W-Ray: A Strategy to Publish Deep Web Geographic Data

Helena Piccinini[1], Melissa Lemos[1],
Marco A. Casanova[1], Antonio L. Furtado[1]

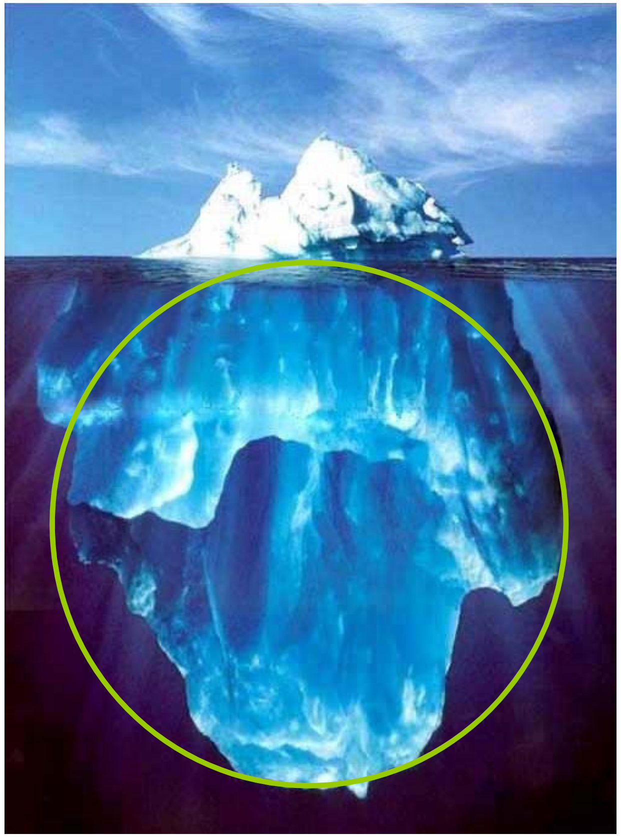
[1]Departamento de Informática, PUC-Rio
[2]Diretoria de Informática, IBGE

Topics

- Motivation
- Conventional Data
- Geographical Data – Vectorial Format
- Geographical Data – Raster Format
- Conclusions

Motivation

- **Deep Web**
  - data stored in databases, dynamic pages, scripted pages, multimedia data,…



Iceberg Photo

Judith Currelly, Diane Farris Gallery

Motivation

- **Problem**
  - traditional search engines cannot discover data stored in databases by following hyperlinks, but rather they have to use query interfaces
  - (traditional search engines are virtually blind to data stored in databases)

Motivation

- **Current solutions**
  - *surfacing* or *Deep Web Crawl*
    - automatically fills HTML forms to query the databases
    - executes queries offline
    - translates results to static Web pages
    - indexes the static Web pages
  - *federated search* or *virtual integration*
    - uses domain-specific mediators to access the databases

Motivation

- **W-Ray**
  - a methodology to publish Deep Web data
    - creates a set of natural language (NL) sentences to describe data in a database
    - publishes the sentences as static Web pages, which are then indexed as usual
- **W-Ray 2.0**
  - uses RDF triples in addition to natural language sentences

