Technology Watch Report

Preserving Geospatial Data

Guy McGarva
EDINA, University of Edinburgh

Steve Morris
North Carolina State University (NCSU)

Greg Janée
University of California, Santa Barbara (UCSB)

DPC Technology Watch Series Report 09-01
May 2009

Executive Summary:
Geospatial data are becoming an increasingly important component in decision making processes and planning efforts across a broad range of industries and information sectors. The amount and variety of data is rapidly increasing and, while much of this data is at risk of being lost or becoming unusable, there is a growing recognition of the importance of being able to access historical geospatial data, now and in the future, in order to be able to examine social, environmental and economic processes and changes that occur over time.

The geospatial domain is characterized by a broad range of information types, including geographic information systems data, remote sensing imagery, three-dimensional representations and other location-based information. The scope of this report is limited to two-dimensional geospatial data and data that would typically be considered comparable to paper maps or charts including vector data, raster data and spatial databases.

There are a number of significant preservation issues that relate specifically to geospatial data, including:

- the complexity and variety of data formats and structures;
- the abundance of content that exists in proprietary formats;
- the need to maintain the technical and social contexts in which the data exists;
- and the growing importance of web services and dynamic (and ephemeral) data.

Standards for geospatial metadata have been defined at both the national and international levels, yet metadata often becomes dissociated from the data, or is incorrect, non-standard in nature, or not created in the first place. Additional considerations to be taken into account in preserving geospatial data include coordinate reference systems, cartographic representations, topology, project files and data packaging.

Standards bodies are in place at the national and international levels to address general geospatial data standardization issues, yet working groups addressing preservation issues have only recently been formed. A number of technologies and tools that are, or may be, of relevance to geospatial data preservation efforts have emerged, although the nature of the problem is such that there is not a single tool or technology that will be relevant in all cases.

A number of projects and activities have been addressing various aspects of geospatial data preservation, creating an initial body of experience from which some initial recommendations can be made. While these recommendations provide a basic checklist of issues to be considered when preserving geospatial data, it must be emphasized that the collective experience in preserving such data is still very much in an early stage and that further investigations are needed.

Keywords:
Geographic Information Systems, geospatial data, preservation, spatial databases, geospatial formats, web mapping services

Contents:

1 Introduction: why preserve geospatial data?

Several influential reports have recently been produced highlighting the current and future importance of geospatial data. In the US the National Geospatial Advisory Committee produced the report "The Changing Geospatial Landscape"[1] which provides a history of the developments in the geospatial industry and possible future directions. In the UK the "Place matters: the Location Strategy for the United Kingdom"[2] report highlighted the importance of geographic information to the UK. Although not directly addressing geospatial data preservation, these reports show the increased awareness and value of geospatial data to a wide range of users.

Geospatial data inherits the preservation challenges inherent to all digital information. This report does not attempt to address the more general aspects of digital preservation. Focus is instead given to significant issues and technologies which relate specifically to preservation and management of geospatial data.

The geospatial domain is characterized by a broad range of information types, with content types such as geographic information systems data and remote sensing imagery increasingly being complemented by three-dimensional representations and other location-based information. The scope of this report is limited to two-dimensional geospatial data and data that would typically be considered comparable to paper maps or charts but which may be supplied in various forms including raster, vector, database or dynamically through web services.

This report is mainly aimed at repository and archive managers and digital preservation specialists who are increasingly dealing with geospatial data; however it will also be useful to geospatial data specialists, academics and researchers who are becoming involved with the preservation challenges of geospatial data.

2 Background: key challenges with geospatial data

Geospatial data, also termed 'geographic information' or 'spatial data' depending on the context, can be defined as data that describe features on the earth. Typically datasets such as transportation networks, property boundaries, coastlines, aerial imagery, or terrain models can all be considered to be geospatial data.

When discussing geospatial data it is useful to have an understanding of Geographic Information Systems (GIS). There are many textbooks[3] available that can provide in-depth information and online resources such as the GIS Files[4] from Ordnance Survey can give a brief introduction to the subject. GIS are the software environments that are commonly used to create, visualise, edit and analyse geospatial data. There are a wide variety of GIS available from commercial suppliers as well as open source projects. They range from simple, general purpose, web-based clients to large, highly complex,

[1] http://www.fgdc.gov/ngac/NGAC%20Report%20-%20The%20Changing%20Geospatial%20Landscape.pdf
[2] http://www.communities.gov.uk/publications/communities/locationstrategy
[3] For example: Geographic Information Systems and Science
by Paul Longley, Michael F. Goodchild, David Maguire, David Rhind (Wiley, 2001) or An Introduction to Geographical Information Systems
by Ian Heywood, Sarah Cornelius, Steve Carver (Prentice Hall, 2006)
[4] http://www.ordnancesurvey.co.uk/oswebsite/gisfiles/

integrated systems aimed at specific application domains. The sheer variety of types of geospatial data, taken together with the variety of GIS used to manipulate that data, is one of the main factors in making the preservation of geospatial data a highly complex issue.

