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

popular web browsers to allow storing a HTML page in self-contained file (.mht extension).

Even if the scope of this format is limited to HTML documents, its generic design allows applying it to geospatial data and was used by the WCS GMLCov GeoTIFF Coverage Encoding Profile (Meissl, 2014) to include the GMLCov metadata and a GeoTIFF image in a single file. Being a text file, common binary files used in GIS (e.g. Shapefiles) should be encoded in text (for instance in base64) resulting in substantially larger files. In addition it lacks a directory structure support, forcing a flat list of parts.

2.2. Compressed KML documents (KMZ)

The KML file is a geospatial data format based on XML. It can embed geographical features and attributes and also supports links to external parts, such as images, models or textures. KML (Wilson, 2008) became an OGC standard in 2008. Google recommends the distribution in KMZ files, which are ZIP files (.kmz). A KMZ is composed by a single root KML document (typically named "doc.kml") and optionally any overlays, images, icons, models, etc, referenced from the KML. The KMZ format was never transferred to OGC but a set of recommendations for being compatible with current implementations can be found on the developers section of the Google site (https://developers.google.com/kml/documentation/kmzarchives). An interesting characteristic is the ability to link to components in other KMZ files by concatenating the URI of a KMZ file with the internal reference of the internal component (e.g., http://www.someserver.com/pictures.kmz/images/photo01.jpg).

2.3. Flexible formats (HDF, NetCDF, GeoPackage, ...)

Specific thematic communities have achieved a certain degree of success by designing flexible common formats. The weather and climate communities often use HDF and NetCDF to represent scientific data. HDF is developed by the National Center for Supercomputing Applications (NCSA, 2001), University of Illinois, and NetCDF is developed by the Unidata Program Center in Boulder, Colorado (Rew et al., 2016). These binary files have an extensible header and internal modules providing a system capable of growth with scientific data needs and allowing for efficient extraction of a subset of a dataset. They can contain geospatial data (mainly in gridded and array formats) but also user annotations, metadata, and specific descriptions. Both NetCDF 4 and HDF 5 support compression (Rew et al., 2016), and access is done by using open libraries (Bunting and Gillingham, 2013).

GeoPackage is a self-contained, cross-platform, open relational database standard designed to simplify the use of geospatial data, originally designed for defence and intelligence applications. GeoPackage is capable of holding multiple vector feature types, rasters from various sources, and multiple tile pyramids (Daisy, 2012). It defines ways to store geospatial information in tables and an SQL syntax to generate and access them. SQLite is the initial reference implementation of a GeoPackage container. The specification is an OGC standard and current implementations are focused on spreading data mainly to mobile devices. GDAL (http://gdal.org/) supports GeoPackage Features, and Luciad (http://www.luciad.com) uses GeoPackage as the core format in some products.

2.4. BagIt file package format

BagIt is a hierarchical file package format designed to support transfer of generalized digital content. A "bag" consists of a base directory containing a set of top-level files and a sub-directory named "data/" that holds the payload. The top level directory contains a "manifest-algorithm.txt", a "bagit.txt", and zero or more additional files. The package can be contained in a single-file archive format such as TAR or ZIP (Kunze et al., 2015). This format is well known in the data preservation community but not much in the geospatial domain, even if some examples exist (Bethune et al., 2009).

2.5. MiraMon compressed map (MMZ)

The MMZ is a binary file able to compress different files and links to Internet resources in a single file. The separated parts are compressed with a gzip algorithm and stored together in a multipart file with a specific header format including metadata about the original files (http://www.miramon.cat/new_note/usa/notes/MMZFormatSpecification%20v1.docx). It was initially intended for compressing MiraMon maps (containing raster, images, vector, tables, symbolization, etc) and for disseminating free environmental data produced by the Department of Environment of the Catalan government. To account for integrity, an optional certification process is available consisting in some encrypted files that include author information that guarantees the integrity of the original parts and allows showing the name of the authors to the MMZ users. The file format is used in the MiraMon GIS and remote sensing software (Pons, 2002) and a free reader exists.

From the user perspective, once the reader is installed in the computer, a single click on an MMZ file in a web page opens the multipart file and gives immediate access to the same data the producer created. Furthermore, the information included can be extracted and aggregated to other GIS data for professional and analytical work with GIS tools.

2.6. Open Packaging Conventions (OPC)

OPC integrates elements of the ZIP compression (PKWARE, 2004), XML documents, and the web MIME types into an open standard that makes easier to organize, store, and transport data packages. It is the ISO 29500-2 and ECMA-376 and is used by Office 2007 and newer versions of Word (.docx), Excel (.xlsx) and PowerPoint (.pptx), along with XPS (.xps), Autodesk AutoCAD (.dwfx), etc. The benefits of these formats for storing scientific data are recognized in the literature (Townsend et al., 2009). An OPC package can contain several files with a directory structure. The format adds extra files to increase interoperability (ISO29500-2, 2008): records file relations, or links, a file with a few elements of metadata (not constraining more detailed metadata information inside) and a thumbnail image for presentation purposes. These extra capabilities allow some basic data maintenance, such as the extraction of a fragment of a package, thus guaranteeing that all the related resources are considered without needing to understand the actual part's encoding (Davis and Shur, 2007).