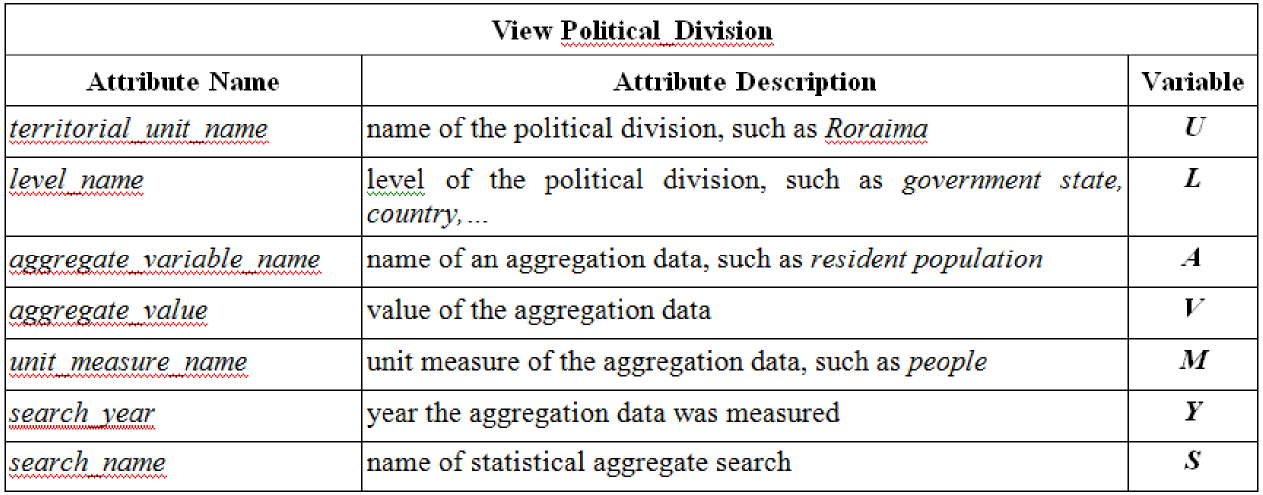
Motivation

- **W-Ray**
  - a methodology that reverses the roles:
    - database administrator:
      - responsible for publishing (selected) data as Web pages
    - search engine
      - responsible just for locating and indexing the Web pages (as usual)

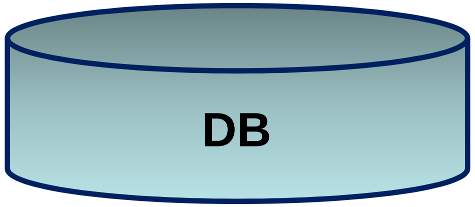
Conventional Data

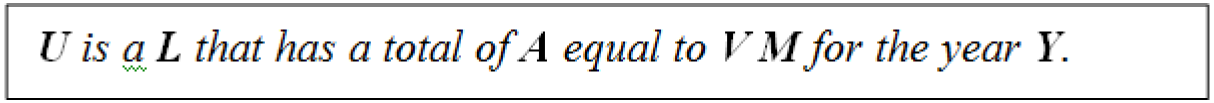
- **Strategy for conventional data**
  - define database views
    - views describe classes of objects in the database
    - views should preferably use a controlled vocabulary
  - create templates to guide the synthesis of NL sentences
  - publish static Web pages
    - materialize the views
    - translate materialized view data to NL sentences
    - publish the sentences as static Web pages

Table 1 - Summary of "*political_division*" view over the SIDRA database.



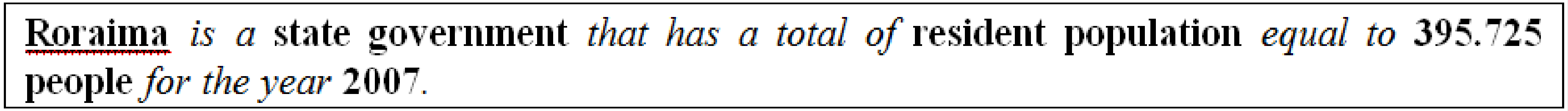
View Political_Division		
Attribute Name	**Attribute Description**	**Variable**
territorial_unit_name	name of the political division, such as *Roraima*	*U*
level_name	level of the political division, such as *government state, country, …*	*L*
aggregate_variable_name	name of an aggregation data, such as *resident population*	*A*
aggregate_value	value of the aggregation data	*V*
unit_measure_name	unit measure of the aggregation data, such as *people*	*M*
search_year	year the aggregation data was measured	*Y*
search_name	name of statistical aggregate search	*S*