GIS applications excel at conflating (or combining) data from multiple sources including data with different accuracy levels and based on different geographic referencing systems, for example combining terrain models, aerial imagery and transportation networks, regardless of whether these combinations are compatible or whether the results obtained are warranted. For these reasons it is important that key metadata such as coordinate referencing systems and accuracy measures are recorded and preserved along with the data.

Additional preservation risk arises from the inherent nature of geospatial data itself:

- Geospatial data spans a wide variety of data structures: vector and raster; unstructured and topological; over domains both discrete and continuous. Geospatial applications and data formats support differing subsets and aspects of these data structures, and to varying degrees. Attempts at defining a universal data model for geospatial data have been made (for example the Spatial Data Transfer Standard (SDTS)[5]) but have not achieved widespread adoption. As a consequence, it is not possible to speak of "geospatial data" as a single type of information that can be handled by multiple, functionally equivalent applications and formats.

- In contrast to textual information, which can be successfully modelled using multi-page (hyper) textual documents as the sole granule size, geospatial data are regularly processed at varying levels of granularity. The granule sizes range from individual features having geographic location, geometry, and related attributes; to homogeneous, thematic layers of features; to integrated, heterogeneous databases. Data can be aggregated, disaggregated, and operated on with fluidity. Each of these granule sizes has its uses, affords different functionalities, and poses different preservation challenges. As a consequence, there is no *single* preservation problem for geospatial data; instead, choosing which level or levels of granularity to address, and therefore identifying the preservation problem(s), is a first step in the process.

- Many, if not most, geospatial formats are proprietary and therefore closely tied to applications, and are frequently subject to backwardly incompatible revisions over time.

The net result of these characteristics is that, today, there is no single, easy or universal solution to the problem of preservation of geospatial data. There are many formats and applications, all of which have overlapping but different capabilities. Because conversion of geospatial data across formats, data structures, and applications often results in loss of data or data alteration, the migration of geospatial formats over time is not easily automated, but instead must be evaluated on a case-by-case basis.

[5] http://mcmcweb.er.usgs.gov/sdts/whatsdts.html

Furthermore, the general preservation problem for geospatial data will simply compound over time with increasing quantities of data being produced by collection systems such as satellites and sensor networks. Historical geospatial data is of great value in understanding and modelling climate and land use change, for example, and hence future users and archivists are likely to want to use and curate increasing quantities of increasingly older geospatial data.

3 Geospatial Data Preservation Issues

The following sections identify and describe in more detail some of the main issues related to the preservation of geospatial data. There are two principal data types associated with geospatial data: vector data which has similarities to Computer Aided Design (CAD) data; and raster data which has similarities to image data. These and a number of emerging data types are introduced in the sections that follow but it is worth noting that most GIS will incorporate both forms of data to a greater or lesser extent. In fact it is in the interaction of different data types that geospatial data finds its greatest value.

3.1 Generic Geospatial Data Issues

This initial section deals with aspects of geospatial data that are common to different types of geospatial data including: coordinate reference systems which define how locations on the earth are described; cartographic representations of data (the equivalent of a paper map); the topological model used to represent vector data relationships; the formats used to define project files used by different systems; and how geospatial data can be 'packaged' so that all elements required for a dataset can be tied together.

3.1.1 Coordinate Reference Systems

A coordinate reference system is a means of identifying the location of features on the earth by coordinates, for instance WGS 84 Latitude/Longitude values or National Grid easting/northing values. A coordinate reference system comprises a number of elements including a spheroid, datum and projection (in the case of a projected coordinate system). Knowledge of the coordinate reference system is important for the accurate use of geospatial data.

