Modelling challenges

when reusing data on Wikimedia Commons

Jarek Tuszynski
user:jarekt

2023-11-30

This session is recorded: Please mute your microphone and camera when you're not speaking.

DATA MODELLING DAYS
November 30
December 2
2023
WIKIDATA

Introductions

User:Jarekt:

- 17 years on Wikipedia and 14 years as admin on Commons
- Wrote and/or maintains most of the most transcluded modules on Commons, including for infoboxes:
  - Module:Information (83M pages) - can pull required fields from SDC (structured data on Commons)
  - Module:Coordinates (47M) – uses SDC for files and Wikidata for categories
  - Module:Artwork (16M), Module:Institution (17M), Module:Creator (18M), Module:Authority control (19M) – use Wikidata
- Can also talk about Template:Wikidata Infobox used on 5M categories and maintained by User:Mike Peel , who was originally scheduled for this ses

Connecting to Wikidata

- Files on Commons or articles on Wikipedia have 1:1 association with SDC or Wikidata item
- Category pages can look up Wikidata item
  - Option 1: direct link if item has sitelink to category
  - Option 2: indirect link category -> category item -> category's main topic (P301) -> item
- Files on Commons which use Module:Artwork, can:
  - Use template field “wikidata”
  - digital representation of (P6243) was proposed as SDC property creating connection to a single item, but it was approved only for 2D, leaving sculptures , books, movies, etc. unconnected
  - main subject (P921) is also used to look up Wikidata item, but leads to many mismatches, as main subject might link to subject of the artwork (for example depicted person)

Challenge example: nationality and occupation of a person

- Creator template always displays translated nationality and occupation of a person, like: “Polish poet and writer”, “Italian sculptor and medalist”, etc.
- Nationality:
  - Wikidata stores nationality mostly as text in item description
  - 3 proposals for Nationality property were rejected
  - ethnic group (P172), and country of citizenship (P27) can be sometimes used but often leads to nonsense
- Occupation: can be too many to use
- Barack Obama:
  - ethnic group -> African Americans,
  - country of citizenship -> US and Kenya,
  - occupation -> 10 different
  - Resulting in “American-Kenyan politician, lawyer, political writer, community organizer, statesperson and jurist”

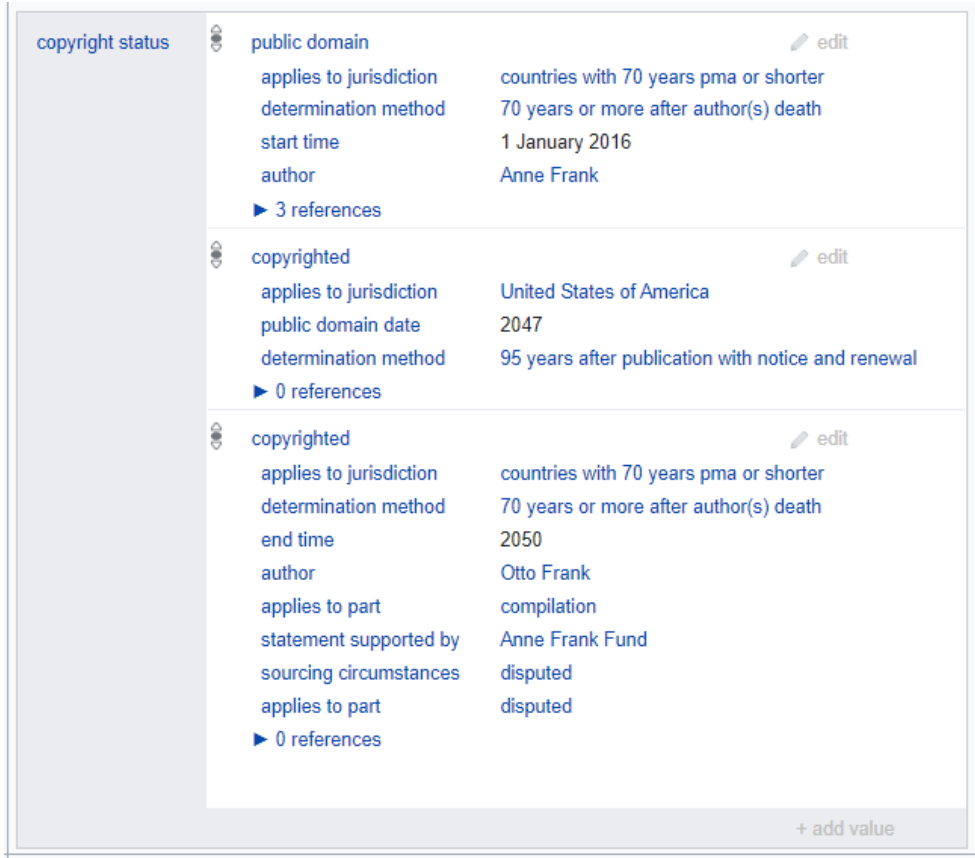
Challenge example: author of files (SDC)

