15th International ISKO Conference
9-11 July 2018
Porto, Portugal

Proposal for the integration of the semantic structure of Wikipedia categories into Wikidata using SKOS

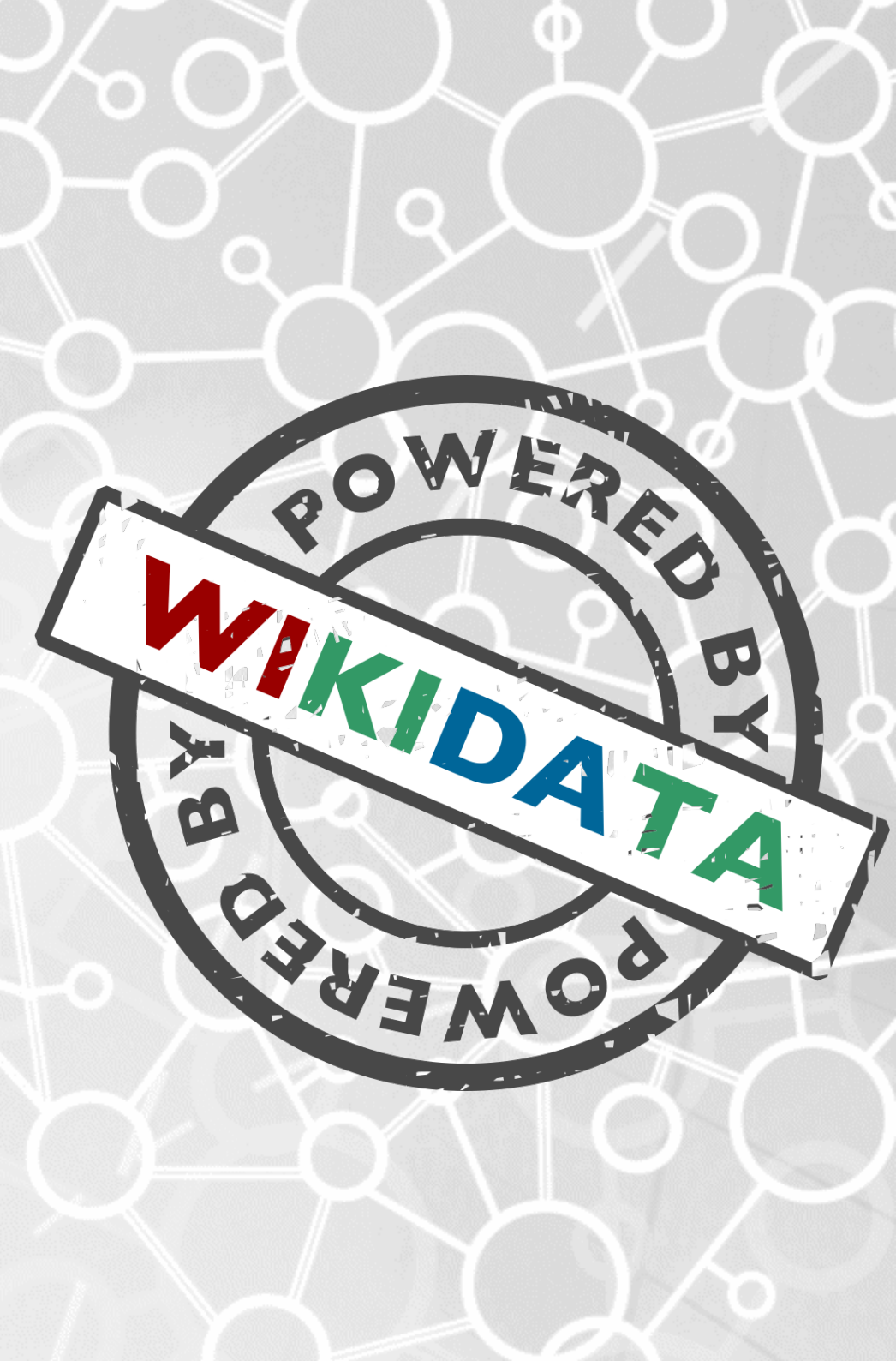
Juan-Antonio Pastor-Sánchez - pastor@um.es
Tomás Saorín Pérez – tsp@um.es
University of Murcia (Spain)

Wikidata in brief

- Wikidata is a project of the Wikimedia Foundation with factual data from Wikipedia articles and other projects.
- Combines the collaborative editing wiki model and structured semantic data to create a knowledge base and structure the encyclopedic information.
- Highly adaptive ad-hoc data model by the community in an open and negotiated way.
- RDF modellization and SPARQL Endpoint.

