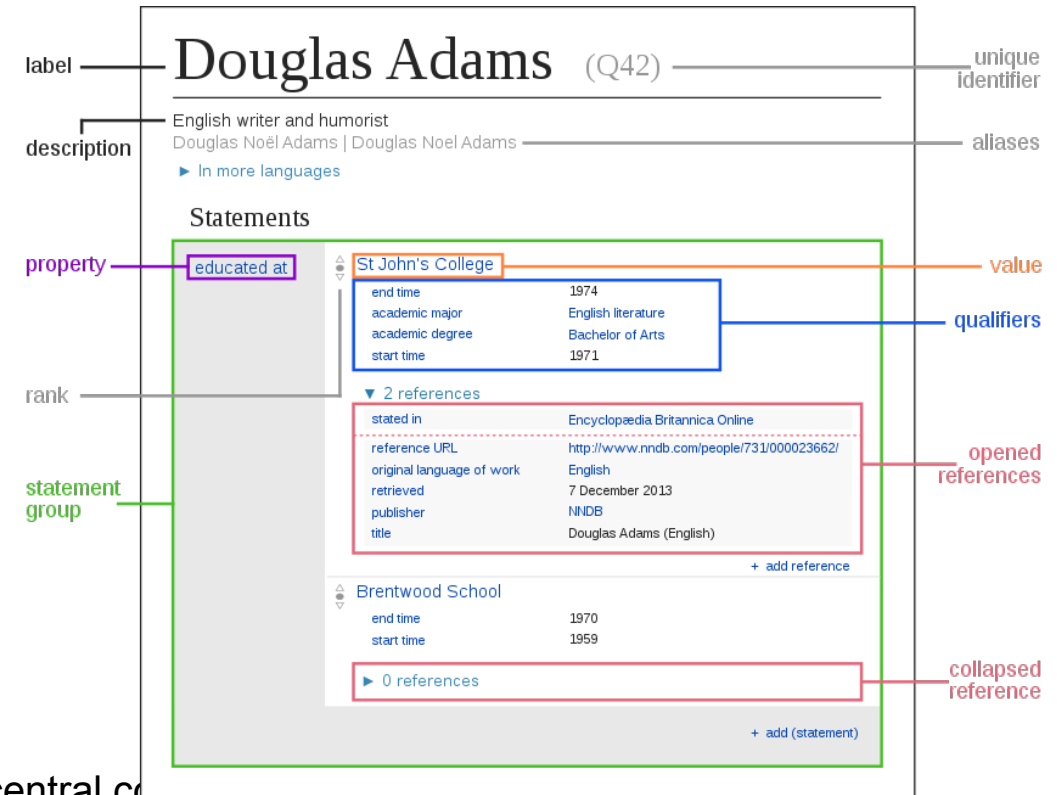
Wikidata item structure

```
wd:Q42
    a wikibase:Item ;
    rdfs:label "Douglas Adams"@en ;
    skos:prefLabel "Douglas Adams"@en ;
    schema:name "Douglas Adams"@en ;
    ...
    rdfs:label " 더글러스 애덤스 "@ko ;
    skos:prefLabel " 더글러스 애덤스 "@ko ;
    schema:name " 더글러스 애덤스 "@ko ;
    ... ;

    skos:altLabel
        "Douglas Noël Adams"@en,
        "DNA"@en,
        "Адамс, Дуглас"@ru,
        "Дуглас Ноэль Адамс"@ru,
        …

    schema:description
        "English science fiction writer
         and humourist"@en,
        "écrivain de science-fiction et
         humoriste anglais"@fr,
        ...
```



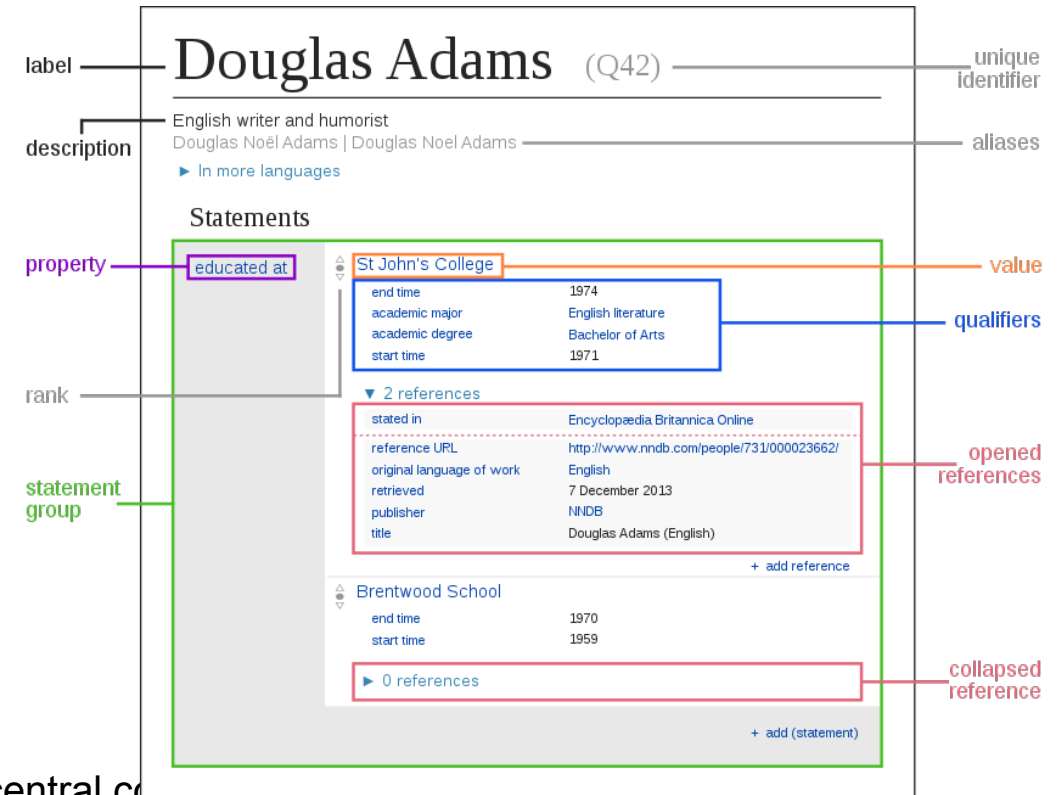
Wikidata item structure

```
wd:Q42
    a wikibase:Item ;
    ...
    wdt:P69 wd:Q691283, ... ;
    … .

wd:P69 a wikibase:Property ;
    rdfs:label "educated at"@en ;
    skos:prefLabel "educated at"@en ;
    schema:name "educated at"@en ;
    schema:description
        "educational institution attended by subject"@en .

wd:Q691283 a wikibase:Item ;
    rdfs:label "St John's College"@en ;
    skos:prefLabel "St John's College"@en ;
    schema:name "St John's College"@en ;
    schema:description
        "constituent college of the
         University of Cambridge"@en .
```

