# Ontologies, Knowledge Bases, Wikidata

MPRI 2.26.2: Web Data Management

Antoine Amarilli

Friday, January 11th

# Reminder

- Ontology: vocabulary (classes and relations) to describe things
- Knowledge base: set of facts in one or several ontologies

→ Focus on Wikidata: a general-purpose knowledge base and ontology

# Ontologies

# Ontologies

- Various domain-specific vocabularies used across knowledge bases
- One general-purpose ontology used by Google, Microsoft, Yahoo, Yandex: `schema.org`
- Other ontologies that come together with a knowledge base

# Friend of a friend (FOAF)

Describe people, relationship, profiles, activities (social network)

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<#JW>
    a foaf:Person ;
    foaf:name "Jimmy Wales" ;
    foaf:mbox <mailto:jwales@bomis.com> ;
    foaf:homepage <http://www.jimmywales.com> ;
    foaf:nick "Jimbo" ;
    foaf:depiction <http://www.jimmywales.com/aus_img_small.jpg>
    foaf:interest <http://www.wikimedia.org> ;
    foaf:knows [
        a foaf:Person ;
        foaf:name "Angela Beesley"
    ] .
```

# Creative Commons

Describe the license and rights on documents

```
<div about="http://lessig.org/blog/"
  xmlns:cc="http://creativecommons.org/ns#">
  This page, by <a property="cc:attributionName"
      rel="cc:attributionURL"
      href="http://lessig.org/">Lawrence Lessig</a>,
  is licensed under a <a rel="license"
      href="http://creativecommons.org/licenses/by/3.0/">
      Creative Commons Attribution License</a>.
</div>
```

- Many content providers add this kind of markup (e.g., Flickr)
- Search engines can use it (e.g., Google)

# Other domain-specific ontologies

- Dublin Core (DC): Describe digital resources (videos, images, etc.) and physical resources (books, CDs, etc.)
- Simple knowledge organization system (SKOS): describe thesauri, taxonomies, etc.
- Open Graph Protocol: metadata for Web pages to be integrated in Facebook's social graph; also Twitter Cards for Twitter
- DOAP (Description of a Project): describe software projects
- VoID (Vocabulary of Interlinked Datasets): describe a linked dataset
- Countless others

# Schema.org: a general-purpose ontology

- General-purpose ontology: 598 types and 862 properties in version 3.5
- Intended to be used on Web pages to annotate the semantics of elements
- Used by search engines for rich search results
- Used in over 10 million sites[1]

[1] Source: `https://schema.org/`

# Format: Microdata

```
<div class="event-wrapper" itemscope itemtype="http://schema.org/Event">
  <div class="event-date" itemprop="startDate"
      content="2013-09-14T21:30">Sat Sep 14</div>
  <div class="event-title" itemprop="name">
      Typhoon with Radiation City</div>
  <div class="event-venue" itemprop="location"
      itemscope itemtype="http://schema.org/Place">
    <span itemprop="name">The Hi-Dive</span>
    <div class="address" itemprop="address" itemscope
        itemtype="http://schema.org/PostalAddress">
      <span itemprop="streetAddress">7 S. Broadway</span><br>
      <span itemprop="addressLocality">Denver</span>,
      <span itemprop="addressRegion">CO</span>
      <span itemprop="postalCode">80209</span>
    </div>
  </div>
  <div class="event-time">9:30 PM</div>
</div>
```

- `itemscope` creates an item and `itemtype` gives its type
- `itemprop` gives values for properties of the item

# Format: RDFa

Competing format to Microdata, seems less common[2]

```
<div vocab="http://schema.org/" class="event-wrapper" typeof="Event">
  <div class="event-date" property="startDate"
      content="2013-09-14T21:30">Sat Sep 14</div>
  <div class="event-title" property="name">
      Typhoon with Radiation City</div>
  <div class="event-venue" property="location" typeof="Place">
    <span property="name">The Hi-Dive</span>
    <div class="address" property="address" typeof="PostalAddress">
      <span property="streetAddress">7 S. Broadway</span><br>
      <span property="addressLocality">Denver</span>,
      <span property="addressRegion">CO</span>
      <span property="postalCode">80209</span>
    </div>
  </div>
  <div class="event-time">9:30 PM</div>
</div>
```

[2] http://webdatacommons.org/structureddata/index.html#toc2

# Format: JSON-LD

Alternative approach: give the structured data separately in JSON

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Event",
  "location": {
    "@type": "Place",
    "address": {
      "@type": "PostalAddress",
      "addressLocality": "Denver",
      "addressRegion": "CO",
      "postalCode": "80209",
      "streetAddress": "7 S. Broadway"
    },
    "name": "The Hi-Dive"
  },
  "name": "Typhoon with Radiation City",
  "startDate": "2013-09-14T21:30"
}
</script>
```