Wikipedia: Knowledge in “silos”

- Different editions of Wikipedia in different languages, with their own contents, categories, editorial policies, etc.
- Wikidata centralizes knowledge distributed in different Wikipedias, generating unique identifiers for each element.



Main components of Wikidata

Items (Q): Wikipedia pages (articles, categories)

London (Q84) → **Name of the item**

capital of England and the United Kingdom

London, UK | London, United Kingdom | London, England

Labels, descriptions and aliases (multilingual)

In more languages Configure

Language	Label	Description	Also known as
English	London	capital of England and the United Kingdom	London, UK London, United Kingdom London, England
Spanish	Londres	capital de Inglaterra y del Reino Unido	London
Portuguese	Londres	capital do Reino Unido	
French	Londres	capitale du Royaume-Uni	London

Identifiers: Links to external data value (controlled, vocabularies, authorities).

MeSH ID	D008131
	0 references

Claims (statement about items)

population → **Property**

8,787,892 → Value / Item

point in time	2016
determination method	estimation process

→ **Qualifiers (Reification)** Properties of statements

1 reference

reference URL	http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-england-and-wales/mid-2012/mid-2012-population-estimates-for-england-and-wales.html

→ **References** (Verifiability policy)

Sitelinks: links to the equivalents articles in Wikipedia, Wikinews, Wikiquote, etc...

Wikipedia (241 entries)

ab	Лондан
ace	London
ady	Лондон
af	Londen
als	London
am	ለንደን
ang	Lunden
an	Londres
arc	ܠܘܢܕܘܢ
ar	لندن

Categories: The wiki way

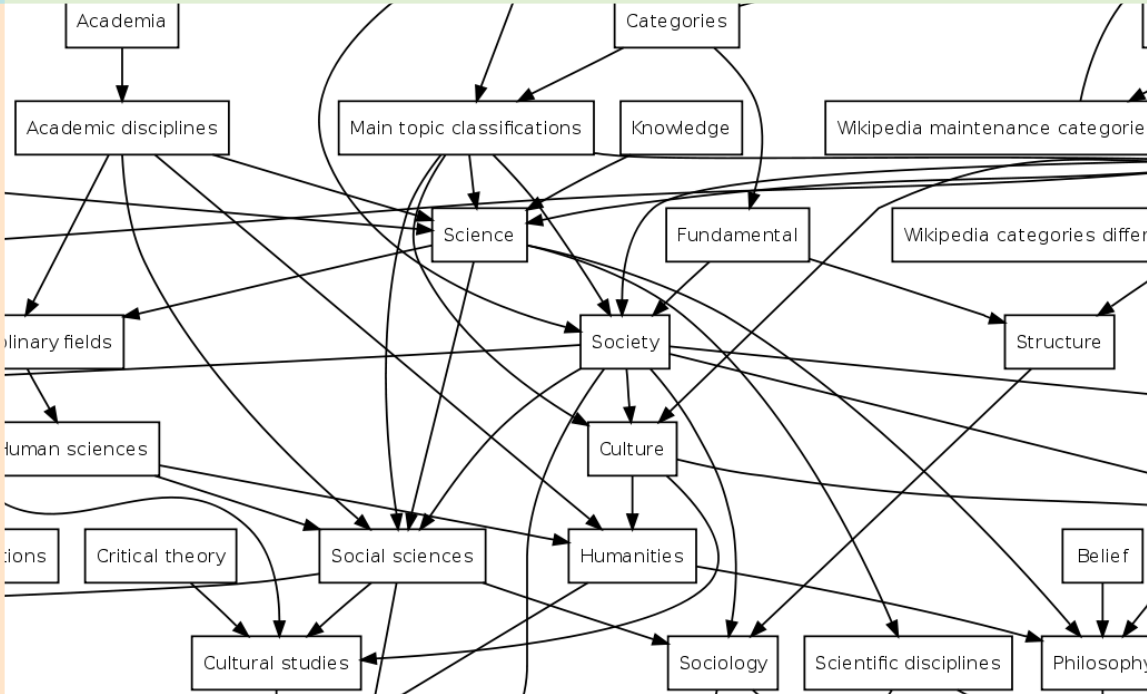
- Wiki categories are the only instrument (quite limited) for the KO in MediaWiki.
- The community of editors defines the categories while editing the articles.
- The categories scheme is a poly-hierarchical directed graph that evolves organically together with the content.
- Creators of the categories in Wikipedia is a clear case of "accidental taxonomist" (Hedden, 2016).

