Computers & Geosciences 97 (2016) 89–97

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

A comprehensive open package format for preservation and distribution of geospatial data and metadata

X. Pons[a], J. Masó[b,*]

[a] Grumets Research Group, Dep Geografia, Edifici B. Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain
[b] Grumets Research Group, CREAF, Edifici C. Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain

ARTICLE INFO

Keywords:
Data standard
Internet GIS
Metadata
Data model
Preservation
Package
MMZX

ABSTRACT

The complexities of the intricate geospatial resources and formats make preservation and distribution of GIS data difficult even among experts. The proliferation of, for instance, KML, Internet map services, etc, reflects the need for sharing geodata but a comprehensive solution when having to deal with data and metadata of a certain complexity is not currently provided. Original geospatial data is usually divided into several parts to record its different aspects (spatial and thematic features, etc), plus additional files containing, metadata, symbolization specifications and tables, etc; these parts are encoded in different formats, both standard and proprietary. To simplify data access, software providers encourage the use of an additional element that we call generically “map project”, and this contains links to other parts (local or remote). Consequently, in order to distribute the data and metadata refereed by the map in a complete way, or to apply the Open Archival Information System (OAIS) standard to preserve it for the future, we need to face the multipart problem. This paper proposes a package allowing the distribution of real (comprehensive although diverse and complex) GIS data over the Internet and for data preservation. This proposal, complemented with the right tools, hides but keeps the multipart structure, so providing a simpler but professional user experience. Several packaging strategies are reviewed in the paper, and a solution based on ISO 29500-2 standard is chosen. The solution also considers the adoption of the recent Open Geospatial Consortium Web Services common standard (OGC OWS) context document as map part, and as a way for also combining data files with geospatial services. Finally, and by using adequate strategies, different GIS implementations can use several parts of the package and ignore the rest: a philosophy that has proven useful (e.g. in TIFF).

1. Introduction

Although the usage of GIS often requires only the last updated version of the data and metadata, other studies may also require historical data and/or time series to be included in analyses as urbanization dynamics, environmental and climate change, land cover change, etc. Unfortunately, essential data for such applications are often lost; either through data overwrite with an updated version, by technological obsolescence or simply by not preserving comprehensive versions at the adequate time frames (Morris, 2009). Two incidents are often cited to illustrate the importance of data preservation: the National Aeronautic and Space Administration (NASA)’s lost data from the 1976 Viking mission, and the BBC's 1986 digital Domesday Project data that was almost lost (Duerr, 2009). Digital geospatial data producers are recently considering the need for addressing data preservation (Bethune, 2009). The Open Archival Information System (OAIS) (CCSDS, 2002) proposes the use of the information package (IP) to prevent digital information losses by grouping all parts that form the data, metadata, context, semantics, etc in a package format. OAIS recognizes three types of information packages: submission (SIP), archive (AIP) and delivery (DIP). The OAIS conceptual model does not provide any concrete format for packaging, and leaves the implementation to communities. McDonough (2010) describes how to use the Functional Requirements for Bibliographic Records (FRBR) information model to create packages containing realizations of the same videogames together with the representation information. Waugh (2006) proposes the VERS Encapsulated Object to be applied to the Australian Public Record Office Victoria: an XML document that encapsulates data and metadata that can be digitally signed. The European Space Agency (ESA) developed the Standard Archive Format for Europe as an extension of XML Formatted Data Unit, as a common format for archiving ESA remote sensing data. It is not limited to the geospatial domain but covers all kinds of digital information. Unfortunately, and despite the adequacy of OAIS for

[*] Corresponding author.
E-mail address: Joan.Maso@uab.cat (J. Masó).

http://dx.doi.org/10.1016/j.cageo.2016.09.001
Received 14 September 2015; Received in revised form 26 August 2016; Accepted 6 September 2016
Available online 17 September 2016

our purposes, it does not propose any concrete packaging format and no current format seems to cover the geospatial community requirements.

Currently, instead of providing a consistent multipart format composed by all the interrelated parts, producers are forced to oversimplify data. Indeed, commonly used data access services such as those conformal to WFS, SOS and WCS do not always correctly link to metadata about the data and rarely provide any form of symbolization information making data difficult to interpret by non-expert users. In some cases, producers opt to go back to formats like GeoPDF that contains a visual representation of the project as a printable map (Cervantes, 2009). Not preserving the real data, this solution is only useful for displaying and printing purposes (e.g., raster data types are lost, and there are no references to data dictionaries and metadata, Morris and Tuttle, 2008). Other common alternatives are geospatial web services (such as WMS, WMTS, WFS, SOS, WCS, etc) are difficult to preserve due to their dynamic nature and the complexities of transferring them to the archives.

The ISO Technical Committee 211 has started to draft ISO19165, a geospatial data and metadata preservation standard based on OAIS, acknowledging the need for standardized multipart information packaging. The need for collecting related parts in a single packaging file, as well as the relations between these parts in a way that can be easily distributed to others offline or on the Internet has been identified by Figueroa and Abergel (2011) and others, but this important problem has not been solved yet. A multipart file can be also used in other applications, such as publishing maps in web pages (Masó and Pons, 2011), send them by email, or be uploaded as a complete dataset in a Web Processing Service (WPS) (Schut, 2007): in this case both as process inputs or to deliver the results of WPS processes. It could be possible to expose or send the individual parts as individual attachments in an email or as separate WPS input, but this requires that the user know the relations between them all. In the WPS case the number of inputs of a process will depend on the MIME file type.

