Wikidata use in KDE’s travel apps

Wikidata Data Reuse Days 2022

Volker Krause

vkrause@kde.org
@VolkerKrause
- No online access for data
  - Public transport line logos are loaded on demand and then locally cached
- Data is pre-processed and shipped with the apps
  - Primary motivation: privacy
  - Side-effect: extra sanity checks/validation of the data
  - Allows expensive merging with OSM data
Airport Identification

- **IATA airport code (P238)**
  - Change over time and can be re-allocated
  - Exceptions to uniqueness...
  - Import errors in northern America and Argentina
  - Use for train stations

- **Name**
  - Can require complex disambiguation logic
Civil vs Military Airports

- Only airports with passenger services are relevant for us
- Irrelevant airports can impede name-based identification
- Civil, military and civil/military hybrid uses can change over time
  - Complex modeling
  - Name-based heuristic turned out to be easier
Train Stations

• Numerous identifier systems
  – Often operator- or country-specific
  – Specializations of P296
  – Newly added: P8181 (SNCF), P8448 (Benerail)
  – Proposed: [https://www.wikidata.org/wiki/Wikidata:Property_proposal/Via_Rail_station_code](https://www.wikidata.org/wiki/Wikidata:Property_proposal/Via_Rail_station_code)

• Name-based identification not reliable due to many (globally) ambiguous names and/or missing context
Train Stations - Open Issues

- IATA airport code (P238)
  - Separate property for train stations or constraint violation?
  - https://www.wikidata.org/wiki/Property_talk:P238#IATA_codes_for_train_stations

- Modeling multi-part / “virtual” stations
  - Technically separate stations considered as one by routing/ticketing
  - “Virtual” stations have their own identifiers
Country Information

- Country codes
  - ISO 3166-1 mappings moved to iso-codes data set
  - UIC country code
- Used power plugs
- Driving side
- The usual fun with countries...
Timezones

- Highly important for our use cases
- Specified in different formats in Wikidata (P421)
  - IANA timezone ids
  - UTC offsets for normal time/DST
- UTC offsets lack information about DST transitions
- Using a coordinate-based index rather than Wikidata now
Public Transport Lines

- Line logo and color (P154, P465)
  
  ![U1](image1.png)  ![U1](image2.png)  ![U1](image3.png)  ![U1](image4.png)

- Mode or product logo
  
  ![U](image5.png)  ![U](image6.png)  ![M](image7.png)  ![M](image8.png)  ![M](image9.png)
Public Transport Lines - Identification

- Names are highly ambiguous
- Use OSM to obtain a bounding box for each line
- Use name + coordinate to identify a line
- Relies on Wikidata <-> OSM mapping (P402)
  - Added/extended for a number of cities, in both directions
Public Transport Lines - Challenges

- Complex modeling
  - Distinction between modes of transport is very blurry
  - Unexpected things like historical lines or amusement rides
  - Level of detail varies

- Correct attribution for CC-BY licensed image assets
Thanks for your attention! Questions?
References

- KDE Itinerary: https://apps.kde.org/itinerary
- KTrip: https://apps.kde.org/ktrip
- Travel document extraction: https://invent.kde.org/pim/kitinerary
- Airport locations: https://volkerkrause.eu/2020/05/02/kde-itinerary-airport-locations.html
- Public transport line meta data:
  - https://invent.kde.org/libraries/kpublictransport
- Timezone lookup: