GEOSPATIAL METADATA RETRIEVAL FROM WEB SERVICES

Recuperação de metadados geoespaciais a partir de serviços web

IVANILDO BARBOSA

Instituto Militar de Engenharia
Praça General Tibúrcio, 80 Praia Vermelha
Rio de Janeiro – RJ Brasil
ZIP Code: 22290-270
ivanildo.barbosa@gmail.com

ABSTRACT

Nowadays, producers of geospatial data in either raster or vector formats are able to make them available on the World Wide Web by deploying web services that enable users to access and query on those contents even without specific software for geoprocessing. Several providers around the world have deployed instances of WMS (*Web Map Service*), WFS (*Web Feature Service*) and WCS (*Web Coverage Service*), all of them specified by the Open Geospatial Consortium (OGC). In consequence, metadata about the available contents can be retrieved to be compared with similar offline datasets from other sources. This paper presents a brief summary and describes the matching process between the specifications for OGC web services (WMS, WFS and WCS) and the specifications for metadata required by the ISO 19115 – adopted as reference for several national metadata profiles, including the Brazilian one. This process focuses on retrieving metadata about the identification and data quality packages as well as indicates the directions to retrieve metadata related to other packages. Therefore, users are able to assess whether the provided contents fit to their purposes.

Keywords: Geospatial Metadata; OGC Web Services; Data Base Matching.

RESUMO

Nos dias atuais, os produtores de dados geoespaciais, tanto em formato matricial como vetorial, podem disponibilizá-los pela internet, por meio de serviços *web* que permitem aos usuários acessar e consultar esse conteúdo sem a necessidade de aplicativos especializados de geoprocessamento. Diversos produtores de dados geoespaciais ao redor do mundo implementaram serviços como o WMS (*Web Map*

Service), WFS (*Web Feature Service*) e WCS (*Web Coverage Service*), todos especificados pelo *Open Geospatial Consortium* (OGC). Sendo assim, é possível extrair metadados do conteúdo disponibilizado de modo a permitir compará-lo com dados *offline* de outras fontes. Ao longo deste trabalho serão apresentados um breve resumo e a descrição do processo de comparação entre as especificações de serviços OGC e os metadados requeridos pela norma ISO 19115, adotada como base para elaboração de perfis nacionais de metadados como o Perfil Brasileiro de Metadados Geoespaciais. Essa comparação privilegia a recuperação de metadados relacionados aos pacotes de identificação e qualidade, ao mesmo tempo em que direciona o processo para a recuperação de metadados de outros pacotes. Desta maneira, os usuários poderão avaliar se o conteúdo disponibilizado se adapta às suas necessidades.

Palavras-chave: Metadados Geoespaciais; OGC Web Services; Alinhamento de bancos de dados.

1. INTRODUCTION

Maps are not frames, pictures or drawings. They are a way to represent geospatial information, with defined (or controlled) meaning, precision and accuracy over time and space. Their contents are intended to be designed to support decision-making for several activities even if it is not always possible to display all of them in a single map.

Many producers have made available a huge amount of data (and respective metadata) over Spatial Data Infrastructures (SDI) either by using independent web services or by lists of links for direct downloading. Despite this huge amount of data available on the internet, choosing the best map for some application requires an adequate analysis of quality indicators, spatial coverage, represented feature types and other metadata.

Metadata profiles have been developed to describe geospatial data and most of them are based on ISO (2003b). It implies the interoperability among distinct profiles even in different languages. With a complete profile fulfilled with reliable metadata, planners are able to better select the products. The most complete and correct geospatial metadata are supposed to support decision makers to select the geospatial data that better fits to the application requirements. However many products have published only mandatory fields, leaving a lack in useful information. One of the explanations for missing metadata in some datasets is the difficulty to gather them due to previous metadata registering policies. For example, some details about lineage (process steps and data sources).

