Comparative Integration Potential Analyses of OSM and Wikidata

The case study of Railway Stations

Alishiba Dsouza, Data Science & Intelligent Systems
University of Bonn

Moritz Schott
Institute of Geography, GIScience
Heidelberg University

Sven Lautenbach
Institute of Geography, HeiGIT
Heidelberg University

State of the Map 2022
Florence, Italy
August 21, 2022
Knowledge Graphs (KGs)

• Rich source of semantic information
• Contain semantic information regarding real-world entities, their types and properties
  • Generic KGs: Wikidata, DBpedia, Yago
  • Geographic KGs: LinkedGeoData, Yago2Geo, WorldKG

• Problem:
  • Few geographic entities are present in generic KGs
  • Few geographic classes are present in specialized geographic KGs
Wikidata Knowledge Graph

• Wikidata: Open Source General purpose KG of Wikimedia foundation
• Edited and used by Humans and Machines
  • Eg: “CyclingInitBot”: bot for initializing cycling related items
• Provides Semantic Representation
• Represented in the triple format
  • Subject – Predicate – Object
  • Eg: Florence – capital of – Tuscany

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>country</td>
<td>Tuscany</td>
</tr>
<tr>
<td></td>
<td>start time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 reference</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>18 June 1946</td>
</tr>
<tr>
<td>capital of</td>
<td>Tuscany</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 references</td>
<td></td>
</tr>
</tbody>
</table>
OpenStreetMap VS Wikidata

- Rich but heterogeneous schema
- No fixed tags for a type
- Not directly accessible for semantic applications

- Fixed Schema
- Class hierarchy
OpenStreetMap linking Wikidata

• OSM links to Wikidata with “wikidata” tag
  • Over 2.5 million entities linked from OSM to Wikidata

• Wikidata links to OSM with OpenStreetMap object (P10689) property
  • Only ~1000 entities linked from Wikidata to OSM

Entities linked from OSM to Wikidata i.e. linking from geodatabase (OSM) to an information source (KG)
Integrating OSM and KGs

• Linking schema elements
  • Align OSM tags to KG classes [1]
  • Eg: “natural”=“peak” (OSM) → “mountain” (Wikidata)
• Linking entities
  • Already existing links between OSM and KGs
  • Find new links using existing links [2, 3]
• Integration
  • Integrate the schema and entities
  • OSM can benefit from wide semantic information
    • Geographic information retrieval, Question Answering, Visualization
  • Wikidata can benefit from the precise geoinformation
  • Beneficial for both sources in terms of completeness and correctness
WorldKG Knowledge Graph

- OSM data in a knowledge graph format [3]
  - Semantic representation
- Overcomes the class hierarchy issue
- Currently contains Nodes from OSM
- Accessible at: www.worldkg.org
Goal of the analyses

• OSM and Wikidata are comparable
  • Community structure
  • Free and open
  • Simple contribution

• Comparative data insights
  • Potential and implications of integration between KGs and OSM

• Integration of OSM and KGs:
  • Closer step toward completeness and correctness
  • Integration of data also means integration of communities and working styles
Case Study of Railway Stations

• Comparable definition in both datasets
  • 'railway=station' or 'railway=halt'
  • 'instance of Q55488' (railway station)

• Well represented in both datasets
  • ~130,000 objects in OSM and ~100,000 objects in Wikidata
    • Indicates integration potential
General Comparison Statistics

• OSM contains 26% more entities
• Division into 6 categories  
  • Not all wikidata=* tags refer to railway stations  
  • wikidata without geometry can only be linked manually (wikidata tag) or semantically (e.g. name)
• High linking potential  
  • Necessary for “safe” integration
Growth Rate Analysis

• OSM is reaching a saturated state
• Wikidata sees steady growth
• No obvious correlation between OSM and Wikidata
  • Independent communities!?  
• Links to Wikidata added much later than the launch of Wikidata
  • Integration potential is rising
Regional Distribution (log)

- OSM overabundance for countries with little to medium railway infrastructure
  - Wikidata requires more data before linking is possible
Regional Distribution (linear)

- Discrepancy for large railway infrastructures
  - UK, Poland
  - China, Russia
- Sources of discrepancy
  - Unequal completeness
  - Historic elements in Wikidata
- Data errors (e.g. mistagged tram stations)
- Good (India) does not equal linkage
Linking Potential OSM

- Especially high for many small railway infrastructures
- Russia, China show low linkage
- High potential/low linking percentage hinders integration
Linking Potential Wikidata

- Quasi independent of railway infrastructure size
  - Many “unmapped” countries
Semantic Information

- Wikidata
  - Average: 30
  - Potential multiplication through KG links
- OSM
  - Average: 7.6
- linked objects
  - "Main Stations"
- Low quality of non-geographic Wikidata and unlinked entities
  - Automated integration may overcome this problem
Geometric Information

• Share of polygons
  • 4% for linked elements
    • Despite being “main” stations?
    • Mapping scheme continues to evolve/disputed
    • Point location may be arbitrary
  • 8% for unlinked elements
• “no” polygons in Wikidata
  • Integration potential reduced by OSM mapping scheme
Object History

• Very high number of revisions in Wikidata
  • Data maintenance
  • More tags = more revisions
  • Developing scheme -> is subject to changes

• OSM
  • Data creation may take priority over data maintenance
  • Little real world changes (stable tags and geometry)

• Up-to-dateness?
Community Size

- Relatively small Wikidata Community
- Limited to railway “station” mappers
  - Wikidata users/bots edit multiple topics
  - RS makes up only small part of OSM
User Activeness

• Wikidata
  • High number of “power users”
  • Multiple RS from one source

• OSM
  • Many one-time users
  • Possibly limited to a certain area (only one RS present)
  • Localised mapping styles

• OSM community wary towards bots

![Diagram showing number of created objects per user for OSM and Wikidata]
User Diversity

- **Wikidata**
  - Relatively many users with high specialisation
  - Possibly topic-dependent bots
- **OSM**
  - Railway stations are only one topic of many
Outlook

• Manual and automated linking are progressing well
  • Still much work to do
• If you map, think of wikidata=*!
  • If Wikidata is missing: you are welcome to add data to Wikidata!

• Open Questions
  • Regional data trends
  • Integration potential of other classes

• Future Work
  • Extend schema alignment to keys and properties
  • Actual integration of OSM and Wikidata
References