Some geospatial data formats (both raster and vector) directly contain information about the coordinate reference system the dataset is based on, either embedded as part of the file itself (e.g. a GeoTIFF[6] file) or as an additional, tightly-bundled file (e.g. a ".prj" projection metadata file[7] stored as part of an ESRI Shapefile). However, some other formats that are spatially referenced do not directly contain coordinate reference system information, for example plain TIFF image files. A plain TIFF file may have an associated TIFF World File[8] (TFW), but the association is loose and prone to breakage.

[6] http://www.remotesensing.org/geotiff/spec/contents.html – see section on Coordinate Systems
[7] ESRI WKT as used in .prj file and used by ESRI Projection Engine is described at: http://support.esri.com/index.cfm?fa=knowledgebase.techArticles.articleShow&d=14056
[8] http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?id=3046&pid=3034&topicname=World_files_for_raster_datasets

There are several ways of specifying a coordinate reference system including using EPSG[9] codes or OGC Well Known Text Representation for Spatial Reference Systems (WKT)[10]. Regardless of which format is used to specify a coordinate reference system, this information should be included in the dataset's metadata record, especially if the format that the dataset is in does not explicitly provide this information in a readily accessible way. It should be noted that TFW files on their own do not specify a coordinate reference system. They should be accompanied by information (such as WKT or an EPSG code) that specifies the coordinate reference system.

Metadata profiles that have been designed to describe geospatial data will generally include an element to describe the coordinate reference system (if applicable), however in some cases this may be included as part of another element. It should also be noted if the coordinate reference system applies both to the metadata (e.g. bounding box coordinates) as well as to the actual dataset.

3.1.2 Cartographic Representation

Cartographic representations are a common output from geospatial data, often taking the form of simple digital maps in image format. Such maps may in some cases be geo-referenced to permit use of the map as an overlay in geospatial applications. Cartographic output may also take the form of more complex documents such as PDF or GeoPDF files which support more advanced options for user interaction with the resulting digital map. These finished information products typically don't include the actual data that was used to make the document, though some formats, such as PDF or GeoPDF, support the retention of some amount of data intelligence derived from the original data.

Preservation of data and preservation of documents derived from the data comprise two separate and non-exclusive objectives. The data itself must be saved to allow for re-creation of earlier analyses or to engage the data in some new project work. The finished information product, whether a map, chart, or other output, is an information product that is very different from the data and includes synthesized information that is not a part of the underlying data. Decision-makers often interact with end product representations rather than directly with the underlying data, bolstering the importance of these end products as records for preservation.

3.1.3 Topology

Topology is the spatial relationship between features such as their connectivity (e.g. a road network) or adjacency (e.g. countries sharing a boundary). Vector datasets can model these relationships in different ways depending on the software being used and the data model being implemented. Topological data models can range from what are termed 'spaghetti' or 'unstructured' datasets where there are no explicit relationships between vector features, to 'fully' or 'structured' topological datasets in which every feature has a relationship. The preservation issue here is that topology information is often stored in proprietary vector formats and data structures, and any conversion

[9] http://www.epsg.org/ and the online EPSG Geodetic Parameter Registry: http://www.epsg-registry.org/

[9] For a description of the OGC WKT format see the OGC specifications for Simple Feature Access – Part 1: Common Architecture http://www.opengeospatial.org/standards/sfa

process or data transfer may result in a loss of information. It is thus important when looking to preserve data that contains such information to identify and pay particular attention to the data structures that the target data format supports.

3.1.4 Project Files

GIS software "project files" are complex digital documents that tie together a wide variety of components including: data, instructions on how the data will be presented, metadata, data models, scripts and other ancillary components. One typical feature of a project file is some manner of data view in which a combination of data is presented in a tailored manner that involves classification, symbolization, and annotation based upon the data content. These data views typically appear as maps, charts, or tables, or some combination thereof. In order for an end user to render this content it is necessary to have the project file, the software that supports the project file, related components (possibly including software add-ons or extensions), as well as the actual data. The required use of specific software, the complexity of the project file formats, and the tenuous links to the actual data, which is often simply pointed to, put these project files at high risk for failure over time.

Examples of project files include the ESRI ArcView .apr file, the ESRI ArcGIS .mxd file, and the MapInfo Workspace (.wor). It should be noted however, that simply preserving a project file does not preserve the underlying data or auxiliary files that are needed to display and use the data.

There is a growing recognition in the GIS community of the need to be able to archive not just data but also projects and their various components in order to preserve the ability to revisit how different data processes and analyses were carried out. Yet project file incompatibilities with software upgrades point to possible future preservation challenges in maintaining this content, and vendor commitment to forward compatibility of current project files with future software releases remains unclear.

