An Instance-based Approach for Matching Export Schemas of Geographical Database Web Services

Daniela F. Brauner, Chantal Intrator, João Carlos Freitas, Marco A. Casanova
{dani, cintrator, jcsfreitas, casanova}@inf.puc-rio.br

Pontifical Catholic University of Rio de Janeiro (PUC-Rio)
Department of Informatics
Summary

- Motivation
- Related Work
- Instance-based Schema Matching
- Experimental Approach
  - Global Schema
  - Global Instances
  - Geographical Databases Web Services
  - Results
  - Further considerations
- Conclusion
Summary

- Motivation
- Related Work
- Instance-based Schema Matching
- Experimental Approach
  - Global Schema
  - Global Instances
  - Geographical Databases Web Services
  - Results
  - Further considerations
- Conclusion
**Motivation**

- *Database Web service:*
  - a Web service interface with operations to access a backend database
  - returns results following a given export schema

- *Export schema:*
  - a subset of the backend database schema visible to the clients
Motivation

- Our goal:
  - match **export schemas** with a **global schema** of geographical database Web services using a small set of **typical instances**
Summary

- Motivation
- **Related Work**
- Instance-based Schema Matching
- Experimental Approach
  - Global Schema
  - Global Instances
  - Geographical Databases Web Services
  - Results
  - Further considerations
- Conclusion
Related work

• An instance-based schema matching technique using domain-specific query probing, applied to Web databases

• A Web database is composed of a query interface and a backend database
  – Interface schema: what can be queried
  – Result schema: what is shown to users

• Using:
  – a global schema (GS) for Web databases of the same domain
  – a set of global instances

Related work

Related work

Related work

Summary

- Motivation
- Related Work
- **Instance-based Schema Matching**
- Experimental Approach
  - Global Schema
  - Global Instances
  - Geographical Databases Web Services
  - Results
  - Further considerations
- Conclusion
How to match database Web services export schemas?
- Using the schema matching process:
Summary

- Motivation
- Related Work
- Instance-based Schema Matching

**Experimental Approach**
- Global Schema
- Global Instances
- Geographical Databases Web Services
- Results
- Further considerations

- Conclusion
Global Schema

- Captures the essential characteristics of a gazetteer
- Based on the ISO 19112:2003, a model for geographic information

Summary

- Motivation
- Related Work
- Instance-based Schema Matching
- **Experimental Approach**
  - Global Schema
  - Global Instances
  - Geographical Databases Web Services
  - Results
  - Further considerations
- Conclusion
Global Instances

- A set of 36 objects representing famous geographic places
- Collected from Geonames.org Web service
- Stored in a local database, following the global schema

Global Instances Fragment

<table>
<thead>
<tr>
<th>idInstance</th>
<th>name</th>
<th>lat</th>
<th>lon</th>
<th>idType</th>
<th>adminkd1</th>
<th>adminkd2</th>
</tr>
</thead>
<tbody>
<tr>
<td>175</td>
<td>Galapagos Islands</td>
<td>0.0</td>
<td>-90.5</td>
<td>4</td>
<td>73</td>
<td>-</td>
</tr>
<tr>
<td>52</td>
<td>Alps</td>
<td>46.416667</td>
<td>10.0</td>
<td>15</td>
<td>165</td>
<td>-</td>
</tr>
<tr>
<td>149</td>
<td>Atlantic Ocean</td>
<td>10.0</td>
<td>-25.0</td>
<td>9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>90</td>
<td>Niagara Falls</td>
<td>43.083416155</td>
<td>-79.06627052</td>
<td>21</td>
<td>123</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>Pão de Açúcar</td>
<td>-22.9472</td>
<td>-43.1561</td>
<td>14</td>
<td>101</td>
<td>-</td>
</tr>
<tr>
<td>34</td>
<td>Mississippi River</td>
<td>29.1510582</td>
<td>-89.2533842</td>
<td>19</td>
<td>109</td>
<td>-</td>
</tr>
</tbody>
</table>
Summary

• Motivation
• Related Work
• Instance-based Schema Matching
• Experimental Approach
  – Global Schema
  – Global Instances
  – Geographical Databases Web Services
  – Results
  – Further considerations
• Conclusion
In our experiments we use the following gazetteers available through Web Services:

- Geonames.org
  http://www.geonames.org

- Alexandria Digital Library (ADL) Gazetteer
  http://www.alexandria.ucsb.edu/gazetteer
**Export Schemas**