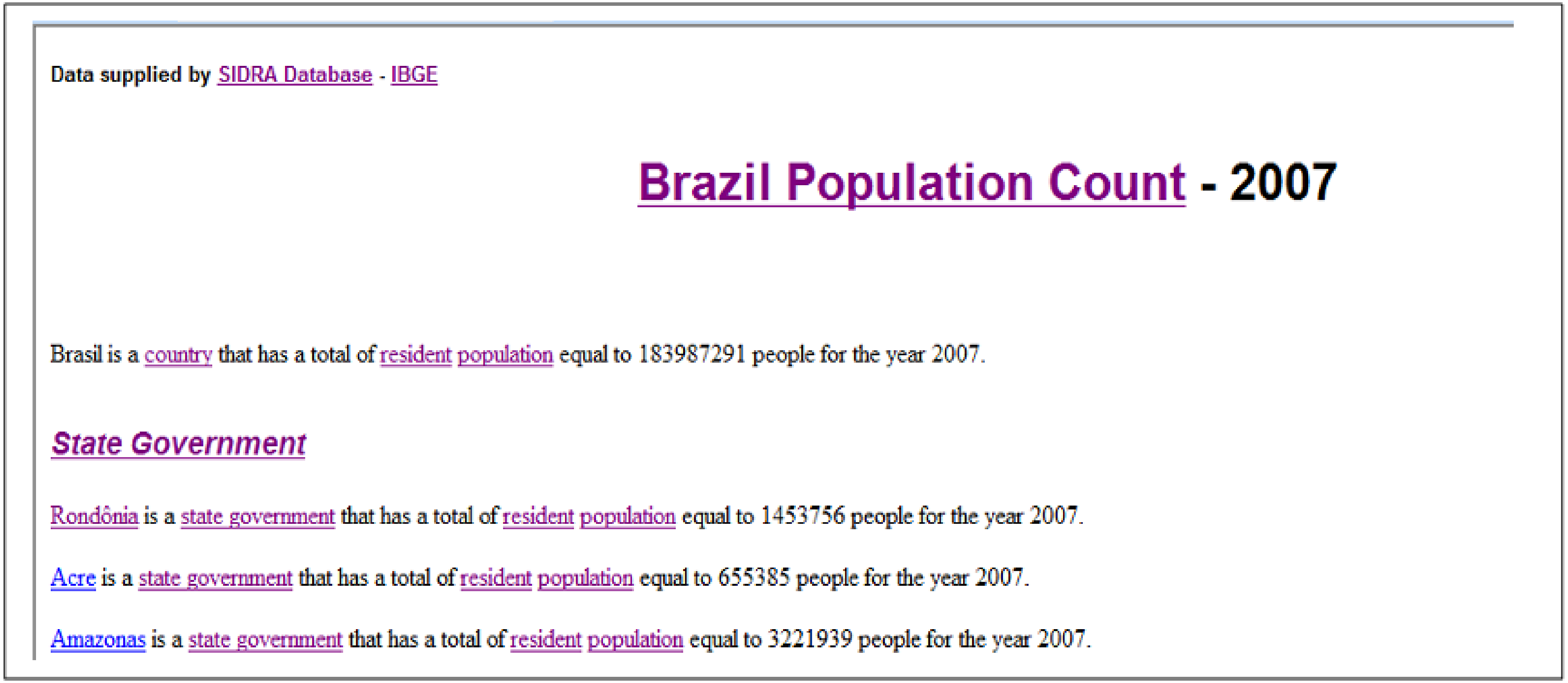


U is a L that has a total of A equal to V M for the year Y.

Name	Level	…
Roraima	State government	…
Acre	State government	…



Roraima *is a* state government *that has a total of* resident population *equal to* 395.725 people *for the year* 2007.



Data supplied by SIDRA Database - IBGE

Brazil Population Count - 2007

Brasil is a country that has a total of resident population equal to 183987291 people for the year 2007.

State Government

Rondônia is a state government that has a total of resident population equal to 1453756 people for the year 2007.

Acre is a state government that has a total of resident population equal to 655385 people for the year 2007.

Amazonas is a state government that has a total of resident population equal to 3221939 people for the year 2007.

Conventional Data

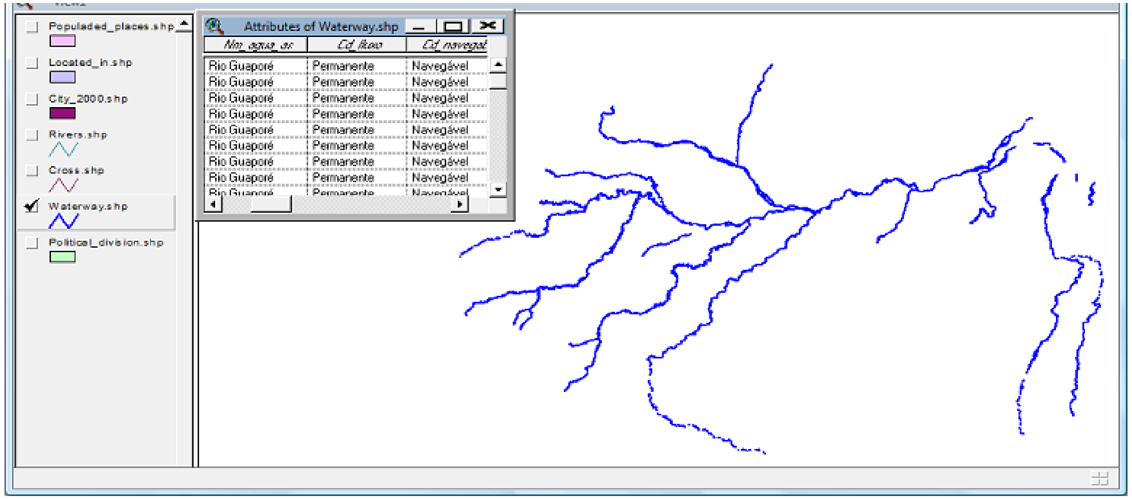
- **Strategy for conventional data – final remarks**
  - conventional objects are described by their attribute values
  - a judicious choice of the vocabulary is essential