3.1.5 Data Packaging

Geospatial data frequently consists of complex, multi-file, multi-format objects, including one or more data files as well as: geo-referencing files, metadata files, licensing information, and other ancillary documentation or supporting files. The absence of a standard scheme for content packaging can make transfer and management of these complex data objects difficult both for archives and for users of the data. Some other information industries have complex XML-based wrapper formats or content packaging standards, including METS (digital libraries), MPEG 21 DIDL (multimedia), XFDU (space data), and IMS-CP (learning technologies), yet no similar activity has occurred in the geospatial industry.

In practice, in the geospatial community archive formats such as Zip[11] commonly function as rudimentary content packages for multi-file datasets or groups of related datasets. Such archive files typically lack data intelligence about file relationships and functions within the data bundle. In some cases formalized approaches to the use of Zip files are beginning to appear, for example KMZ files that are used to package

[11] http://www.pkware.com/support/zip-application-note

KML files and their ancillary components. The Metadata Exchange Format (MEF)[12], developed for use in the open source GeoNetwork[13] catalogue environment, uses Zip as the basis for a formalized packaging of geospatial metadata as well as associated data and ancillary components. MEF, which is explicit in packaging of metadata but non-explicit in packaging of the actual data and ancillary components, might provide a starting point for exploration of geospatial data packaging solutions.

3.2 Vector Data

A common form of geospatial data is vector data, which models features on the earth's surface as points, lines, and polygons. Attribute data is often associated with vector data, carrying values for individual characteristics of the data features. For example, a line section representing a portion of a street might have attribute information for characteristics like "street name", "number of lanes", "speed limit" etc. Attribute data may either be stored directly within the vector dataset or stored externally in a spreadsheet or database.

Changes in vector data

Real world features that are represented by vector data are typically subject to change and the corresponding data may be updated accordingly. The updated dataset typically replaces the previous version and, unless a snapshot of that earlier dataset is set aside and archived, it becomes impossible to look at historical changes in the data.

In some cases the dataset itself may be designed to store the historical changes within the active dataset. In some spatial databases, previous versions of a vector dataset may be stored along with the current dataset but with a different date attribute. There is also the possibility that only changes to features in a dataset are provided and the receiving system is required to apply updates to the dataset.

Vector formats

Geospatial vector data formats tend to be specific to the geospatial industry. These formats can be highly complex and are extremely sensitive to both format migration and software environment changes. The absence of vector data formats that are both non-commercial and widely supported has led to a preponderance of vector data that is available only in commercial or proprietary formats.

3.2.1 Commercial Vector Data Formats

A range of commercial vector data formats exists, each of which is most directly associated with a particular commercial software environment. Options for conversion between common commercial formats exist as built-in features within desktop GIS software, as a function provided by open source conversion tools such as Geospatial Data Abstraction Library (GDAL/OGR)[14] or as a service provided by specialized commercial tools and services that focus on data conversion such as the Feature Manipulation Engine (FME) from Safe Software[15]. Due to the complexity of the data, migration from a proprietary or poorly-supported data format into another more preservation-friendly format can lead to unacceptable distortion or loss of data.

[12] http://www.fao.org/geonetwork/docs/Manual.pdf
[13] http://www.fao.org/geonetwork/srv/en/main.home
[14] http://www.gdal.org/
[15] http://www.safe.com/

Shapefiles
One commercial format in particular, the ESRI Shapefile[16], has come into wide use as a distribution format that is supported by a range of both commercial and open source tools. Since Shapefiles do not support advanced features such as topology (the spatial relationships between features), they have a simple data structure and lend themselves to rapid drawing and analysis. While the Shapefile format is owned by ESRI[17], it is openly documented, making it feasible to support Shapefiles in a variety of software tools.

A Shapefile consists of at a minimum three files, a .shp file (feature geometry), an .shx file (index of the feature geography), and a .dbf file (a dBASE database file that stores the attribute information of the features). Additional files can also be included, including projection files (.prj), metadata files (.xml) and spatial index files (.sbx and .sbn). Although the Shapefile format is by today's standards 'old' it is still widely used and supported.

In GIS operations that use ESRI software it is increasingly common for vector data to be created and managed within the ESRI Geodatabase spatial database format, with Shapefile versions of the data used for distribution and access outside of the maintaining organization. Since the ESRI Geodatabase format is proprietary and not openly documented, many commercial and open source software tools still use the Shapefile for both data management and data distribution.