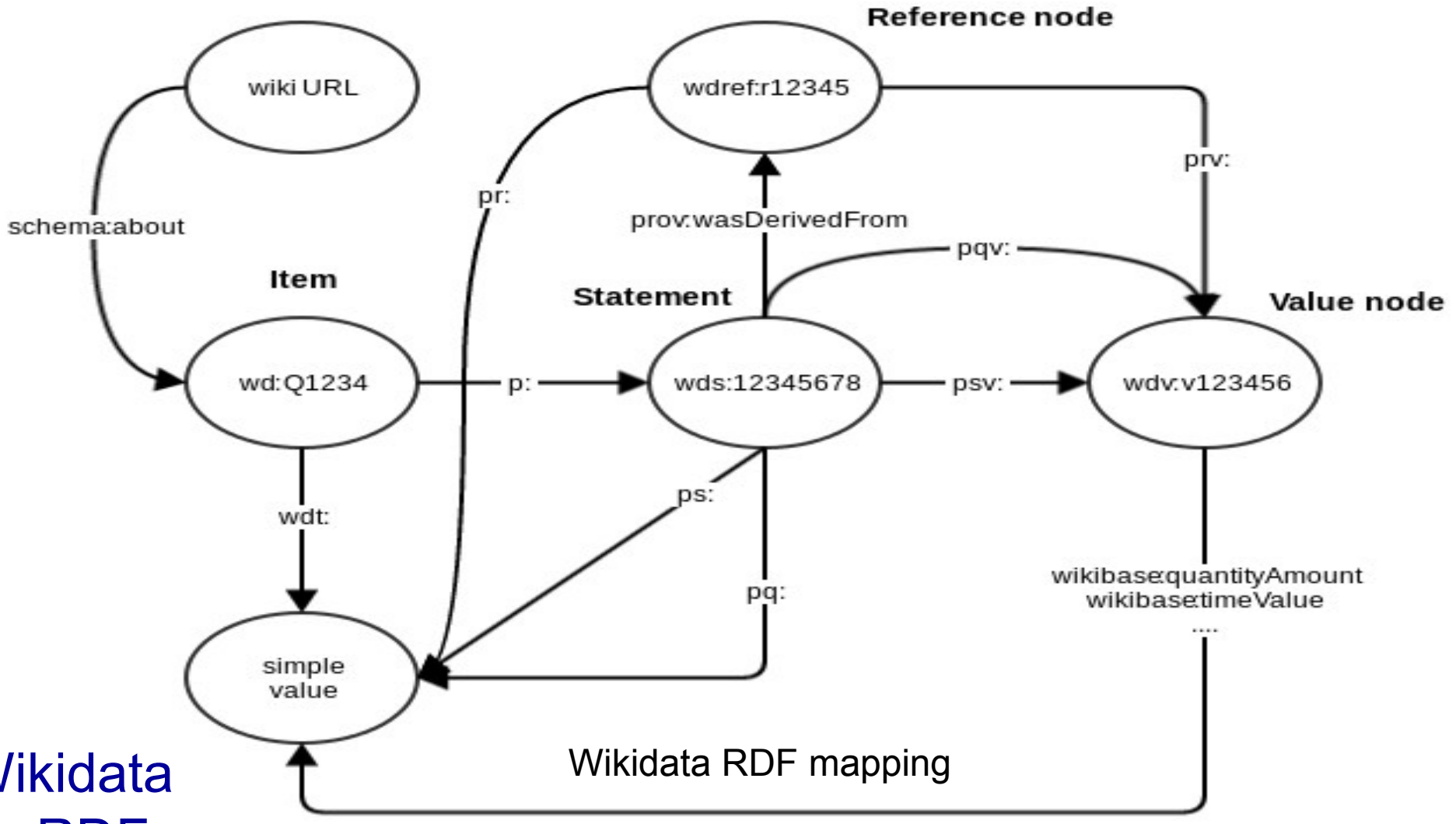


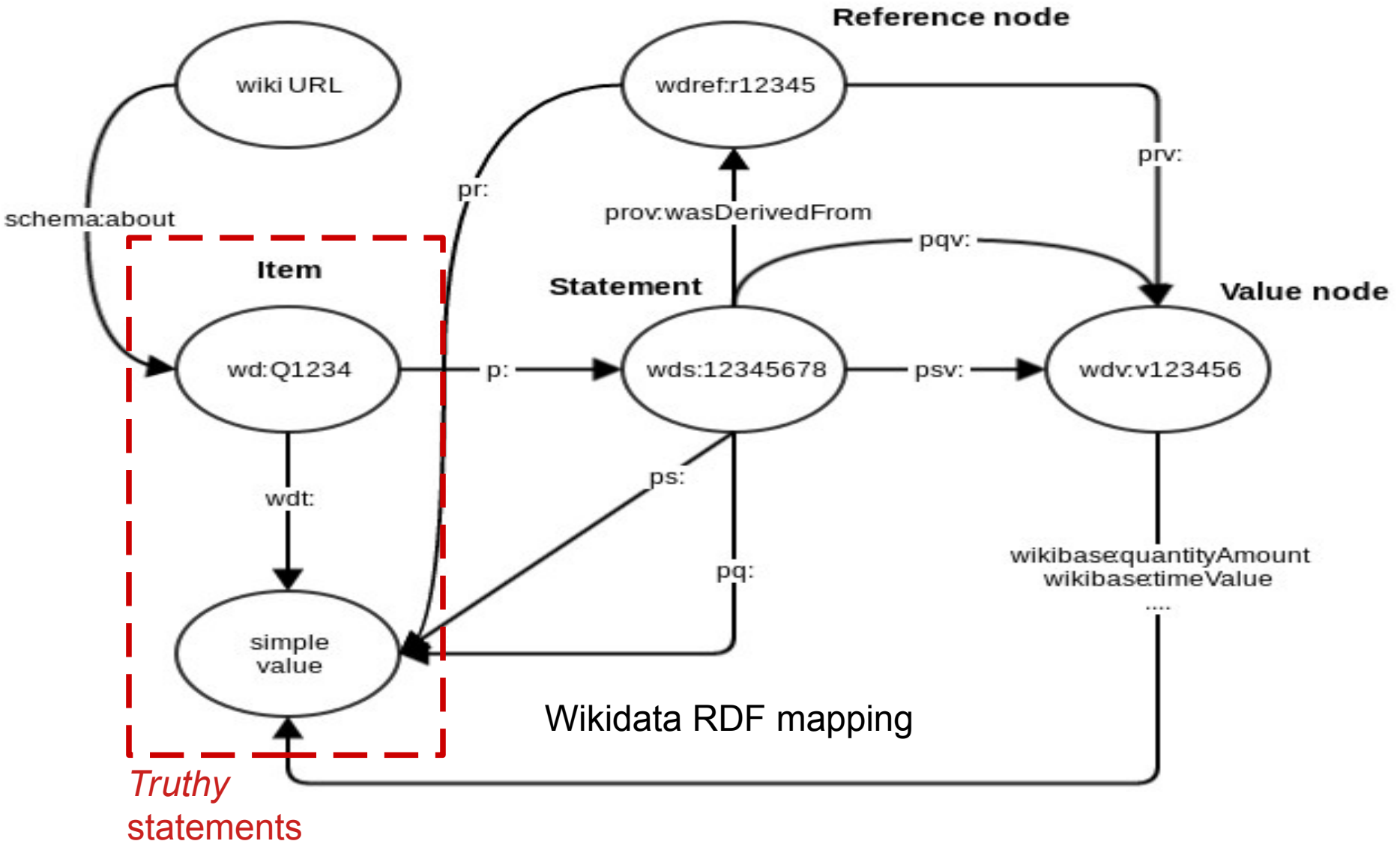
Wikidata item structure

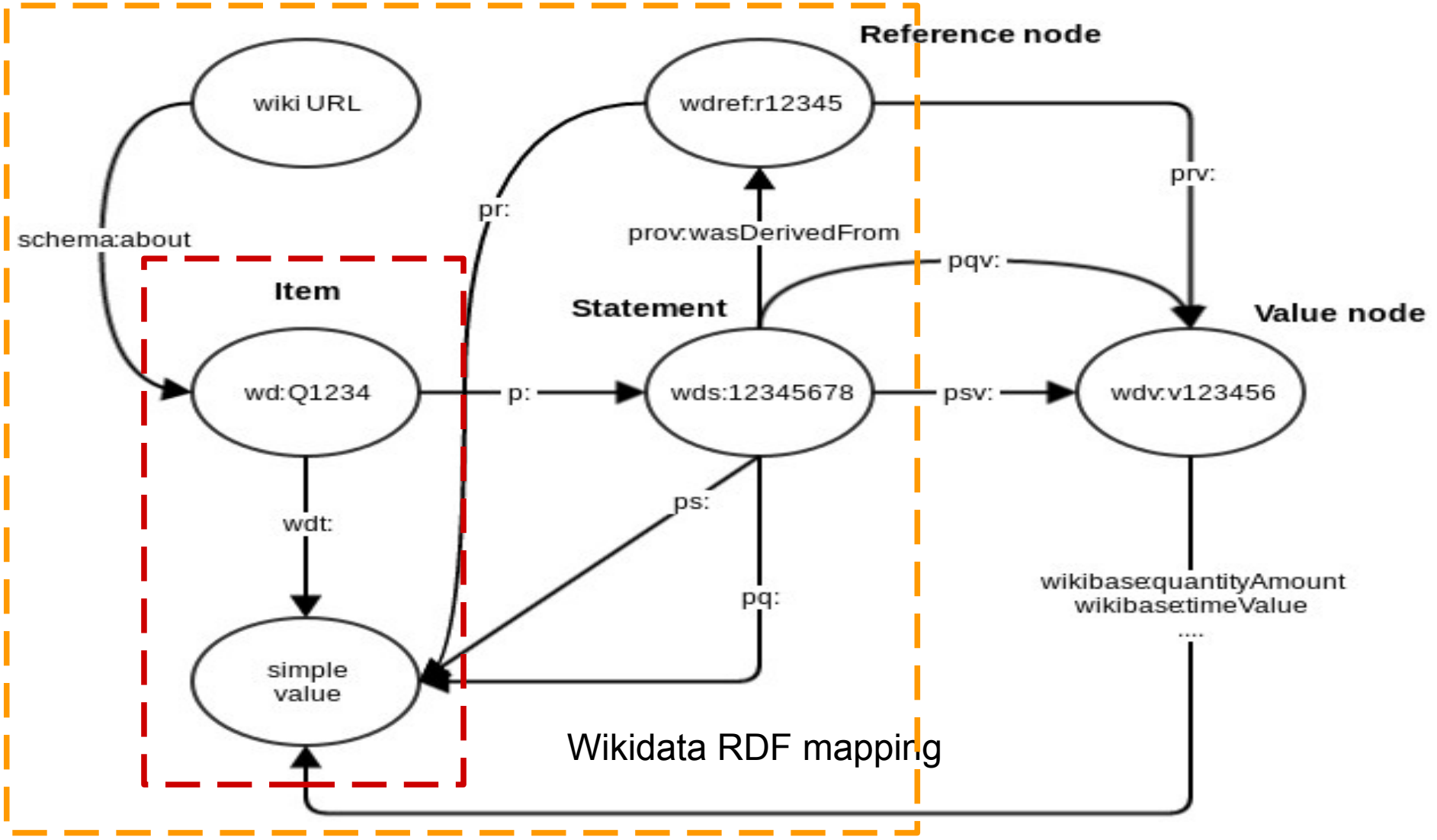
```
wd:Q42
    a wikibase:Item ;
    ...
    wdt:P69 wd:Q691283, ... ;
    … ;
    p:P69 s:q42-0E9C4724-C954-4698-84A7-5CE0D296A6F2 .

s:q42-0E9C4724-C954-4698-84A7-5CE0D296A6F2
    a wikibase:Statement,
        wikibase:BestRank ;
    wikibase:rank wikibase:NormalRank ;
    ps:P69 wd:Q691283 ;
    pq:P580 "1971-01-01T00:00:00Z"^^xsd:dateTime ;
    pq:P582 "1974-01-01T00:00:00Z"^^xsd:dateTime ;
    pq:P812 wd:Q186579 ;
    …. .

wd:P812 a wikibase:Property ;
    rdfs:label "academic major"@en ;
    … .

wd:Q186579 a wikibase:Item ;
    rdfs:label "English literature"@en ;
    … .
```