The OPC standard introduces the possibility of relating files outside the package (external relations). A practical application of this is to exclude from the package some elements that are considered too remote in the relation tree, or that are too big, or that are completely out of a bounding box. The files filtered out will be left as a remote URI for further download by the OPC-enabled client application. It is possible to link to a file that is in another package by combining the package name with the part name (e.g.: http://www.someserver.com/folder/map.mmzx#internal_folder/file.ext).

3. Choosing a format for geospatial data packaging

The selection of KMZ, and flexible formats like HDF, NetCDF, GML or GeoPackage will require to transform the original data if it is in another geospatial format before packaging. There is a risk of losing some data and metadata during this process (violating the identity and completeness criteria); for instance, KML does not support multiple attributes and/or one to many relationships, or complex symbolization. MHTML has been discarded because it is a text file, too bulky for storage or network transmitting purposes (not complying with the compression and direct access criteria). It also lacks an integrity check

Table 1
Summary of the extended St-Denis et al. (2000) and Hoebelheinrich (2012) criteria for the analyzed multipart file formats. The symbol ✓ indicates criteria compliance, ½ needs to be interpreted as compliance *in some aspect* or *partially* with the criteria.

Criteria:	MHTML	KMZ	HDF NetCDF	Geo Package	BagIt	MMZ	OPC
Identity	✓	✓	✓		✓	✓	✓
Scalability	✓	✓	✓	✓	✓	✓	✓
Simplicity	✓	✓			✓	✓	✓
Neutrality	✓		✓	✓	✓		✓
Formality	✓		✓	½	½		✓
Flexibility			✓		✓	✓	✓
Evolvability			✓	✓	✓	✓	✓
Adoption	½	✓	✓		½		✓
Completeness					✓	✓	✓
Metamodel identity	✓		½	½	✓	✓	✓
Solution reuse	✓				✓	✓	✓
Readability	✓		½				½
Integrity		✓	✓	✓	✓	✓	✓
Disclosure	✓		✓	✓	✓		✓
Transparency	✓	✓	½		✓	½	✓
Self- documentation	½	✓	✓			½	✓
Compression		✓	½	✓	✓	✓	✓
Entry point	✓	½				✓	✓
Direct access		½	✓	✓	½	½	½
Open libraries	✓	½	✓	½	✓		✓

mechanism, and does not respect the flexibility, evolvability and adoption criteria. BagIt, MMZ and OPC formats, are similar packaging solutions in their conception but only OPC offers the advantage to store the relations between the parts of the package in a separate format, making possible to understand them without even knowing the format of the internal parts (self describing criteria). OPC is an open standard well documented and widely implemented in other domains (disclosure criteria). Table 1 summarizes how the different formats comply with the different criteria. Similar reasoning is used by Phillips and Allemang (2010) to justify their preference for OPC as a file format. The proposed approach is fully compatible with the OPC standard but, due to the nature of the geospatial information; it incorporates a few extensions to accommodate some extra requirements.

3.1. Open Packaging Conventions for exchanging geospatial data

To facilitate the explanation of the proposed solution, we will use a simple map consisting of a 1:1 000 000 country boundaries (vector), (from FAO [United Nations – FAOStat; http://geodata.grid.unep.ch/options.php?selectedID=2135]) on top of a digital elevation model (raster) with 5′ of arc pixel size (from NOAA and NGDC; http://geodata.grid.unep.ch/options.php?selectedID=1414). Data is from the UNEP EDE Data Portal (UNEP, 2013). Vector data is in a Shapefile format (a closed de-facto standard) (ESRI, 1998), while raster consists of either a raw 16-bit data or a TIFF file (Perkins, 1995), a format that was originally published in 1986 and that has been revised several times (Adobe, 1992). Relations between different parts are stated in XML files with the same name as that of its respective source part, adding the extension.rels and placed in a.rels subfolder. They list the target parts related to the source and the semantics of these relations. Fig. 1 illustrates the example data map shown in ArcGIS and MiraMon. Metadata and symbolization files are included in ISO19115 XML files and in SLD files, respectively, as well as in proprietary formats.

OPC can define entry points to the data by listing them in a.rels part in the root rels folder (Fig. 2). In our example, we provide three map files: for ESRI software a world.mxd map, for MiraMon software a world.mmm map, and a world.xml map following the OWS Context standard (OGC OWS) document (Brackin and Gonçalves, 2014). A geospatial application reading the package will determine which entry point it better supports to start reading the data. Fig. 3 provide a representation of the package structure.

Additionally, the OPC standard has two special parts: a metadata file including some OPC metadata elements (mainly following Dublin Core [ISO 15836:2009]) and a thumbnail image. These two parts facilitate a better representation of the file in the operating system file browser. Table 2 lists the OPC core metadata elements mapped to ISO 19115 metadata elements as part of the adaptation of OPC to geospatial data needs.

The description of the OPC file is complemented by a [Content_Types].xml part, stating the MIME types present in the package parts normally associated with a file extension. This metadata is useful for compatibility among operating systems that do not use file extensions, as well as for Internet communications (“/Types/Default” elements) (see Fig. 4).

Fig. 5 represents all the relations involved in the example package. Some part names are repeated, showing that they are reused in more than one entry point.

Current map files are mainly written in proprietary formats. The use of Open Geospatial Consortium (OGC) OWS Context document, as one of the entry points to the package increases its interoperability. An OWS context document provides a way to reference different units of information (such features types, coverages, etc) that resides on services, external files, or embedded data. Its XML encoding is based on the Atom W3C standard (Nottingham and Sayre, 2005). One of the OWS context document use cases described in the standard is to share the common operational picture. In this case, this is achieved by providing the necessary references to the elements stored as parts of the OPC information package.