Coverage Files
The ESRI coverage file[18] is a proprietary vector data format that preceded the Shapefile. Coverage files include some information, such as topology and annotation, which is not directly transferred to Shapefiles in data conversions. Coverage files are more awkward to manage than Shapefiles since coverages have a multi-file, multi-directory structure that makes the data susceptible to corruption in data transfers. The .e00 format is a coverage export format that can more easily be transferred because it can be contained in a single file. However, the specifications for these formats are not publicly available.

Other Commercial Vector Formats
A wide range of additional commercial vector formats are available and in use. Notable examples include MapInfo's[19] TAB and MIF/MID formats and Autodesk's[20] DXF/DWG formats. These tend to be used in specialist markets such as for business market analysis in the case of MapInfo or planning and design for AutoCAD.

The MapInfo MIF/MID[21] format is a relatively simple published format with the graphics stored in the MIF file and attributes in the MID file. To use MIF/MID files in MapInfo they need to be imported and converted to TAB files. The MapInfo TAB format is the native format used by MapInfo and allows data to be read directly. The TAB format is proprietary to MapInfo and is a logical file made up of a number of

[16] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
[17] http://www.esri.com/
[18] http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=What_is_a_coverage
[19] http://www.mapinfo.com/
[20] http://www.autodesk.com/
[21] http://www.directionsmag.com/mapinfo-l/mif/AppJ.pdf

different physical files with different file extensions (in the same way that an ESRI Shapefile is composed of a number of files).

The Autodesk DXF (Drawing eXchange Format) is a common vector format used for geospatial data, primarily in CAD environments. There are many versions of the format in use, some dating back many years. It is a proprietary format controlled by Autodesk; however the latest DXF reference documentation[22] is available from the Autodesk web site. Documentation for some previous versions[23] can also be found on the Autodesk website. The DWG format is a binary, proprietary format maintained by Autodesk but also used in a number of other software systems. There is also an organisation called the Open Design Alliance[24] that provides software libraries to read and write DWG and publishes an open version of the specification.

3.2.2 Open Vector Data Formats

While commercial vector data formats may dominate the global market space, there are a number of "open" options for vector data creation, management, and distribution.

SDTS

An outcome of an early effort to define a means for open exchange of data was the Spatial Data Transfer Standard (SDTS)[25], which was created in the 1990s by the U.S. Geological Survey (USGS) to support exchange of geospatial data. While SDTS was used extensively by USGS and other US government agencies to distribute vector as well as raster digital elevation model data, the format has not gained traction in the wider geospatial data community.

GML

Geography Markup Language (GML)[26] is a standard first introduced in 2000 by the Open Geospatial Consortium (OGC)[27] and subsequently published as ISO standard 19136. The GML specification declares a large number of elements and attributes intended to support a wide range of capabilities. Since the scope of GML is so wide, profiles of GML that deal with a restricted subset of GML capabilities have been created in order to encourage interoperability within specific domains that share those profiles. While GML can be used for handling file-based data, it has wider use in web services-oriented environments.

While GML would appear to provide a promising alternative for data preservation, there are a number of complicating factors. GML is not so much a single format as it is an XML language for which there are a wide range of different community implementations as embodied by specific GML profiles associated with specific GML versions, and for which different application schemas might be available.

The GML specification is highly complex, and that complexity, combined with the diversity of profiles and application schemas, can present a barrier to vendor and tool

[22] http://usa.autodesk.com/adsk/servlet/ps/item?siteID=123112&id=2882295&linkID=9240617
[23] http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=12272454&linkID=10809853
[24] http://www.opendwg.org/
[25] http://mcmcweb.er.usgs.gov/sdts/whatsdts.html
[26] http://www.opengeospatial.org/standards/gml
[27] http://www.opengeospatial.org/

support. In light of these problems, in 2006 the OGC released the Simple Features Profile which, as a constrained set of GML, was designed to lower the barrier to implementation. While the Simple Features Profile might provide the basis for creation of a supportable archival profile of GML, something roughly analogous to PDF/A[28], there would still be the question of quality and functionality tradeoffs, including data loss that might comprise the cost of transferring data into a sustainable GML-based archival format.