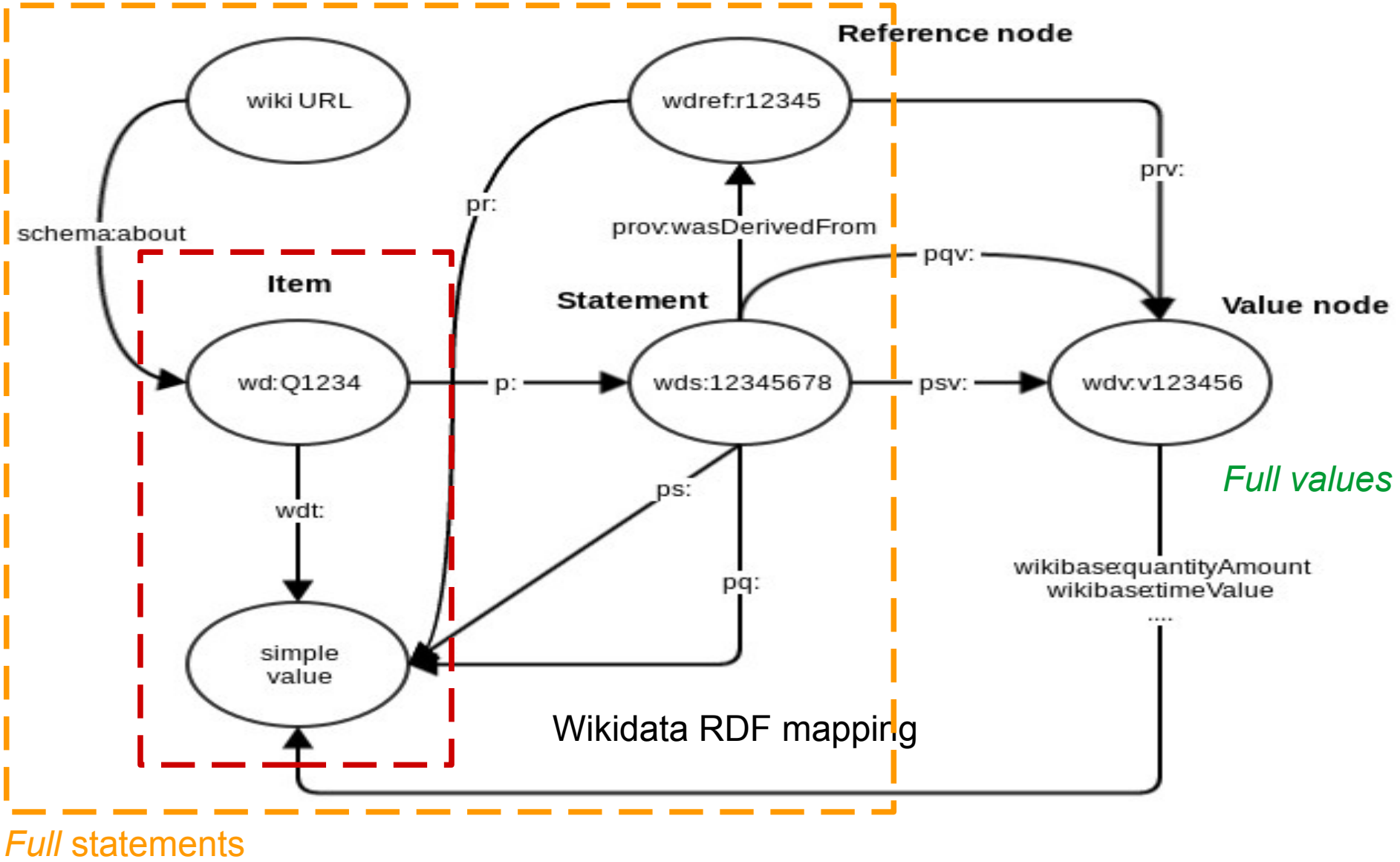
Wikidata as RDF



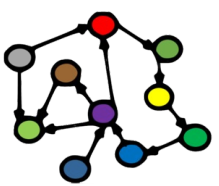

Reference node
wiki URL
wdref:r12345
prv:
pr:
schema:about
prov:wasDerivedFrom
pqv:
Item
Statement
Value node
wd:Q1234
p:
wds:12345678
psv:
wdv:v123456
wdt:
ps:
pq:
wikibase:quantityAmount
wikibase:timeValue
....
simple value
Wikidata RDF mapping
Truthy statements



Full statements


Reference node
wiki URL
wdref:r12345
prv:
pr:
schema:about
prov:wasDerivedFrom
pqv:
Item
Statement
Value node
wd:Q1234
p:
wds:12345678
psv:
wdv:v123456
Full values
wdt:
ps:
pq:
wikibase:quantityAmount
wikibase:timeValue
....
simple value
Wikidata RDF mapping
Full statements

```
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" .
      }
}
```

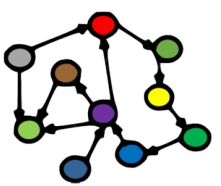


(c) Andreas L Opdahl, 2023

http://query.wikidata.org

```
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" .
      }
}
```

Lots of prefixes built-into the query interface



(c) Andreas L Opdahl, 2023

http://query.wikidata.org

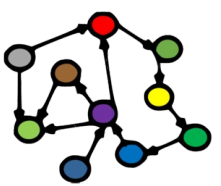
```
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
#defaultView:BubbleChart
```

Lots of prefixes built-into the query interface

Built-invisualisations

```
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" .
      }
}
```

http://query.wikidata.org



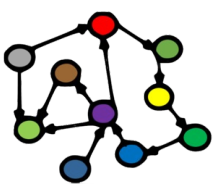
(c) Andreas L Opdahl, 2023

```
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" .
      }
}
```

Lots of prefixes built-into the query interface

Built-invisualisations

Automatic multi-language labelling of resources

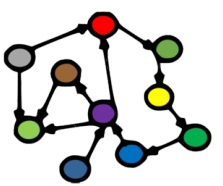


(c) Andreas L Opdahl, 2023

http://query.wikidata.org

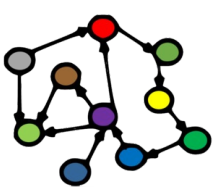
Wikidata Query Service (WDQS)

- SPARQL wrapper for Wikidata (http://query.wikidata.org)
  - based on Blazegraph (and the OpenRDF/RDF4J Java API)
  - lots of built-in prefixes
  - generate query URIs
  - various entity/ontology explorers, e.g.,
    - SQID (https://tools.wmflabs.org/sqid/#/)
  - GraphBuilder
  - built-in visualisations
  - automatic multi-language labelling SERVICE (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

