NLGIS: A USE CASE IN LINKED HISTORIC GEODATA

Wouter Beek wouter@triply.cc, VU University Amsterdam, Triply
Richard Zijdeman richard.zijdeman@iisg.nl, International Institute for Social History
June 3rd, 2018
Example of an artifact with poor geodata: link

Geodata in today's LOD Cloud...

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>geo:geometry</td>
<td>POINT(24.909999847412 35.209999084473)</td>
</tr>
<tr>
<td>geo:lat</td>
<td>35.209999 (xsd:float)</td>
</tr>
<tr>
<td>geo:long</td>
<td>24.910000 (xsd:float)</td>
</tr>
<tr>
<td>skos:exactMatch</td>
<td><a href="http://rfddata.eionet.europa.eu/ramon/nuts/EL43">http://rfddata.eionet.europa.eu/ramon/nuts/EL43</a></td>
</tr>
</tbody>
</table>
## NLGIS DATASETS

[View Dataset](https://druid.datalegend.net/nlgis)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Nº statements</th>
<th>Main concepts</th>
<th>Nº geometries</th>
<th>Timeframe</th>
</tr>
</thead>
<tbody>
<tr>
<td>CShapes</td>
<td>6,120</td>
<td>countries, cities</td>
<td>510</td>
<td>1920-present</td>
</tr>
<tr>
<td>Mint Authorities</td>
<td>6,987</td>
<td>authorities, houses</td>
<td>950</td>
<td>565-present</td>
</tr>
<tr>
<td>Gemeentegeschiedenis</td>
<td>46,929</td>
<td>municipalities, provinces</td>
<td>3,219</td>
<td>1813-present</td>
</tr>
<tr>
<td>nlGis</td>
<td>60,036</td>
<td>features, geometries</td>
<td>4,679</td>
<td></td>
</tr>
</tbody>
</table>
LESSONS LEARNED

1. Combine what belongs together
2. Do not use ambiguous ‘null’ values
3. No perfect tool for data transformation
4. No perfect triple store for geo
5. Direct feedback helps a lot
6. Use interoperable representations
[1] COMBINE WHAT BELONGS TOGETHER

DATE/TIME

:Greece iisg:cowStartYear "1946"^^xsd:gYear ;
iisg:cowStartMonth "1"^^xsd:gMonth ;
iisg:cowStartDay "1"^^xsd:gDay .

:Greece iisg:cowStart "1946-01-01"^^xsd:date

(In CShapes, ‘cow’ stands for Correlates of War.)
COMBINING WHAT BELONGS TOGETHER PREVENTS BUGS

Longitude/latitude

:somewhere wgs84:lat "..."
wgs84:lat "..." ;
wgs84:long "..." ;
wgs84:long "..." .

There is also wgs84:lat_long, but it is almost never used.
There are many instances of this!

OCLC VIAF example

:EmmaGoldman schema:givenName "Ema" ;
   schema:givenName "Ėmma" ;
   schema:familyName "Gol'dman" ;
   schema:familyName "Gōrudoman" .

There are many other instances of this problem, e.g., foaf:firstName and foaf:lastName.
CShapes uses -1 to denote an unknown year. In the context of CShapes (countries after 1920) this makes sense.

But on the web we can query CShapes and Pleiades.
[3] NO PERFECT TOOL FOR DATA TRANSFORMATION

Requirements:

1. Support multiple source formats
2. Scale to datasets of arbitrary size

No currently available data transformation tool implements these two core requirements.
SUPPORT MULTIPLE SOURCE FORMATS

- CSV
- (Geo)JSON
- XML
  - (GML/MARCXML/EAD)
- relational DB
- RDF

Proprietary formats can sometimes be transformed into open formats, e.g., ESRI ShapeFile.
### Scale to Datasets of Arbitrary Size

Be able to stream through the data at the required granularity level.

<table>
<thead>
<tr>
<th>#</th>
<th>name</th>
<th>population</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Amsterdam</td>
<td>1.3M</td>
<td>MultiPolygon(....)</td>
</tr>
<tr>
<td>2</td>
<td>Athens</td>
<td>3.1M</td>
<td>MultiPolygon(....)</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GeoSPARQL support is either absent, not standards-compliant, or not performant.