**XML response fragment of Geonames.org**

```xml
<?xml version="1.0" encoding="UTF-8"?>
<geonames style="FULL">
<totalResultsCount>1</totalResultsCount>
<placename>
<name>Amazon River</name>
<lat>0.166666</lat>
<geonameId>3407729</geonameId>
<countryCode>BR</countryCode>
<featureClass>ST</featureClass>
<featureCode>SC</featureCode>
<name>stream, lake, ..., in this Name>
<population />
<alternateNames>Rio Amazonas, Río Amazonas, Río Maranón, Río Solimões, Río Solimões, Río de Amazonas, Río Amazonas, Río Marañón, Río de Amazonas, Salimoes River, Solimoes</alternateNames>
<elevation />
<adminCode1>00</adminCode1>
<adminName1 />
<adminCode2 />
<adminName2 />
<timezone dateOffset="-3.0" gmtOffset="-3.0">America/Belem</timezone>
</placename>
</geonames>
```

**XML response fragment of ADL Gazetteer**

```xml
<?xml version="1.0" encoding="UTF-8"?>
<gazetteer-service xmlns="http://www.alexandria.ucsb.edu/gazetteer"
xmlns:gml="http://www.opengis.net/gml"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.alexandria.ucsb.edu/gazetteer
 http://www.alexandria.ucsb.edu/gazetteer/protocol/gazetteer-service.xsd" version="1.2">
<query-response>
<standard-reports>
<gazetteer-standard-report>
<identifier>adigaz:1-1410143-3a</identifier>
<place status="current">Amazon River - Brazil</place>
<name primary="true">Amazon River</name>
<name primary="false">Salimoes River</name>
<name primary="false">Maranão</name>
<name primary="false">Amazonas, Río</name>
<name primary="false">Amazonas, Rio de Janeiro</name>
<name primary="false">Solimões</name>
<bounding-box>
<gml:point x="-99.0" y="-0.1667"/>
</bounding-box>
<footprints>
<footprint primary="true">
<gml:point x="-99.0" y="-0.1667"/>
</footprint>
</footprints>
</gazetteer-standard-report>
</standard-reports>
</query-response>
</gazetteer-service>
```
Summary

- Motivation
- Related Work
- Instance-based Schema Matching
- **Experimental Approach**
  - Global Schema
  - Global Instances
  - Geographical Databases Web Services
  - **Results**
  - Further considerations
- Conclusion
Results

• 36 global instances submitted as queries
  – 459 registries returned from ADL Gazetteer
  – 703 registries returned from Geonames.org

Ex:
  – “Mount Everest” submitted to Geonames.org
Results

- Occurrences matrices:
  
  **(a)** Geonames.org Export Schema $\times$ Global Schema
  
  **(b)** ADL Gazetteer Export Schema $\times$ Global Schema

<table>
<thead>
<tr>
<th>geonames</th>
<th>$I_{GS}$</th>
<th>$N_{GS}$</th>
<th>$A_{GS}$</th>
<th>$O_{GS}$</th>
<th>$T_{GS}$</th>
<th>$A1_{GS}$</th>
<th>$A2_{GS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>lat</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lon</td>
<td>0</td>
<td>0</td>
<td>188</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>name</td>
<td>0</td>
<td>551</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>countryCode</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>countryName</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>id</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>thecode</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>theName</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>population</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>alternateName</td>
<td>0</td>
<td>156</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>elevation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminCode1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminCode2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminName1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminName2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GeoNames</th>
<th>$I_{GS}$</th>
<th>$N_{GS}$</th>
<th>$A_{GS}$</th>
<th>$O_{GS}$</th>
<th>$T_{GS}$</th>
<th>$A1_{GS}$</th>
<th>$A2_{GS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>identifier</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>placeStatus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>name</td>
<td>0</td>
<td>459</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>displayName</td>
<td>0</td>
<td>352</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>footprintX</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>134</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>footprintY</td>
<td>0</td>
<td>0</td>
<td>94</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>class</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>thesaurus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>names</td>
<td>0</td>
<td>435</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>relationships</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(b)
Results

- Given an occurrence matrix: “an attribute of the export schema matches an attribute of the global schema iff the normalized value is greater than 0.2”

