



GEOINFO 2007

IX Brazilian Symposium on GeoInformatics



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

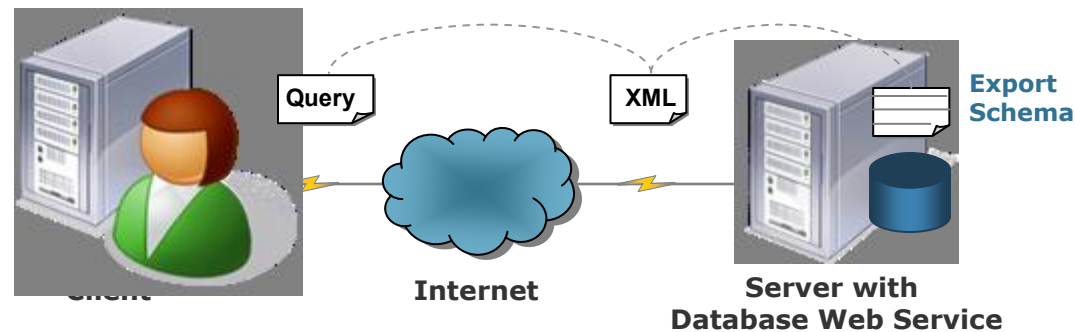


- **Motivation**
- Related Work
- Instance-based Schema Matching
- Experimental Approach
 - Global Schema
 - Global Instances
 - Geographical Databases Web Services
 - Results
 - Further considerations
- Conclusion

Motivation



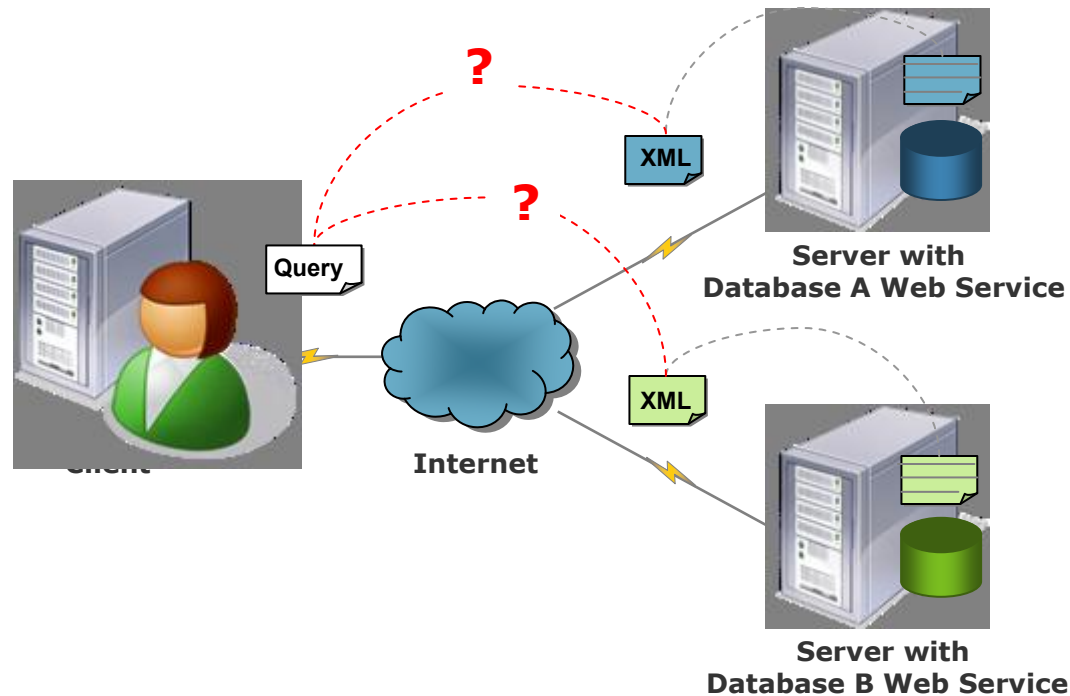
- *Database Web service:*
 - a Web service interface with operations to access a backend database
 - returns results following a given export schema
- *Export schema:*
 - a subset of the backend database schema visible to the clients





