

# Publishing Deep Web Geographic Data

Helena Piccinini<sup>1,2</sup>, Marco A. Casanova<sup>1</sup>, Luiz André P. P. Leme<sup>3</sup>,  
Antonio L. Furtado<sup>1</sup>

<sup>1</sup>Departamento de Informática – PUC-Rio, Rio de Janeiro, RJ – Brazil

<sup>2</sup>IBGE, Departamento de Informática, Rio de Janeiro, RJ – Brazil

<sup>3</sup>Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ – Brazil

hpiccinini@inf.puc-rio.br, casanova@inf.puc-rio.br, lapaesleme@ic.uff.br, furtado@inf.puc-rio.br

## Summary

This article introduces a design process, called W-RayS, to describe Deep Web geographic data and to publish the descriptions both on the Web of Data and on the Surface Web. The article also outlines a toolkit that supports the process and discusses an experiment in which the toolkit was used to publish data stored in a large map server. Briefly, to describe geographic data in vector format, the designer should first specify views over the underlying geographic database that capture the basic characteristics of the geographic objects and their topological relationships represented in the vector data. The same idea is applied to raster data, but using a gazetteer or any other geographic database that covers the same area as the raster data. Then, the designer should map the view definitions to an RDF schema, following the Linked Data principles. The descriptions of the geographic data are therefore formalized as sets of RDF triples synthesized from the conventional data. To publish geographic data descriptions on the Web of Data, the designer may decide to materialize the RDF triples and store them in a repository or create a SPARQL endpoint to access the triples on demand. To publish geographic data descriptions on the Surface Web, W-RayS offers the designer tools to transform the RDF triples to natural language sentences, organized as static Web pages with embedded RDFa. The inclusion of RDFa preserves the structure of the data and allows more specific queries, processed by engines that analyze Web pages with RDFa.

**Keywords:** Deep Web, Linked Data, Geographic Data, Natural Language Processing.

## 1 Introduction

Unlike the Surface Web of static pages, the Deep Web [1, 2] comprises data stored in databases, dynamic pages, scripted pages and multimedia data, among other types of objects. Deep Web databases are typically under-represented in search engines due to the technical challenges of locating, accessing, and indexing the databases. Indeed, since Deep Web data is not available as static Web pages, traditional search engines cannot discover data stored in the databases through the traversal of hyperlinks, but rather they have to interact with (potentially) complex query interfaces.

In particular, the Deep Web includes geographic data in vector format, usually made available as dynamic HTML pages [3, 4], and geographic data in raster format, including satellite images, described by their metadata. Hence, search engines do not properly index such data. Furthermore, Web-based Geographic Information Retrieval (GIR) systems [5, 6] do not address this limitation.

Two basic approaches to access Deep Web data have been proposed. The first approach, called *surfacing* or *Deep Web Crawl* [7-10], tries to automatically fill HTML forms to query the databases, execute offline queries and translate the results to static

Web pages, which are then indexed. The second approach, called *federated search* or *virtual integration* [11-15], suggests the use of domain-specific mediators to facilitate access to the databases. Hybrid strategies, which extend the previous approaches, have also been proposed [16].

Despite recent progresses, accessing Deep Web data is still a challenge, for several reasons. First, there is the question of scalability. Since the Deep Web is orders of magnitude larger than the Surface Web, it may not be feasible to completely index the Deep Web. Second, databases typically offer interfaces designed for human users, which complicates the development of software agents to interact with them. Third, multimedia data, including geographic data, have no straightforward description that can be indexed, apart from their metadata. Fourth, for security and performance reasons, many organizations do not allow their databases to be indexed without their consent.

In parallel with efforts to harness the Deep Web, Berners-Lee (2006) [17] proposed a set of best practices for publishing and connecting structured data on the Web, which came to be known as the ‘Linked Data Principles’. According to these principles, a dataset must comply with the following requirements: (i) be available on the Web; (ii) be available as machine-readable structured data; (iii) be in a non-proprietary format; (iv) use open standards from W3C (i.e. RDF and SPARQL); and (v) be linked to other datasets to provide additional data. We will use the term Web of Data<sup>1</sup> to refer to data published on the Web according to the Linked Data principles.

In the context of publishing data on the Web, we also mention RDFa (RDF-in-attributes) [18], a W3C Recommendation that adds a set of attribute-level extensions for embedding RDF within XHTML documents, thereby enabling the extraction of RDF triples by software agents. The option to publish Web pages with RDFa is justified by statistics released by Yahoo! showing that RDFa demonstrated a growth around 510% in 2010. This explosive growth is credited to the fact that, since 2009, Web search engines such as Google and Yahoo! started to process RDFa tags [19, 20].

In this article, we introduce the W-RayS design process to describe Deep Web geographic data and to publish the descriptions both on the Web of Data and on the Surface Web. More precisely, the W-RayS design process addresses three issues:

- (1) *“How to create descriptions of geographic data”*.
- (2) *“How to publish the descriptions on the Web according to the Linked Data principles”*.
- (3) *“How to publish the descriptions on the Web so that search engines can find the data”*.

We first discuss in detail how the W-RayS design process addresses these issues. Then, we outline a toolkit that supports the design process. Finally, we present a case study, developed to assess the usefulness of the process. The case study is based on real-world data maintained by the Brazilian Institute of Geography and Statistics (IBGE). A step-by-step guide indicating how to use the toolkit, together with other realistic examples, is available at the W-RayS Web site<sup>2</sup>.

Briefly, to describe geographic data in vector format, we adopt views, defined over the underlying geographic database, that capture the basic characteristics of the geographic objects and their topological relationships represented in the vector data.

---

<sup>1</sup> <http://www.w3.org/standards/semanticweb/data>

<sup>2</sup> [www.inf.puc-rio.br/~hpiccinini/](http://www.inf.puc-rio.br/~hpiccinini/)

Indeed, geographic data in vector format is often associated with conventional data that capture the properties of (geographic) objects. We therefore explore such conventional data to create descriptions of the vector data. We apply the same idea to raster data, but using a gazetteer or any other geographic database that covers the same area as the raster data. In their final format, we express the view definitions as an RDF schema, following the Linked Data principles. The descriptions of the geographic data are therefore formalized as sets of RDF triples synthesized from the conventional data stored in a database associated with the geographic data.

We note that this strategy encompasses more than just publishing conventional metadata. It involves creating meaningful descriptions for opaque (vector or raster) geographic data, a point we consider the first contribution of the article.

To publish geographic data descriptions on the Web of Data, the designer may decide to materialize the RDF triples and store them in a repository or create a SPARQL endpoint to access the triples on demand.

To publish geographic data descriptions on the Surface Web, W-RayS offers the designer tools to transform the RDF triples to natural language sentences, organized as static Web pages with embedded RDFa, also generated from the RDF triples. We observe that the use of natural language sentences is convenient for three reasons. First, they lead to Web pages that are acceptable to Web search engines that consider words randomly distributed in a Web page as an attempt to manipulate page rank. Second, they facilitate the task of more sophisticated engines that support semantic search based on natural language features [21]. Lastly, the descriptions thus generated are minimally acceptable to human users. The Web pages are then easily indexed by traditional Web search engines, as well as by engines that support RDFa-based semantic search.

In this aspect, W-RayS may be understood as an alternative to the surfacing approach, combined with a strategy to describe geographic data in a way that makes the data visible to search engines. It places on the (database) designer the responsibility for deciding which data should be exposed on the Web, and how it should be published, thereby relieving the search engines from probing the database through HTML forms. This paradigm shift represents a departure from the usual Deep Web surfacing approach. We consider it the second contribution of the article.

The remaining of the text is organized as follows. Section 2 summarizes related work. Section 3 discusses in detail how to describe Deep Web geographic data and to publish the descriptions both on the Web of Data and on the Surface Web. Section 4 outlines a toolkit that supports the design process. Section 5 describes an experiment carried out to evaluate the design process. Section 6 presents the conclusions.

## **2 Related Work**

The W-RayS approach involves three research topics – Deep Web, mapping relational databases to RDF, and natural language generation – separately discussed in this section.

### **2.1 Deep Web**

As mentioned in the introduction, two basic approaches to access Deep Web data have been proposed: *federated search* or *virtual integration* [11-15]; and *surfacing* or *Deep Web Crawl* [7, 8, 10].

Virtual integration is basically a data integration solution aiming at accessing Deep Web data. In other words, it is based on building mediation systems, potentially one for each information domain, such as used car sales or job offers. The mediated schema of each domain can be created manually or by means of the semi-automatic analysis of the data sources of the domain. Queries are executed on the mediated schema, reformulated and passed to each underlying data source. The results returned by each source are combined and presented to the user.

An example of a tool to help the virtual integration approach is the PruSM (Prudent Schema Matching) system [22], proposed for schema alignment, which combines multiple sources of information by similarity without the need for pre-processing forms.

This approach has the following problems: (1) there are millions of data sources belonging to countless domains on the Web; building and managing mappings on such scale would be an epic challenge, and should be done in over 100 languages; (2) the creation and maintenance of each mapping requires a large human involvement; (3) the response times are long.