Geographical Data – Vectorial Format

- **Strategy for geographical data in vectorial format**
  - similar to the strategy used for conventional data
    - define views + define templates + publish static Web pages
  - geographic data is described by the objects they depict
    - typically geographic data in vectorial format is already stored with the objects they depict

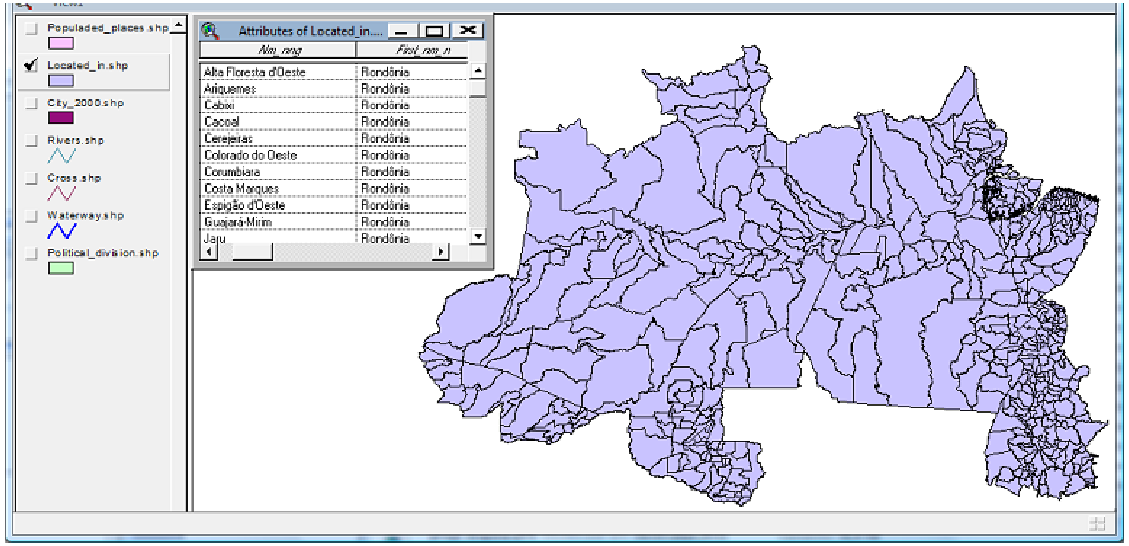
Geographical Data – Vectorial Format

- **Example of view definition**
  - layers:
    - political division + populated places + waterways
  - selected attributes
    - “political division”: name, abbreviate name, …
    - …
  - spatial restriction
    - all located in the north region
  - spatial relationships
    - populated places X political division: ‘is located in’
    - waterways X political division: ‘crosses’