- The `@context` attribute gives the namespace for the `@type`.
- No longer gives any link to the page contents
- Also `@id` to give an URI to a node
- Many other features (editor's draft of the spec is 167 pages)

# Web Data Commons Structured Data

- Extraction of semantic content from the Common Crawl
- Also useful to measure usage of structured data:
  - In November 2017, the Common Crawl contained 66 TB (compressed), 260 TB (uncompressed), 3.2G pages
  - 39% of pages (and 28% of domains) contained semantic data
  - 9G entities and 38G triples
  - `http://webdatacommons.org/structureddata/`

# Knowledge bases

# Common Knowledge bases

- Generalistic: DBpedia, YAGO, Freebase (defunct), Wikidata
- Proprietary: Google Knowledge Graph, Bing Knowledge Graph (aka Satori)
- Domain-specific
- We will focus afterwards on Wikidata

# DBpedia


- Started in 2007
- License: CC-BY-SA
- Code license: GPLv2
- Actors: Leipzig University, University of Mannheim, Open Link Software
- Latest release: 2016-10
- Extracted from Wikimedia projects
- 6M entities and 10G triples in 2016-04[3],

[3] https://blog.dbpedia.org/2016/10/19/yeah-we-did-it-again-new-2016-04-dbpedia-release/

# YAGO


- Started in 2008
- License: CC-BY
- Code license: GPLv3
- Actors: Max Planck Institute for Informatics, Télécom ParisTech
- Latest release: YAGO 3.1 (2017)
- Extracted from Wikipedias and other sources; manual evaluation
- 10M entities and 120M triples[4],

[4] http://yago-knowledge.org/

# Freebase


- Started in 2007, discontinued in 2016
- License: CC-BY
- Code license: Apache2 (provided after-the-fact by Google)
- Actors: Metaweb, acquired by Google in 2010
- Initially imported from various sources
- Could be edited by anyone
- Partially imported into Wikidata (but not completely)
- Last release: 2016
- Last dump has 1.9G triples