The surfacing approach focuses on pre-computing the most relevant executions of HTML forms considered to be interesting. The HTML forms are automatically filled and the queries are executed offline. The results of these executions are translated into static HTML pages, which are then indexed by a search engine. This approach involves two main technical challenges: (1) defining which values must be selected to fill the text boxes in a form; and (2) defining the maximum number of input value combinations for forms with multiple inputs, so that the executed queries return a reasonable number of distinct and useful data—without blank responses—to be indexed.

Madhavan et al. (2008) [8] describe a solution implemented by Google. The authors attempt to tackle the first problem mentioned above by filling the keyword search field—which is provided in most HTML forms—with words that were previously classified according to the domain. The results returned are analyzed and new words are extracted, identified in a domain, and then different keywords are selected to be tested again. Regarding the second problem, the authors present an algorithm that is able to select, from the set formed by the Cartesian product of the candidate values to fill a form with multiple entries, only one subset that provides a reasonable number of distinct and useful results.

Despite the progress made by tools that follow the surfacing approach, they still suffer from the following limitations: (1) they are unable to cover all types of data that exist in the Deep Web, such as geographic data; (2) they lose the semantics of the data when publishing static HTML pages; (3) users cannot define which data in their databases should be indexed, which can lead, for security and performance reasons, to a complete denial-of-access to search engines. W-RayS addresses the first issue by creating descriptions for the geographic data from related conventional data. It solves the second issue by embedding RDFa in the Web pages, based on an RDF schema whose vocabulary is aligned with well-known vocabularies. It circumvents the third issue by placing on the (database) designer the responsibility of deciding which data should be exposed on the Web, and how it should be published, thereby relieving the search engines from probing the database through HTML forms.

## **2.2 Generation of Natural Language Sentences**

We first summarize three approaches involving the verbalization of structured data in OWL [23], which inspired some of the solutions we adopted in W-RayS.

ACE (Attempto Controlled English) [24] is a subset of English such that each sentence in the chosen subset is interpreted unambiguously, relating the sentence to a unique logical form. The intention behind the design of ACE is to offer the expressivity required for knowledge engineering tasks, but retaining a natural subset of English. The goal of an ACE View is to simplify viewing and editing expressive and syntactically complex OWL knowledge bases by making most of the interaction with the knowledge base happen via ACE. Its mapping is based on the OWL syntax, where properties are mapped to verbs and classes to proper nouns.

The Swoop approach [25] uses a Part-of-Speech Tagger<sup>3</sup> to automatically detect the linguistic categories of words and to generate the corresponding sentences in natural language. It also proposes a fixed set of expansion rules based on linguistic templates. Swoop is able to connect all datatype properties and object properties of a given subject in one sentence. While this fact reduces the number of clauses related to the same subject, it creates a very long sentence that is difficult to understand. To tackle this problem, the Swoop approach presents the sentence in the form of a numbered list of clauses nested under the subject of the sentence.

The NIBA (Natural Language Information Requirements Analysis) approach [26] for verbalizing OWL has the following goals: to develop a set of recommendations for standardizing labels in OWL, to filter linguistic standards, and to create a set of rules for the creation of sentences in natural language which make the OWL concepts explicit. Sentence generation is executed after the analysis and rule application steps. A set of rules with its own grammar is defined in Prolog using Definite Clause Grammar (DCG).

A serious problem faced by approaches for mapping structured data into natural language lies in balancing accuracy and readability. Because the ACE approach also works as a first-order language, its sentences have no ambiguities, and therefore it has limited malleability regarding human readability. Swoop offers more flexible solutions and ensures that the subject is always present even when all of the properties are aggregated into a single sentence, providing better readability for humans. Another significant issue is that these approaches consider only the English language, using its specificities to synthesize sentences.

Hollink et al. (2003) [27] show how structured sentences, supported by a controlled vocabulary, can be generated with the purpose of describing images. The authors present a tool that assists in the manual generation of image descriptions using, whenever possible, one or more interconnected ontologies. The content of the descriptions is structured. For example, when describing an art painting, the user employs a set of declarations of the form “*agent+action+object+recipient*”. Each declaration must have at least one agent and one object. The terms used in the sentences are selected from different thesauri or ontologies. Several clauses can be used to describe a single painting.

W-RayS adopts a solution inspired on these approaches. It synthesizes Web pages in much the same way as ACE and NIBA generate sentences, using the sentence simplification strategy suggested by Swoop. The W-RayS toolkit has native support for WordNet and some geospatial vocabularies, inspired on the tool described in [27].

---

<sup>3</sup> A Part-of-Speech Tagger marks (“tags”) a word in a text (within a corpus) with its corresponding part of speech, based on its definition and its relationship with adjacent and related words or phrases in a clause or paragraph. Part-of-speech is a linguistic category of words or lexical items that is usually defined by the syntactic or morphologic behavior of the lexical item in question. Common linguistic categories include nouns and verbs.

Furthermore, W-RayS includes RDFa in the XHTML markup, for the reasons already outlined in the introduction.

### 2.3 Mapping Relational Databases to RDF

The process of transforming a relational database (schema and data) to RDF triples (or simply *triples*) is known as *RDB-to-RDF* or *triplification* [28]. Currently there are several RDB-to-RDF mapping tools (see Sahoo et al. 2009 [29] for an early survey) and a standard mapping language, R2RML [30]. We briefly summarize some of the triplification tools in what follows.

Triplify [31] is able to recognize views of a relational database and to convert the results of these views to triples. Triplify uses the SQL language, which allows creating aliases to solve RDB-to-RDF mapping problems, as well as to align other existing ontologies.

The D2Rserver [32] has a declarative language for the automatic generation of the mapping files. This approach offers the possibility of generating either virtual or materialized triples. The user may edit the mapping file, if he decides to reuse ontologies in the mapping process or detects incorrect mappings.

The DB2OWL tool [33] is able to generate a class hierarchy when a table  $T$  is related to another table  $U$  and all primary keys in  $T$  are foreign keys pointing to  $U$  that have a referential integrity restriction. It also creates inverted object properties in the case of  $n$ - $m$  relationships.

RDBtoOnto [34] is a tool for designing mapping projects. Apart from the rules that usually exist in databases, it creates constraint rules, which can be suggested by the system or created and changed by the user to influence the mapping process. It is able to automatically categorize instances of database columns by means of an automatic learning method that takes advantage of the metadata in the database schema and the instances themselves.

Karma [35] is an example of a tool that combines virtual integration with triplification. Karma helps users to integrate data from several data sources according to an ontology and then publish the integrated data as RDF or store it in a triplestore.

We remark that well-engineered triplification tools should help publish triplesets according to the Linked Data principles [36]. In special, the tool should support the reuse of vocabularies that are popular for the application domain in question.

The W-RayS toolkit includes a triplification module that supports the definition of RDF schemas and RDB-to-RDF mappings. The module follows the Linked Data principles, in the sense that it helps the designer select vocabularies that are popular to describe geographic data. Furthermore, unlike typical triplification tools, the module also captures additional information to help generate the natural language templates, especially templates for reified relationships, as discussed in Section 3.2.1.

## 3 The W-RayS Design Process

In this section, we discuss in detail the three issues stated in the introduction: (1) How to create descriptions of geographic data; (2) How to publish the descriptions on the Web according to the Linked Data principles; (3) How to publish the descriptions on the Web so that search engines can find the data.

### 3.1 Creating Descriptions of Geographic Data

We first address how to create descriptions of geographic data in vector format and then expand the ideas to raster data.

Very briefly, we propose to describe geographic data stored in a database by sets of RDF triples, defined in two stages: (1) specification of conventional, alphanumeric views over the original database or another geographic database; (2) mapping of the views to an RDF schema, following the Linked Data principles. Descriptions of geographic data are then sets of triples synthesized from the underlying data, according to the view specifications and the mappings to RDF.

#### 3.1.1 Creating Descriptions of Geographic Data in Vector Format

We begin by observing that geographic data in vector format, stored in a geographic database, is commonly organized by layers and is associated with conventional data that capture the properties of the geographic objects depicted in the layer. Such conventional data is typically stored in the same database as the vector data.

The first step to create a description of the vector data is to specify, over the underlying database, one or more views that return conventional, alphanumeric data. We offer the following guidelines for this manual process:

- Views should combine a small number of layers that represent interrelated geographic objects.
- For each layer, the corresponding view should select the most relevant attributes of the geographic objects and include restrictions that filter out unimportant objects.
- When a view combines several layers, it should specify the priority between the layers and which topological relationships between the geographic objects of different layers should be materialized.

One last view should be defined to associate each instance  $V$  of vector data with a query  $Q_V$  to the underlying geographic database server where  $V$  is stored. The exact syntax of  $Q_V$  is entirely dependent on the server's interface. The description of  $V$  will be exposed on the Surface Web as a set of Web pages  $W_V$  (or on the Web of Data as a set of triples  $T_V$ ). Hence, a user will locate  $V$  with the help of  $W_V$  (or  $T_V$ ) and execute  $Q_V$  to effectively surface  $V$ . The user will have access to  $Q_V$  via a link in the Web pages (or as the value of a datatype value property), as explained in the examples in this section.