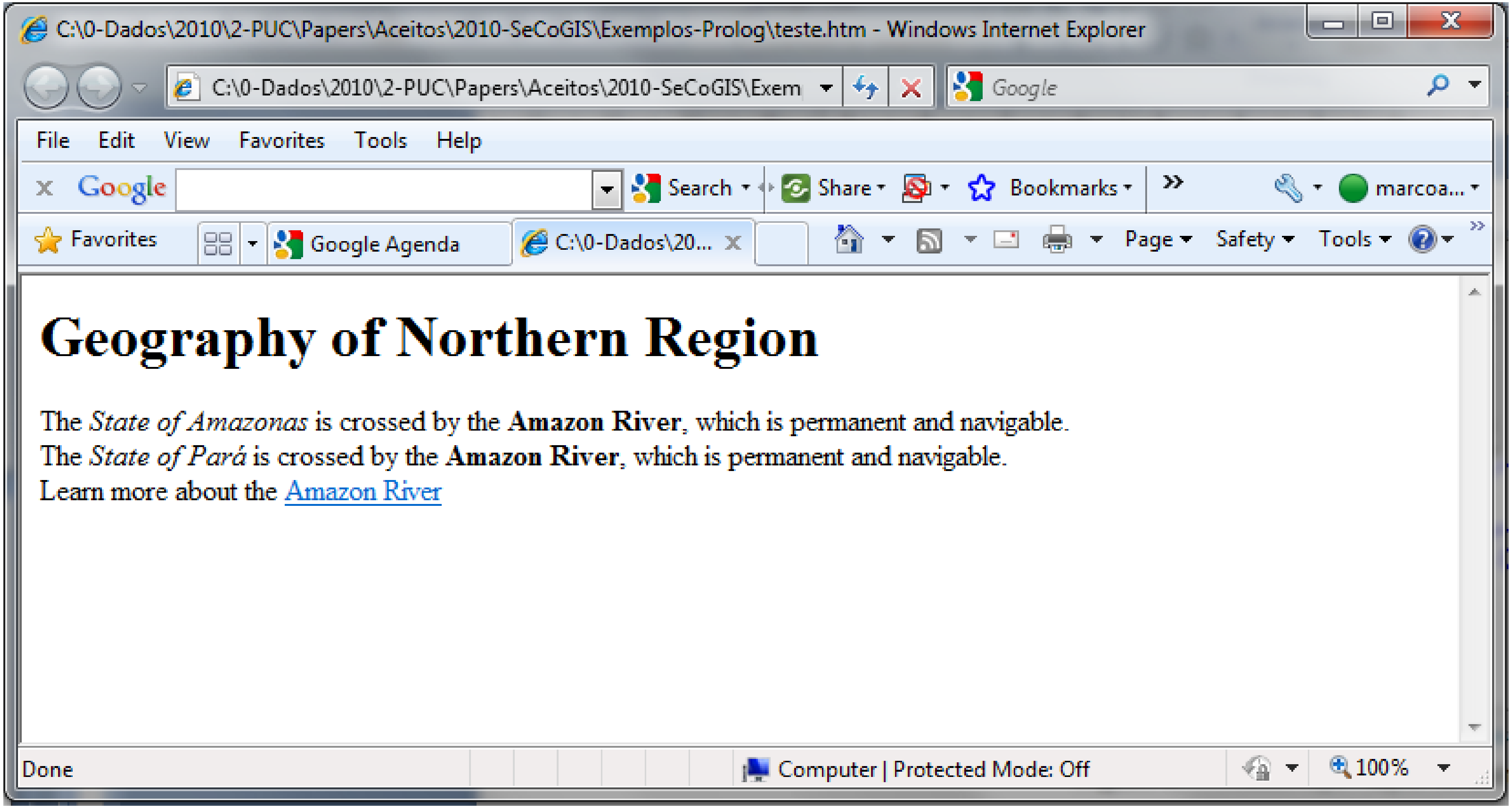
Nm_agua_ac	Cd_fluxo	Cd_navegal
Rio Guaporé	Permanente	Navegável
Rio Guaporé	Permanente	Navegável
Rio Guaporé	Permanente	Navegável
Rio Guaporé	Permanente	Navegável
Rio Guaporé	Permanente	Navegável
Rio Guaporé	Permanente	Navegável
Rio Guaporé	Permanente	Navegável
Rio Guaporé	Permanente	Navegável
Rio Guaporé	Permanente	Navegável

Figure 6 – "waterways" layer of north region of Brazil. BCIM – 2009 Database.



Nm_mng	First_nm_n
Alta Floresta d'Oeste	Rondônia
Ariquemes	Rondônia
Cabixi	Rondônia
Cacoal	Rondônia
Cerejeiras	Rondônia
Colorado do Oeste	Rondônia
Corumbiara	Rondônia
Costa Marques	Rondônia
Espigão d'Oeste	Rondônia
Guajará-Mirim	Rondônia
Jaru	Rondônia

Figure 7 – *"Located_in"* topological relationship between populated places and political division.