Categories relations

- Broader relationship is the only one available, which is the result of classifying category pages.
- Any other type of relationship is incorporated through an artifice: wiki links in a specific template.
- Categories have correspondence between languages (interwiki), frequently with specific articles (Main topic) and with categories of the Commons.
- Associative relationships are very sparse.
- There are container categories, which only contain other categories and do not index items.

Categories as silos

- Each Wikipedia has a scheme of categories and therefore produces its own system of knowledge organization.
- There aren't a single Taxonomy in Wikipedia categories, but more than 200.



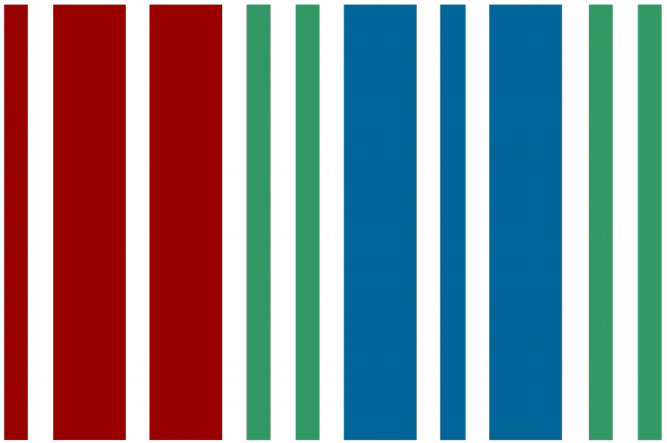
Wikipedia Category oddities

- Unusual and weird uses because it is the only KO instrument available in MediaWiki.
- They are used both to describe/organize encyclopedic knowledge (articles) and for coordination/administration functions.
- Articles do not have descriptive metadata, and categories are used to create geographical, chronological, typological and onomastic subdivisions ("diffusing large categories").
- There are many container-categories or meta-categories, which are not used to index articles, but to organize the category network.
- Combine heterogeneous subdivision criteria: it is a mixture of classes, parts, instances, etc.

How Wikidata community see categories?

- *“WP categories are a kind of classification without global overview. So instead of using an empirical classification, just start the work correctly.”*
- *“Category is no knowledge but an arbitrary way to sort items [...] It would not be possible to source that kind of statement and it would generate a lot of repetitions with the other statements.”*
- *“WP categories are outdated: categories are a specific way to group articles which is not common knowledge but particular point of view.”*
- *“Wikipedia use categories for classification. Wikidata use something else”*
- *“The category system is pointless because there is no unique system, but different ways to classify articles.”*

https://www.wikidata.org/wiki/Wikidata:Property_proposal/Archive/30



Categories describe aspects of the articles, which Wikidata models as properties that connect items-resources with items-classes.