Web map services have been developed to exhibit geospatial data based on more or less interactive queries. Most of these results are only considered partial ones instead of products, so metadata storing may be not interesting. However, the underlying datasets, either individually either combined, may provide useful information to decision support. Therefore, the metadata about each layer and about the whole map should be available for helping users to decide whether they are fit to their expectations.

This paper aims to introduce an approach to select and retrieve a subset of metadata for geospatial datasets at the context of the OGC web services to support users to assess (manual or automatically) their usability. It implies a dynamic analysis of XML documents which describe capabilities and contents of the mentioned services to compose another document containing metadata according to ISO (2003b) specifications. For the sake of simplicity, the scope of analysis is restricted only to the following OGC services: (1) Web Map Services (WMS); (2) Web Feature Services (WFS); e Web Coverage Services (WCS). The main interest is to retrieve metadata either for identification either for the data quality packages.

After this introduction, the Section 2 summarizes OGC web services specifications. In the Section 3, a set of metadata specified at ISO (2003b) is compared to the fields provided by OGC web services descriptions (OGC, 2004; OGC, 2005; OGC, 2006). Guidelines to retrieve metadata from web services descriptions are presented in Section 4. The fifth section is reserved to conclusions and suggestions for future works.

2. OGC WEB SERVICES

The Open Geospatial Consortium (OGC) defines itself as a *non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services*. Among its contributions, let’s focus the specifications for WMS (OGC, 2004), WFS (OGC, 2005) and WCS (OGC, 2006), widely adopted by producers to publish raster and vector geospatial data to users over the WWW. For all of these services, users send HTTP requests and server returns either a message in XML (*eXtended Markup Language*) format or the bitmap to be parsed by the user application. Operations’ requests must contain the adequate parameters and values according to the respective specifications. Data are arranged as layers and user may combine them even coming from different services from different providers.

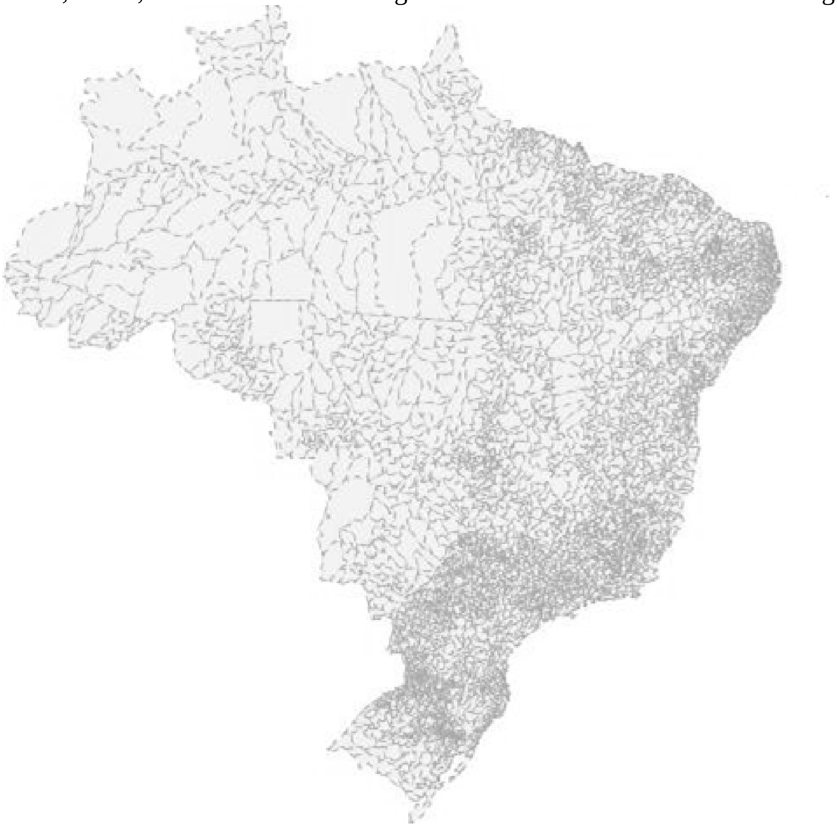
The approach used at this paper will emphasize operations and contents useful to compose the metadata set for each provided product.