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



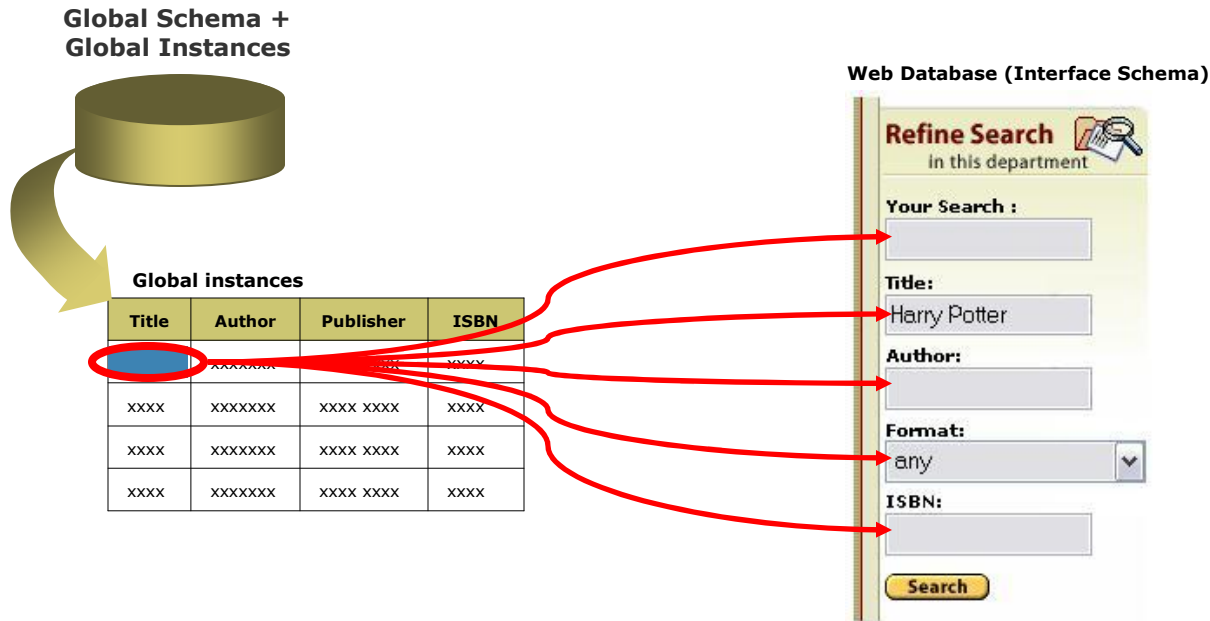
Related work

- 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



Wang, J., Wen, J. Lochofsky, F.H. and Ma, W. (2004). **Instance-based schema matching for web databases by domain-specific query probing**, In Proceedings of 30th Intl. Conference on Very Large Data Bases, pp. 408-419.

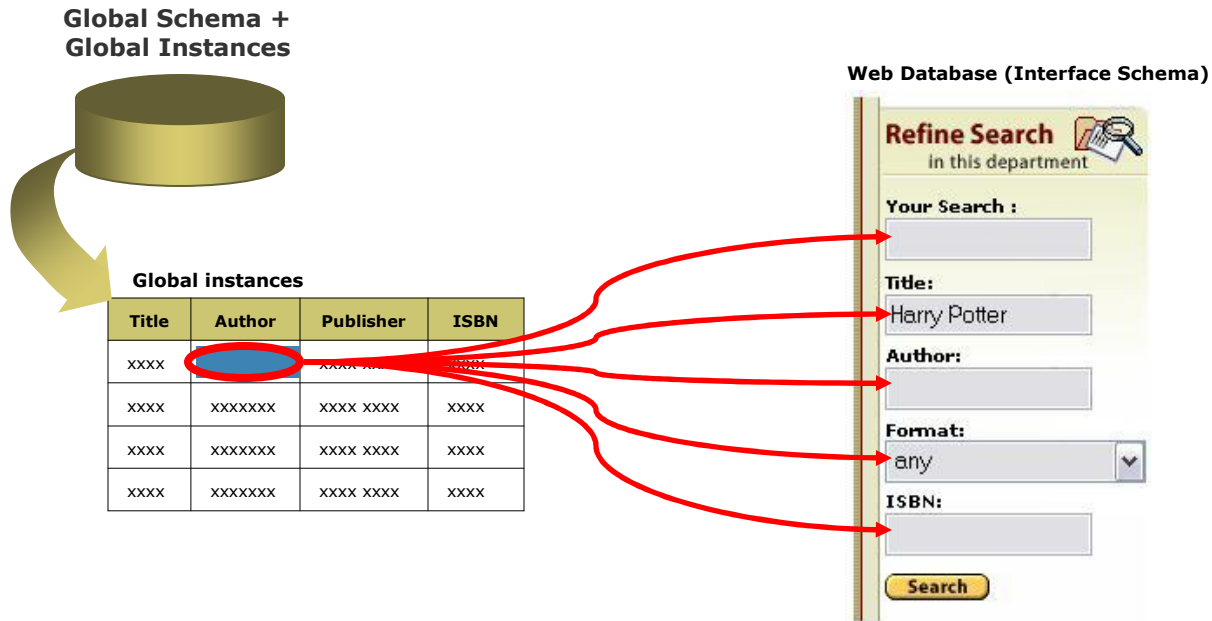
Related work



Wang, J., Wen, J. Lochovsky, F.H. and Ma, W. (2004). **Instance-based schema matching for web databases by domain-specific query probing**, In Proceedings of 30th Intl. Conference on Very Large Data Bases, pp. 408-419.