The OPC package can link to external files (in a similar way BagIt uses the fetch.txt) but does not natively include a useful capability that MHTML format supports: the inclusion of external files while keeping the original URI. This capability is particularly useful for long term preservation where there is no guarantee that external links will be preserved in the long future. To do this, we propose to extend the ISO 29500-2 to include a new internal part called rels/{filename}.{extension}.urls that lists URI, where the file can be externally accessed and the date when the part was downloaded and included in the package is recorded. In environments with high speed connectivity, tools can ignore internal parts and negotiate with the web server if a newer version of the linked data is available, and can therefore download it instead of using the embedded part; in environments with low bandwidth or non-existing connection, internal parts are used instead. This extension is useful for linking to data dictionaries and vocabularies that evolve in time with corrections or extensions that will become

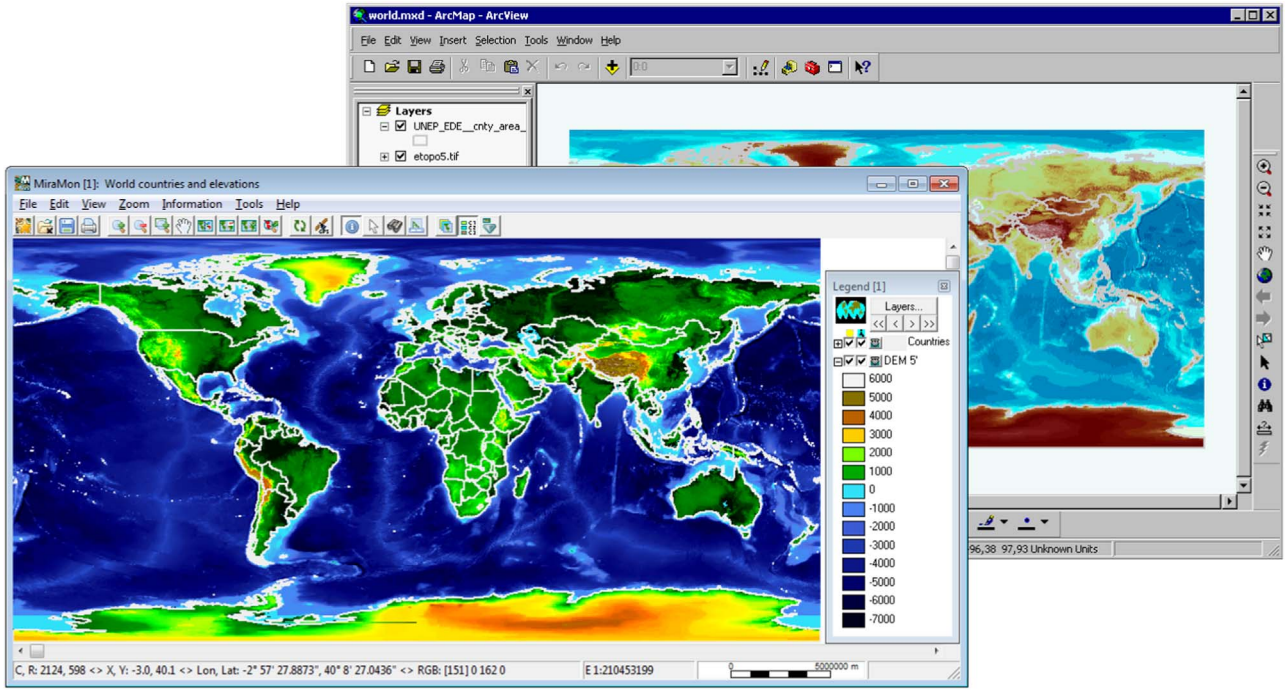

Fig. 1. : The example data map shown in ArcGIS (background) and in MiraMon (foreground). ArcGIS is showing a TIFF representation while MiraMon is showing the raw raster data. Example data is available at https://zenodo.org/record/56047 and also in the supplementary material section of the journal.

available, but ensuring that a version of the dictionary is always present in the package. For now, we propose a simple version negotiation based on the date of the part. Fig. 6 shows a.urls part example for the etopo5 TIFF part.

3.2. Requirements for tools to create and read OPC for geospatial data

In order to implement a packaging strategy in a GIS system, we need a tool that generates the package (Geospatial OPC generator) and an un-packaging routine that can be incorporated into GIS software (Geospatial OPC extractor library). A Geospatial OPC generator has to be able to understand the formats that are involved in the package. It will receive one or a few file entry points and, it will elaborate a list of parts that are going to be packaged. Then, it will add determine the relative path in a way that the packaging could be decompressed in a single file folder. It will replace the original linked paths by the equivalent relative part path in each part in the package. In doing so, the resulting package will not require linked path modifications during the extraction. In addition, the Geospatial OPC generator has to create the extra part described before.