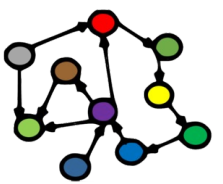


(c) Andreas L Opdahl, 2023

DBpedia

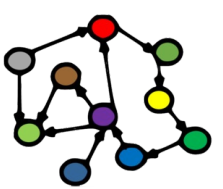
DBpedia

- Extracting structured information from Wikipedia
  - a crowd-sourced community effort
  - making Wikipedia information available as a semantic knowledge graph
  - central source of URIs (in particular before Wikidata):
    - 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)



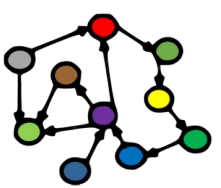
DBpedia: old extraction

- 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

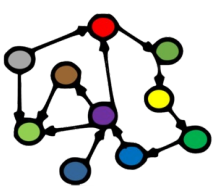
- Wikipedia's *infoboxes* are central
  - raw, automatic transformation from *infoboxes* to triples:
    - language-specific property names
    - infobox templates may be badly defined and used
    - inconsistent properties
    - no literal types, units
  - hand-written scripts from *infoboxes* to triples: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



(c) Andreas L Opdahl, 2023

DBpedia: ontology and identities

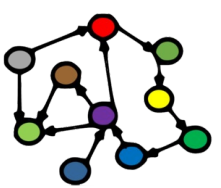
- URIs derived from Wikipedia, e.g.:
  - *http://en.wikipedia.org/wiki/Bergen* →
  - *http://dbpedia.org/resource/Bergen*
  - English, canonical, always *dereferencable URIs*
  - localised/national:
    - *http://no.dbpedia.org/resource/Bergen*
    - not always *dereferencable*
      - …they are *URNs*,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



(c) Andreas L Opdahl, 2023

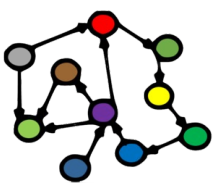
DBpedia: new extraction

- 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



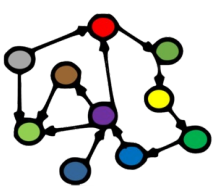
DBpedia ↔ Wikidata

- 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
- Currently, DBpedia *also extracts data directly 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 (seed) for
  - Google's KG
  - Wikidata

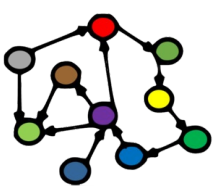


(c) Andreas L Opdahl, 2023

GeoNames

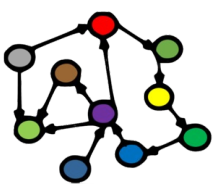
GeoNames

- *Adding geospatial semantic information to the web*
  - a geographical database: *http://www.geonames.org*
  - collected from a large number of sources
  - > 27M geographical names (*toponyms*, Norway 68k),
    > 12M unique features, ~ 4.8M populated places,
    ~ 15M 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 accessibility

- 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 Triple Pattern Fragments:
  - http://data.linkeddatafragments.org/geonames



(c) Andreas L Opdahl, 2023

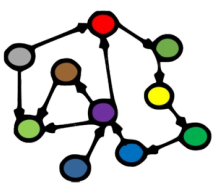
GeoNames ontology

- 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:parentCountry, gn:population, gn:wikipediaArticle
  - also uses properties from *geo, foaf, dcterms, cc, rdfs...*

The Global Database of Events, Language, and Tone (GDELT)

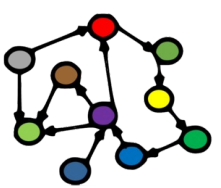
The GDELT project

- Global Database of Events, Language, and Tone (GDELT)
  - free open platform
  - monitors and analyses the world's broadcast, print, and web news
  - identifies people, locations, organizations, themes, sources, emotions, counts, quotes, images, events
  - global, covers over 100 languages, 65 fully translated (incl nb & nn)
  - focus on crises, but much broader in practice
  - *"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?"*



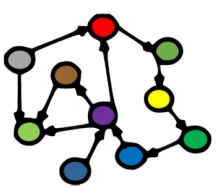
The GDELT project

- Archives back to 1979 (expanding back to 1800)
- Increasingly integrating social media
- Increasingly sophisticated analyses
- *Enormous data quantity – varying and sometimes poor quality*
- Supported by Google Jigsaw
  - runs in the Google Cloud
  - available on Google BigQuery
- Almost a knowledge graph, but
  - not native RDF
  - not fully linked
  - no ontology



The GDELT project

- Archives back to 1979 (expanding back to 1800)
- Increasingly integrating social media
- Increasingly sophisticated analyses
- *Enormous data quantity, limited quality*
- Supported by Google Jigsaw
  - runs in the Google Cloud
  - available on Google BigQuery
- Almost a knowledge graph, but
  - not native RDF
  - not fully linked
  - no ontology

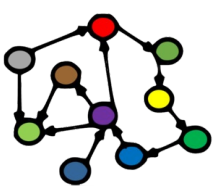


(c) Andreas L Opdahl, 2023

https://www.gdeltproject.org/

The GDELT project: data streams

- Downloadable CSV files (every 15 minutes)
  - http://data.gdeltproject.org/gdeltv2/lastupdate.txt
  - *Global Knowledge Graph (...gkg.CSV, ~15M)*
    - which *people, locations, organizations, themes, sources, emotions, counts, quotes, images, events are mentioned where?*
  - *Events* (...export.CSV, ~350k)
    - low-level actor - event type – actor triples
  - *Mentions* (...mentions.CSV, ~600k)
    - where in and which source is each event mentioned?
- Lots of other datasets and streams, raw and analysed, native language or translated to English



Example

- A 7.8 magnitude tremor struck Turkey on 2023-02-06T0117

<https://ground.news/article/turkey-earthquake-anguished-pm-modi-says-india-ready-to-provide-assistance>

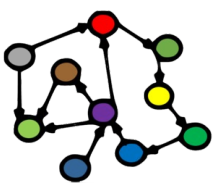
India to send disaster relief teams to quake-hit Turkey

Summary by Ground News

The decision to help Turkey was taken in a meeting led by P.K. Mishra, Principal Secretary to the Prime Minister. The meeting decided to send search and rescue teams of the National Disaster Response Force (NDRF) and medical professionals. Two teams of NDRF comprising 100 personnel with specially trained dog squads and medical personnel have been prepared to be deployed.