Related work



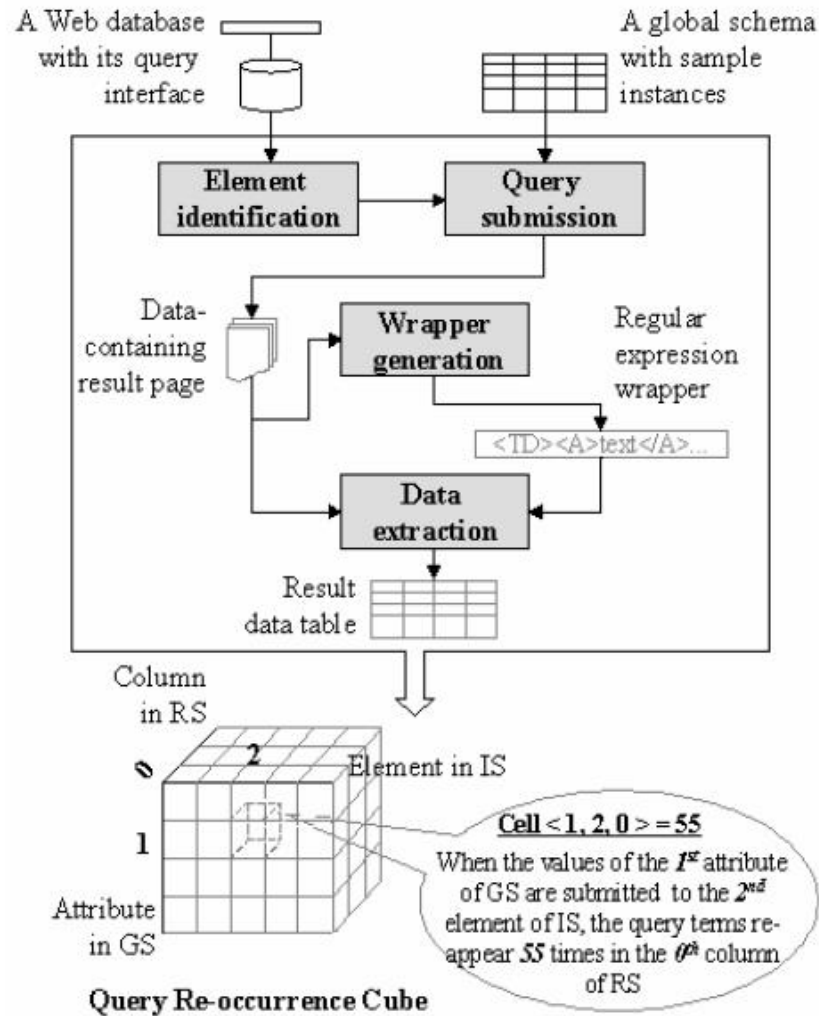
Wang, J., Wen, J. Lochovsky, F.H. and Ma, W. (2004). **Instance-based schema matching for web databases by domain-specific query probing**, In Proceedings of 30th Intl. Conference on Very Large Data Bases, pp. 408-419.

Related work



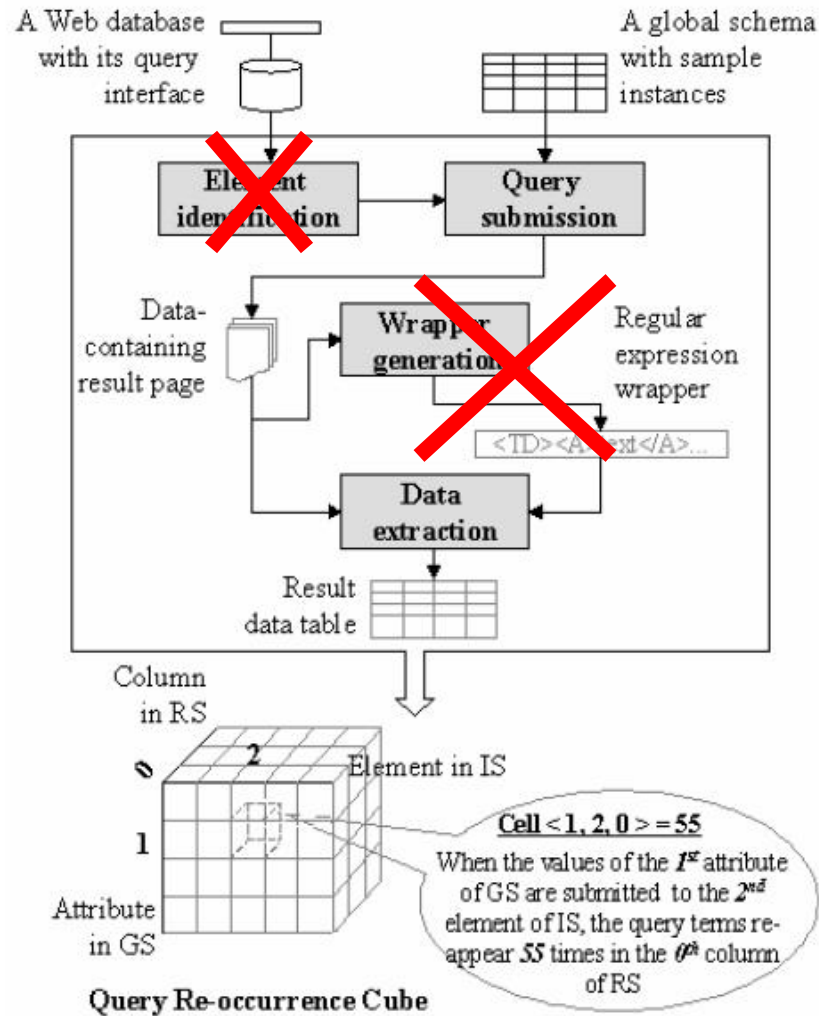
Wang, J., Wen, J. Lochofsky, F.H. and Ma, W. (2004). **Instance-based schema matching for web databases by domain-specific query probing**, In Proceedings of 30th Intl. Conference on Very Large Data Bases, pp. 408-419.

Related work



Wang, J., Wen, J. Lochofsky, F.H. and Ma, W. (2004). **Instance-based schema matching for web databases by domain-specific query probing**, In Proceedings of 30th Intl. Conference on Very Large Data Bases, pp. 408-419.

Related work



Wang, J., Wen, J. Lochovsky, F.H. and Ma, W. (2004). **Instance-based schema matching for web databases by domain-specific query probing**, In Proceedings of 30th Intl. Conference on Very Large Data Bases, pp. 408-419.