2.1 Web Map Services

The *GetCapabilities* operation is specified for all the services approached. It returns general descriptions about data layers such as its provider, the range of scales to represent each data layer, their respective *Spatial Reference Systems* (SRS) and projections. Most of identification metadata may be retrieved from *GetCapabilities* contents.

The WMS *GetMap* operation works by accessing the underlying data source to create a *bitmap* (usually in JPEG or PNG format) representing the geospatial data according to the parameters requested to the server such as geographical extents, layers, and scale. Some values passed as parameters are directly assigned as metadata. Other ones must be adapted to the whole request. Users may insert the parameters manually (respecting the restrictions declared at the *GetCapabilities* document) or just interact visually by using control tools such as zoom and pan if available. The output bitmap illustrated at the figure 1 is the response to the *GetMap* operation based on the request for the layer *IBGE:e1000_municipio*.

Figure 1 – Layer *IBGE:e1000_municipio* created dynamically by WMS.
Source:
http://www.geoservicos.ibge.gov.br/geoserver/IBGE/wms?service=WMS&version=1.1.0&request=GetMap&layers=IBGE:e1000_municipio&styles=&bbox=-73.999,-33.75,-28.84,5.277&width=512&height=442&srs=EPSG:4326&format=image/jpeg



2.2 Web Feature Services

The WFS works directly by accessing features in a remote layer and by returning its schema and the values for their attributes. The layers accessed in a server may compose a dataset with other layers at the user desktop and/or layers got from others servers. Schemas may be retrieved at the correspondent field returned by *DescribeFeatureType* operation as well as other definitions about the dataset. This way, contents beyond the geometry may be retrieved. However the main source of metadata is the document provided by *GetCapabilities* operation. Such as the WMS *GetMap* operation, the WFS *GetFeature* operation accesses the dataset based on the parameters provided by the user.

2.3 Web Coverage Services

The WCS functionalities are useful to grant access to phenomena represented by values at each measurement point (e.g. digital elevation data). In addition to be exhibited as a georreferenced image, those values may be used directly or may be processed to obtain some information. The first source of metadata may be retrieved from the *GetCapabilities* operation results (figure 3 exhibits a sample). Data description containing data such as geographical extents, spatial resolution and interpolation methods is obtained by *DescribeCoverage* operation. Figure 4 exhibits a sample of *DescribeCoverage* operation results. The WCS *GetCoverage* operation retrieves the information contained at the dataset according to the specified parameters.

Due to the size of the response files in XML format provided by WMS, WFS and WCS operations, we indicate, in Table 1, examples of valid requests to be performed by a web browser.

Table 1 – Valid requests for OGC web services.

Operation	Request
WMS GetCapabilities	http://www.geoservicos.ibge.gov.br/geoserver/ows?service=wms&version=1.1.1&request=GetCapabilities
WFS GetCapabilities	http://www.geoservicos.ibge.gov.br/geoserver/ows?service=wfs&version=1.1.1&request=GetCapabilities
WFS DescribeFeatureType	http://www.geoservicos.ibge.gov.br/geoserver/IBGE/ows?service=WFS&version=1.0.0&request=DescribeFeatureType&typeName=IBGE:e1000_ilha
WCS GetCapabilities	http://www.geoservicos.ibge.gov.br/geoserver/ows?service=wcs&version=1.1.1&request=GetCapabilities
WCS DescribeCoverage	http://www.geoservicos.ibge.gov.br/geoserver/ows?service=wcs&version=1.1.1&request=DescribeCoverage&identifiers=world

3. GEOSPATIAL METADATA

The ISO 19115 International Standard aims to provide data producers with *appropriate information to characterize their geographic data* properly, facilitate the *organization and management of metadata* for geographic data, enable users to apply geographic data in the most efficient way by *knowing its basic characteristics*, *facilitate data discovery*, *retrieval and reuse* (users will be better able to locate, access, evaluate, purchase and utilize geographic data) and enable users to *determine whether geographic data in a holding will be of use to them*. Hence, spatial data may be described by the set of metadata provided by their producers or managers. Geospatial metadata are grouped in 12 different role packages represented as classes.

