Importing the Digital Guide to the Cultural Heritage of Andalusia into Wikidata, experiences and lessons learned

Andalusian heritage initial state in Wikidata

- Andalusian cultural heritage is old and huuuuuge.
- There were a previous massive data import from the Spain ministry through the Wiki Loves Monuments database:
  - is really good to have a dataset imported
  - is pretty sad to instance (P31) using:
    - monument (Q4989906): «imposing structure created to commemorate a person or event, or used for that purpose»
    - or an heritage designation element like Monument (Q907116): «Spanish heritage structures of cultural interest»
- but that database is pretty old:
  - it is really outdated:
    - the WLM database is outdated but the ministry database is too!
  - since 2007? national ministry doesn't manage the heritage designations, it's done, in this case, government of Andalusia Autonomous Community.
  - the Spanish Asset of cultural interest code (P808) values aren't even recognized (neither managed) by the regional government, so it's impossible to use to match elements;
  - worst: you'll find a single heritage monument in a source modeled with two codes in the other 🙄
- Available the P3318 property (Guía Digital del Patrimonio Cultural de Andalucía ID) being using by a lot related elements (3K? don't remember), but not all.

Goal

To update in Wikidata the Andalusian heritage.

Researching

We find an **excellent** source from the government:

- Digital Guide to the Cultural Heritage of Andalusia (Q5758805);
- published by the Andalusian Institute of Historical Heritage (Q5917182), agency of the ministry of Culture of Andalusia;
- a informative product with records of:
  - immovable cultural heritage (>24 000)
  - movable heritage (> 100 000)
  - intangible heritage (>1 000)
  - cultural landscapes (>100)
  - cultural routes

Unfortunately it export open data in an INFERNAL format
(because you can use open formats in evil ways).
(no, you don't want to see an example, believe me).

Anyhow, as a government source, it is considered the sensible choice as a sole reference for a massive data import.

Challenges

(Challenges to me, because I'm an heritage n00b 🥸)

How many elements?

- reduced the scope to immovable cultural heritage
- about 4000 current WD elements
- more the 24000 elements in the Guide
- ⇒ we want it ALL

How to instance?

- we should not to use monument/Monument but the typology of the source
  - example Hacho bridge in Wikidata, in the Digital Guide.
- we find the Digital Guide uses «typologies» recorded in their own thesaurus (Andalusian Historical Heritage Thesaurus, Q97322747) (> 18000 entries)
- we find the AHHT thesaurus has some mappings:
  - National Library of Spain ID (P950) (good)
  - DBpedia-es (wot!?)
- ⇒ we need to map AHHT thesaurus with Wikidata:
  - and unknown size task o_0

Tools

Main tool is OpenRefine (Q5583871) to the rescue. If you are an advanced Wikidatista you NEED it:

- manages really big files
- matches with Wikidata
- data transform features:
  - example: mapping between different OR tables (solution for our instancing work)
- it's free software

And other helpers to download and preprocess data:
- curl
- bash / python scripts

The work

The mapping

- manual mapping comparing AHHT, BNE id's, DBpedia-es elements and present Wikidata elements,
- and sometimes checking other sources to figure out what is what.

The matching

- Match the first 4 thousands using the P3318 property
  - hurrah, something easy to do!
  - woah, that dirty OpenRefine trick.
- And now you just only need to figure out which other Wikidata elements matches to the pending 20 thousands records in the Digital Guide 🥳🥳🥳

Uploading

- most of the time using OpenRefine directly
- sometimes exporting/importing to Quickstatements

Handicaps

Duplicated elements

- some elements imported by WLM
- some elements imported from a Wikipedia
  - (not using P31, who cares)
- other causes, so
- you can *easily* find three elements describing a single item
- philosophical disquisition: are two Wikidata elements describing the same concept duplicated if you can't find both with a single query? Discuss.

Wikidata editors not always share modelling conventions

- sometimes they share their OWN modelling conventions (hi User:Romaine!)
- ⇒ educate the community through assertive undo changes comments.
- AFAIK we don't have explicit conventions to model monuments
  - maybe ShEx to the help?

Wikidata editors not always share meaning conventions

- the strange case of the megalith wikidata element.

How to manage overlapping elements?

- because they are are conceptually overlapped
- because they have been imported from different Wikipedias articles with different scopes
  - «easy» to manage if you understand **that** Wikipedia language.
- because Wikipedia articles describes two or more concepts related but different.

Wikimedia Commons

- Welcome the Commons' categories hell
- BTW, does anyone how to properly use the instances of Wikimedia category Q4167836, asking for a friend 😺
- Hey, you can find Wikidata duplicated elements browsing Commons categories too!

Reusing Wikidata: the unexpected

- Example case: elements with heritage designation (P1435) equal to UNESCO World Heritage Site (Q9259):
  - Vatican City (Q237) (WIN!)
  - Yellowstone National Park (Q351) (W0T!?)
- Andalusian examples:
  - Cathedral, Alcázar and Archivo de Indias in Seville (Q9709444)
  - Doñana National and Natural Park (Q463141) 😢😢😢
- (Other problems, TBD)

Do it wrong

- The Conqueror Syndrome:
  - «Now all this land is mine»
  - «I have OpenRefine and I know how to use it, muahahah!»
  - coined by User:Maria zaos.
- Bad testing: firsts upload batches bigger than necessary:
  - At that time OpenRefine could'n modify/remove statements
    - plus, I+don't+know+what+I'm+doing
  - Added wrong statements to be manually fixed

        - by the Future Me 👨‍🦱
- Conqueror Syndrome + bad testing:
    - 🔥💥🔥 Fire, walk with me 🔥💥🔥
- using Spanish label in the other languages:
    - /me because that label is a proper name
    - /restoftheworld: dude, are you sure «*Hacho bridge*» is a proper name?
    - Pissing off some colleagues 🤦‍♀️🤦‍♂️
- some AHHT thesaurus-wikidata mapping is wrong
    - seems easy to fix
    - seems I wasn't as smart as I thought 👀
- Considering the 27000 elements
    - do we really needed/wanted all? Open question.
- Added apparently confusing values to heritage designation (P1435):
    - procedure date of opening of file
    - legal approval date
    - (my bad)

Lessons learned

- Seems I'm not as smart as I thought,
    - at least I serve as a bad example 🥸
- Devil is in the details 👺
- Ask the government institution:
    - most of the time they don't answer or don't really (can/want) help
    - but when they can, everything is really better 😻

Future challenges

- Movable heritage (>40000)
    - do we really want/need all?
    - probably better ask institution first.
- Inmaterial heritage
    - on work in the frame of Wiki Loves Living Heritage
- To fix my immovable heritage elements errors
- To synchronize the immovable records with the last changes
    - import made in 202007;
    - how to distinguish the statements changes;
    - how to apply statement changes without pissing-off other Wikidatistas work ;
    - figure out what to fix from the previous upload, if any.
- Don't piss off others 💩