Modeling challenge faced by Module:Information

- Photographs:
  - Photograph taken by person with Wikidata item: creator (P170)->person item
  - Photograph taken by the Commons user: creator (P170)->somevalue (Wikimedia username->username)
  - Photograph taken by the external user: creator (P170)->somevalue (URL (P2699) ->url, author name string (P2093)-> name)
- 2D Artworks:
  - Author with Wikidata Item:
    - digital representation of (P6243) -> artwork item -> creator (P170)->person item
    - creator (P170)->person item (object has role (P3831) -> painter)
  - Named or described author without Wikidata item, (ex. “John Smith”, “16-th century Flemish painter”)
  - Litograph of a painting: use object has role (P3831) to specify roles
- Photographs of artworks:
  - Photograph taken by the Commons user of an 2D artwork with known creator: photographer as above, digital representation of (P6243) -> artwork item
  - Photograph taken by the Commons user of an 3D artwork with known creator : photographer as above, unclear how to link to item

Challenge example: copyrights

- Copyrights:
  - May differ by jurisdiction (US vs. Europe)
  - May have multiple authors
  - Example Anne Frank diary ->
- Two styles:
  - For complex, use bundle of qualifiers form: one bundle per coauthor and jurisdiction
  - For simple case, like {{CC-by-3.0}} use copyright status (6216), copyright license (P275) properties

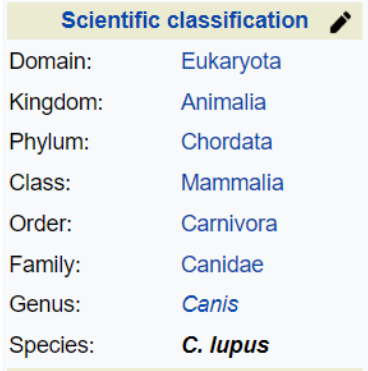


Other complex modeling challenges

- **Dates** (well modeled): “19-th century” , “spring of 1999”, “before 1914”, “11/12 May 1910”, “1921-1925”. Parsed by Module:Wikidata date, I18n by Module:Complex date
- Provenance, or history of an artwork (not well modeled):
  - properties
    - owned by (P127)
    - significant event (P793)
    - collection (P195)
  - Easy to model: current ownership, auctions and sales
  - Hard to model:
    - “previously in the collection of …”
    - “in years … to … on permanent loan from … to …”

How many items have to be accessed?

- Typical artwork will have 3 infoboxes accessing 3 items:
  - Module:Artwork based on artwork item
  - Module:Institution based on museum item
  - Module:Creator based on artist item
- Best practice for infobox is to access limited number of items.
  - Depicted people: might need to loop through all depicted items to determine which one are people items
  - Country or city of birth: when birth location is a hospital or building item
    - Query for Tchaikovsky 's city of death: SELECT DISTINCT ?city { ?city ^(wdt:P20/wdt:P131*) wd:Q7315; wdt:P31/wdt:P279* wd:Q515 . }
  - scientific classification of animals or plants: require traversing a chain of parent taxon (P171) properties



Scientific classification	
Domain:	Eukaryota
Kingdom:	Animalia
Phylum:	Chordata
Class:	Mammalia
Order:	Carnivora
Family:	Canidae
Genus:	*Canis*
Species:	***C. lupus***