The root entity which defines metadata about a resource or resources is called *Metadata entity set information* package. The *Identification* package presents basic information required to uniquely identify a resource or resources. The *Constraint information* package refers to restrictions on the access and use of a resource or metadata. The *Data quality information* is about quality information for the data specified by a data quality scope (including events and source data used in constructing and quantitative quality information). *Maintenance Information* package gathers information about the scope and frequency of updating, and the *Spatial Representation information* package describes digital mechanism used to represent spatial information (vector and raster data definitions). The *Reference system information* package defines references for coordinates systems (geodetic references and projections) while *Content information* includes feature catalogue and coverage descriptions. The *Portrayal catalogue information* package is about information identifying the portrayal catalogue used. The *Distribution information package* provides information about the distributor as well as the options for obtaining the resource.

The structure proposed by ISO (2003b) may be used with new datasets, adapting the methodologies to automatically or manually register the metadata, minimizing the lack of information. Old fashioned datasets may not have all the available metadata, reducing the data quality information about the dataset. Other ones have custom metadata profiles, designed for private control, so they have to match them to ISO profile.

Previously, at the Section 1, it was mentioned that the main interest is to retrieve metadata either for identification either for the data quality packages. Among metadata specified for the Identification package, we mention the producer, spatial references and extents, title and scale. These data are specified at the response to the *GetCapabilities* operation for WMS, WFS and WCS, as well as the responses for WFS *DescribeFeatureType* and WCS *DescribeCoverage* operations. Metadata specified for data quality are organized on two groups: *Lineage* (*Process Steps* and *Source Data* for each of them) and *Results* comprising test results of data quality elements (ISO, 2002) and the *Conformance* of these results for a predefined purpose.

Considering that producers usually are not able to retrieve or produce the whole quality related metadata, they usually fulfill the *LI_Lineage statement* with general details – usually by summarizing processes steps and data sources – about methods used to generate that dataset, processes and data sources according to the available data. In these cases, the retrieval of quality references requires a deep analysis to choose the best dataset to support user application. It may be done by reading (human processing) or by automatic processing (robots, crawlers and other Artificial Intelligence resources for natural language analysis). Human processing may be more efficient if the processor is an expert, the language is well known and the amount of datasets to process is not too high. On the other hand, search engines designed to process text may not be so effective to assign the real meaning of terms in order to properly convert the free text into a structured quality report.

At the *Result* package, test results based on the specifications of ISO (2003a) must be registered. References about the closure to a well defined (and accessible) conceptual schema (*logical consistency*) and *attributes filling completeness* are important to evaluate a geospatial data set to support an application and may be computed based on WFS operations (*DescribeFeatureTypes* and *GetFeature*) results.

The qualitative description of quality evaluations – registered as *Conformance*, indicates whether the quantitative results evaluated previously fits to the stated specification. These results are meaningful only when both quality results and specifications (a set of bounding values for quantitative results) are compared.

4. METADATA OF WEB SERVICES RESULTS

The response of an OGC web service is another product with individual set of metadata. Users may retrieve a lot of relevant metadata for these new products. As proposed, let’s focus on both identification and quality metadata, organized as presented at the Section 3: *Lineage* and *Results*.

4.1 Identification

Some identification metadata, such as title and abstract, must be inserted manually due to the flexible arrange of input parameters. Users customize the queries so title and abstract must be composed to better represent the purposes that the map was created for. Other metadata may be retrieved automatically from the system, such as date and time.

