Data coherence between OSM and Wikipedia

Cristian Consonni
Fondazione Bruno Kessler

State of the Map 2013 - Birmingham
September 2013

Outline

Collecting Information About the Real World

Collecting Information About the Real World

Collecting Information About the Real World

Wikipedia and OpenStreetMap are:

- collaborative
- volunteer-driven
- free (as in freedom *and* as in beer)

Both projects collect information about the real world.

Different Processes and Communities

Wikipedia

OpenStreetMap

- anonymous users can edit
- entries consist in *text* (or media)
- only encyclopedical subjects
- content can be protected from editing in case of problems

- only registered users can edit
- entries consist in *data*
- everything can be described
- content is always editable

Inconsistencies in the data

Data in Wikipedia can be *inconsistent* with data from OpenStreetMap.
We should *compare* the data and *reconcile* the differences.

Inconsistencies in the data

Data in Wikipedia can be inconsistent with data from OpenStreetMap.
We should compare the data and reconcile the differences.
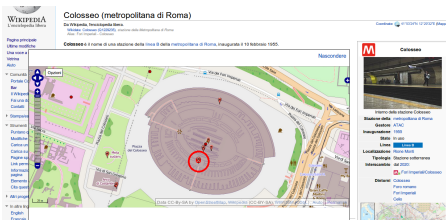
On Wikipedia the metro station "Colosseum" is inside the Colosseum itself.

Inconsistencies in the data

Data in Wikipedia can be inconsistent with data from OpenStreetMap.
We should *compare* the data and *reconcile* the differences.
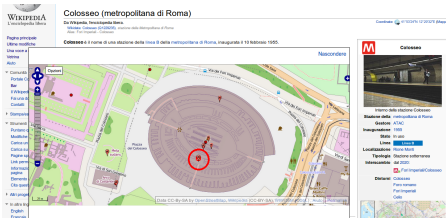
On Wikipedia the metro station
"Colosseum" is inside the Colosseum
itself.
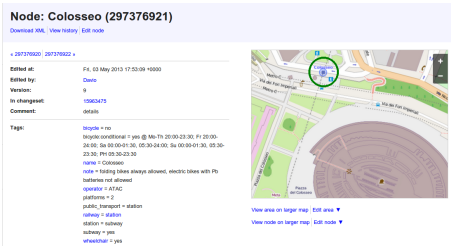
On OpenStreetMap the metro station is
correctly placed outside the monument.

OpenStreetMap maps on Wikipedia provided by **WIWOSM** tool by User:Master and User:Kolossos, check it out on:

`http://wiki.openstreetmap.org/wiki/WIWOSM`

Proposal of the Solution

Two steps towards a solution:

1. Compare the data
   - Identify links between Wikipedia pages and OSM entities
   - Extract all the available geographical information
   - Define metrics to calculate if the data are "close" or not

2. Reconcile the differences
   - Provide the communities with the result of previous analysis
   - Creating tools to facilitate the reconciliation

Proof-of-concept: comparing data about churches in Italy:

<div align="center">

Wikipedia-OpenStreetMap comparator

source code: https://github.com/CristianCantoro/WOcomparator

</div>

Easy case:

- pre-defined category of items (selection on a set of features in OSM, articles with a given template in Wikipedia)
- only entities with a (it:)Wikipedia attribute were selected

⇒ linking is straightforward.

Comparing the data

Wikipedia-OpenStreetMap comparator

http://it.wikipedia.org/wiki/Utente:CristianCantoro/Georeferenziazione

Comparing the data
nuts4nuts

For the hard case (try to link every possible thing), another tool:

<div align="center">

Nuts4Nuts

source code: https://github.com/SpazioDati/Nuts4Nuts

</div>

```
▼ {
  ▼ "q0": {
    ▼ "result": [
      ▼ {
          "score": 1,
        ▼ "type": [
          ▼ {
              "id": "/LAU2",
              "name": "Basic service"
            }
          ],
          "id": "http://data.spaziodati.eu/resource/1b3843bb1cdfc5b342fefa9f76aa5caf0eb2ddb3",
          "match": true,
          "name": "Firenze"
        }
      ]
    }
  }
}
```

http://nuts4nutsrecon.spaziodati.eu/reconcile?queries={%22q0%22:%20{%22query%22:%20%22Palazzo%20Vecchio%22}}

Known limitations:

- limited to Italy
- uses of external services

grab the source code: https://github.com/SpazioDati/Nuts4Nuts

Dandelion

Nuts4Nuts is built using the infrastracture provided by

Dandelion (http://dandelion.eu)

a datamarket by SpazioDati srl.

Future Work

Nuts4nuts is a step to find geographical information for Wikipedia article that have no explicit coordinates in them.

Future work:

- study new approaches to link entities between Wikipedia and OpenStreetMap
- an application to fix inconsistencies or fill in missing data, like this:

Conclusions

- Wikipedia and OSM collect information about the real world

Conclusions

- Wikipedia and OSM collect information about the real world

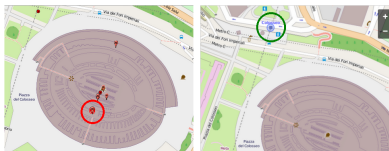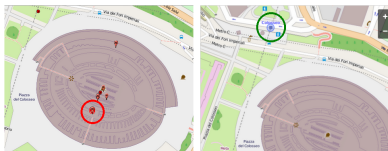- Comparing data among the two project can highlight inconsistencies

Conclusions

- Wikipedia and OSM collect information about the real world

- Comparing data among the two project can highlight inconsistencies

- We should **fix** them

im in ur serverz making thingz better!!!

Questions?

mail: consonni@fbk.eu
twitter: @CristianCantoro
github: https://github.com/CristianCantoro

Thank you!

This work was supported by:

T2DataExchange

A project by:

- *SpazioDati srl*
- *Edizioni Curcu & Genovese*

with funds from the *European Regional Development Fund*.
More information: http://trentino.dandelion.eu

Copyright notice

The following presentation is realeased under the licence CC3.0-BY-SA.