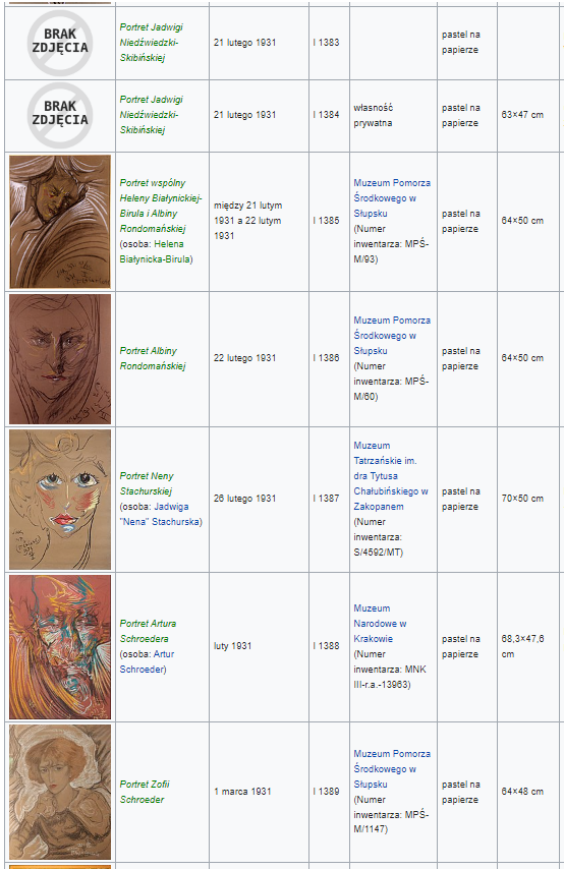
Cashing rarely changing Wikidata results

- Problem: Module:Authority control (used 19M times) might need to access tens of items to look up needed properties of each identifier
- Solution: cashing of data for list of supported identifier properties:
  - Template:Authority_control/IdentifierList generates lua code based on list of identifiers
  - Purging the page and copying results into Module:Authority control/conf makes it official sourcecode
  - Module:Authority control lua code calls “properties = require('Module:Authority control/conf')” to access the data

Authority control	: Q79822 · VIAF: 64009368 · ISNI: 0000 0001 2136 3458 · LCCN: n80050378 · NLA: 35350725 · MusicBrainz: 9441fd78-1ea5-4f8b-9db3-b19f52632ffd · Open Library: OL114808A · DBNL: mick004 · GND: 11873377X · SELIBR: 195101 · SUDOC: 027028828 · BNF: 11916045g · NDL: 00755816 · BIBSYS: 90189864 · HDS: 041485 · NKC: jn19990005678 · SBN: RAVV023387 · RSL: 000083572 · BNE: XX852201 · CiNii: DA04111433 · NLP: a0000001004227 · J9U: 987007265465905171 · Koninklijke: 069802181 · WorldCat

Auto-generating tables for Wikipedia articles

- Lista dzieł malarskich Stanisława Ignacego Witkiewicza on Polish Wikipedia, catalogs all artworks of famous Polish artist, photographer, dramatist and novelist (1885–1939): about 2500 of them
- Approach:
  - Each artwork has a Wikidata item with image, metadata and sources
  - SPARQL query can find all his artworks and order them
  - A spreadsheet operations described here, can convert results of the query into a wikicode
  - Wikicode calls Module:User:Jarekt/table row which creates Polish Wikipedia wikicode for a row in a table with metadata about a single artwork
  - Polish Wikipedia wikicode is copied into the article
- One challenge: list ordering – sorting by complex dates is hard



BRAK ZDJĘCIA	*Portret Jadwigi Niedźwiedzki-Skibińskiej*	21 lutego 1931	I 1383		pastel na papierze	
BRAK ZDJĘCIA	*Portret Jadwigi Niedźwiedzki-Skibińskiej*	21 lutego 1931	I 1384	własność prywatna	pastel na papierze	63×47 cm
	Portret wspólny Heleny Białynickiej-Birula i Albiny Rondomańskiej (osoba: Helena Białynicka-Birula)	między 21 lutym 1931 a 22 lutym 1931	I 1385	Muzeum Pomorza Środkowego w Słupsku (Numer inwentarza: MPŚ-M/93)	pastel na papierze	64×50 cm
	Portret Albiny Rondomańskiej	22 lutego 1931	I 1386	Muzeum Pomorza Środkowego w Słupsku (Numer inwentarza: MPŚ-M/60)	pastel na papierze	64×50 cm
	Portret Neny Stachurskiej (osoba: Jadwiga "Nena" Stachurska)	26 lutego 1931	I 1387	Muzeum Tatrzańskie im. dra Tytusa Chałubińskiego w Zakopanem (Numer inwentarza: S/4592/MT)	pastel na papierze	70×50 cm
	Portret Artura Schroedera (osoba: Artur Schroeder)	luty 1931	I 1388	Muzeum Narodowe w Krakowie (Numer inwentarza: MNK III-r.a.-13963)	pastel na papierze	68,3×47,6 cm
	Portret Zofii Schroeder	1 marca 1931	I 1389	Muzeum Pomorza Środkowego w Słupsku (Numer inwentarza: MPŚ-M/1147)	pastel na papierze	64×48 cm