Some parts in formats that are not commonly used or in proprietary formats may be inappropriate for long term data preservation and additional format transformation could be required. For example, proprietary database table formats could be converted to SQLite or CSV files before being incorporated into a package. Original and open

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Relationships xmlns="http://schemas.openxmlformats.org/package/2006/relationships">
       <Relationship Id="rId0"
Type="http://schemas.openxmlformats.org/package/2006/relationships/metadata/core-
properties" Target="mmzProps/core.xml"/>
       <Relationship Id="rId1"
Type="http://schemas.openxmlformats.org/package/2006/relationships/metadata/thumbnail"
Target="mmzProps/thumbnail.jpeg"/>
    <Relationship Id="rId2"
Type="http://www.miramon.uab.cat/schemas/mmzDocument/2013/relationships/rootMiraMon"
Target="world.mmm"/>
    <Relationship Id="rId3"
Type="http://www.miramon.uab.cat/schemas/mmzDocument/2013/relationships/rootArcGis"
Target="world.mxd"/>
    <Relationship Id="rId4"
Type="http://www.miramon.uab.cat/schemas/mmzDocument/2013/relationships/rootOWSContext"
Target="world.xml"/>
</Relationships>
```

Fig. 2. Root.rel file that follows the ISO 29500-2. It is read by the applications to determine the names of the package entry points and to select the appropriate one. The XML schema of this file is included in Annex D of the ISO 29500-2 standard document.

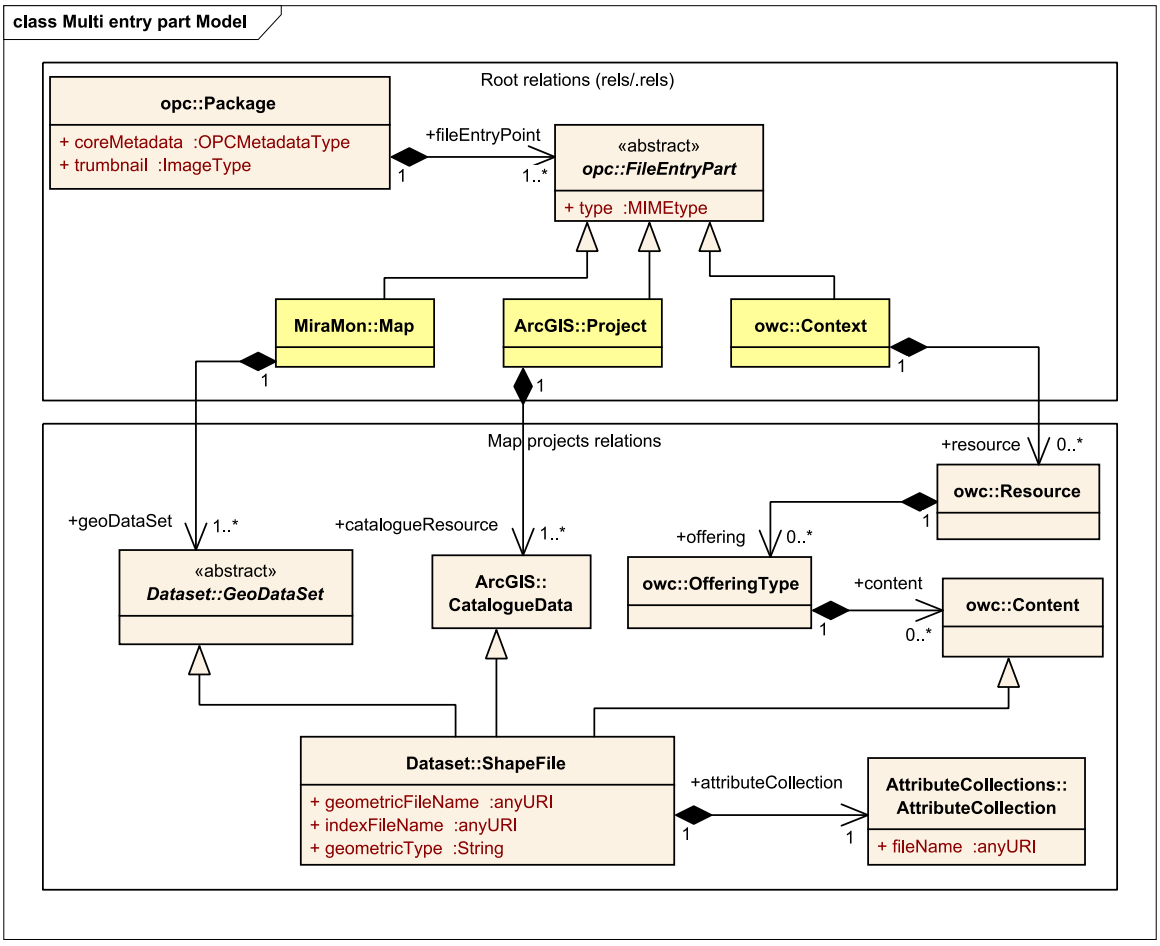

Fig. 3. : Several types of file entry points can be used in OPC format. In this diagram, ShapeFile: geometricFileName, Resouce: offering:content and CatalogueData: file can point to the same Shapefile part. Raster parts of the map have been hidden for clarity.

Table 2
Mapping between OPC core metadata elements and ISO 19115 metadata.

OPC core metadata	ISO 19115 metadata
revision	N/A
language	MD_DataIdentification/language
description	MD_DataIdentification/supplementalInformation
category	MD_DataIdentification/topicCategory
subject	MD_Identification/abstract
created	MD_Identification/citation/CI_Citation.date[CI_Date/dateType=creation]/CI_Date/dateType
lastPrinted	MD_Identification/citation/CI_Citation.date[CI_Date/dateType=publication]/CI_Date/dateType
modified	MD_Identification/citation/CI_Citation.date[CI_Date/dateType=revision]/CI_Date/dateType
identifier	MD_Identification/citation/CI_Citation/identifier/MD_identifier/code
version	MD_Identification/citation/CI_Citation/identifier/MD_identifier/version
title	MD_Identification/citation/CI_Citation/title
keywords	MD_Identification/descriptiveKeywords