Prominent examples of national GML implementations are UK Ordnance Survey MasterMap, based on GML 2.1.2, and TIGER/GML, which has been in development in the U.S. for use with census geography datasets released by the Census Bureau. A notable domain implementation of GML is CityGML[29], implemented as an application schema for the representation, storage and exchange of virtual 3D city and landscape models. Each different GML implementation will raise its own preservation challenges in terms of schema evolution, ongoing tool support, and dependencies on any data resources or content that might be externally referenced.

NTF

The format is officially British standard BS 7567 "Electronic transfer of geographic information (NTF)"[30] and is primarily used by Ordnance Survey (UK)[31]. NTF defines a number of levels of differing complexity that support different types of features and data, from Level 1 for simple vector features, to Level 5 which allows users to define their own data model and is used to transfer data such as Digital Terrain Models (DTM). Although still used by Ordnance Survey for a number of products, newer products such as OS MasterMap vector layers are supplied in GML format only. The use of NTF outside of Ordnance Survey is limited and is mostly used by consumers who typically convert the data into other formats for use in their GIS.

OS MasterMap® GML

Ordnance Survey uses GML as the data format for the transfer of its OS MasterMap[32] product (except for the Imagery Layer which is supplied in common raster formats). The OS MasterMap Topography Layer[33] is a large-scale continuous dataset covering all of Britain and is continually maintained and updated. A key aspect of OS MasterMap is the ability to supply Change Only Updates (COU) to users. Using COU, only features that have changed since a specified date are supplied to a user. It is then up to the user's GIS to process these COU and apply them to a data holding.

Another key feature of OS MasterMap is the ability to associate and integrate a user's own data with features in OS MasterMap using the TOID® (a unique 16-digit string identifier) as a reference. As features in OS MasterMap can be updated and modified without changing their TOID, it is necessary in some circumstances to know and access not only the TOID but the specific version of the TOID (and by implication the version of the dataset) that is being referred to. For future use it will be necessary for organisations to ensure they have sufficient archives of their own OS MasterMap

[28] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51502
[29] http://www.citygml.org/
[30] http://www.standardsuk.com/shop/products_view.php?prod=6534
[31] http://www.ordnancesurvey.co.uk/
[32] http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/
[33] http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/layers/topography/index.html

data, at sufficiently frequent intervals, so that they can recreate the correct associations with their own data.

3.3 Raster Data

Raster geospatial data is organized on a regularly-spaced, multidimensional grid of cells or lattice points. Such data can generally be characterized by the number of dimensions (most often two, but occasionally three); the number of bands (i.e. the number of coincident layers); and the data type of the cell values in each layer (whether drawn from a continuous or discrete domain, or categorical in nature). For continuous and discrete data types, the data can be further characterized by the range (or "depth") of the data values. For categorical data types, the category semantics (at minimum, the category labels such as "desert" or "ocean") are critical metadata without which the data loses meaning. Depending on the file format, such metadata may be stored with the raster data, in a separate data dictionary, or along with external metadata. Topology is not a concern with raster data since the relationships between cells are inherent in the raster itself.

Raster data is closely related to image data and many of the issues associated with the use and preservation of images[34] pertain also to raster geospatial data. In some cases raster data *is* imagery, that is, the raster bands contain colorimetric information such as visible wavelength radiances. In other cases, such as with elevation or bathymetric data and many other examples that have no direct visual interpretation, image formats nevertheless provide a natural and convenient way to store such raster data. As a consequence, many of the issues that arise with the preservation of image data apply to geospatial raster data as well. The evaluation factors identified[35] for still images—clarity (resolution and bit depth) and colour maintenance—apply to raster data as well, though colour maintenance generalizes here to maintenance of the semantics of the data.

3.3.1 Georeferencing and Rectification

As with vector data, an issue in preserving geospatial raster data is the need to maintain coordinate reference system information. Because raster data has a regular organization, it is sufficient to describe the geospatial reference of the raster grid only, and not individual data points or features therein.

Raster data may undergo a process of georeferencing and rectification to bring the data into a known coordinate system (projection and datum). For imagery data (e.g. satellite or aerial photography) a further process of ortho-rectification corrects for scale differences due to surface topography and requires a digital elevation model. In the latter case, since the accuracy of the data is dependent on the accuracy of the elevation model, the source elevation model and ortho-rectification software should be recorded as part of the data's lineage.

3.3.2 Compression

An issue that arises in preserving geospatial raster data specifically is sensitivity to lossy compression. Lossy image compression techniques subtly change data values.