Furthermore, after matching the specified XML tags of OGC web services and the identification metadata, it is feasible to retrieve:

- *scale* (*scaleDenominator*): the web service provides automatically the current scale for the displayed map;
- *sourceCitation* (specially, *date* element): describes the service provider with data such as name, address and contacts; and

- *reference system* (*sourceReferenceSystem*): necessary to correctly place coordinates and guide the on-the-fly SRS conversion to gather layers from different sources or to display the map in another SRS.

Consider the metadata for the example illustrated by figure 1. The title should be inserted by the user or the system could suggest the retrieved layers' titles to be edited by the user to improve the meaning for the texts (e.g. "*malha municipal na escala 1:1.000.000*" instead of "*e1000_municipio*"). The same procedures could be implemented to insert abstract and keywords. For more than one layer or empty abstracts, a text template may be suggested from the tags such as title, service description or service provider (e.g. "*malha municipal na escala 1:1.000.000, fornecida por IBGE - Diretoria de Geociências, cobrindo a area delimitada pelos paralelos [minlat] e [maxlat] e pelos meridianos [minlong] e [maxlong]*").

The Spatial Reference System (parameter *SRS* – "*EPSG:4326*") and the geographic boundaries (parameter *bbox* containing both minimum and maximum values for latitude and longitude – "*-73.999,-33.75,-28.84,5.277*") may be retrieved from the request parameters. If these parameters are not specified, default values may be retrieved from the server. Date and time are extracted from the HTTP response time stamp.

4.2 Data Quality – Lineage

At the *Lineage* package, the complete fulfilling of *Source* and *Process Steps* data allows to distinguish different datasets. The broader *process step* used to create a new map is to *merge selected layers* (performed by the service provider). The time reference for this process may be the system timestamp on loading data or compiling de map for print.

The respective data sources are the underlying databases (*layers*) described at the respective service capabilities. Theses contents are useful to guide the user to portrait the geospatial context at a specific date or to choose the most recent source, according to its needs. Because each layer is an individual product, their metadata comprises data about identification, lineage and quality reports. However the services capabilities usually provide only a brief identification.

It seems to be a simple information, but layers descriptions returned by WMS and WFS *GetCapabilities* operation do not mention the equivalent scale (it is different from the bitmap scale). However, some implementations of these services return compute the scale value by fitting the geographical extents to the map dimensions (width and height, in pixels). The WCS *DescribeCoverage* operation return *GridOffsets* information, meaning the smallest spatial extent represented. In this case, the scale value may be computed based on the smallest graphic dimension.

On the other hand, the service description provides a lot of metadata about each layer citation (Table 2). Even quite relevant, *date* is not retrieved easily from layers descriptions, due to *time* is considered another dimension in WMS and it is not specified in WFS and WCS. Most of times, producers include date of creation or last update at the layers description text – so it is necessary to analyze the text searching for time references – or as a field in schema (verifiable with WFS *DescribeFeatureType* operation).

The spatial reference in WMS and WFS is provided by tags shown at Table 2. Layers in WCS are defined by the list of *supportedCRS* tags.

Table 2 – Relationship between metadata fields and layers description tags.

ISO 19115 metadata	WMS service description	WFS service description
Title	<layer> <Title>	<*FeatureTypeList*> <Title>
Date	<Dimension name="time" units="ISO8601" />	-
citedResponsibleParty (organisationName)	<Attribution>	<ows:ProviderName>
referenceSystemIdentifier (code)	<layer> <CRS>	<FeatureTypeList> <DefaultSRS>
LI_Source (description)	<layer> <abstract>	<FeatureTypeList> <abstract>
EX_GeographicBoundingBox	EX_GeographicBounding Box	WGS84BoundingBox

4.3 Data Quality – Results

As mentioned before, this paper considers relevant measures able to be retrieved from OGC web services operations the *logical consistency* and *attributes filling completeness*. This section will deal the results that may be computed only using WFS operations, due to the characteristics of this service.