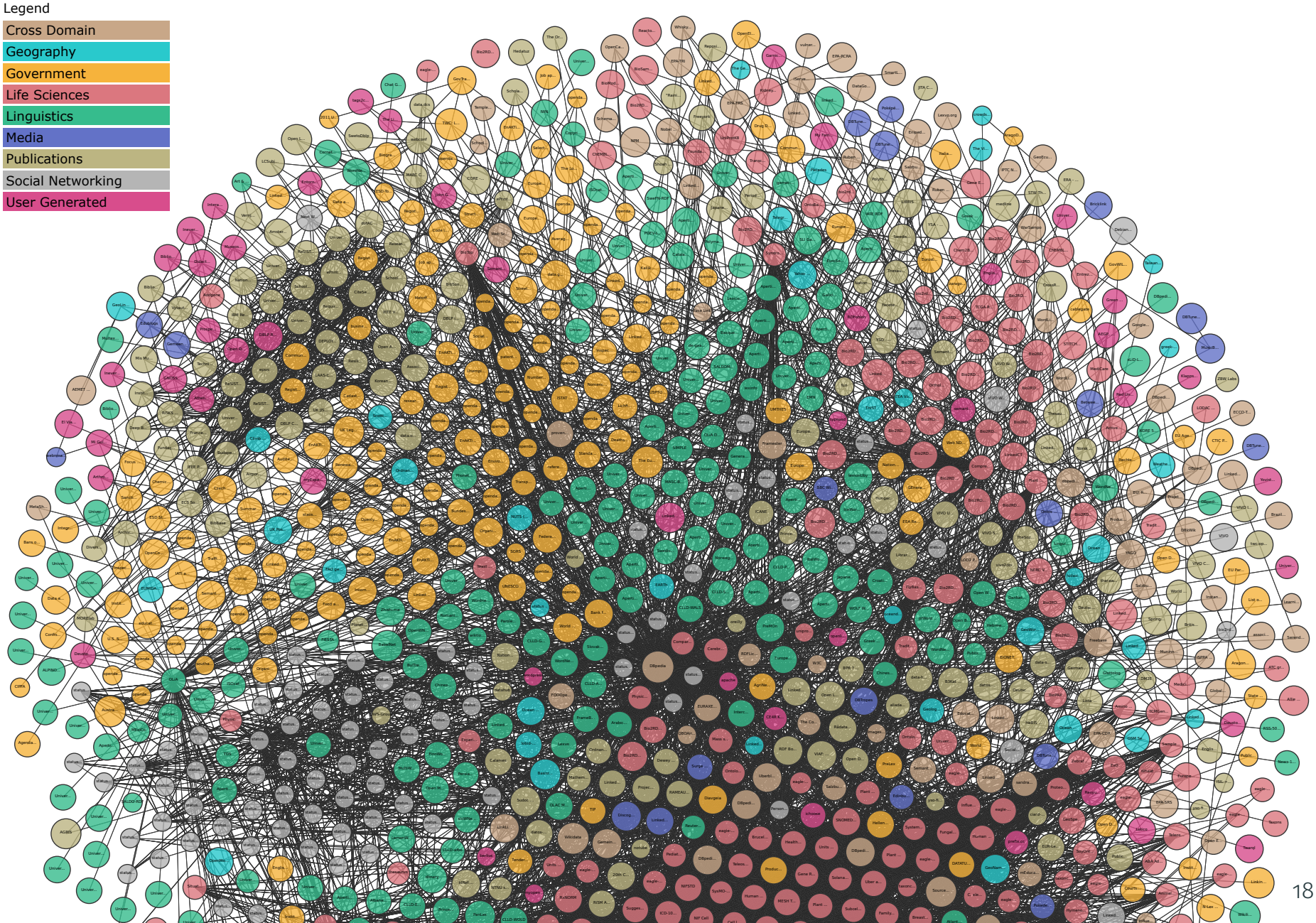
# Wikidata


- Started in 2012
- License: public domain
- Code license: GPLv2
- Actors: Wikimedia Deutschland, Wikimedia
- Last release: weekly
- Around 650M statements and 54M items
- Can be edited by anyone! Around 20k active users.

# Domain-specific

- MusicBrainz, for CDs and music in general (20 million recordings)
- British National Bibliography: bibliographic details about books published in the UK since 1950
- data.bnf.fr, data from the French national library
- OpenStreetMaps, and Geonames
- Medicine and chemistry with SNOMED CT, and other databases: DrugBank, KEGG, UniProt, ChEMBL, etc.
- Linguistic resources, e.g., Babelnet
- Bibliography, e.g., DBLP, Crossref

# Linked Open Data



# Gathering Semantic Web Data

- Browsing online versions of KBs
- Using ad-hoc APIs to retrieve relevant triples
- Using a SPARQL endpoint
- Downloading a dump
- Crawling other knowledge bases, e.g., dereferencing Cool URIs

# Systems

- RDF stores (triplestores) with relational or native backend, open-source or commercial, related to graph databases
  - Apache Jena
  - Virtuoso
  - Blazegraph, essentially acquired by Amazon
  - Amazon Neptune
- SPARQL engines, usually on top of a triplestore. `http://en.wikipedia.org/wiki/SPARQL`
- Tool to view semantic data in Web pages: `http://www.google.com/webmasters/tools/richsnippets`

# Semantic Web challenges

- Complexity:
  - Writing structured content is harder than writing text!
  - Using structured content (with heterogeneous schema) is complicated!
  - Discoverability problem for knowledge bases, vocabularies
- Performance:
  - Data is large
  - Running queries on graphs is tricky
  - Reasoning makes it even worse
  - Federation makes things worse again

# Semantic Web challenges, cont'd

- Data quality:
  - Vagueness and modeling issues
  - Trust (anyone can add a triple)
  - Canonicity and alignment
  - Temporality, sources often complicated to represent
  - Open-world semantics: missing values vs no values
- Incentives: many data providers do not want to be eaten by others

# Wikidata

# Why Wikidata matters


- Backed by the Wikimedia foundation: credible and noncommercial
- Not run by academics, but some academics are involved
- Genuine uses on Wikipedia (to some extent)
- Centralized model, which is a good idea for now
- Good tradeoffs in terms of expressiveness, scope…
- Uses the successful wiki model

# Wikidata basics

- Entities: Q1, Q2, Q3, ..., Q60527475 and beyond
- Properties: P1, P2, P3, ..., P6343 and beyond

# Wikidata basics

- Entities: Q1, Q2, Q3, ..., Q60527475 and beyond
- Properties: P1, P2, P3, ..., P6343 and beyond
- Entities and properties have a label and short description in each language, along with aliases (search engine)
- Entities can also have sitelinks to Wikimedia projects (e.g., the corresponding Wikimedia pages)

# Wikidata basics

- Entities: Q1, Q2, Q3, ..., Q60527475 and beyond
- Properties: P1, P2, P3, ..., P6343 and beyond
- Entities and properties have a label and short description in each language, along with aliases (search engine)
- Entities can also have sitelinks to Wikimedia projects (e.g., the corresponding Wikimedia pages)
- For each entity and property, we can have facts (or claims) with different objects