lastModifiedBy	MD_Identification/pointOfContact[CI_ResponsibleParty/role=processor]/CI_ResponsibleParty/organisationName|MD_DataIdentification/environmentDescription
creator	MD_Identification/pointOfContact[CI_ResponsibleParty/role=resourceProvider]/CI_ResponsibleParty/organisationName
contentStatus	MD_Identification/status

formats can be included as part of the package, and the respective.rels can record the relation to both of them.

A Geospatial OPC extractor library needs to decompress the rels/.rels part. It will get the list of entry points and will look for its suitable map entry point. Then, it will explore the.rels files, looking for other linked files to be decompressed. In the presence of .url files, the extractor library will select to download the corresponding external file or to decompress the internal part; the decision should be based on the version available on the net and on the bandwidth availability.

4. Reference implementation for geospatial packaging

Office works with the packaged format in direct way. It is so transparent to users that most of them ignore they are using a multipart file. In GIS applications, a different strategy is suggested: When the user decides to share a map, he/she requests a package generation where every file becomes a part in the package. The package file can be directly opened in the GIS (that internally decompresses it in a temporary space). The difference in the way a package file is used is due to the intrinsic nature of the GIS application. In fact, many maps

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Types xmlns="http://schemas.openxmlformats.org/package/2006/content-types">
    <Override PartName="/mmzProps/core.xml" ContentType="application/vnd.openxmlformats-
package.core-properties+xml"/>
    <Override PartName="/world.xml" ContentType="application/atom+xml"/>
    <Default Extension="rels" ContentType="application/vnd.openxmlformats-
package.relationships+xml"/>
    <Default Extension="jpeg" ContentType="image/jpeg"/>
    <Default Extension="xml" ContentType="application/xml"/>
    <Default Extension="dbf" ContentType="application/x-dbf"/>
    <Default Extension="img" ContentType="application/x-img"/>
    <Default Extension="mmm" ContentType="application/x-mmm"/>
    <Default Extension="mxd" ContentType="application/x-mxd"/>
    <Default Extension="rel" ContentType="application/x-rel"/>
    <Default Extension="shp" ContentType="application/x-shp"/>
    <Default Extension="shx" ContentType="application/x-shx"/>
    <Default Extension="tif" ContentType="image/tiff"/>
    <Default Extension="tfw" ContentType="application/x-tfw"/>
</Types>
```