The second step is to map the conventional views to an RDF schema. The mapping to RDF should abide to the Linked Data principles. In particular, since view and attribute names are typically inappropriate to externalize data on the Web, the designer should align the view and attribute names to pre-existing vocabularies as much as possible. The tool described in Section 4 has built-in support to help the designer to align the vocabulary of an RDF schema with: the Dublin Core Metadata Element Set (with the extensions defined in the January 2008 version); WordNet, available through SUMO [37]; the ADL Feature Type Thesaurus in the RDF version [38]; the BuildingsAndPlaces and SpatialRelations [39], listed in the W3C Geospatial Ontologies Web site [40]; and the Geonames vocabulary.

We illustrate the view specification process with the help of the Mural Maps, used in the experiment discussed in Section 5. The Mural Maps are a collection of thematic

maps available at the IBGE Web site<sup>4</sup> in PDF format or through the ArcGIS Online server. Each map is represented in the geographic database as a different layer of vector data, together with a minimum set of metadata. For example, the political limits map comes with the names of the states, their acronyms and capital cities.

The experiment focused on the biome, political limits and vegetation maps. We first specified conventional views that describe the geographic objects represented in such maps, as shown in Figure 1 and Table 1. The last column of Table 1 indicates that the values of most attributes were taken from the metadata associated with the maps, the key values were defined as the polygon centroids of the geographic objects and the values of the *percentageOfStateArea* attribute of the *coverage* view were computed with the help of a plugin that determines the area of the intersection of two polygons. Furthermore, the *occupiedBy* and *coverage* relationships were computed based on the fact that all maps have the same projection and use the same cartographic base.

We then mapped the conventional views to the RDF schema<sup>5</sup> summarized in Table 2. The mapping of the views and their attributes to RDF classes and properties is one-to-one in most cases, which is indicated by using the same names both in the views and in the RDF schema. However, since the binary relationship *coverage* has an attribute, *percentageOfStateArea*, we reified it [41] with the help of a new class, *Coverage*, and two object properties, *encloses* and *locatedIn*. We also aligned some of the terms of the RDF schema with terms of other vocabularies as follows: *stateName* with *officialName* of Geonames; *biomeName* with *name* of Geonames; *encloses* with *encloses* of SpatialRelations; and *locatedIn* with *locatedIn* of SpatialRelations.

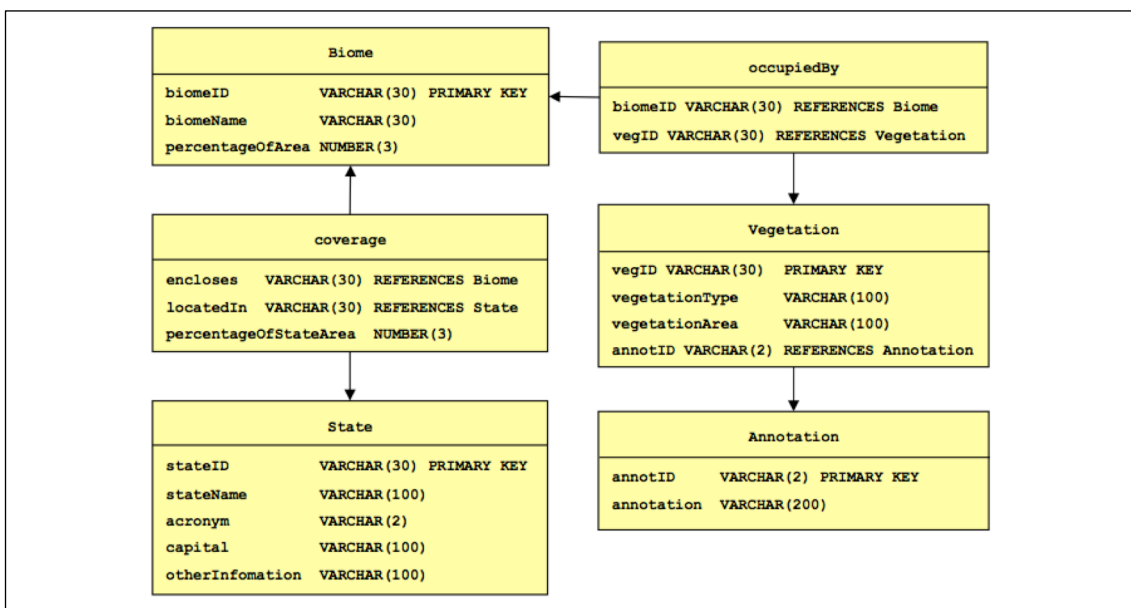


Figure 1 - Conventional views for the *Mural Maps of Brazil*.

<sup>4</sup> <http://mapas.ibge.gov.br>

<sup>5</sup> The definition of the RDF schema is available at [www.inf.puc-rio.br/~hpiccinini/wray/biome.owl](http://www.inf.puc-rio.br/~hpiccinini/wray/biome.owl)

**Table 1** – Description of the conventional views for the *Mural Maps of Brazil*.

View Name	Attribute Name	Description	Attribute Origin
<i>Biome</i>		Biomes located in Brazil	
	<i>biomeID</i>	Primary key	Polygon centroid
	<i>biomeName</i>	Biome name	Map metadata
	<i>percentageOfArea</i>	% of Brazilian territory covered by the biome	Map metadata
<i>State</i>		States of Brazil	
	<i>stateID</i>	Primary key	Polygon centroid
	<i>stateName</i>	State name	Map metadata
	<i>acronym</i>	Official acronym of the state	Map metadata
	<i>capital</i>	Capital city of the state	Map metadata
	<i>otherInformation</i>	Link to information about the state	Map metadata
<i>Vegetation</i>		Brazilian vegetation	
	<i>vegID</i>	Primary key	Polygon centroid
	<i>vegetationType</i>	Vegetation type	Map metadata
	<i>vegetationArea</i>	Area of the Brazilian territory covered by the vegetation type	Map metadata
	<i>annotID</i>	Foreign key to <i>Annotation</i>	(foreign key)
<i>Annotation</i>		Annotations about vegetation	
	<i>annotID</i>	Primary key	(auto number)
	<i>annotation</i>	Text annotation	Map metadata
<i>occupiedBy</i>		Relationship between <i>Biome</i> and <i>Vegetation</i>	
	<i>biomeID</i>	Foreign key to <i>Biome</i>	(foreign key)
	<i>vegID</i>	Foreign key to <i>Vegetation</i>	(foreign key)
<i>coverage</i>		Relationship between <i>Biome</i> and <i>State</i>	
	<i>encloses</i>	Foreign key to <i>Biome</i>	(foreign key)
	<i>locatedIn</i>	Foreign key to <i>State</i>	(foreign key)
	<i>percentageOfStateArea</i>	% of the state territory covered by the biome	(computed value)

**Table 2** – RDF schema for the *Mural Maps of Brazil*.

Type	Name	Domain	Range
Class	<i>Biome</i>		
	<i>State</i>		
	<i>Vegetation</i>		
	<i>Coverage</i>		
Inverse functional datatype property	<i>biomeID</i>	<i>Biome</i>	
	<i>biomeName</i>	<i>Biome</i>	
	<i>stateID</i>	<i>State</i>	
	<i>stateName</i>	<i>State</i>	
	<i>acronym</i>	<i>State</i>	
	<i>vegID</i>	<i>Vegetation</i>	
	<i>vegetationType</i>	<i>Vegetation</i>	
Functional datatype property	<i>percentageOfArea</i>	<i>Biome</i>	
	<i>capital</i>	<i>State</i>	
	<i>otherInformation</i>	<i>State</i>	
	<i>vegetationArea</i>	<i>Vegetation</i>	
	<i>percentageOfStateArea</i>	<i>Coverage</i>	
Multivalued datatype property	<i>annotation</i>	<i>Vegetation</i>	
Object property	<i>encloses</i>	<i>Biome</i>	<i>Coverage</i>
	<i>locatedIn</i>	<i>Coverage</i>	<i>State</i>
	<i>occupiedBy</i>	<i>Biome</i>	<i>Vegetation</i>

Using the view specifications and the mapping to RDF outlined in Table 1 and Table 2, we generated the RDF triples that describe the Mural Maps. As an example, Figure 2 shows the RDF triples that describe the “Caatinga” biome in the State of Alagoas, where:

- Lines 1, 4 and 7 define individuals of the *Biome*, *Coverage* and *State* classes, respectively.
- Lines 2, 5 and 8 indicate the values of properties *name*, *percentageOfStateArea* and *officialName* respectively for the individuals defined in Lines 1, 4 and 7.
- Line 3 indicates that property *encloses* relates the individual of the *Biome* class defined in Line 1 to the individual of the *Coverage* class defined in Line 4.
- Line 6 indicates that property *locatedIn* relates the individual of the *Coverage* class defined in Line 4 to the individual of the *State* class defined in Line 7.
- Line 13 indicates that property *represents* (of the ontology *XRAY*, not shown in Table 2) associates the individual of the *Biome* class defined in Line 1 with a query that retrieves it from the ArcGIS Online server.
- Line 14 schematically indicates the query.