(a) Given an occurrence matrix:

```
<table>
<thead>
<tr>
<th></th>
<th>IGS</th>
<th>NGS</th>
<th>AGS</th>
<th>OGS</th>
<th>TGS</th>
<th>A1GS</th>
<th>A2GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>geoNameId</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lat</td>
<td>0</td>
<td>0</td>
<td>188</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lon</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>name</td>
<td>0</td>
<td>551</td>
<td>78%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>countryCode</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>countryName</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>id</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fcId</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fcName</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fccode</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fcName</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>population</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>alternatename</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>elevation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminCode1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminCode2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminName1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminName2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```

(b) Given an occurrence matrix:

```
<table>
<thead>
<tr>
<th></th>
<th>IGS</th>
<th>NGS</th>
<th>AGS</th>
<th>OGS</th>
<th>TGS</th>
<th>A1GS</th>
<th>A2GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>idInstance</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>name</td>
<td>0</td>
<td>435</td>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>display</td>
<td>0</td>
<td>435</td>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>footprintX</td>
<td>0</td>
<td>435</td>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>footprintY</td>
<td>0</td>
<td>435</td>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>class</td>
<td>0</td>
<td>435</td>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>thesaurus</td>
<td>0</td>
<td>435</td>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>names</td>
<td>0</td>
<td>435</td>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>relationships</td>
<td>0</td>
<td>435</td>
<td>95%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```

Global Schema:

- IGS: idInstance
- NGS: name
- AGS: lat
- OGS: lon
- TGS: idType
- A1GS: adminId1
- A2GS: adminId2
Results

Occurrences of the value of the attribute *name* ($N_{GS}$) as *adminName1* and *adminName2* from Geonames.org

<table>
<thead>
<tr>
<th>geonameID</th>
<th>lat</th>
<th>lng</th>
<th>name</th>
<th>countryCode</th>
<th>fcode</th>
<th>adminName1</th>
<th>adminName2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2524810</td>
<td>37.75</td>
<td>15.0</td>
<td>Mount Etna</td>
<td>IT</td>
<td>MT</td>
<td>Sicily</td>
<td></td>
</tr>
<tr>
<td>2523116</td>
<td>37.5</td>
<td>14.0</td>
<td>Sicily</td>
<td>IT</td>
<td>ISL</td>
<td>Sicily</td>
<td></td>
</tr>
<tr>
<td>2523119</td>
<td>37.75</td>
<td>14.25</td>
<td>Sicily</td>
<td>IT</td>
<td>ADM1</td>
<td>Sicily</td>
<td></td>
</tr>
<tr>
<td>6517485</td>
<td>33.99157</td>
<td>15.07467</td>
<td>Noto Sicily B&amp;B</td>
<td>IT</td>
<td>HTL</td>
<td>Sicily</td>
<td>Provincia di Siracusa</td>
</tr>
<tr>
<td>6499018</td>
<td>37.97624</td>
<td>14.94922</td>
<td>B&amp;B Holiday in Sicily</td>
<td>IT</td>
<td>HTL</td>
<td>Sicily</td>
<td>Provincia di Catania</td>
</tr>
<tr>
<td>6491033</td>
<td>37.4816</td>
<td>15.0834</td>
<td>Le Dune Sicily Hotel</td>
<td>IT</td>
<td>HTL</td>
<td>Sicily</td>
<td>Provincia di Catania</td>
</tr>
<tr>
<td>6486041</td>
<td>38.31382</td>
<td>15.03599</td>
<td>Dancing Flamingo Sicily Country Resort</td>
<td>IT</td>
<td>HTL</td>
<td>Sicily</td>
<td>Provincia di Siracusa</td>
</tr>
<tr>
<td>6409363</td>
<td>39.03131</td>
<td>14.02366</td>
<td>B &amp; B Ma &amp; Mi Cefalu Palermo Sicily Italy</td>
<td>IT</td>
<td>HTL</td>
<td>Sicily</td>
<td>Provincia di Palermo</td>
</tr>
<tr>
<td>3393171</td>
<td>-8.68353</td>
<td>-37.4543</td>
<td>Pão de Açúcar</td>
<td>BR</td>
<td>ADM2</td>
<td>Pão de Açúcar</td>
<td>Pão de Açúcar</td>
</tr>
</tbody>
</table>
Summary

• Motivation
• Related Work
• Instance-based Schema Matching
• **Experimental Approach**
  – Global Schema
  – Global Instances
  – Geographical Databases Web Services
  – Results
  – **Further considerations**
• Conclusion
Further considerations

- The **design** of the global schema influences the matching process.
- The **selection** of the global instances influences the performance of the instance-based matching approach.
Further considerations