Fig. 4. [Content_Types].xml part reflecting the MIME types of the parts for the example (some elements hidden for clarity).

share some files (layers, palettes, thesaurus, etc) with other maps and a extreme case is a cartographic series where there is a common structure that is shared by all sheets composing the series (Morris and Tuttle, 2008). Furthermore, currently GIS is working on huge datasets that will experiment performance degradation if forced to work on the compressed package.

MiraMon implementation of the Geospatial OPC (called MMZX) is described next. The typical MiraMon entry part to the data is the map project.

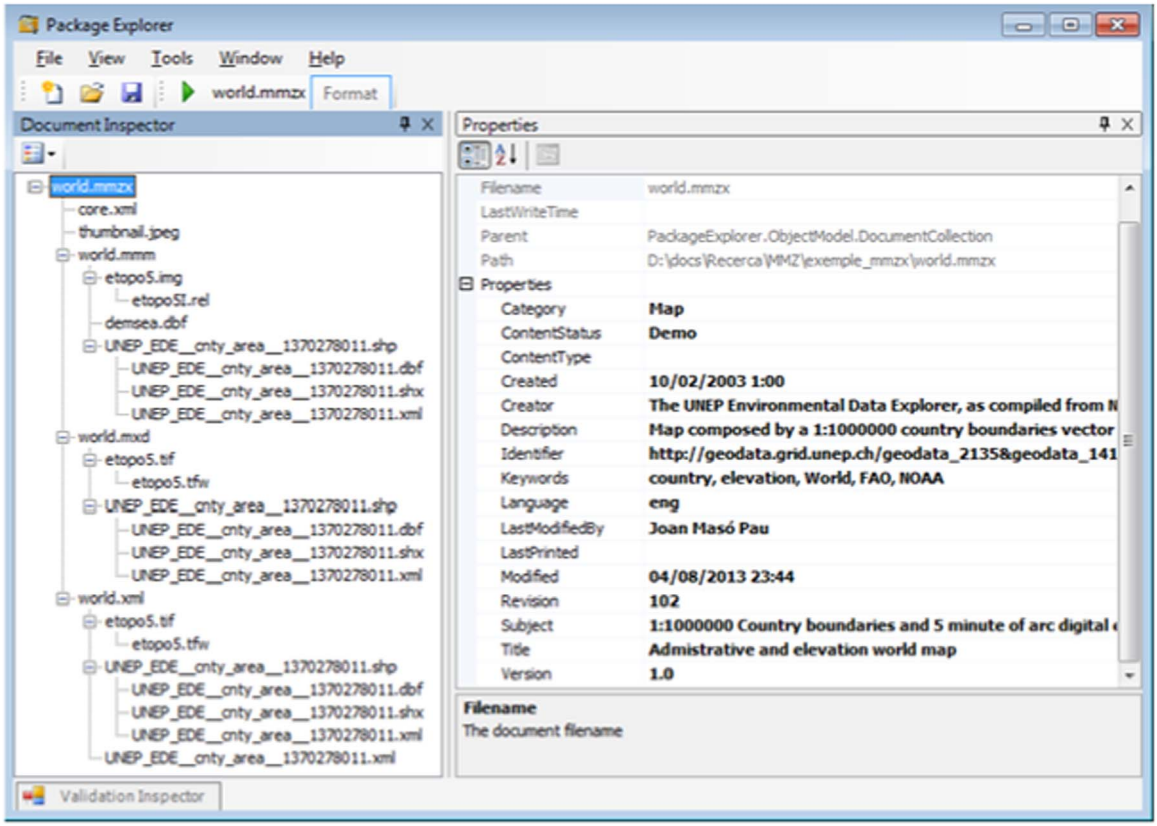

Fig. 5. Package Explorer 3.0 RC1 by Wouter van Vugt view showing the package relations structure of the use case example for the three software implementations and the reuse of elements.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ExternalReferences
xmlns="http://www.miramon.uab.cat/schemas/package/2013/externalreferences">
        <ExternalReference Id="url0" Date="2013-02-12T11:18:00"
            Url="http://ede.grid.unep.ch/download/etopo5_all_tif.zip"
            Part="etopo5_all_tif"/>
</ExternalReferences>
```

Fig. 6. etopo.tif.urls part that follows the proposed extension for ISO 29500-2. It enumerates Internet resources that can contain an updated version of the same resource that is internally stored as etopo.tif part.

4.1. Reference implementation MMZX creation and decompression tool

A reference tool for creating, decompressing and exploring the content of a package has been created and can be downloaded as a no cost executable for the Windows operating system (http://www.miramon.cat/USA/Prod-LectorUniversal.htm). In the creation mode, it receives a maps name and is able to generate a list of all files needed. To do so, it reads all parts containing hyperlinks, generates a copy of those files, converts internal references and paths to a URI compatible syntax (OPC requirement), and changes the internal path to a new path structure that can be decompressed without path regeneration. It also converts any databases and geodatabases into DBF files and Shapefiles, respectively. After creating all the extra parts, it submits the list to a standard ZIP compression library that will generate an OPC compliant package.

4.2. Opening an MMZX file in the reference GIS implementation

The GIS opens an MMZX file by decompressing it first in a temporary and private space, opening the suitable entry point. This procedure is invisible to the user. At the end of a session files that were decompressed are deleted.

During the setup of the software, MMZX extension is registered in the Windows registry, and MMZX is also registered as a trusted format in Internet browsers. This way, a single click in a web page that contains an MMZX link will open the file in the GIS automatically without user intervention.

For most users that explore or print the data and this schema works perfectly. Others edit the data, change symbolization or complement metadata. When the session is closed the software checks if parts have been modified in the temporary copy and the package is updated before freeing or deleting the temporary space.

4.3. Map integrity: certification

Certifying is useful to ensure integrity of contents of the MMZX. The OPC standard incorporates an internal certification approach that is able to certify authorship and guarantee integrity for the package if the producer is interested in this extra capability. In fact, other certification approaches can coexist in the MMZX format, such as the ones based on certifying each file having sensible data; by accompanying it with a certification file that contains a stream of bytes based on the original content and the certification entity encrypted by a proper algorithm.

5. Conclusions

There has been an absence of a general solution for a geospatial package that avoids the complexities of the geospatial data and extends the use of the GIS data to more general public or other professionals, and replaces less informative alternatives. This gap became more obvious in the first drafts of the newly proposed ISO 19165 “Geospatial data and metadata preservation” standard. We analyzed some geospatial file formats such as netCDF, KML and ESA SAFE; even when they allow for the integration of data and metadata in a single format, they require transforming data structures into their internal data models, making them not appropriate for data preservation. More generic multipart formats, such as MHTML and BagIt were studied but are not self-describing in terms of part relations.

This paper proposes a solution based on the Open Packaging Convention for a comprehensive preservation and exchange of real and heterogeneous multipart GIS data, while hiding the complexities of multipart structures.