We stress that the RDF triples are not meant to be read by human users, but rather by software agents. They may be stored in a repository or exposed as a SPARQL endpoint. Alternatively, they may be used to generate static Web pages containing readable descriptions (with embedded RDFa) of the Mural Maps, as discussed in Section 3.2.

```

1. <biome:Biome rdf:about="http://www.inf.puc-rio.br/~hpiccinini/wray/biome.owl#-39.84049008,-9.44877857">
2.   <geonames:name xml:lang="pt">Caatinga</geonames:name>
3.   <biome:encloses>
4.     <biome:Coverage rdf:about="http://www.inf.puc-rio.br/~hpiccinini/wray/biome.owl#-39.84049008,-
36.71609664,-9.44877857,-9.64509767">
5.       <biome:percentageOfStateArea
   _rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">48</biome:percentageOfStateArea>
6.       <biome:locatedIn>
7.         <bcim:State rdf:about="http://www.inf.puc-rio.br/~hpiccinini/wray/biome.owl#-36.71609664,-
9.64509767">
8.           <geonames:officialName xml:lang="pt">Alagoas</geonames:officialName>
9.           </bcim:State>
10.        </biome:locatedIn>
11.       </biome:Coverage>
12.     </biome:encloses>
13.   <xray:represents>
14.     (A query to the ArcGIS Online server to retrieve the individual being described goes in here)
15.   </xray:represents>
16. </biome:Biome>

```

**Figure 2** – Sample RDF triples describing the “caatinga” biome in the State of Alagoas.

### 3.1.2 Creating Descriptions of Geographic Data in Raster Format

To describe raster data, the designer should select geographic objects contained within the bounding box of the raster data and use them to generate a description of the raster data, much in the same way as for vector data. The geographic objects might be obtained, for instance, from a gazetteer or a geographic database that covers the same area as the raster data.



**Figure 3** – An image fragment of the State of Rio de Janeiro (Source: INPE).

As a concrete example, consider the image of the State of Rio de Janeiro shown in Figure 3 and define a conventional view over the Geonames gazetteer [42] such that a geographic feature  $F$  is in the view if and only if  $F$  has *Feature Code* equal to “Lake” or “Stream” and the centroid of  $F$  is inside the bounding box of the image, defined by:

$$((-22.15320, -44.17330), (-24.07070, -42.71030))$$

Examples of geographic features that satisfy the view restrictions are:

- “Lagoa Rodrigo de Freitas”, marked (a) in Figure 3, of the category “Lake” (or *H.LGN*).
- “Rio Comprido” and “Rio Maracanã”, respectively marked (b) and (c) in Figure 3, of the category “Stream”(or *H.STM*).

As for vector data, the view definition is then mapped to an RDF schema<sup>6</sup>. Finally, Figure 4 shows a sample set of RDF triples describing the image, where:

- The image is an individual of the class *SatelliteImage*, of the Image vocabulary created for this experiment, with ID “L7ETM21707820030220”.
- The geographic features are individuals of the class *Feature* of Geonames.
- The line before the last indicates that property *represents* (of the ontology *XRAY*) associates the image with a query that retrieves it from the image server.

<sup>6</sup> The complete RDF schema is available at [www.inf.puc-rio.br/~hpiccinini/wray/image.owl](http://www.inf.puc-rio.br/~hpiccinini/wray/image.owl)

```

<image:SatelliteImage xml:lang="pt"
  rdf:about="http://www.inf.puc-rio.br/~hpiccinini/wray/image.owl#L7ETM21707820030220">
  <image:contains>
    <geonames:Feature rdf:about="http://sws.geonames.org/3450970/about.rdf">
      <geonames:featureCode rdf:resource="http://www.geonames.org/ontology#H.LGN"/>
      <geonames:name xml:lang="pt">Lagoa Rodrigo de Freitas</geonames:name>
    </geonames:Feature>
  </image:contains>
  <image:contains>
    <geonames:Feature rdf:about="http://sws.geonames.org/3465821/about.rdf">
      <geonames:featureCode rdf:resource="http://www.geonames.org/ontology#H.STM"/>
      <geonames:name xml:lang="pt">Rio Comprido</geonames:name>
    </geonames:Feature>
  </image:contains>
  <image:contains>
    <geonames:Feature rdf:about="http://sws.geonames.org/3457855/about.rdf">
      <geonames:featureCode rdf:resource="http://www.geonames.org/ontology#H.STM"/>
      <geonames:name xml:lang="pt">Rio Maracanã</geonames:name>
    </geonames:Feature>
  </image:contains>
  <xray:represents>(A query to the image, stored in an image server, goes in here</xray:represents>
</image:SatelliteImage>

```

**Figure 4** – Sample RDF triples describing the image of the City of Rio de Janeiro.

## 3.2 Publishing Descriptions of Geographic Data

Recall that the descriptions of geographic data are sets of RDF triples. To publish geographic data descriptions on the Web of Data, the designer may decide to materialize the RDF triples and store them in a repository or to create a SPARQL endpoint to access the triples on demand. This is straightforward use of standard technology and will not be further discussed in this paper.

To publish geographic data descriptions on the Surface Web, W-RayS offers the designer tools to transform RDF triples to natural language sentences, organized as static Web pages with embedded RDFa, generated from the RDF triples. W-RayS also helps include links, in the static Web pages, that express the queries that retrieve the geographic data (in vector or raster format) from the underlying database. Such links effectively connect the Surface Web to the Deep Web and are synthesized from the values of the *xray:represents* property, explained in the examples of Sections 3.1.1 and 3.1.2.

In this section, we explore the second alternative in more detail.

### 3.2.1 Transforming RDF Triples into Sentences

A set of RDF triples is transformed into static Web pages with embedded RDFa with the help of (natural language) templates, created from the RDF schema. Briefly, nouns, verbs, adjectives or adverbs in the templates correspond to the classes and properties of the RDF schema.

The templates are categorized as follows: a *simple template* corresponds to a functional datatype property; a *multivalued template* to a multivalued datatype property or to an object property derived from an 1-n relationship; a *binary relationship template* to an object property derived from a n-m relationship; and a *reification template* to a

class and object properties introduced by the reification [41] of an n-ary relationship or the reification of a relationship with attributes.

Finally, in the templates, geographic objects are identified by externally recognized names, rather than artificially generated key values. For example, returning to Table 2, *biomeName* values are adopted, rather than *biomeID* values.

We illustrate how we generate templates using the following notational conventions:

- **boldface**: property values
- **<boldface within angle brackets>**: a placeholder for property values
- *Italics*: classes and properties
- plain text: additional words that the tool automatically inserts

Consider again the RDF schema of Table 2. The first example refers to the simple template associated with the functional datatype property *vegetationArea*:

**<vegetationType>** is a *vegetation* that has **<vegetationArea>** as *vegetation area*.

Note that the template uses the value of the inverse functional datatype property *vegetationType*, rather than *vegID*, since the former has an externally recognized semantics. A sample sentence generated from the above template is:

**Savana Estépica** is a *vegetation* that has **Região Fitoeológica** as *vegetation area*.

Multivalued templates are more elaborate because all values of a multivalued datatype property should be represented in the same sentence. The second example shows a template associated with the multivalued datatype property *annotation*:

**<vegetationType>** is a *vegetation* that has *annotations*: **<annotations>**,... and **<annotations>**.

A sample sentence generated from the above template is:

**Savana Estépica** is a *vegetation* that has *annotations*: **Caatinga do Sertão Árido, Campos de Roraima, Chaco Sul-Matogrossense** and **Parque do Espinho da Barra do Rio Guaraí**.

The designer may combine templates, reorder the template components and alter the default property names to improve readability, resulting in a new template:

The **<vegetationType>** is a *vegetation* that has **<vegetationArea>** as *vegetation area and annotations*: **<annotations>**,... and **<annotations>**.

A sample sentence generated from the new template is:

The **Savana Estépica** is a *vegetation* that has **Região Fitoeológica** as *vegetation area and annotations*: **Caatinga do Sertão Árido, Campos de Roraima, Chaco Sul-Matogrossense** and **Parque do Espinho da Barra do Rio Guaraí**.

The next example depicts the binary relationship template associated with the object property *occupiedBy*:

**<biomeName>** is a *biome* that is *occupied by* the *vegetation type* **<vegetationType>**.

A sample sentence generated from the above template is:

**Amazonas** is a *biome* that is *occupied by* the *vegetation type* **Floresta Estacional Decidual**.

The final example shows the reification template associated with *Coverage*, *encloses* and *locatedIn*:

**<biomeName>** is a *biome* that *encloses* **<percentOfTheArea>** percent of the area located in the state of **<stateName>**.