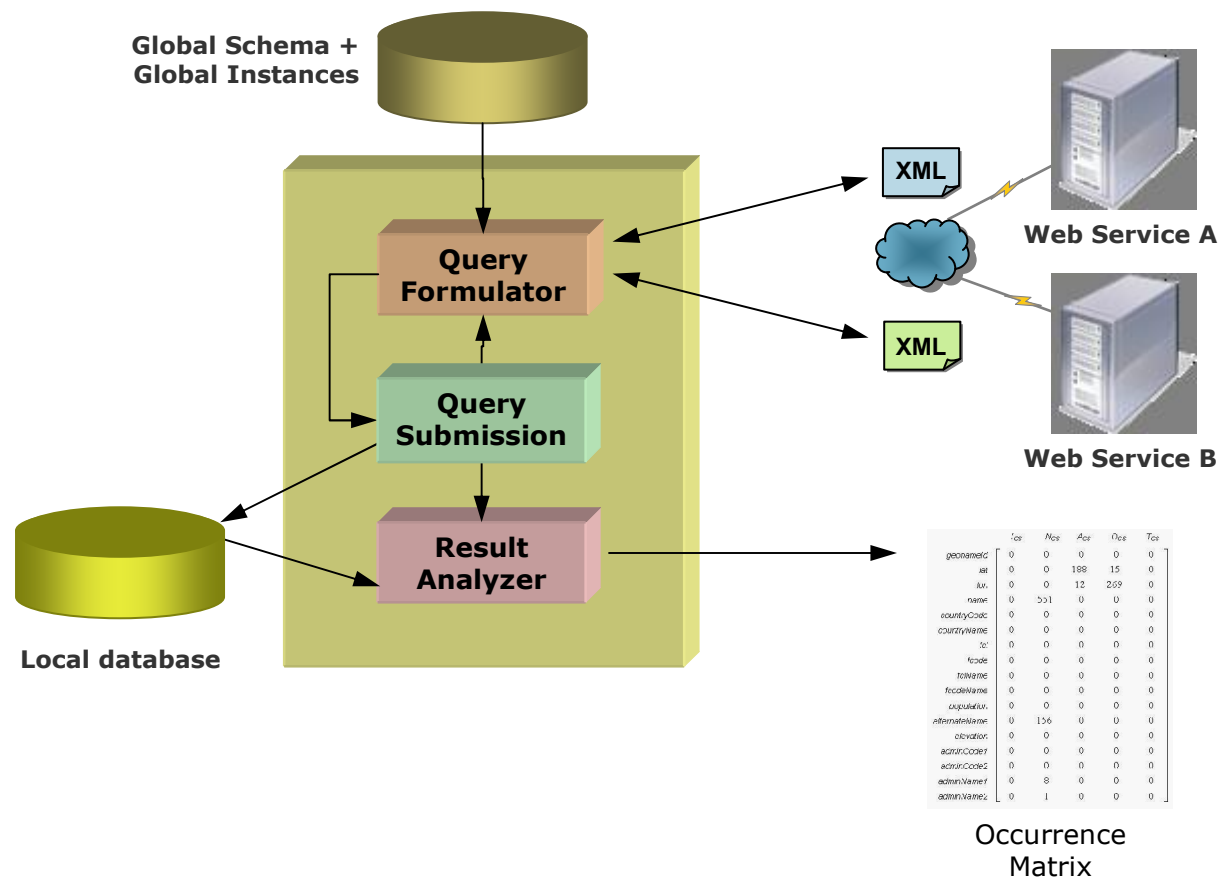


- Motivation
- Related Work
- **Instance-based Schema Matching**
- Experimental Approach
 - Global Schema
 - Global Instances
 - Geographical Databases Web Services
 - Results
 - Further considerations
- Conclusion



Instance-based Schema Matching

- How to match database Web services export schemas?
 - Using the schema matching process:





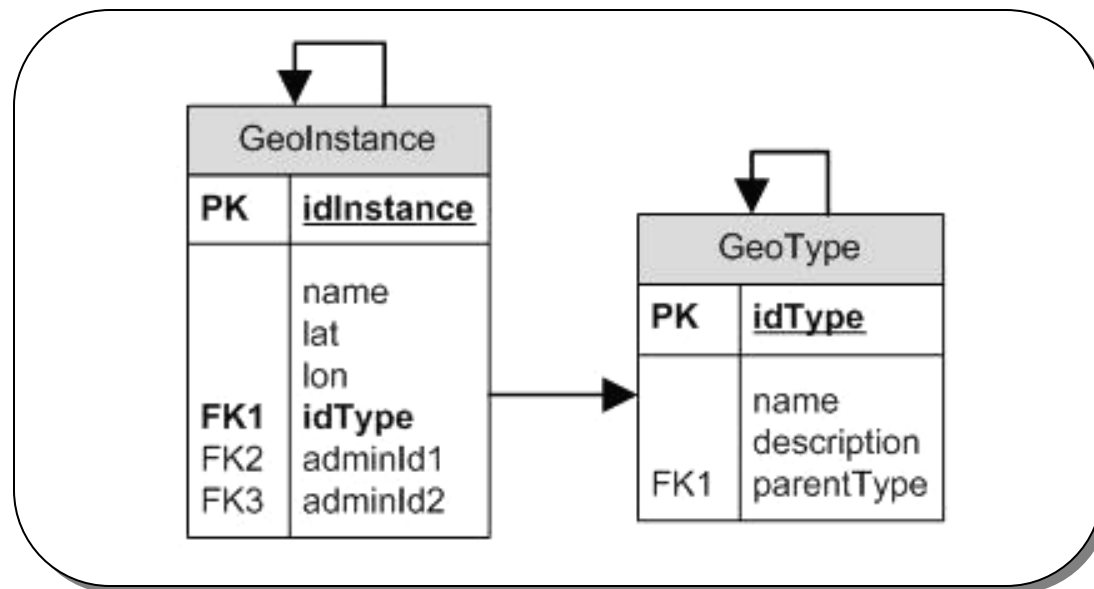
- Motivation
- Related Work
- Instance-based Schema Matching
- **Experimental Approach**
 - **Global Schema**
 - Global Instances
 - Geographical Databases Web Services
 - Results
 - Further considerations
- Conclusion



Global Schema

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

E-R Model of the proposed Geographical Global Schema



ISO/TC211 (2003). **ISO 19112:2003 Geographic information — Spatial referencing by geographic identifiers**. International Standard 19112. Technical report of Technical Committee ISO/TC 211.