Several current packaging formats have been analyzed. The Open Packaging Convention (OPC) offers several advantages compared to other alternatives. OPC is based on the consolidated ZIP compression format, –now extended to support files and contents larger than 4 Gbyte– and OPC libraries are available both as open source software (http://libopc.codeplex.com) for any operating system and integrated in modern Windows operating systems (e.g. in a COM based API https://msdn.microsoft.com/en-us/library/windows/desktop/dd742822.aspx). The proposed format adapts the OPC standard to the geospatial needs and includes several extensions: it permits keeping the separation between the geospatial data format elements (geometric data, attribute tables, symbolization, metadata, etc), supports different entry points (map projects), while allowing combined use of common data files. One of the proposed entry points is the new OGC OWS context document format: a standard map that also provides a way to combine geospatial data files with geospatial service accesses. The self contained package can be easily exchanged or linked on the Internet. An extension for ISO 29500-2 is also proposed, whereby external files can also be compressed in the package, for intended use in environments of no Internet connectivity (remote areas, disaster management), low bandwidth or long term data preservation.

An adaptation of the MiraMon software is also presented as a reference implementation for the new Geospatial OPC package. The original MMZ already had many of the needed properties for preservation and distribution, but it did not accept several entry points, it was not developed according to open specifications, etc. Lessons learnt in the past have been now rethought into an open solution that can be applied to other GIS software tools for creating managing and reading this multipart format, making preservation and distribution of spatial data open and standardized, easier, more flexible and robust.

Acknowledgements

This paper has been written thanks to the support of the European Commission through the H2020- 641538- ConnectinGEO (SC5-18a-2014) and the H2020- 641762- ECOPotential (SC5-16-2014), to a Grant to Consolidated Research Groups given by the Catalan Government (2014 SGR 1491) and to the Spanish Ministry of Economy and Competitiveness in cooperation with the European Regional Development Fund (ERDF) under Grant CGL2015-69888-P (ACAPI). Xavier Pons is recipient of an ICREA Academia Excellence in Research grant.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cageo.2016.09.001.

References

Adobe, 1992. TIFF Revision 6.0 Final — June 3 [online]. Available from: ⟨http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf⟩ (accessed 04.09.15).

Berners-Lee, T., 2006. Linked Data Design Issues. W3C.. [online]. Available from: ⟨http://www.w3.org/DesignIssues/LinkedData.html⟩ (accessed 14.07.15).

Bethune, A., Lazorchak, B., Nagy, Z., 2009. GeoMAPP: A Geospatial Multistate Archive and Preservation Partnership. J. Map Geogr. Libr.: Adv. Geospatial Inf., Collect. Arch. 6 (1), 45–56.

Borkar, V., Carey, M.J., Li, C., 2012. Inside "Big Data Management": Ogres, Onions, or Parfaits? In: Proceedings of the 15th International Conference on Extending Database Technology (EDBT). March 27–30.

Bowman, I.T., Godfrey, M.W., Holt, R.C., 2000. Connecting architecture reconstruction frameworks. Inf. Softw. Technol. 42, 91–102.

Brackin, R., Gonçalves, P., 2014. OGC OWS Context Atom Encoding Standard. OGC 12-084r2. Available from: ⟨https://portal.opengeospatial.org/files/?artifact_id=55183⟩ (accessed 04.09.15).

Bunting, P., Gillingham, S., 2013. The KEA image file format. Comput. Geosci. 57, 54–58.

CCSDS, 2002. CCSDS 650.0-M-2: Reference model for an Open Archival Information System (OAIS). Magenta Book. (Issue 1). January 2002. [online] Available from: ⟨http://public.ccsds.org/publications/archive/650×0m2.pdf⟩ (accessed 04.09.15).

Cervantes, D., 2009. Using GIS to Create an Interactive GeoPDF Mapbook for the Big Island of Hawaii. PHD thesis. [online]. Available from: ⟨http://www.nwmissouri.edu/library/theses/2009/CervantesDanielle.pdf⟩ (accessed 04.09.15).

Daisy, P., 2012. GeoPackage Encoding Standard – With corrigendum, Version 1.0.1. OGC 12-128r11 [online]. Available from: ⟨https://portal.opengeospatial.org/files/?artifact_id=63378⟩ (accessed 04.09.15).

Davis, J., Shur, A., 2007. A New Standard for Packaging Your Data. MSDN Magazine [online]. Available from: ⟨http://blogs.msdn.com/b/msdnmagazine/archive/2007/08/08/4277558.aspx⟩ (accessed 04.09.15).

Duerr, R.E., Cao, P., Crider, J., Folk, M., Lynnes, C., Yang, M.Q., 2009. Ensuring Long-Term Access to Remotely Sensed Data with Layout Maps.

ESRI, 1998. Shapefile technical description. An ESRI White Paper. July [online]. Available from: ⟨http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf⟩ (accessed 04.09.15).

Figueroa, J., Abergel, M.E., 2011. Capturing and combining media data and geodata in a composite file. US patent 7921114 B2.

Hoebelheinrich, N.J., 2012. An aid to analyzing the sustainability of commonly used geospatial formats: the library of congress sustainability website. J. Map Geogr. Libr.: Adv. Geospatial Inf. Collect. Arch. 8 (3), 242–263.

Holt, R.C., Winter, A., Schurr, A., 2000. GXL; Toward a Standard Exchange Format. Seventh Working Conference on Reverse Engineering (WCRE). Brisbane. November.

Horak, P., Charvat, K., Vlk, M., 2010. In: Papadopoulos, G.A., Wojtkowski, W., Wojtkowski, G., Wrycza, S., Zupancic, J. (Eds.), Web Tools for Geospatial Data Management. Information Systems Development. Springer, 793–800.

ISO 15836:2009 Information and documentation — The Dublin Core metadata element set [online]. Available from: ⟨http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52142⟩ (accessed 04.09.15)].