A sample sentence generated from the above template is:

**Caatinga** is a *biome* that *encloses* **48** percent of the area located in the state of **Alagoas**.

### 3.2.2 Transforming Sentences into Web pages

The design of the Web pages containing the sentences synthesized from the RDF triples is based on the W3C recommendations [43], Google recommendations [19], linked data recommendations [36] and the following guidelines:

- The Web site should describe the ontology used to publish the data; each Web page should contain hyperlinks from the terms present in the sentences to the corresponding ontology concepts.
- Sentences generated from the same view should be grouped into one or more Web pages, hyperlinked from the home page and among themselves; data coming from domains defined by enumeration should be hyperlinked to the corresponding terms of the external vocabularies.
- Sentences generated from reification are grouped into one Web page that points to other Web pages containing the descriptions formed by their respective relationships.
- The sentences should be linked to the queries that return the geographic data described from the underlying geographic database.

Figure 5 illustrates how Web pages are interlinked: (1) the Home page points to the other Web pages; (2) a Web page with sentences generated from an n-ary relationship template; (3) a Web page with sentences generated from a simple and a multivalued template; (4) *Deep Web* data accessed via queries over the geographic database.

Finally, we observe that the Web pages might have embedded RDFa, generated with the help of the RDF schema. For example, consider again the following sentence:

**Caatinga** is a *biome* that *covers* **48** percent of the area located in the State of **Alagoas**.

Figure 6 shows the source code corresponding to this sentence (compare it with example in Figure 2 and the explanation thereof). The following comments apply:

- Lines 1, 8 and 14 define individuals of the *Biome*, *Coverage* and *State* classes, respectively.
- Line 2 defines a link to the Web page that contains descriptions of the biome; this Web page in turn has a link expressing a query to the map stored in the ArcGIS Online server; the link connects the Surface Web with the Deep Web.
- Lines 3, 9 and 15 indicate the values of properties *name*, *percentageOfStateArea* and *officialName* respectively for the individuals defined in Lines 1, 8 and 14.
- Line 7 indicates that property *encloses* relates the individual of the *Biome* class defined in Line 1 to the individual of the *Coverage* class defined in Line 8.
- Line 13 indicates that property *locatedIn* relates the individual of the *Coverage* class defined in Line 8 to the individual of the *State* class defined in Line 14.

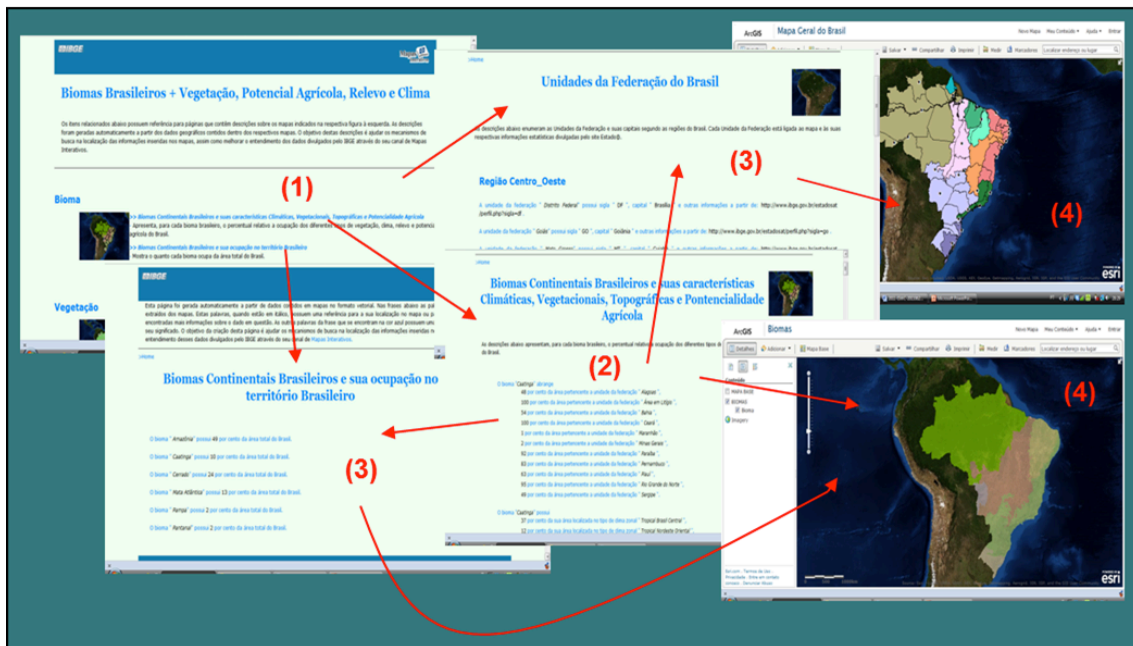


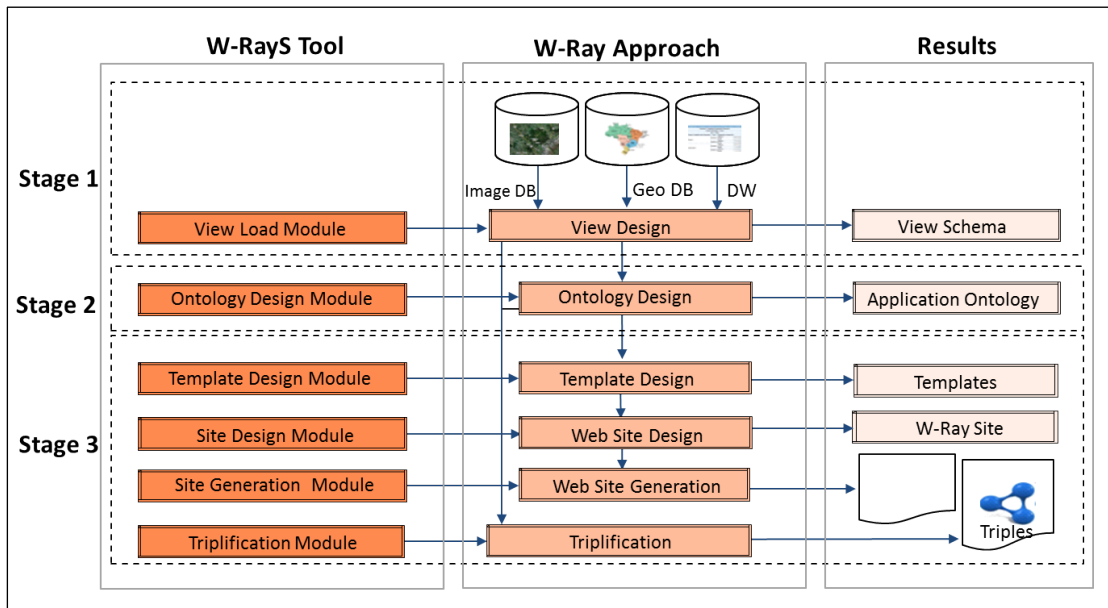
Figure 5 – Sample hyperlinked Web pages.

```

1. <div about="http://www.inf.puc-rio.br/~hpiccinini/wray/biome.owl#-39.84049008,-9.44877857"
   typeof="biome:Biome">
2.   <i><ahref="bioma.html#a-39.84049008-9.44877857">Caatinga</a></i>
3.   <span property="geonames:name" content="Caatinga" >
4.     is a <a href=#biome">biome</a>
5.   </span>
6.   that <a href=#covers">covers</a>
7.   <dl rel="biome:encloses">
8.     <dd about="http://www.inf.puc-rio.br/~hpiccinini/wray/biome.owl#
       -39.84049008,-36.71609664,-9.44877857,-9.64509767" typeof="biome:Coverage">
9.       <span property="biome:percentageOfStateArea" datatype="xsd:decimal">48</span>
10.      <a href=#percent_of_the_area">percent of the area</a>
11.      <a href=#located_in">located in</a>
12.      <a href=#state_of_the">state of the</a>
13.      <i>
14.        <span rel="biome:locatedIn">
15.          <span about="http://www.inf.puc-rio.br/~hpiccinini/wray/biome.owl#-36.71609664,-9.64509767"
16.            typeof="bcim:State">
17.              <span property="geonames:officialName" content="Alagoas">
18.                <a href="uf.html#a-36.71609664-9.64509767">Alagoas</a> (16)
19.              </span>
20.            </span>
21.          </span>
22.        </i>
23.      </dd>
24.    ...

```

Figure 6 – Example of the RDFa generated.



**Figure 7** – Overview of the W-RayS Design process, toolkit and data output.

#### 4. The W-RayS Toolkit

Sections 3.1 and 3.2 covered the central strategies of the W-RayS design process to address the problem of describing Deep Web geographic data and publishing the descriptions both on the Web of Data and on the Surface Web. In this section, we very briefly enumerate the modules of a toolkit implemented to support the process. A step-by-step description of how to use the toolkit is available at the W-RayS Web site.

As summarized in Figure 7, the main modules of the toolkit are:

*View Load Module:* handles the database views that specify which data should be published.

*Ontology Design Module:* helps the designer map the database view definitions to an RDF schema, following the Linked Data principles.

*Template Design Module:* synthesizes templates to generate natural language sentences that describe the materialized data of an RDF schema.

*Site Design Module:* helps the designer organize the sentences as static HTML pages with embedded RDFa.

*Site Generation Module:* publishes the static HTML pages with embedded RDFa.