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

idInstance	name	lat	lon	idType	admink1	admink2
175	Galapagos Islands	0.0	-90.5	4	73	-
52	Alps	46.4166667	10.0	15	165	-
149	Atlantic Ocean	10.0	-25.0	9	-	-
90	Niagara Falls	43.083416155	-79.06627052	21	123	-
16	Pão de Açúcar	-22.9472	-43.1561	14	101	-
34	Mississippi River	29.1510582	-89.2533842	19	109	-



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

Geographical Databases Web Services (I/II)



- 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>

Geographical Databases Web Services (II/II)



- Export Schemas

XML response fragment of Geonames.org

```
<?xml version="1.0" encoding="UTF-8" ?>
- <geonames style="FULL">
  <totalResultsCount>1</totalResultsCount>
  - <geoname>
    <name>Amazon River</name>
    <lat>-0.1666667</lat>
    <lng>-49.0</lng>
    <geonameId>3407729</geonameId>
    <countryCode>BR</countryCode>
    <countryName>Brazil</countryName>
    <fcl>H</fcl>
    <fcode>STM</fcode>
    <fclName>stream, lake, ...</fclName>
    <fcodeName>stream</fcodeName>
    <population />
    <alternateNames>Orellana,Rio Amazonas,Rio Maranon,Rio Solimoies,Rio
    Solimões,Rio el Amazonas,Rio Amazonas,Rio Marañón,Rio el
    Amazonas,Salimoies River,Solimoies</alternateNames>
    <elevation />
    <adminCode1>00</adminCode1>
    <adminName1 />
    <adminCode2 />
    <adminName2 />
    <timezone dstOffset="-3.0" gmtOffset="-3.0">America/Belem</timezone>
  </geoname>
</geonames>
```

XML response fragment of ADL Gazetteer

```
<?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>adlgaz-1-1410143-3a</identifier>
      <place-status>current</place-status>
      <display-name>Amazon River - Brazil</display-name>
      - <names>
        <name primary="true" status="current">Amazon River</name>
        <name primary="false" status="current">Solimoies</name>
        <name primary="false" status="current">Salimoies River</name>
        <name primary="false" status="current">Orellana</name>
        <name primary="false" status="current">Maranon, Rio</name>
        <name primary="false" status="current">Amazonas, Rio</name>
        <name primary="false" status="current">Amazonas, Rio el</name>
        <name primary="false" status="current">Solimoies, Rio</name>
      </names>
      - <bounding-box>
        - <gml:coord>
          <gml:X>-49.0</gml:X>
          <gml:Y>-0.1667</gml:Y>
        </gml:coord>
        - <gml:coord>
          <gml:X>-49.0</gml:X>
          <gml:Y>-0.1667</gml:Y>
        </gml:coord>
      </bounding-box>
      - <footprints>
        - <footprint primary="true">
          - <gml:Point>
            - <gml:coord>
              <gml:X>-49.0</gml:X>
              <gml:Y>-0.1667</gml:Y>
            </gml:coord>
            </gml:Point>
          </footprint>
        </footprints>
      - <classes>
        <class thesaurus="ADL Feature Type Thesaurus" primary="true">streams</class>
        <class thesaurus="NIMA Feature Designation" primary="false">STM (stream)</class>
      </classes>
      - <relationships>
        <relationship relation="part of" target-name="UTM grid GE28" />
        <relationship relation="part of" target-name="JOG Sheet Number SA22-0" />
        <relationship relation="part of" target-name="Brazil" target-identifier="adlgaz-1-19-19" />
      </relationships>
    </gazetteer-standard-report>
  </standard-reports>
</query-response>
</gazetteer-service>
```



- 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

geonameid	lat	lng	name	country Code	fcode	...
1283416	27.9833	86.9333	<u>Mount Everest</u>	NP	MT	
1004850	-28.15	29.16667	<u>Mount Everest</u>	ZA	MT	
4122419	33.78733	-93.3804	<u>Mount Everest</u> Church	US	CH	
4334114	29.94326	-90.0904	<u>Mount Everest</u> Baptist Church	US	CH	
4341122	29.94104	-90.089	Second <u>Mount Everest</u> Baptist Church	US	CH	
4694788	32.70374	-96.7881	Greater <u>Mount Everest</u> Baptist Church	US	CH	

Results



- Occurrences matrices:

(a) Geonames.org Export Schema \times Global Schema

(b) ADL Gazetteer Export Schema \times Global Schema

Global Schema:

I_{GS}	idInstance
N_{GS}	name
A_{GS}	lat
O_{GS}	lon
T_{GS}	idType
$A1_{GS}$	adminId1
$A2_{GS}$	adminId2