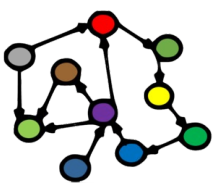
Published 21 days ago · New Delhi, India

- Downloaded GDELT files from 1200 the same day:
  - http://data.gdeltproject.org/gdeltv2/20230206120000.gkg.csv.zip
  - http://data.gdeltproject.org/gdeltv2/20230206120000.export.CSV.zip
  - http://data.gdeltproject.org/gdeltv2/20230206120000.mentions.CSV.zip



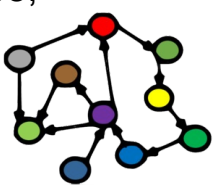
The GDELT project: GKG (the Global Knowledge Graph)

- For each document:
  - record id and datetime
  - source and document identifier (e.g., a URL)
  - keywords/themes (GKG has 120k entries, GCAM 1.2M entries)
  - person and organisation names and types
  - locations, they 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



The GDELT project: GKG (line 670 in ...gkg.csv)

- **GKGRECORDID, V2.1DATE:**
  20230206120000-669, 20230206120000
- **V2SOURCECOLLECTIONIDENTIFIER:**
  1 (WEB)
- **V2SOURCECOMMONNAME:**
  ground.news
- **V2DOCUMENTIDENTIFIER:**
  https://ground.news/article/...
- **V1COUNTS, V2.1COUNTS:**
  KILL#300##1#Turkey#TU#TU#39.059012#34.911546#TU#1025;
  WOUND#200##1#Syria#SY#SY#35#38#SY#1052;
  ...
- **V1THEMES, V2ENHANCEDTHEMES:**
  GENERAL_GOVERNMENT,1313;
  NATURAL_DISASTER_RICHTER_SCALE,1767;
  TAX_FNCACT_PRIME_MINISTER,1614;
  ...
- **V1LOCATIONS, V2ENHANCEDLOCATIONS:**
  1#India#IN#IN##20#77#IN#14;
  1#Turkey#TU#TU##39.059012#34.911546#TU#1788;
  1#Syria#SY#SY##35#38#SY#1013;
  ...
- **V1.5TONE:**
  3.536,0.321,3.858,4.180,21.864,0.643,276
- **V2.1GCAM** (Global Content Analysis Measures)**:**
  wc:276,c12.1:7,c12.10:38,c12.12:11,c12.13:24...
- **V2.1ALLNAMES:**
  Prime Minister Narendra Modi,146;
  New Delhi,747; New Delhi,1060; ...
- **V2.1AMOUNTS:**
  8,magnitude on Richter scale,1445;
  500,people have been killed,358;
  8,struck central Turkey,799; ...
- ...and a lot more:
  images, videos, quotations...



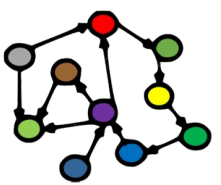
The GDELT project: events

- 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

https://www.gdeltproject.org/

The GDELT project: events (line 372 in export.CSV)

- **GlobalEventID:**
  1083595865
- **Day, MonthYear, Year, FractionDate:**
  20230206, 202302, 2023, 2023.0986
- **Actor1Code, Actor1Name:**
  GOV, PRIME MINISTER
- **Actor2Code:** GOV
- **IsRootEvent:** 0 (not in first sentence)
- **EventCode, EventBaseCode, EventRootCode:**
  13 (Make optimistic comment)
  13 (Make optimistic comment)
  1 (MAKE PUBLIC STATEMENT)
- **QuadClass:**
  1 (verbal cooperation)
- **GoldsteinScale:**
  0.4 (potential impact on stability [-10, +10])
- **NumMentions, NumSources, NumArticles:**
  10, 1, 10 (first 15 min)
- **AverageTone:**
  -3.858 (first 15 min, [-100, +100])
- **Actor1Geo_Type, _Fullname, _CountryCode, _ADM1Code, _Lat, _Long, _FeatureID:**
  1 (COUNTRY), Turkey, TU,
  TU, 39.059, 34.911, TU
- **Actor2Geo_...:**
  (empty)
- **Action_Type, _Fullname, _CountryCode, _ADM1Code, _Lat, _Long, _FeatureID:**
  1 (COUNTRY), Turkey, TU,
  TU, 39.059, 34.911, TU
- **Dateadded:**
  20230206120000
- **SourceURL:**
  https://ground.news/article/...



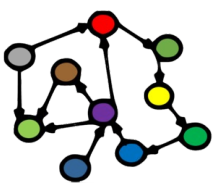
(c) Andreas L Opdahl, 2023

https://www.gdeltproject.org/

The GDELT project: mentions

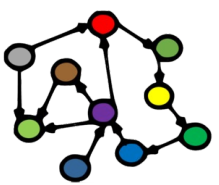
GDELT mentions connect events to documents mapped into the GKG

- 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



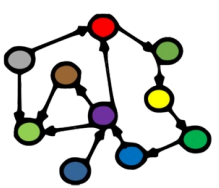
The GDELT project: mentions (line 1042 in mentions.CSV)

- **GlobalEventID**:
  1083595865
- **EventTimeDate, MentionTimeDate:**
  20230206120000, 20230206120000
- **MentionType:**
  1 (WEB)
- **MentionSourceName:**
  ground.news
- **MentionIdentifier:**
  https://ground.news/article/...
- **MontionDocLen:**
  1896
- **MentionsDocTone:**
  -3.85852090032154
- **SentenceID:**
  11
- **Actor1CharOffset:**
  1664
- **Actor2CharOffset:**
  -1
- **ActionCharOffset:**
  1684
- **InRawText:**
  1 (found in original unaltered raw text)
- **Confidence:**
  100



The GDELT project: additional 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