- Most stores do not implement GeoSPARQL syntax, but some do.
- Most stores have miserable/unusable performance, but some have good performance.
- Some stores change the data merely by loading it.
- Some stores cannot load larger shapes.
- Commercial stores are not necessarily better than FOSS (if fact: they are very often worse).
When writing GeoSPARQL queries, a table of results is not enough.
[6] USE INTEROPERABLE REPRESENTATIONS
OPTIONS FOR REPRESENTING GEODATA IN LOD

- WGS84 Geo Positioning Vocabulary (W3C)
- GeoSPARQL (OGC)
  - Well-Known Text (WKT)
  - Geography Markup Language (GML)
- Make up your own vocabulary
- GeoJSON + JSON-LD
Without interoperable representations:

- clients do not know what to do with your data
- triple store cannot index your geometries
- reasoners arrive at contradictions

After applying standards...
LOD-A-LOT

Large-scale empirical analyses

lod-a-lot.lod.labs.vu.nl
## Linked GeoData Vocabulary Use

<table>
<thead>
<tr>
<th>Property</th>
<th>Nº statements</th>
<th>Nº documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>wgs84:alt</td>
<td>2,349,607</td>
<td>9,843</td>
</tr>
<tr>
<td>wgs84:lat</td>
<td>42,883,363</td>
<td>11,134</td>
</tr>
<tr>
<td>wgs84:lat_long</td>
<td>283</td>
<td>173</td>
</tr>
<tr>
<td>wgs84:location</td>
<td>14,688,561</td>
<td>117</td>
</tr>
<tr>
<td>wgs84:long</td>
<td>42,916,785</td>
<td>11,134</td>
</tr>
<tr>
<td>geo:asGML</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>geo:asWKT</td>
<td>188,427,329</td>
<td>50</td>
</tr>
<tr>
<td>geo:hasGeometry</td>
<td>28,366,268</td>
<td>7</td>
</tr>
</tbody>
</table>

Based on the LOD-a-lot data collection (Fernández et al. 2017).
GEOJSON + JSON-LD

Unfortunately, these two popular formats are incompatible:

- GeoJSON uses square brackets to denote (nested) lists of geographic coordinates.
- JSON-LD uses square brackets as summarized syntax for repeated property assertion.

This may be fixed in future a version of the JSON-LD standard.
## NLGIS DATASETS

https://druid.datalegend.net/nlgis

<table>
<thead>
<tr>
<th>Dataset</th>
<th>№ statements</th>
<th>Main concepts</th>
<th>№ geometries</th>
<th>Timeframe</th>
</tr>
</thead>
<tbody>
<tr>
<td>CShapes</td>
<td>6,120</td>
<td>countries, cities</td>
<td>510</td>
<td>1920-present</td>
</tr>
<tr>
<td>Mint Authorities</td>
<td>6,987</td>
<td>authorities, houses</td>
<td>950</td>
<td>565-present</td>
</tr>
<tr>
<td>Gemeentegeschiedenis</td>
<td>46,929</td>
<td>municipalities, provinces</td>
<td>3,219</td>
<td>1813-present</td>
</tr>
<tr>
<td>nlGis</td>
<td>60,036</td>
<td>features, geometries</td>
<td>4,679</td>
<td></td>
</tr>
</tbody>
</table>

Dutch Cultural Heritage institutes already use this to annotate their collection with (example).
FUTURE WORK

- Cover more places & times.
- Annotate Cultural Heritage objects with detailed geographic information.
- Create a standardized vocabulary for how geolocations change through time.
- Improve GeoSPARQL support in triple stores.
- Explore new ways of displaying Cultural Heritage objects in space and time (example).
THANK YOU!

nlGis: https://druid.datalegend.net/nlgis/

Wouter Beek wouter@triply.cc, VU University Amsterdam, Triply
Richard Zijdeman richard.zijdeman@iisg.nl, International Institute for Social History

June 3rd, 2018