	I_{GS}	N_{GS}	A_{GS}	O_{GS}	T_{GS}	$A1_{GS}$	$A2_{GS}$
<i>geonameId</i>	0	0	0	0	0	0	0
<i>lat</i>	0	0	188	15	0	0	0
<i>lon</i>	0	0	12	269	0	0	0
<i>name</i>	0	551	0	0	0	0	0
<i>countryCode</i>	0	0	0	0	0	0	0
<i>countryName</i>	0	0	0	0	0	0	0
<i>fcl</i>	0	0	0	0	0	0	0
<i>fcode</i>	0	0	0	0	0	0	0
<i>fclName</i>	0	0	0	0	0	0	0
<i>fcodeName</i>	0	0	0	0	0	0	0
<i>population</i>	0	0	0	0	0	0	0
<i>alternateName</i>	0	156	0	0	0	0	0
<i>elevation</i>	0	0	0	0	0	0	0
<i>adminCode1</i>	0	0	0	0	0	0	0
<i>adminCode2</i>	0	0	0	0	0	0	0
<i>adminName1</i>	0	8	0	0	0	0	0
<i>adminName2</i>	0	1	0	0	0	0	0

(a)

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

(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”

Global Schema:

I_{GS}	idInstance
N_{GS}	name
A_{GS}	lat
O_{GS}	lon
T_{GS}	idType
$A1_{GS}$	adminId1
$A2_{GS}$	adminId2

	I_{GS}	N_{GS}	A_{GS}	O_{GS}	T_{GS}	$A1_{GS}$	$A2_{GS}$
geonameId	0	0	0	0	0	0	0
lat	0	0	188	15	0	0	0
lon	0	0	12	269	0	0	0
name	0	551	0	0	0	0	0
countryCode	0	0	0	0	0	0	0
countryName	0	0	0	0	0	0	0
fcl	0	0	0	0	0	0	0
fcode	0	0	0	0	0	0	0
fclName	0	0	0	0	0	0	0
fcodeName	0	0	0	0	0	0	0
population	0	0	0	0	0	0	0
alternateName	0	156	0	0	0	0	0
elevation	0	0	0	0	0	0	0
adminCode1	0	0	0	0	0	0	0
adminCode2	0	0	0	0	0	0	0
adminName1	0	8	0	0	0	0	0
adminName2	0	1	0	0	0	0	0

(a)

	I_{GS}	N_{GS}	A_{GS}	O_{GS}	T_{GS}	$A1_{GS}$	$A2_{GS}$
identifier	0	0	0	0	0	0	0
placeStatus	0	0	0	0	0	0	0
name	0	459	0	0	0	0	0
displayName	0	352	0	0	0	0	0
footprintX	0	0	14	134	0	0	0
footprintY	0	0	94	12	0	0	0
class	0	0	0	0	0	0	0
thesaurus	0	0	0	0	0	0	0
names	0	435	0	0	0	0	0
relationships	0	24	0	0	0	0	0

(b)

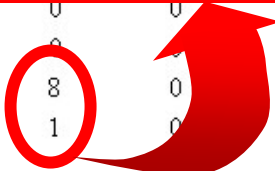
Results



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

geonameId	lat	lng	name	countryCode	fcode	adminName1	adminName2
2524810	37.75	15.0	Mount Etna	IT	MT	Sicily	
2523118	37.5	14.0	Sicily	IT	ISL	Sicily	
2523119	37.75	14.25	Sicily	IT	ADM1	Sicily	
6517485	36.89157	15.07487	Noto Sicily B&B	IT	HTL	Sicily	Provincia di Siracusa
6489618	37.87824	14.94922	B&B Holiday in Sicily	IT	HTL	Sicily	Provincia di Catania
6491033	37.4816	15.0834	Le Dune Sicily Hotel	IT	HTL	Sicily	Provincia di Catania
6488041	36.81382	15.03599	Dancing Flamingo Sicily Country Resort	IT	HTL	Sicily	Provincia di Siracusa
6489363	38.03131	14.02396	B & B Ma & Mi Cefalu Palermo Sicily Italy	IT	HTL	Sicily	Provincia di Palermo
3393171	-9.68353	-37.4543	Pão de Açúcar	BR	ADM2	Pão de Açúcar	Pão de Açúcar

adminCode1	0	0	0	0	0	0	0
adminCode2	0	0	0	0	0	0	0
adminName1	0	8	0	0	0	0	0
adminName2	0	1	0	0	0	0	0



(a)

(b)



- 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

	<i>I_{GS}</i>	<i>N_{GS}</i>	<i>A_{GS}</i>	<i>O_{GS}</i>	<i>T_{GS}</i>	<i>A1_{GS}</i>	<i>A2_{GS}</i>
<i>geonameId</i>	0	0	0	0	0	0	0
<i>lat</i>	0	0	188	15	0	0	0
<i>lon</i>	0	0	12	269	0	0	0
<i>name</i>	0	551	0	0	0	0	0
<i>countryCode</i>	0	0	0	0	0	0	0
<i>countryName</i>	0	0	0	0	0	0	0
<i>fc1</i>	0	0	0	0	0	0	0
<i>fcode</i>	0	0	0	0	0	0	0
<i>fc1Name</i>	0	0	0	0	0	0	0
<i>fcodeName</i>	0	0	0	0	0	0	0
<i>population</i>	0	0	0	0	0	0	0
<i>alternateName</i>	0	156	0	0	0	0	0
<i>elevation</i>	0	0	0	0	0	0	0
<i>adminCode1</i>	0	0	0	0	0	0	0
<i>adminCode2</i>	0	0	0	0	0	0	0
<i>adminName1</i>	0	8	0	0	0	0	0
<i>adminName2</i>	0	1	0	0	0	0	0

Global Schema:

<i>I_{GS}</i>	idInstance
<i>N_{GS}</i>	name
<i>A_{GS}</i>	lat
<i>O_{GS}</i>	lon
<i>T_{GS}</i>	idType
<i>A1_{GS}</i>	adminId1
<i>A2_{GS}</i>	adminId2