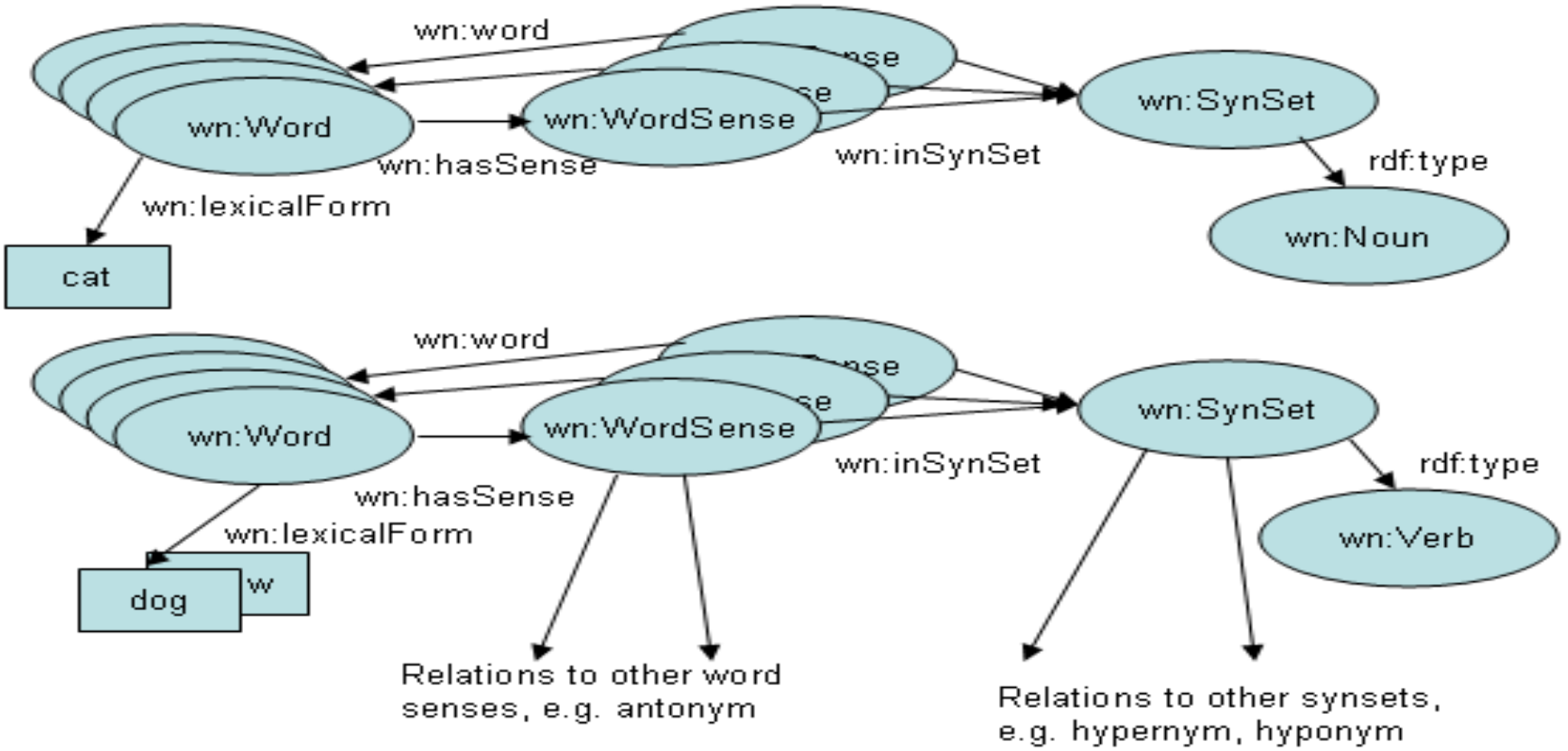
(c) Andreas L Opdahl, 2023

https://www.gdeltproject.org/

WordNet, BabelNet, and ConceptNet

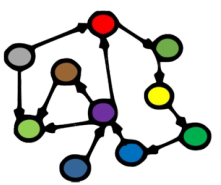
WordNet

- An electronic open-source dictionary (Miller, 1985-)



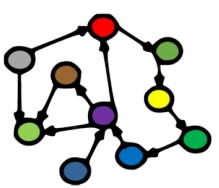
WordNet

- 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: synset structure

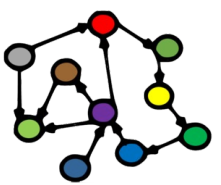
- Different *concept relations* for each *Part of Speech (PoS)*
- Nouns:
  - hyponyms/hypernyms
    bat-n-1 is-kind-of placental_mammal-n-1
  - type / instance
    Norway-n-1 instance-of Scandinavian_country-n-1
  - holonyms/meronyms
    bat-n-1 has-part wing-n-1
  - *antonyms*
    birth-n-1 has-antonym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



(c) Andreas L Opdahl, 2023

WordNet: synset structure

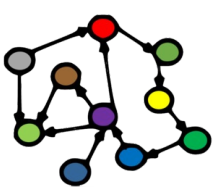
- Different *concept relations* for each *Part of Speech (PoS)*
- Verbs:
  - troponyms/hypernym
    communicate-v-2 has-troponym talk-v-2
    talk-v-2 has-troponym whisper-v-1
    - depending on semantic field:
      run-v-1 has-troponym jog-v-3
      like-v-2 has-troponym love-v-2
  - verb groups
  - antonyms
    love-v-1 has-antonym hate-v-1
  - similarity, sister terms
    bat-v-1 has-sister swat-v-1



(c) Andreas L Opdahl, 2023

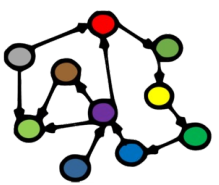
WordNet: synset structure

- Different *concept relations* for each *Part of Speech (PoS)*
- Adjectives:
  - semantic, similarity, antonyms, indirect antonyms
- Adverbs:
  - similar to adjectives
- Also cross-PoS:
  - island – islander (derived from)
  - talk – speak for (phrasal)...
  - ...and others



BabelNet

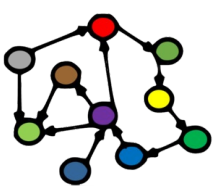
- A multilingual encyclopedic dictionary and a semantic network of concepts and named entities
  - integrates data from *WordNet, Open Multilingual Wordnet, Wiktionary, Wikidata, Wikipedia, Wikiquotes, GeoNames* and several others
  - both *lexicographic* and *encyclopedic* coverage
  - 16 million Babel synsets
  - > 800 million word senses
  - > 280 languages
- *Open* alternatives:
  - DBpedia Wiktionary (http://data.linkeddatafragments.org/wiktionary)
  - Dbnary (http://kaiko.getalp.org/about-dbnary/)
  - Global Wordnet Grid (http://globalwordnet.org/)
  - Open Multilingual WordNet (compling.hss.ntu.edu.sg/omw/)