---

[34] See the JISC Digital Media for resources and advice on still images http://www.jiscdigitalmedia.ac.uk/stillimages/

[35] http://www.digitalpreservation.gov/formats/content/still_quality.shtml

For example, JPEG changes data values in ways that are not readily noticeable to human vision because the changes are designed to exploit limitations and characteristics of human vision. As a consequence, formats such as JPEG are most suitable for images intended for human consumption. However, such changes may be very significant to analytic functions the data is intended to support. As a general rule, if the data is to support analysis, only lossless compression should be used.

3.3.3 Raster Formats

As with vector data, there are a number of formats in common use for raster data.

Simple Raster Formats

A number of simple raster formats, some dating back to the days when data was read directly from tape drives, remain in active use today. BIL (band interleaved by line), BIP (band interleaved by pixel), and BSQ (band sequential) are formats for multi-band raster data, though it would be more accurate to describe these as generic data organization techniques that can be employed by formats. For example, colour USGS digital orthophotos were initially organized as BIP, divided into fixed-length records with an ASCII header (the USGS has since switched to GeoTIFF).

Arc/Info ASCII GRID[36] and USGS DEM[37] are simple, open, ASCII formats for single-band raster data. Each simply lists raster cell values in left-to-right and top-to-bottom order, augmented with georeferencing information in the header and/or trailer records. These formats still find use in converting and processing raster data.

From a preservation perspective, these simple raster formats pose little curation difficulty due to their open standards, widespread support, and ease of transformability.

More Complex Raster Formats

TIFF[38] has emerged as a common format for storing and delivering raster data owing to its open standard (the standard is controlled by Adobe, but openly published and not subject to license), its flexibility in describing multiple bands and data types, its extensible framework for embedded metadata ("tags"), and its popularity in the desktop publishing world. TIFF itself defines the semantics of a few tags; GeoTIFF[39] is an open standard that defines additional tags applicable to geospatial raster data, including complete coordinate reference information.

JPEG 2000[40] is a relatively new standard that supports progressive, wavelet-based compression. It offers a wealth of other features, including lossy and lossless compression techniques, selective and adaptive compression, etc. JPEG 2000 also allows arbitrary XML metadata to be embedded in image files, and the OGC has defined a standard for embedding GML documents in JPEG 2000[41]. By exploiting the full capabilities of GML, this opens up the possibility of embedding in image fields not just coordinate reference system information, but also coverage metadata,

[36] http://docs.codehaus.org/display/GEOTOOLS/ArcInfo+ASCII+Grid+format
[37] http://rmmcweb.cr.usgs.gov/nmpstds/demstds.html
[38] http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf
[39] http://trac.osgeo.org/geotiff/
[40] http://www.jpeg.org/jpeg2000/
[41] http://www.opengeospatial.org/standards/gmljp2

annotations, and even vector features. More detailed information about JPEG 2000 from a preservation point of view can be found in the relevant DPC Technology Watch Report[42].

A number of proprietary formats have also emerged for handling very large geospatial imagery datasets including ECW[43] from ER Mapper (now part of ERDAS) and MrSID[44] from LizardTech, which use wavelet compression methods to reduce file sizes.

Support for JPEG 2000 is increasing, but today GeoTIFF is arguably the most survivable format for geospatial raster data due to its widespread use and support.

3.3.4 Mosaicked Raster Data

Raster data is often used to represent continuous phenomena (e.g., surface elevation), but for convenience of data management and delivery it is packaged into fixed-size tiles divided along arbitrary tile boundaries. This means that it is often desirable to mosaic the tiles back together into a seamless whole, and to thereby allow users to browse and crop out just the portion of the entire dataset that is of interest to them.

The ramifications of mosaicking for preservation purposes depend greatly on the implementation specifics. If the raster tiles are stored as files in a filesystem, for example as GeoTIFFs, each independently carrying metadata and georeferencing information, and if the mosaicking system is entirely automated, then the preservation problem may be no more difficult than the problem of preserving a collection of files. In this case, the mosaic can be viewed purely as an access mechanism. Preservation of the raster tile files alone is sufficient to recreate the mosaic in the future, but only if the coordinates of the tiles are preserved.

However, sophisticated mosaicking systems often perform edge alignment and colour balancing across tile boundaries, and even allow for fully manual and/or manually directed adjustments. In this case, the mosaicked image may effectively become a new data product derived from source raster tiles, and as such it may merit preservation independent of the source tiles.