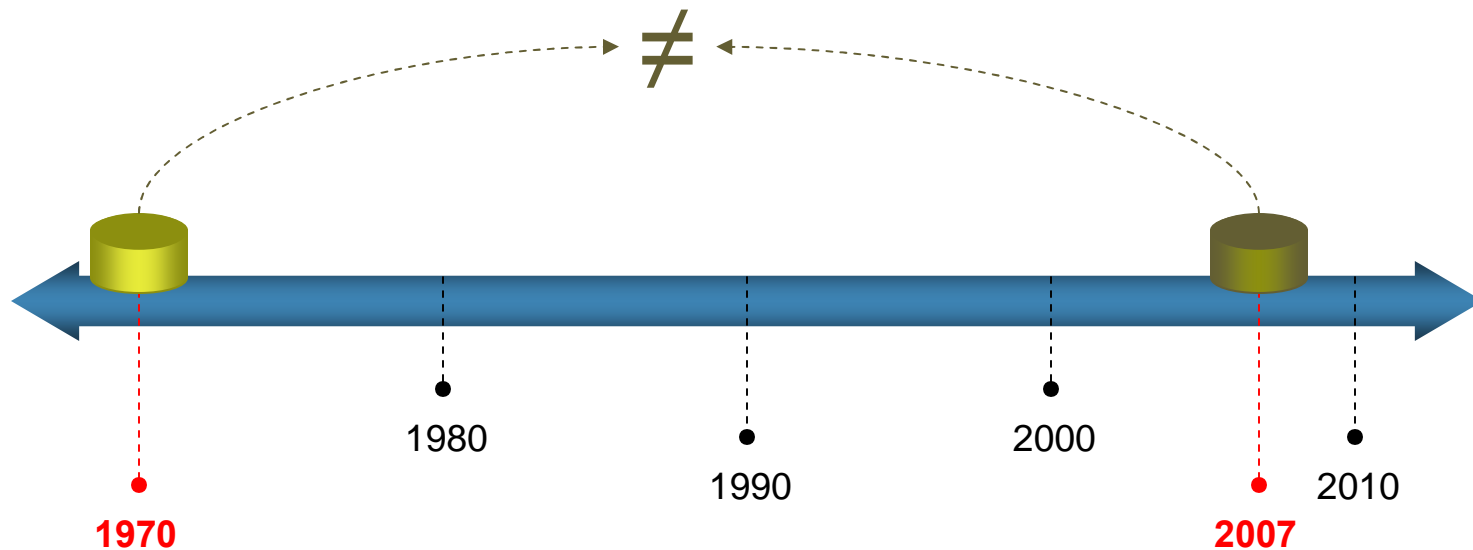
Occurrence Matrix: Geonames.org Export Schema x Global Schema

Further considerations



2) be careful with the temporal aspects of the global schema attributes, because they could be useless in this context

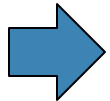
Ex: supposes that the global instance set holds data from 2007, but a specific Web service provides data from 1970. In this case, the values of attribute *population*, say, would never re-occur on the returned data.





Further considerations

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



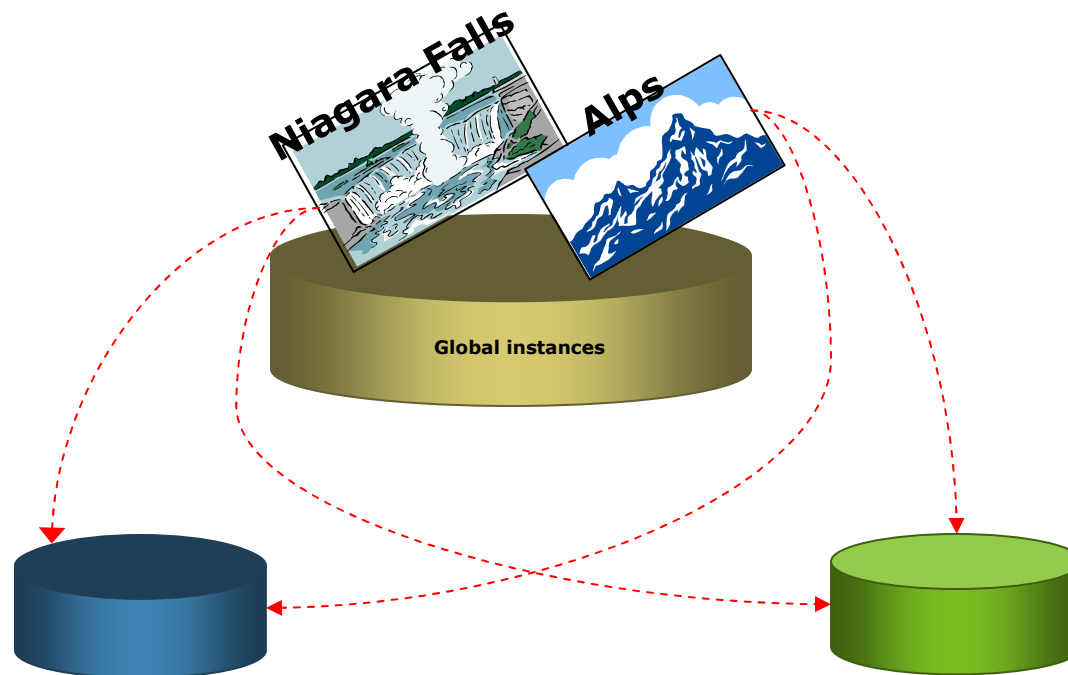
The **selection** of the global instances influences the performance of the instance-based matching approach