*Triplification Module:* materializes the triples of an RDF schema or creates a SPARQL endpoint to access the triples on demand.

We stress that the Web Site Publication and the Triplification stages are alternatives to each other, but they are not mutually exclusive, in the sense that the designer may choose to publish a Web site describing the geographic data, or to triplify the view data, or both.

We observe that the View Design stage is manual and the designer should use DBMS tools to browse the database schema and define the views. However, it requires no more user intervention than similar triplification tools (see Section 2.3). Furthermore, differently from some triplification tools, W-RayS does not require specific knowledge

of RDF. The other W-RayS modules do not require any user intervention. However, we remark that, to improve the readability of the sentences, the designer may choose to calibrate the templates before generating Web pages.

Finally, we observe that Web Site Generation module uses the RDFa Developer 1.1.1, an add-on that can be easily used with the Firefox browser. It allows identifying all RDFa triples embedded in a Web page and running SPARQL queries on the RDFa content.

## 5 A Case Study

The W-RayS Web site describes experiments conducted to publish descriptions of the Aggregate Database System (SIDRA) [44, 45], the Continuous Vectorial Cartographic Base of Brazil (BCIM) and the Murals Maps of Brazil, available at the Web site of the Brazilian Institute of Geography and Statistics (IBGE)<sup>7</sup> as Deep Web opaque databases. The W-RayS Web site also contains an experiment to publish descriptions of satellite images available at the Web site of the Brazilian Institute for Space Research (INPE).

In this section, we describe an experiment to partly publish the Mural Maps of Brazil on the Surface Web and discuss the results obtained.

### 5.1 Setup of the Mural Maps Experiment

As mentioned in Section 3.1.1, the Mural Maps are a pedagogic collection available at the IBGE Web site in PDF format or through the ArcGIS Online map server. The collection includes thematic maps about biome, vegetation, relief, climate and agricultural potential, among others. IBGE does not publish summaries of the maps on the Web, except for certain manually defined metadata related to the major products. However, this manual procedure demands considerable time and does not cover the vast majority of the available data.

We therefore proceeded to apply the W-RayS design process to generate descriptions of some of the Mural Maps and to publish the descriptions on the Surface Web. The experiment permitted us to test: (1) the scalability of the W-RayS toolkit; (2) the sentence generation process for relationships with attributes; (3) the readability of long sentences; (4) the publication of sentences based on different templates in the same Web page; and (5) the benefits of surfacing Deep Web data using W-RayS.

The last point is the key issue. For example, before publishing the W-RayS Web pages, when the user submitted to Google the following keyword search:

*tipos +vegetação + bioma + caatinga*

(In English: *types + vegetation + biome + caatinga*)

Google returned hits to the IBGE Web pages that contained descriptions of the biomes and vegetation maps in PDF format. Interactive maps available through the ArcGIS Online map server would not be returned. This is obviously because conventional Web crawlers do not index data in vector format.

On the other hand, with the W-RayS Web pages, for the same keyword search, Google returned a hit to the Web page

<http://tomcat.inf.puc-rio.br:8080/muralmaps/vegetacao.html>

which points to the vegetation map available at the map server.

---

<sup>7</sup> <http://mapas.ibge.gov.br>

The key point of the experiment was to measure how many more hits to the Mural Maps could we induce by publishing W-RayS Web pages.

There were essentially two alternatives to proceed. First, we could have copied some of the Mural Maps to a separate map server, publish Web pages with metadata about the Mural Maps (as in the IBGE Web site) and publish W-RayS Web pages pointing to this map server. Then, we would measure the number of hits to the Mural Maps coming from the metadata Web pages and from the W-RayS Web pages. However, this experiment would be biased in favor of the W-RayS Web pages, since the map server (and the associated metadata) would be completely unknown to the users and would not be properly indexed by the search engines.

We opted for a more neutral, albeit complex experimental setup. We added W-RayS Web pages to the IBGE Web portal and had the Web pages point to the IBGE ArcGIS Online map server. Furthermore, we instrumented the IBGE Web site to count the number of hits to the Mural Maps coming from users that directly accessed the IBGE Web site and from users that reached the Mural Maps through the W-RayS Web pages.

This more complex setup accounts for the fact that, over the years, a few thematic maps in the Mural Maps collection became fairly popular and gained a large population of regular users. The experiment described in this section was defined to answer the question of how many more users, if any, could the W-RayS Web pages attract.

## 5.2 Evaluation of the Mural Maps Experiment

To evaluate the effectiveness of the W-RayS Web pages describing Mural Maps, we proceeded as follows. We created a log registering each click on a hyperlink in the W-RayS Web pages leading to a mural map on the IBGE map server. We then periodically analyzed this log, comparing our results with the total number of accesses to the Mural Maps in the IBGE map server (which included the accesses originating in the W-RayS Web pages).

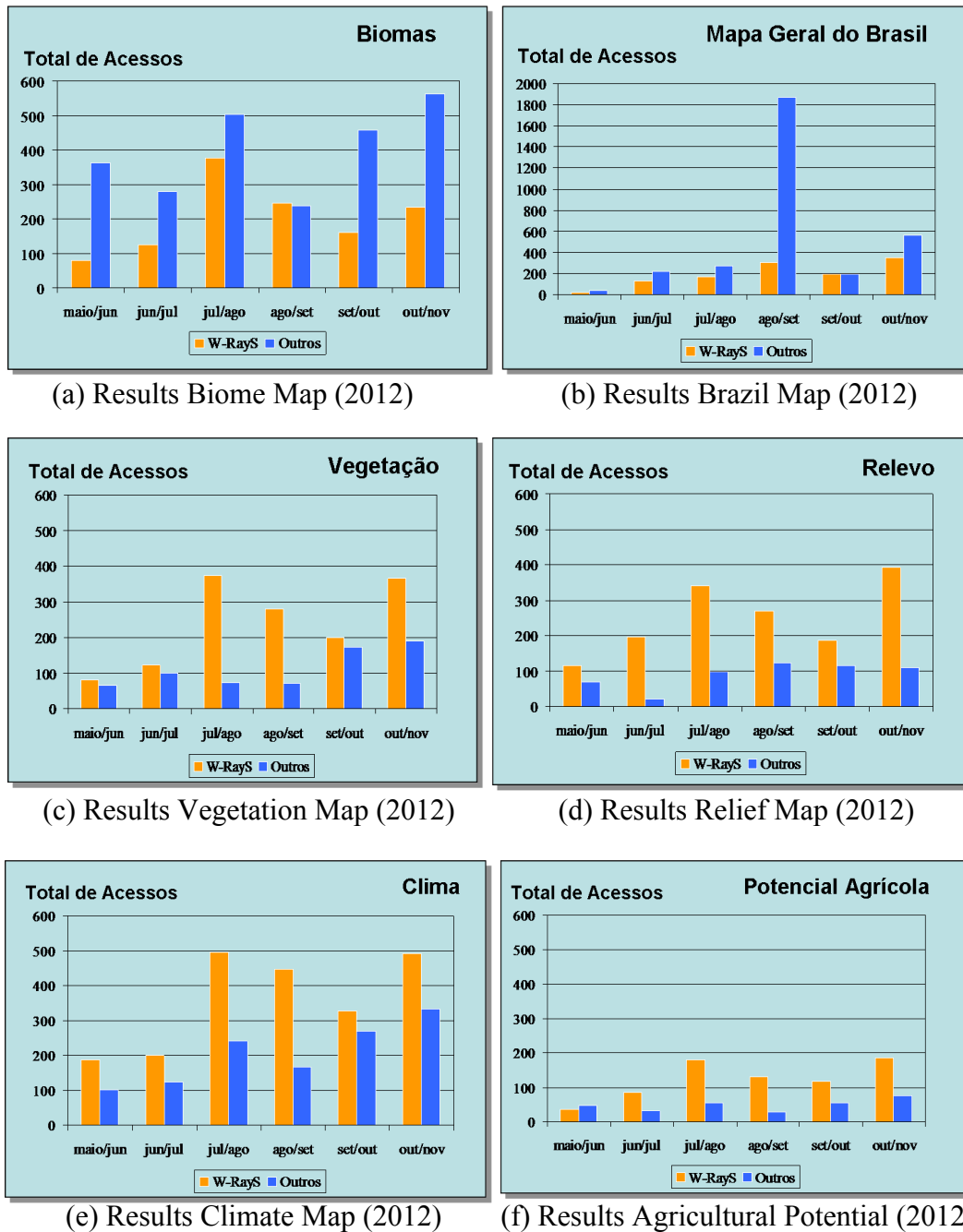
Figure 8 shows, for each mural map, the number of accesses during six months, where: the blue (or dark grey) bars show the number of *direct accesses*, that is, accesses which did not originate from the W-RayS Web pages; the yellow (or light grey) bars show the number of *W-RayS mediated accesses*, that is, accesses which originated from the W-RayS Web pages.

Figure 8(c) to (f) show that the number of W-RayS mediated accesses to the Vegetation, Relief, Climate, and Agriculture Potential maps surpasses the number of direct accesses. Intuitively, the W-RayS Web pages increased the visibility of these maps on the Web, as expected.