The **design** of the global schema influences the matching process

- The **selection** of the global instances influences the performance of the instance-based matching approach
Further considerations

1) some attributes of the export schemas have no direct correspondence with any of the attributes of the global schema

<table>
<thead>
<tr>
<th>Attribute</th>
<th>$I_{GS}$</th>
<th>$N_{GS}$</th>
<th>$A_{GS}$</th>
<th>$O_{GS}$</th>
<th>$T_{GS}$</th>
<th>$A1_{GS}$</th>
<th>$A2_{GS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>geonameId</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lat</td>
<td>0</td>
<td>0</td>
<td>188</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lon</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>269</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>name</td>
<td>0</td>
<td>551</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>countryCode</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>countryName</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>foce</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>foceName</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>foceName</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>population</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>alternateName</td>
<td>0</td>
<td>155</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>elevation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminCode1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminCode2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminName1</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>adminName2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Occurrence Matrix: Geonames.org Export Schema x Global Schema
2) be careful with the temporal aspects of the global schema attributes, because they could be useless in this context.

**Ex:** supposes that the global instance set holds data from 2007, but a specific Web service provides data from 1970. In this case, the values of attribute *population*, say, would never re-occur on the returned data.
Further considerations

- The **design** of the global schema influences the matching process
- The **selection** of the global instances influences the performance of the instance-based matching approach
Further considerations

1) they are representative of the overall application domain

Ex: the global instance set must cover, as much as possible, the variety of types of geographic features, and it must contain “famous” places (w.r.t. the region considered)
Further considerations

2) Global instances have attribute values that do not match with too many attribute values of an export schema.

**Ex:** If we have "United States" as a global instance, it could reoccur as:
- `countryName` on Geonames instances
- `displayName` on ADL Gazetteer instances

---

**Geonames Instances Fragment**

<table>
<thead>
<tr>
<th>geonameid</th>
<th>name</th>
<th>lat</th>
<th>lng</th>
<th><code>countryName</code></th>
<th>fcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>5128722</td>
<td>Niagara Falls</td>
<td>43.083416155</td>
<td>-79.06627052</td>
<td>United States</td>
<td>OVF</td>
</tr>
<tr>
<td>4333587</td>
<td>Mississippi River</td>
<td>29.1510582</td>
<td>-89.2533842</td>
<td>United States</td>
<td>STM</td>
</tr>
</tbody>
</table>

**ADL Gazetteer Instances Fragment**

<table>
<thead>
<tr>
<th>adlid</th>
<th>name</th>
<th><code>displayName</code></th>
<th>footprintX</th>
<th>footprintY</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>adlgaz-1-6463903-80</td>
<td>Alps</td>
<td>Swiss Alp - Fayette County - Texas - United States</td>
<td>-96.9092</td>
<td>29.7822</td>
<td>populated places</td>
</tr>
<tr>
<td>adlgaz-1-6840132-1a</td>
<td>Mount Etna</td>
<td>Etna, Mount - Lyon County - Nevada - United States</td>
<td>-119.1436</td>
<td>38.6972</td>
<td>mountains</td>
</tr>
</tbody>
</table>
3) data quality is essential
   - errors in attribute values (or in its interpretation)
     may create false matchings

 **Ex:** "Niagara Falls" occurred 81 times as *alternateNames*
 in Geonames.org
Summary

- Motivation
- Related Work
- Instance-based Schema Matching
- Experimental Approach
  - Global Schema
  - Global Instances
  - Geographical Databases Web Services
  - Results
  - Further considerations
- Conclusion
Conclusion

• a semantic approach, using instances, for matching export schemas of geographical database available through Web services

• experiments using two real Web gazetteers services

• important issues that must be considered when designing the global schema and selecting the global instance set
Conclusion

• Future work:
  – Improve the instance-based schema matching process
  – Improve the re-occurrence detection method
  – Execute a validation step to formally define a threshold to the ratio between reoccurrence values
  – Investigate the automatic generation of the global schema
Conclusion

- Future Work
  - Prototype a **Web databases services mediator** as a proof of concept
An Instance-based Approach for Matching Export Schemas of Geographical Database Web Services

Daniela F. Brauner, Chantal Intrator, João Carlos Freitas, Marco A. Casanova
{dani, cintrator, jcsfreitas, casanova}@inf.puc-rio.br

Pontifical Catholic University of Rio de Janeiro (PUC-Rio)
Department of Informatics