WIKIPEDIA

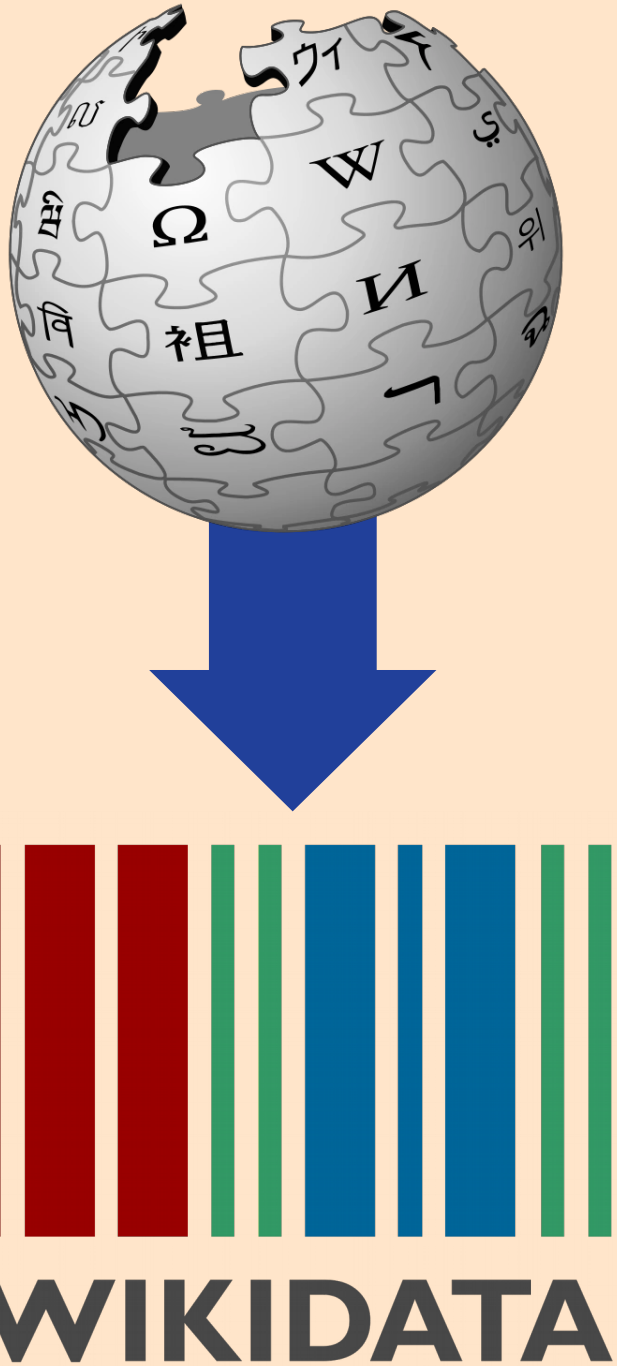
Category:Films set in Stockholm

WIKIDATA

P31 (Instance of) > Q11424 (film);
P915 (filming location) Q506250 (Stockholm Municipality).

How categories are imported in Wikidata

- Categories form a set of elements in the Wikidata graph.
  - P31 (Instance of) Q4167836 (Wikimedia Category)
- Wikidata brings together all the categories of all the Wikimedia projects (including all Wikipedias).
- Other properties used in categories:
  - P301, article for the main topic of that category.
  - category combine topics (P971)
  - category of associated people (P1792)
- IRIs with numerical identifiers: eliminate the linguistic variability (alias) of the categories.
- References in statements:
  - Imported from (P143)> Q8449 (Wikipedia in spanish)



What is missed in Categories in Wikidata

- Wikidata collects only the labels of the categories, does not include the hierarchical relationships.
- They are a mess, but exists in Wikipedia.
- This implies a missrepresentation of the editors' activiy.
- It does not include information on the articles indexed by the categories and neither the categories associated with the articles.
- Categories do not include links to external vocabularies (Library of Congress authority, BNCF Thesaurus, MeSH, ...) but these links are included in the "Main topic" articles.

Problems to obtain category relationships from Wikipedia

- Categories from different editions of Wikipedia are unified when imported in Wikidata (acting as interwiki links).
- No common concepts scheme is defined.
- There is many Wikimedia projects and there is many category hierarchies as there is Wikipedia:
- Must be sourced (verificability os statements)
- Great increase of total data amount of knowledge base.
- Queries by categories are much more inexact than queries by properties.
- Waiting for a bot that transforms their compound meanings into statements.
- Useful properties are yet well established for content ítems but not for categories.

Is it worth to include the hierarchical relationships of the categories in Wikidata?

- “Categories are most assuredly knowledge. A bit messy, lacking in organization (true) but very comprehensive. Wikipedians are very serious about their categorization […] We also don't source Labels and Descriptions, but do you think you could live without them?”
- Categories relations reflect editorial decisions about how encyclopedic knowledge is organized.
- Since Wikidata unifies the categories of different editions of Wikipedia, relationships must also be unified.
- Suggestions of new ways of navigate between categories with a multicultural or multinlingual approach?
- Start point for a new scenario Wikidata is the infrastructure to organize categories in Wikipedia.
- Decisions about provenance: Group the categories into one or several concept schemes? Represent provenance using reification or named graphs?

Skosifying Wikidata categories

- SKOS elements used: skos:Concept, skos:inScheme, skos:prefLabel, skos:altLabel, skos:broader, skos:narrower

Issues

- Single or many “inScheme” for each category …?
- What is the context of each relation?
- What is the case for administrative categories?
- Are useful skos:topConceptOf?
- Not available data for relations with linked vocabularies (skos:broadMatch, skos:exactMatch, etc.)

Sources

- Wikidata for labels.
- Wikipedia for relationships.

Skosification process

- The process must be applied to every category.
- <http://www.wikidata.org/categories> is defined as skos:ConceptScheme
- Every category is defined as skos:Concept and linked to the concept scheme with skos:inScheme
- Get **labels, aliases and sitelinks** of the category from the Wikidata API
  - Labels → skos:prefLabel
  - Aliases → skos:altLabel
- Every sitelink is used to get the **broader and narrower** categories and their equivalent ID entities in Wikidata.
  - Broader → skos:broader
  - Narrower → skos:narrower
- Output in Turtle format with RDF statement.