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



Global instances

Geonames Instances Fragment

geonameId	name	lat	lng	countryName	fcode	...
5128722	Niagara Falls	43.083416155	-79.06627052	United States	OVF	
4333587	Mississippi River	29.1510582	-89.2533842	United States	STM	

ADL Gazetteer Instances Fragment

adId	name	displayName	footprintX	footprintY	class	...
adlgaz-1-6463903-80	Alps	Swiss Alp - Fayette County - Texas - United States	-96.9092	29.7822	populated places	
adlgaz-1-6840132-1a	Mount Etna	Etna, Mount - Lyon County - Nevada - United States	-119.1436	38.6972	mountains	

Further considerations



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



geonameid	lat	lng	name	countryllame	adminllame1	fcode	alternatellames	...
6501773	43.0942	-79.0851	Glengate Hotel	Canada	Ontario	HTL	Niagara Falls	
6468151	43.0835	-79.0828	Hilton Niagara Falls	Canada	Ontario	HTL	Niagara Falls	
6466074	43.0848	-79.0586	Holiday Inn at the Falls	United States	New York	HTL	Niagara Falls	
6485423	43.0895	-79.0818	Howard Johnson Hotel By The Falls	Canada	Ontario	HTL	Niagara Falls	

⋮

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

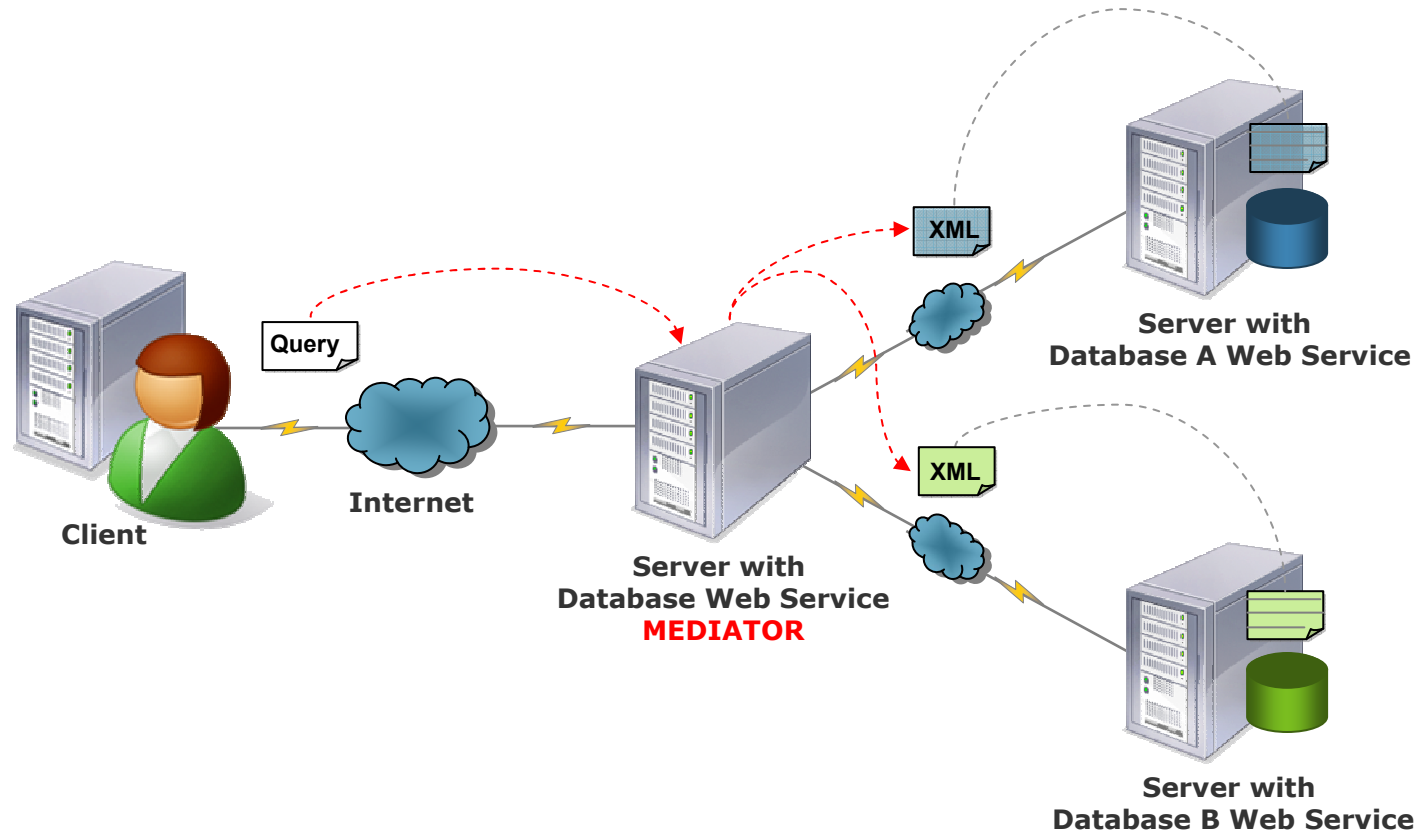


- Future work:
 - Improve the instance-based schema matching process
 - Improve the re-occurrence detection method
 - Execute a validation step to formally define a threshold to the ratio between reoccurrence values
 - Investigate the automatic generation of the global schema

Conclusion



- Future Work
 - Prototype a **Web databases services mediator** as a proof of concept





GEOINFO 2007

IX Brazilian Symposium on GeoInformatics



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