Measuring logical consistency is useful to establish the *degree of adherence to logical rules of data structure, attribution and relationships* (ISO, 2002). It comprises measurements of *conceptual*, *domain*, *format* and *topological consistencies*. The next paragraphs will deal one way to retrieve an indicator for *conceptual* and *domain consistencies*.

Despite WMS enables users to invoke queries to layers (*GetFeatureInfo* operation) the schema is not introduced on descriptions. The WFS *DescribeFeatureType* operation points to schema locations and these points to feature attributes descriptions such as name, type and restrictions. The WCS *DescribeCoverage* operation does not also present a schema but values for dimensions in a raster structure. After schema retrieving, next step tries to address syntactic and semantic matching between the given schema and another one adopted as reference (Mata, 2007). The names given to fields are not always intuitive and there is no description provided on schema. It may be solved by designing schemas for the service based on well-known schemas or mapping the relationship among them manually (Casanova *et al*, 2007).

Domain consistency checking (to verify unusual occurrences) is the only option for data available in a WCS (*range* field in *CoverageDescription* operation). In WFS, servers should to implement queries to find maximum and minimum values for each field and offer to the user a default consistent (range of) value(s).

Attribute completeness aims to compute the percent of features in a particular layer has any fields fulfilled with "*null*" or some default value. This measurement is oriented only to WFS due to the access to the schema and the attributes value. Lack of values in some fields may impact queries involving the correspondent layer. However, depending on the amount of features and the complexity of its schema, to check database completeness every time the layer is invoked may be inappropriate. So it's recommended that servers pre-compute this item when publishing the data or deploy web applications to compute it when user requires it. Moreover, developers of API client may increase these applications to perform this operation, in order to avoid queries over an empty database.

5. CONCLUSIONS

This paper presented a brief summary and described the matching process between the specifications for OGC web services (WMS, WFS and WCS) and the specifications for metadata required by the ISO 19115. This process was focused on retrieving metadata about the identification and data quality packages, either for lineage either for results for quality measurements. The analysis of the whole map and its respective data sources supports the users to compare the services capabilities to their applications specifications and to identify (and discard) inappropriate sources.

For better use of data brought by OGC map services, providers should improve the service descriptions to help users to achieve reliable results by viewing and querying such data. For future works, it is suggested:

a) the creation and upload of pre-processed metadata packages to facilitate the compilation whenever it is requested. The location containing the compiled metadata contents are supposed to be indicated in the *metadataURL* tag on service description for all the layers. It seems to be the easiest way for users and servers, avoiding to repeat the whole process every time the layers are loaded;

b) the development of both solutions for natural language processing and adequate dictionaries to enhance the metadata retrieval in either *abstract* and *statement* fields (by syntactic and semantic matching) at both service descriptions and metadata;

REFERENCES

CASANOVA, M. A., BREITMAN, K. K., BRAUNER, D. F., MARINS, A. L. A. *Database Conceptual Schema Matching*, IEEE Computer, Vol: 40 Issue: 10 Pages 102-104 , 2007.

ISO, *ISO 19113:2002 – Geographic information -- Quality principles*, 2002.

ISO, *ISO 19114 – Geographic information -- Quality evaluation procedures*. 2003a.

ISO, *ISO 19115 – Geographic information – Metadata*. 2003b.

MATA, F. *Geographic Information Retrieval by Topological, Geographical, and Conceptual Matching*. Lecture Notes in Computer Science, Volume 4853/2007, Pages 98-113. 2007.

OGC. *Web Coverage Service | OGC*®. 2006. Accessed 26/11/12. http://www.opengeospatial.org/standards/wcs.

OGC. *Web Feature Service | OGC*®. 2005. Accessed 26/11/12. http://www.opengeospatial.org/standards/wfs.

OGC. *Web Map Service | OGC*®. 2004. Accessed 26/11/12. http://www.opengeospatial.org/standards/wms.

(Recebido em novembro de 2012. Aceito em dezembro de 2012.)