PHP Script

The Wikidata ID entity of the category is used as parammeter

skos:Concept
skos:inScheme

Wikidata API
action=wbgetentities
prop=labels, aliases, sitelinks

skos:Concept
skos:inScheme

Wikipedia API
action=query
generator=categories
prop=pageprops|categoryinfo

skos:broader

Wikipedia API
action=query
generator=categorymembers
gcmtype=subcat
prop=pageprops|categoryinfo

sskos:narrower

PHP script

php concepts.php Q8310847

https://github.com/j-pastor/wdc

Concept definition

```
...
wd:Q8310847 rdf:type skos:Concept ;
   skos:inScheme <http://www.wikidata.org/categories> ;
```

Labels

```
   skos:prefLabel "رده:بریکینگ بد"@fa ;
   skos:prefLabel "Catégorie:Breaking Bad"@fr ;
   skos:prefLabel "Category:Breaking Bad"@en ;
   skos:prefLabel "Categorie:Breaking Bad"@nl ;
   skos:prefLabel "Категорія:Пуститися берега"@uk ;
   skos:prefLabel "Thể loại:Breaking Bad"@vi ;
   skos:prefLabel "تصنيف:اختلال ضلال"@ar ;
   skos:prefLabel "Categoría:Breaking Bad"@es ;
   skos:prefLabel "Categoria:Breaking Bad"@pt ;
   skos:prefLabel "Kategorie:Breaking Bad"@de ;
   skos:prefLabel "Kategori:Breaking Bad"@tr ;
```

Broader categories

```
...
   skos:broader wd:Q47344019; # Category:2000s American crime drama television series
   skos:broader wd:Q47343067; # Category:2010s American crime drama television series
   skos:broader wd:Q9085046; # Category:AMC (TV channel) network shows
   skos:broader wd:Q7484881; # Category:Methamphetamine
   skos:broader wd:Q8836953; # Category:Television shows set in New Mexico
   skos:broader wd:Q8922775; # Category:Wikipedia categories named after American television series
   skos:broader wd:Q47463906; # Category:Wikipedia categories named after media franchises
```

Narrower categories

```
...
   skos:narrower wd:Q19364091; # Category:Better Call Saul
   skos:narrower wd:Q8310844; # Category:Breaking Bad characters
   skos:narrower wd:Q8310845; # Category:Breaking Bad episodes
   skos:narrower wd:Q42300471; # Categoría:Better Call Saul
   skos:narrower wd:Q21708777; # Categoría:Reparto de Breaking Bad
   skos:narrower wd:Q8310846; # Catégorie:Saison de Breaking Bad
   skos:narrower wd:Q28610898 . # Categoria:Stagioni di Better Call Saul
```

To do list...

- Equivalence of Wikimedia category item (Q4167836) to skos:Concept. Creation of a Wikidata item for concept scheme (skos:ConceptScheme).
- Discussions in Wikidata community about the creation of new items for SKOS class properties "in scheme", "broader category" and "narrower category" and the equivalences with SKOS element.
- Skosify labels.
- Wikidata API and BOT for mass import and daily synchronization relations changes.
- Represent the provenance of relations as references in statements (property P143, Imported from)
- Mapping mechanism to external vocabularies.
- Working group for category refactoring in Wikipedia.
- Development / maintenance of specialized vocabularies and domain ontologies using the Wikidata categories and their hierarchical relationships.
- … and so on...

References

- Akdag Salah, A., Gao, C., Suchecki, K., & Scharnhorst, A. (2011). Generating Ambiguities: Mapping Category Names of Wikipedia to UDC Class Numbers. http://www.networkcultures.org/_uploads/%237reader_Wikipedia.pdf

- Hedden, H. (2016). The accidental taxonomist (2a). Information Today.

- Lambe, P. (2007). Organising knowledge: taxonomies, knowledge and organizational effectiveness. Oxford: Chandos Publishing.

- Nastase, V., Strube, M. (2008). Decoding Wikipedia Categories for Knowledge Acquisition http://www.aaai.org/Papers/AAAI/2008/AAAI08-193.pdf

- Voss, J. (2006) Collaborative thesaurus tagging, the Wikipedia way. https://arxiv.org/abs/cs/0604036

THANK YOU

Any questions?