BabelNet availability

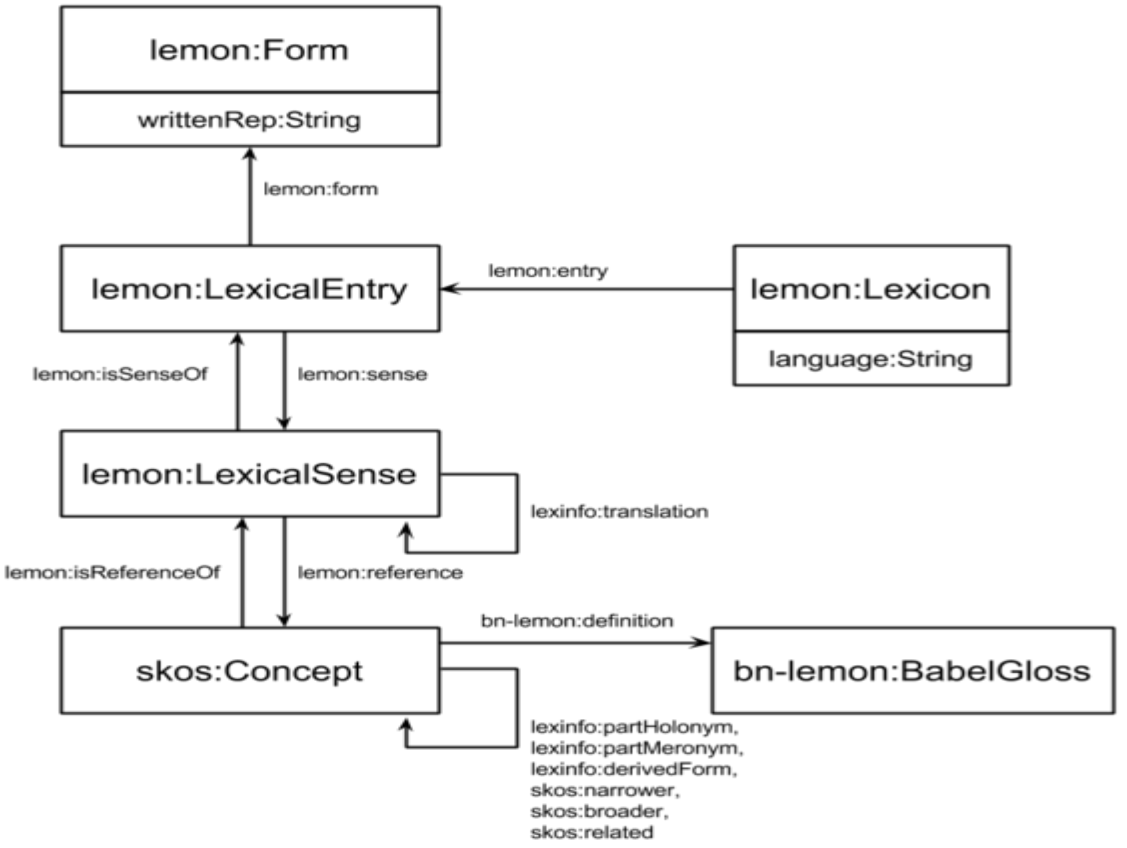
- Available as:
  - web lookup service
  - web translation service
  - web API (JSON) with Java library
  - SPARQL endpoint
  - linked data interface
    - http://babelnet.org/rdf/page/
    - the Linguistic LOD (LLOD) cloud
  - Attribution-NonCommercial-ShareAlike 3.0
    - *but they do not give it out to everyone*

- *Properly open* alternatives:
  - DBpedia Wiktionary (http://data.linkeddatafragments.org/wiktionary)
  - Dbnary (http://kaiko.getalp.org/about-dbnary/)
  - Global Wordnet Grid (http://globalwordnet.org/)
  - Open Multilingual WordNet (compling.hss.ntu.edu.sg/omw/)



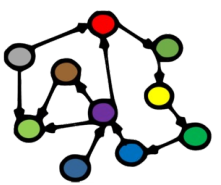
BabelNet conceptual model

- “Making BabelNet part of the LLOD cloud”
- Similar structure to WordNet
- Standard LLOD vocabularies:
  - Lemon
  - BabelNet-lemon
  - LexInfo
  - SKOS
  - RDFS
  - DC elements
  - DC terms
- Lemon is the backbone



ConceptNet

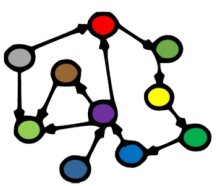
- An open, multilingual knowledge graph
- Designed to help computers understand the meanings of words
- Open, freely-available:
  - web interface (https://conceptnet.io), web API (JSON-LD), downloadable files
- Data sources:
  - DBpedia
  - Wiktionary
  - Open Multilingual WordNet
  - OpenCyc
  - Word associations from “games with a purpose”



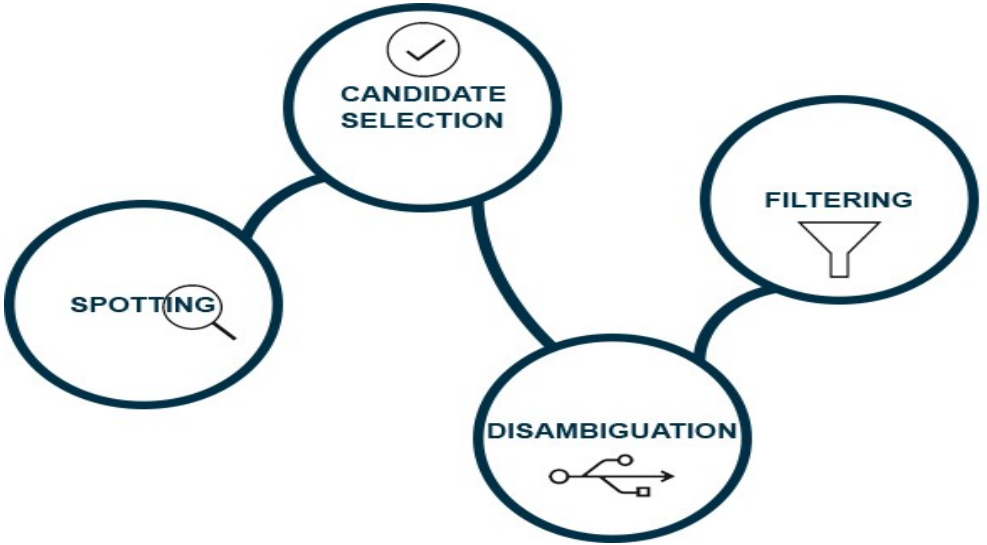
DBpedia Spotlight

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/annotate?text="…"
    - download:
      - open source code
      - Docker



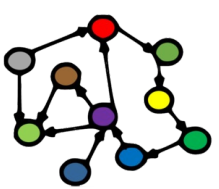
DBpedia Spotlight



- 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



Next week:
Enterprise KGs

Guest Lecture by
Sindre Asplem, CapGemini