C:\0-Dados\2010\2-PUC\Papers\Aceitos\2010-SeCoGIS\Exemplos-Prolog\teste.htm - Windows Internet Explorer
C:\0-Dados\2010\2-PUC\Papers\Aceitos\2010-SeCoGIS\Exem
Google
File Edit View Favorites Tools Help
Google
Search
Share
Bookmarks
marcoa...
Favorites
Google Agenda
C:\0-Dados\20...
Page
Safety
Tools
Done
Computer | Protected Mode: Off
100%

Geography of Northern Region

The *State of Amazonas* is crossed by the **Amazon River**, which is permanent and navigable.
The *State of Pará* is crossed by the **Amazon River**, which is permanent and navigable.
Learn more about the Amazon River

Geographical Data – Vectorial Format

- **Strategy for geographical data in vectorial format – final remarks**
  - – the view definition should reflect a strategy
    to list the geographic objects and their relative priority
    (as in a Geography textbook!)
    - north to south + west to east
    - states have priority over cities, etc…

Geographical Data – Raster Format

- **Strategy for geographical data in raster format**
  - similar to the strategy used for conventional data
    - define views + define templates + publish static Web pages
  - geographic data is described by the objects they depict
    - typically geographic data in raster format IS NOT STORED with the objects they depict (!)
    - use a GIS to locate the objects “covered” by the raster data

Geographical Data – Raster Format

- **Example**
  - extract image parameters
  - query a GIS for ‘hydrographic feature’
    - Feature(“Rodrigo de Freitas, Lagoa - Brazil”, lakes, contains)
    - Feature(“Comprido, Rio – Brazil”, streams, contains)
    - Feature(“Maracana, Rio – Brazil, streams, contains)

***The image of* Rio de Janeiro, Brazil, *contains the lake* “Rodrigo de Freitas” *and the streams* “Comprido” and “Maracanã”.**

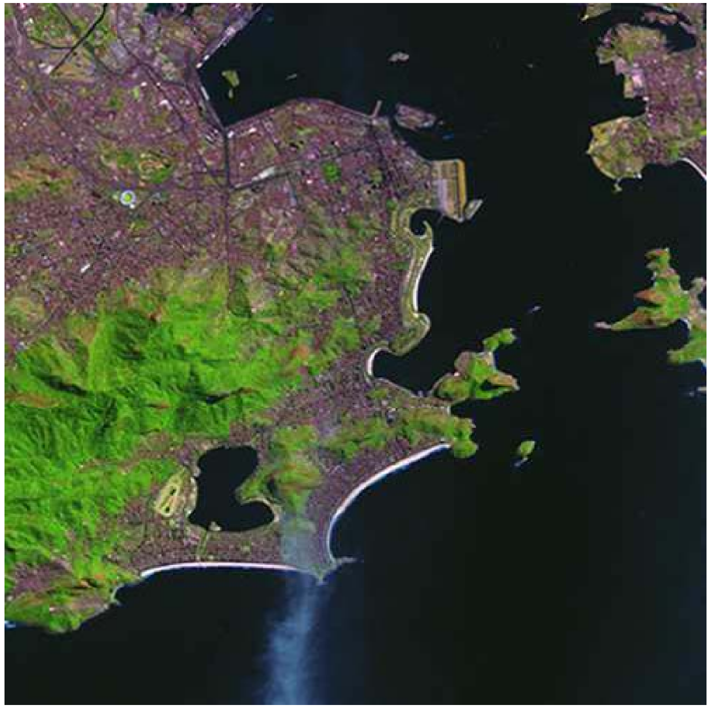


Image fragment of the City of Rio de Janeiro from the Web site “Brazil seen from Space”

Geographical Data – Raster Format

- **Strategy for geographical data in raster format – final remarks**
  - standard “satellite images” are typically repetitive and organized in a grid

⇩

  - it suffices to describe each grid cell, plus
    the specific details of each image of the cell
    (such cloud coverage)

Conclusions

- **Summary**
  - W-ray
    - define views + define templates + publish static Web pages
    - (Web search engines will index the Web pages, as usual)

Conclusions

- **Future (past) work**
  - publish RDF triples (linked data)
    - requires entity identification
  - publish Web pages + RDF triples
    - uses RDFa to label content

Conclusions

- **Future (future) work**
  - massive experiment with IBGE data
    - 180,000 Web pages to describe all published conventional data
    - must be ready by the end of the year!
  - view maintenance
  - database summarization
  - database publishing utility

Thank You!