By contrast, Figure 8(a) and (b) indicate the opposite for the Biome and General Maps. A possible explanation would be that these two themes – biomes and general maps – are much more popular, which increases the chances that experienced or professional users know that IBGE offers these maps. They are, most likely, the users that access such maps directly from the IBGE map server.

During our first tests involving keyword searches, we observed that, for simple keyword searches (for example, “*Brazil+Biome*” or “*Brazil+Vegetation*”), the ranking of the W-RayS Web pages in the result was not particularly good. A plausible explanation is that there are fairly popular (and generic) Web sites about biomes, vegetation, etc... However, when we included “IBGE” to the keyword list (for example, “*IBGE+Biome*” or “*IBGE+Vegetation*”), the ranking of the W-RayS Web pages in the

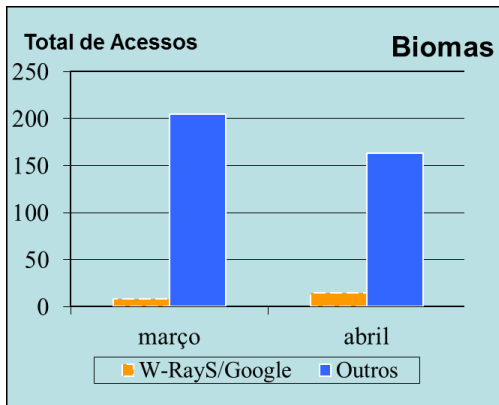
result was excellent. Indeed, since IBGE is the official institution responsible for producing maps in Brazil, users commonly include this acronym among the keywords.



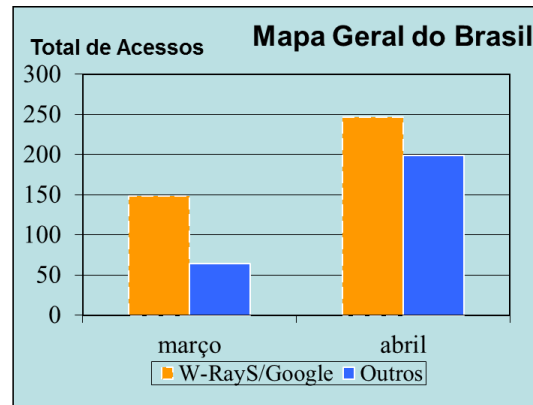
**Figure 8** – Number of accesses during six months in 2012 for each mural map.

Finally, on March and April 2013, we conducted a second experiment that captured keyword searches submitted through Google and automatically redirected the user to the maps. That is, the W-RayS Web pages had the specific purpose of surfacing the geographic data to attract the crawlers, but they were not visible to the users. Figure 9 shows the number of accesses in this case, where: the blue (dark grey) bars show the

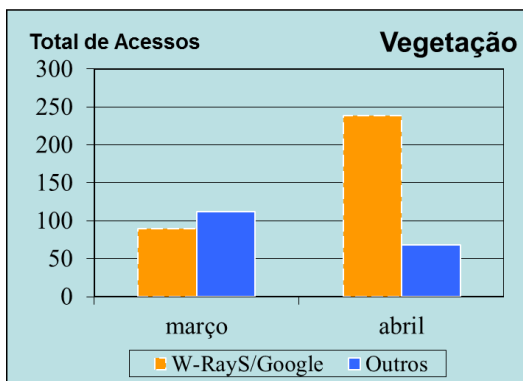
number of accesses directly from the IBGE Web site; the yellow (light grey) bars show the number of accesses which originated from the W-RayS Web pages. The results again show an increase in the number of accesses to the Vegetation, Relief, Climate and Agriculture Potential.



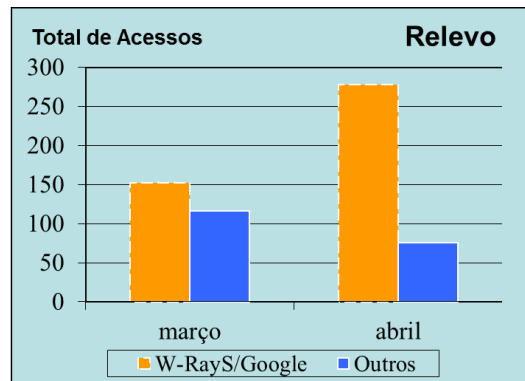
(a) Results Biome Map (2013)



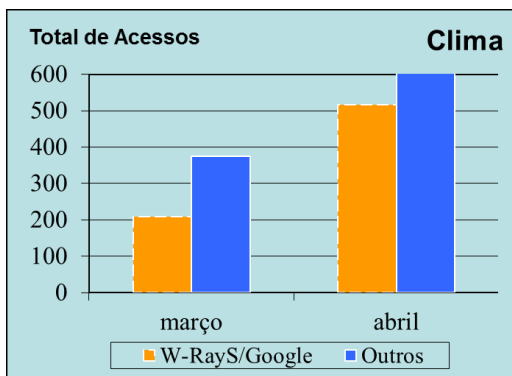
(b) Results Brazil Map (2013)



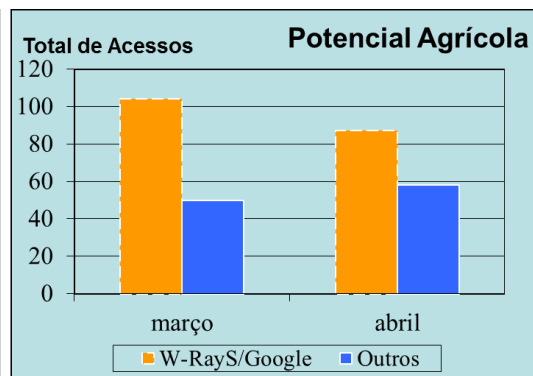
(c) Results Vegetation Map (2013)



(d) Results Relief Map (2013)



(e) Results Climate Map (2013)



(f) Results Agricultural Potential (2013)

**Figure 9** – Number of accesses during two months in 2013 for each mural map (second experiment).

## 6 Conclusions

In this article, we introduced the W-RayS design process and described in detail a use case to illustrate the process. A complete description of the toolkit that supports the process, as well as of other use cases, can be found at the W-RayS Web site.

The main contributions of this article are a sound approach to create descriptions, expressed as sets of RDF triples, of geographic data in vector or raster format, and a strategy to publish the descriptions on the Web of Data and on the Surface Web. In particular, the strategy to surface Deep Web geographic data involves synthesizing natural language sentences, organized as Web pages with embedded RDFa. The inclusion of RDFa exposes the structure of the data and allows more specific queries, processed by engines that analyze Web pages with RDFa.

As future work, we plan to further automate the natural language generation process, while maintaining readability. We are currently working to factor out the Template Design and the Web Site Design modules, so that they can be coupled with triplification tools to publish sets of RDF triples as Web pages with embedded RDFa.

**Acknowledgments.** This work was partly supported by IBGE, for H. Piccinini, and by CNPq under grants 301497/2006-0, 475717/2011-2, and CAPES/PROCAD NF 21/2009, for M.A. Casanova and A.L. Furtado.

## References

1. Bergman, M. K. (2001). The Deep Web: Surfacing Hidden Value. *The Journal of Electronic Publishing*, 7(1), 07–01. doi:10.3998/3336451.0007.104
2. Madhavan, J., Afanasiev, L., Antova, L., & Halevy, A. (2009). Harnessing the Deep Web: Present and Future (Vol. cs.DB). Presented at the Fourth Biennial Conference on Innovative Data Systems Research.
3. Esri. (2013). ArcGIS Online. *esri.com*. Retrieved November 3, 2013, from <http://www.esri.com/software/arcgis/arcgisonline>
4. MapServer open source web mapping. (2013). MapServer open source web mapping. *mapserver.org*. Retrieved November 3, 2013, from <http://mapserver.org>
5. Martins, B., Silva, M. J., & Chaves, M. (2007). *O sistema CaGE no HAREM-reconhecimento de entidades geográficas em textos em língua portuguesa*. Linguateca.
6. Szekely, P., Knoblock, C. A., Gupta, S., Taheriyani, M., & Wu, B. (2011). Exploiting semantics of web services for geospatial data fusion (pp. 32–39). Presented at the Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Semantics and Ontologies, ACM Press. doi:10.1145/2068976.2068981
7. Cafarella, M. J., Halevy, A., & Madhavan, J. (2011). Structured data on the web. *Communications of the ACM*, 54(2), 72–79. doi:10.1145/1897816.1897839
8. Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., & Halevy, A. (2008). Google's Deep Web crawl (Vol. 1, pp. 1241–1252). Presented at the Proceedings of the VLDB Endowment, VLDB Endowment.
9. Maiti, A., Dasgupta, A., Zhang, N., & Das, G. (2009). HDSampler: revealing data behind web form interfaces (pp. 1131–1134). Presented at the Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, ACM.