# Wikidata basics

- Entities: Q1, Q2, Q3, ..., Q60527475 and beyond
- Properties: P1, P2, P3, ..., P6343 and beyond
- Entities and properties have a label and short description in each language, along with aliases (search engine)
- Entities can also have sitelinks to Wikimedia projects (e.g., the corresponding Wikimedia pages)
- For each entity and property, we can have facts (or claims) with different objects
- Everyone can create and edit entities and facts
- Discussion is needed before creating a property

# Wikidata basics

- Entities: Q1, Q2, Q3, ..., Q60527475 and beyond
- Properties: P1, P2, P3, ..., P6343 and beyond
- Entities and properties have a label and short description in each language, along with aliases (search engine)
- Entities can also have sitelinks to Wikimedia projects (e.g., the corresponding Wikimedia pages)
- For each entity and property, we can have facts (or claims) with different objects
- Everyone can create and edit entities and facts
- Discussion is needed before creating a property
- Software: Wikibase, a set of extensions to Mediawiki

# Qualifiers, references, ranks, data types

- Each fact can have qualifiers to indicate things like start/end time, details (e.g., major/degree for P69 “educated at”)
- Each fact can also have sources to indicate where it comes from (a source is a set of key–value pairs)
- Each fact can have a rank among “normal”, “preferred” (e.g., for the current value), or “deprecated”.
- Literal values can have data types

  `https://www.wikidata.org/wiki/Special:ListDatatypes`
- Also two special values
  - “unknown value” (a value exists but is unknown)
  - “no value” (it is known that there is no value)

# Constraints

- Wikidata has constraints which are only advisory (= you can create violations) and are quite simple. Main ones:
  - “single (best) value constraint”
  - “inverse constraint” (mother vs child), “symmetric constraint”
  - “type constraint”, or requiring/disallowing certain facts
  - “range constraint” “contemporary constraint”, “format constraint”
  - “one-of/none-of constraint” (list of allowed/forbidden values)
  - Requiring/allowing qualifiers or units
  - Allowing use as a qualifier/unit
- There is a mechanism for exceptions
- Many constraint violations in practice

# Usage on Wikipedia

- Used for interwiki links, i.e., the links between Wikipedia pages across languages
- Used in some infoboxes on Wikipedia, e.g., to automatically populate some fields
- Can be used for other things, e.g., filling tables, or external links to other sources
- Policy depends on each Wikipedia: some communities are more welcoming than others...

# Ongoing Wikidata discussions

- Project scope: what belongs in Wikidata?
  - The public domain license is a strong requirement
  - Concerns, e.g., about the high number of bibliographic entities (almost half of the entities)
  - Some external datasets are imported, but Wikipedia (historically) gave much importance to human validation of imports
  - Some support for federation in queries; and many external links
- Notability: essentially no policy currently
- Managing vandalism?
- Importance of references?

# Accessing Wikidata data

- Simply by browsing
- Can retrieve in multiple formats, e.g., `https://www.wikidata.org/wiki/Special:EntityData/Q42.json`
- For simple queries (triple patterns), Linked data fragments `https://query.wikidata.org/bigdata/ldf`
- Wikimedia API, e.g., API for recent changes
- SPARQL queries, `https://query.wikidata.org/` (and API)
- Weekly dumps in JSON, RDF, XML (around 50 GB compressed)

# Other cool Wikidata stuff

- Distributed Wikidata Game: crowdsourcing edits on Wikidata `https://tools.wmflabs.org/wikidata-game/distributed/`
- Reasonator: automatically generate a Wikipedia-like page from a Wikidata entity `https://tools.wmflabs.org/reasonator/`
- Lexemes: ongoing effort to add linguistic data to Wikidata
- OWL ontology: `http://wikiba.se/ontology`
- askplatyp.us: natural language question answering tool
- File captions on Wikimedia Commons to have a structured way to give labels to images (deployed on January 10)
- OpenRefine to reconcile datasets with Wikidata and add Wikidata facts `https://www.wikidata.org/wiki/Wikidata:Tools/OpenRefine/Editing/Tutorials/Video`

# Slide acknowledgements


- Many thanks to Thomas Pellissier-Tanon for his helpful feedback
- Slide 4: https://en.wikipedia.org/wiki/FOAF_(ontology)
- Slide 5: https://www.w3.org/Submission/ccREL/
- Slide 8–10: https://schema.org/Event
- Slide 13: https://commons.wikimedia.org/wiki/File:DBpediaLogo.svg
- Slide 14: https://en.wikipedia.org/wiki/File:YAGO.svg
- Slide 15: https://commons.wikimedia.org/wiki/File:Freebase_Logo_optimised.svg
- Slide 16, 23: https://en.wikipedia.org/wiki/File:Wikidata-logo-en.svg