ISO29500-2, 2008. Information technology — Document description and processing languages — Office Open XML File Formats — Part 2: Open Packaging Conventions [online]. Available from: ⟨http://standards.iso.org/ittf/PubliclyAvailableStandards/c051459_ISOIEC%2029500-2_2008(E).zip⟩ (accessed 04.09.15).

Kim, J.W., Park, S.S., Kim, C.S., Lee, Y., 2004. The Efficient Web-Based Mobile GIS Service System through Reduction of Digital Map. Computations Science and Its Applications ICCSA. 2004In: Laganà, A. (Ed.), LNCS 3043. Springer-Verlag Berlin Heidelberg, 410–417.

Kraak, M.J., Ormeling, F., 2003. Cartography: Visualization of Geospatial Data 2nd ed. Longman Group, United Kingdom, 205.

Kunze, J., Littman, J., Madden, L., Summers, E., Boyko, A., Vargas, B., 2015. The BagIt File Packaging Format (V0.97). Network Working Group, Internet-Draft [online]. ⟨http://tools.ietf.org/html/draft-kunze-bagit-11⟩ (accessed 04.09.15).

Laurini, R., and Milleret-Raffort, F., 1990. Principles of geomatic hypermaps. Proc. 4th Intern. Symposium on Spatial Data handling, Zurich, 2, 642–655.

Masó, J.,Pons, X., 2011, The SDI of the data. MMZ. The de-facto standard for distributing multipart data integrated to OGC services. Jornadas Ibéricas de Infraestructuras de Datos Espaciales -JIIDE 2011, Barcelona. Document 52. [online] Available from: ⟨http://grumets.uab.cat/docs/JIIDE2011AbstractMMZ.doc⟩. (accessed 20.06.16).

McDonough, J.P., 2010. Packaging videogames for long-term preservation: integrating FRBR and the OAIS reference model. J. Am. Soc. Inf. Sci. Technol. 62 (1), 171–184.

Meissl, S., 2014. OGC® GML Application Schema - Coverages - GeoTIFF Coverage Encoding Profile, Version 1.0.0, OGC 12-100r1 [online] Available from: ⟨http://portal.opengeospatial.org/files/?artifact_id=54813⟩. (accessed 25.07.15).

Morris, S.P., 2009. The north carolina geospatial data archiving project: challenges and initial outcomes. J. Map Geogr. Libr.: Adv. Geospatial Inf. Collect. Arch. 6 (1), 26–44.

Morris, S.P., Tuttle, J., 2008. Curation and preservation of complex data: The North Carolina geospatial data archiving project. Proc. National Digital Information Infrastructure and Preservation Program (NDIIPP) Conference, (Lib. Congr)., 1–12.

NCSA, 2001. HDF Specification and Developer's Guide. Version 4.1r5. November 2001. [online] Available from: ⟨http://gis-lab.info/docs/hdf/SpecDevG.pdf⟩ (accessed 13.06.16).

Nottingham, M., Sayre, R., 2005. RFC 4287 The Atom Syndication Format. Network Working Group. The Internet Society. December [online]. Available from: ⟨https://tools.ietf.org/rfc/rfc4287.txt⟩ (accessed 04.09.15)

Palme, J., Hopmann, A., Shelness, N., 1999. MIME Encapsulation of Aggregate Documents, such as HTML (MHTML). IETF RFC 2557 [online]. Available from: ⟨http://tools.ietf.org/rfc/rfc2557.txt⟩ (accessed 04.09.15).

Perkins, R., 1995. File formats on the internet. Comput. Geosci. 21 (6), 775–777.

Phillips, A.W., Allemang, R.J., 2010. Requirements for a Long-term Viable, Archive Data Format. T. Proulx (ed.), Structural Dynamics, 3, Conference Proceedings of the Society for Experimental Mechanics Series, 12.

PKWARE, 2004. ZIP File Format Specification from PKWARE, Inc., version 6.2.0, [online]. Available from: ⟨http://www.pkware.com/documents/APPNOTE/APPNOTE_6.2.0.txt⟩ (accessed 04.09.15).

Pons, X., 2002. MiraMon. Geographic Information System and Remote Sensing software, Centre de Recerca Ecològica i Aplicacions Forestals, CREAF. Bellaterra. ISBN: 84-931323-5-7

Rew, R., Davis, G., Emmerson, S., Davies, H., Hartnett, E., Heimbigner, D., Fisher, W., 2016. Appendix B. File Format Specifications. NetCDF User guide, University Corporation for Atmospheric Research (UCAR). [online] Available from: ⟨http://www.unidata.ucar.edu/software/netcdf/docs/file_format_specifications.html⟩ (accessed 13.06.16)

Schut, P., 2007. OGC Web Processing Service (WPS), Version 1.0.0, OGC 05-007r7 [online], Available from: ⟨http://portal.opengeospatial.org/files/?artifact_id=24151⟩ (accessed 04.09.15)

St-Denis, G., Schauer, R., Keller, R.K., 2000. Selecting a Model Interchange Format: the SPOOL Case Study. Proceedings of the 33rd Hawaii International Conference on System Sciences. Island of Maui, (January).

Townsend, J.A., Downing, J., Murray-Rust, P., 2009. CHIC – Converting Hamburgers Into Cows, Fifth IEEE International Conference on e-Science, (December), 9–11.

UNEP, 2013. The UNEP Environmental Data Explorer, as compiled from National Oceanic and Atmospheric Administration (NOAA) and National Geophysical Data Center (NGDC). United Nations Environment Programme [online]. ⟨http://geodata.grid.unep.ch⟩. (accessed 04.09.15).

Vilches-Blázquez, L.M., Villazón-Terrazas, B., Corcho, O., Gómez-Pérez, A., 2014. Integrating geographical information in the Linked Digital Earth. Int. J. Digit. Earth 7 (7), 554–575.

Waugh, A., 2006. The design of the VERS encapsulated object experience with an archival information package. Int. J. Digit. Libr. 6 (2), 184–191. http://dx.doi.org/10.1007/s00799-005-0135-y.

Wessel, P., 2003. Compression of large data grids for Internet transmission. Comput. Geosci. 29, 665–671.

Wilson, T., 2008. OGC KML. Version 2.2.0, OGC 07-147r2 [online], Available from: ⟨http://portal.opengeospatial.org/files/?artifact_id=27810⟩ (accessed 04.09.15).