doi:10.1145/1559845.1560001

10. Raghavan, S., & Garcia-Molina, H. (2001). Crawling the Hidden Web - Stanford InfoLab Publication Server. Presented at the Proceedings of the 27th International Conference on Very Large Data Bases. Retrieved from <http://ilpubs.stanford.edu:8090/725/>
11. Callan, J. (2002). Distributed Information Retrieval. In *Advances in Information Retrieval* (Vol. 7, pp. 127–150). Springer. doi:10.1007/0-306-47019-5\_5
12. Cafarella, M. J., Halevy, A., & Khoussainova, N. (2009). Data integration for the relational web. *Proceedings of the VLDB Endowment (PVLDB)*, 2(1), 1090–1101.
13. He, B., Zhang, Z., & Chang, K. C.-C. (2005). MetaQuerier: querying structured web sources on-the-fly (pp. 927–929). Presented at the Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, ACM Request Permissions. doi:10.1145/1066157.1066291
14. He, H., Meng, W., Yu, C., & Wu, Z. (2005). WISE-Integrator: a system for extracting and integrating complex web search interfaces of the deep web (pp. 1314–1317). Presented at the Proceedings of the 31st International Conference on Very Large Data Bases.
15. Kabisch, T., Dragut, E. C., Yu, C., & Leser, U. (2010). Deep web integration with VisQI (Vol. 3, pp. 1613–1616). Presented at the Proceedings of the VLDB Endowment (PVLDB).
16. Rajaraman, A. (2009). Kosmix: high-performance topic exploration using the deep web (Vol. 2, pp. 1524–1529). Presented at the Proceedings of the VLDB Endowment (PVLDB).
17. Berners-Lee, T. (2006, July 27). Linked Data - Design Issues. *w3.org*. W3C. Retrieved October 31, 2012, from <http://www.w3.org/DesignIssues/LinkedData.html>
18. Herman, I., Adida, B., Sporny, M., & Birbeck, M. (2012). *RDFa 1.1 Primer - Rich Structured Data Markup for Web Documents*. W3C. Retrieved from <http://www.w3.org/TR/rdfa-primer/>
19. Google. (2013). Google Search Engine Optimization Starter Guide. *google.com*. Google. Retrieved November 3, 2013, from [http://books.google.com/books?id=LK\\_ebEqnbzcC&dq=intitle:Google+Search+Engine+Optimization+Starter+Guide&hl=&cd=2&source=gbs\\_api](http://books.google.com/books?id=LK_ebEqnbzcC&dq=intitle:Google+Search+Engine+Optimization+Starter+Guide&hl=&cd=2&source=gbs_api)
20. SearchMonkey. (2013). SearchMonkey Support for RDFa Enabled. *yahoo.com*. Retrieved November 3, 2013, from [http://developer.yahoo.com/blogs/ydn/posts/2008/09/search\\_monkey\\_support\\_for\\_rdfa\\_enabled/](http://developer.yahoo.com/blogs/ydn/posts/2008/09/search_monkey_support_for_rdfa_enabled/)
21. Zheng, Z. (2002). AnswerBus question answering system (pp. 399–404). Presented at the Proceedings of the 2nd International Conference on Human Language Technology Research.
22. Nguyen, T. H., Nguyen, H., & Freire, J. (2010). PruSM: a prudent schema matching approach for web forms (pp. 1385–1388). Presented at the Proceedings of the 19th ACM international conference on Information and knowledge management, ACM Request Permissions. doi:10.1145/1871437.1871627
23. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2012). *OWL 2 Web Ontology Language Primer*. W3C. Retrieved from <http://www.w3.org/TR/owl-primer>
24. Fuchs, N. E., Kaljurand, K., & Kuhn, T. (2008). Attempto Controlled English for

- Knowledge Representation. In *Reasoning Web* (Vol. 5224, pp. 104–124). Springer. doi:10.1007/978-3-540-85658-0\_3
25. Hewlett, D., Kalyanpur, A., Kolovski, V., & Halaschek-Wiener, C. (2005). Effective NL paraphrasing of ontologies on the Semantic Web. Presented at the *Proceedings of the Workshop on End-User Semantic Web Interaction of the 4th International Semantic Web Conference*.
  26. Fliedl, G., Kop, C., & Vöhringer, J. (2010). Guideline based evaluation and verbalization of OWL class and property labels. *Data & Knowledge Engineering*, 69(4), 331–342. doi:10.1016/j.datak.2009.08.004
  27. Hollink, L., Schreiber, G., Wielemaker, J., & Wielinga, B. (2003). Semantic annotation of image collections. *Proceedings of the Workshop on Knowledge Markup and Semantic Annotation of the Second International Conference on Knowledge Capture*.
  28. Auer, S., Feigenbaum, L., Miranker, D., Fogarolli, A., & Sequeda, J. (2010). *Use Cases and Requirements for Mapping Relational Databases to RDF*. W3C. Retrieved from <http://www.w3.org/TR/rdb2rdf-ucr/>
  29. Sahoo, S. S., Halb, W., Hellmann, S., Idehen, K., Thibodeau, T., Auer, S., & Sequeda, J. (2009). *A survey of current approaches for mapping of relational databases to rdf*. W3C RDB2RDF Incubator Group. Retrieved from [http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF\\_SurveyReport.pdf](http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf)
  30. Das, S., Sundara, S., & Cyganiak, R. (2012). *R2RML: RDB to RDF Mapping Language*. W3C. Retrieved from <http://www.w3.org/TR/r2rml/>
  31. Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., & Aumueller, D. (2009). Triplify: light-weight linked data publication from relational databases (pp. 621–630). Presented at the Proceedings of the 18th International Conference on World Wide Web, ACM. doi:10.1145/1526709.1526793
  32. Bizer, C., & Seaborne, A. (2004). D2RQ - Treating Non-RDF Databases as Virtual RDF Graphs. Presented at the Proceedings of the 3rd International Semantic Web Conference.
  33. Cullot, N., Ghawi, R., & Yétongnon, K. (2007). DB2OWL: A Tool for Automatic Database-to-Ontology Mapping. (pp. 491–494). Presented at the Proceedings of the 15th Italian Symposium on Advanced Database Systems.
  34. Cerbah, F. (2008). Learning Highly Structured Semantic Repositories from Relational Databases (Vol. 5021, pp. 777–781). Presented at the Proceedings of the 5th European Semantic Web Conference. doi:10.1007/978-3-540-68234-9\_57
  35. Knoblock, C., Szekely, P., Ambite, J., Goel, A., Gupta, S., Lerman, K., Muslea, M., Taheriyani, M., & Mallick, P. (2012). Semi-automatically mapping structured sources into the semantic web. (pp. 375–390). Presented at the Proceedings of the 9th International Conference on the Semantic Web: Research and Applications, Springer-Verlag. doi:10.1007/978-3-642-30284-8\_32
  36. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. IGI Global. *International Journal on Semantic Web and Information Systems*, 5(3).
  37. SUMO - Suggested Upper Merged Ontology. (2013). SUMO - Suggested Upper Merged Ontology. *ontologyportal.org*. Retrieved November 3, 2013, from <http://www.ontologyportal.org/>
  38. Project, A. D. L., & University of California at Santa Barbara. (Eds.). (2002, July 3). Alexandria Digital Library Feature Type Thesaurus (RDF version). Retrieved November 3, 2013, from

- [http://www.alexandria.ucsb.edu/~lhill/FeatureTypes/FTT\\_metadata.htm](http://www.alexandria.ucsb.edu/~lhill/FeatureTypes/FTT_metadata.htm)
39. Ordnance Survey Ontologies. (2013). Ordnance Survey Ontologies. *data.ordnancesurvey.co.uk*. Retrieved November 3, 2013, from <http://data.ordnancesurvey.co.uk/ontology>
  40. Lieberman, J., Singh, R., & Goad, C. (2007). *W3C Geospatial Ontologies*. W3C Incubator Group. Retrieved from <http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/>
  41. Noy, N., & Rector, A. (2006). *Defining N-ary Relations on the Semantic Web*. W3C. Retrieved from <http://www.w3.org/TR/swbp-n-aryRelations>
  42. GeoNames Gazetteer. (2013). GeoNames Gazetteer. *geonames.org*. Retrieved November 3, 2013, from <http://www.geonames.org/>
  43. Caldwell, B., Chisholm, W., & Slatin, J. (2008). *Web content accessibility guidelines 2.0*. W3C. Retrieved from <http://www.w3.org/TR/WCAG20/>
  44. Figueredo, L. A. G. A., & Masello, J. (2005). *SIDRA - Aggregate Database – Definition and Loading*. Diretoria de Informática, IBGE, Rio de Janeiro, Brazil.
  45. Piccinini, H., Lemos, M., Casanova, M. A., & Furtado, A. L. (2010). W-Ray: A Strategy to Publish Deep Web Geographic Data (Vol. 6413, pp. 2–11). Presented at the Proceedings of the Workshop on Semantic and Conceptual Issues in GIS of the 29th International Conference on Conceptual Modeling. doi:10.1007/978-3-642-16385-2\_2
  46. Google Webmaster Central Blog. (2013). Google Webmaster Central Blog. *google.com*. Retrieved November 3, 2013, from <http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html>