Determining the right multipart package format is not an easy task. St-Denis et al. (2000) and Hoebelheinrich (2012) have independently collected a set of criteria useful to compare formats. They state that a good format does not have to lose, add or alter the original data so it can be reconstructed exactly by the receiver (identity criteria; called "transparency" in the original St-Denis paper), it has to be usable in reality and support big size and complexity (scalability criteria), it has to be as simple as possible addressing the essential problem (simplicity criteria), it has to support multiple platforms, operating systems and programming languages (neutrality criteria), it has to be well-defined and not allow misinterpretations (formality criteria), it has to be adaptable to different scenarios and contexts (flexibility criteria), it has to be adaptable to new requirements and uses (evolvability criteria), it has to be potentially usable by as many people as possible (adoption criteria; called "popularity" in the original St-Denis paper), it has to have all the necessary components to fulfil its purpose or be expanded to new objects and relations [as explained by Bowman et al. (2000) and Holt et al. (2000)] (completeness criteria), it has to use an identical data model representation (metamodel identity criteria), it has to be reusable for similar problems (solution reuse criteria), it has to be easy for a human to read and understand the format (readability criteria; called "legibility" in the original St-Denis paper), and it has to be possible to check that there are no transmission errors (integrity criteria, sometimes related to certification). Hoebelheinrich add the needs of having an open and well documented specification (disclosure criteria), a bit stream easy to decipher (transparency criteria), able to embed metadata and semantics (self-documentation criteria), and free of patents, as fees can be important barriers, especially in long term preservation applications (protection, legal and cost criteria).

Three aspects coming from a recent big data scenario (Borkar et al., 2012) can also be added: the format has to be compact (compression criteria), it needs to have an entry point to the data into the package (entry point criteria), and allow direct access (often named "random access") to any part of the file (direct access criteria). The existence of open source libraries has a big influence on fast adoption of the format (open libraries criteria). Compression often allows faster downloads over the Internet (Wessel, 2003), and is especially beneficial for mobile devices (Kim et al., 2004). Combining compression with identity criteria implies that only lossless compression formats should be used.

The importance of relating geometric data to thematic attributes and data dictionaries, metadata (including data quality information, lineage, etc), symbolization and web services in a seamless environment has been recognized (Horak et al., 2010; Morris and Tuttle, 2008). These components are often stored in separated parts (allowing, for example, that a data dictionary can be used from several datasets) packaging should support their integrated treatment. The idea of having a map that has links to all related parts was introduced in corporate GIS architecture by Laurini and Milleret-Raffort (1990), which defined the hypermap concept as a geo-referenced multimedia system that can hyperstructure individual multimedia components with respect to each other. GIS products hyperlink files are often stored starting by a map files. The map acts as an entry point to the data for easy interaction with a coherent subset of the GIS information. In 2014 the format description document database of the US Congress Library contained 334 formats, of which 34 are geospatially related and 9 are composed by more than one part (http://www.digitalpreservation.gov:8081/formats). This figure grows if we take into consideration that metadata and symbolization instructions are usually included in separated parts (Kraak and Ormeling, 2003). Relations between those parts can eventually be imbricate, resulting in a tree of dependencies that is hard or impossible to remember. Then, it is not longer possible to move or share the dataset without risking the integrity of its relations. In this paper we will call this issue the multipart file problem and we propose a solution for it in the geospatial realm.

One of the problems that the packaging approach is facing is the integration of the distributed GIS into Linked Data (Vilches-Blázquez et al., 2014). Linked Data (Berners-Lee, 2006) is an initiative where *things* in Internet receive a Uniform resource identifier (URI) and a Resource Description Framework (RDF) language is used to relate them to other *things* for a reason. Taken to the extreme, Linked Data leads to a single net where every resource is connected to any other, making a naive application of the OAIS package concept difficult. Consequently, together with a way to link an element in one package to another element in another package, a mechanism to limit each package scope to a convenient size and content is needed.

In 1997, the authors of this paper developed a package format to solve the multipart problem. This paper revisits the original idea and re-masters it using the ISO 29500-2 Open Packaging Conventions (OPC) standard, and proposes improvements and additions, opening the format to allow interoperability. A sound review of several multipart packaging strategies has been done, and considerations are exposed in next section. Afterwards, the paper describes the chosen solution and how it is adapted to the geospatial data needs illustrated by a reference implementation.

2. Current packaging strategies

2.1. MIME encapsulation of aggregate HTML documents

An Hyper Text Markup Language (HTML) page is an example of multipart document, composed by the page itself and linked multimedia, JavaScript, CSS libraries, etc. Local storage of an HTML document is an example of the multipart problem: if we only save the main page all linked contents we still depend on dynamic content that can disappear. The standard MHTML (Palme et al., 1999; RFC 2557) permits to store or transport HTML documents in a MIME multipart document: a single file including the HTML page part and the linked content as additional parts. It is commonly used by some