3.3.5 Stereo, Oblique and Ground-Level Imagery

Our discussion of raster data so far has focused on data that can serve as a representation of the Earth's surface, and hence is suitable for projection and layering. Stereo and oblique imagery are types of imagery that are captured at varying angles to the vertical, and are used to create stereo pair images and 3D models. Such imagery requires additional metadata to describe the 3D spatial orientation of the images.

Ground-level photographs are not suitable for projection, but they can be point georeferenced. The EXIF[45] metadata standard defines a means of capturing coordinate reference system information in JPEG files, and is compatible with GPS systems that are often the source of such metadata.

[42] http://www.dpconline.org/docs/reports/dpctw08-01.pdf
[43] http://www.erdas.com/tabid/84/currentid/1142/default.aspx
[44] http://www.lizardtech.com/
[45] http://www.digicamsoft.com/exif22/exif22/html/exif22_1.htm

3.3.6 Raster Data Size

Because raster data is continuous over an area it can require several orders of magnitude more storage than the equivalent area represented through a vector data source. There is little that can be done about this. While it is possible to convert data from raster to vector representation (and vice versa), doing so is a highly analytic, lossy process that changes the essential character and functionality of the data. Thus, raster data generally must be preserved as raster data.

Compounding the problem of size is that automatic capture methods such as digital aerial photography and satellite remote sensing make it possible to quickly amass volumes of raster data that are large by any measure: MODIS[46], for example, acquires a terabyte of imagery per day. Raster data access mechanisms may impose additional storage requirements. Image tile pyramids that support efficient panning and zooming of large images add at least 30% to the data size.

As a consequence, in comparing raster data to vector, preservation of raster data is a quantitatively larger problem to such a degree that it is a *qualitatively* different problem. Large raster datasets will generally require custom engineered storage and processing systems. If raster data is stored in a spatial database the preservation problems due to size may compound the inherent migration and snapshot problems of preserving spatial databases.

3.4 Emerging Data Formats

Additional geospatial data formats are used for data representation, data visualization, and as network payloads occurring within web-based transfers of information. A number of new formats such as KML, which is used for geographic visualization, annotation, and navigation, and GeoRSS[47], which is used for geographically enabling RSS and Atom feeds, have emerged. These have especially found use in 'Neogeography'[48] applications. These formats might not be used in the creation or management of geospatial information; rather data files occurring in these formats are often created by transforming existing geospatial data. Data in some of these formats might not be obvious targets for archival acquisition since the original data will tend to be more complete. Yet the manner in which such data is represented in visualization environments may be of importance in recording how information has been shown and to record the basis for decision-making.

3.4.1 KML

KML[49], formerly known as Keyhole Markup Language, is an XML language focused on geographic visualization, including annotation of maps or images in digital globe or mapping environments. KML was initially used solely within Google Earth[50] but is now used in a range of software environments, and in April 2008 KML version 2.2 was approved as an international implementation standard by the OGC. KML

[46] http://modis.gsfc.nasa.gov/
[47] http://georss.org/
[48] In, Introduction to Neogeography, by Andrew J. Turner, O'Reilly 2006, Neogeography is described as "'new geography' and consists of a set of techniques and tools that fall outside the realm of traditional GIS".
[49] http://www.opengeospatial.org/standards/kml/
[50] http://earth.google.com/

provides support for both feature data, in the form of points, lines, and polygons, and image data, in the form of ground and photo overlays.

KML files may be associated with images, models, or textures that exist in separate files. KMZ files are archive files which allow one or more KML files to be bundled together along with other ancillary files required for the presentation, allowing for ease of transfer of the entire collection. KMZ files are also compressed in the ZIP archive format, resulting in reduced file size. KML files may refer to external resources and other KML files via "network links" (a link to a local or remote resource), which are used to link related data files and to facilitate data updates. Large data resources such as imagery datasets may be divided into a large number of smaller image files which are then made available via network links on an as needed basis. KML presentations using network links pose a preservation challenge in that any data available via the links may no longer be available in the future.

3.4.2 PDF and GeoPDF

PDF[51] is commonly used to provide end-user representations of data in which multiple datasets may be combined and other value-added elements may be added such as annotations, symbolization and classification of the data according to data attributes. While these finished data views, typically maps, can be captured in a simple image format, PDF provides some opportunity to add additional features such as attribute value lookup and toggling of individual data layers.

GeoPDF[52], which specifies a method for geopositioning of map frames within a PDF document, originated as a proprietary format developed by TerraGo Technologies[53], a strategic partner of Adobe. GeoPDF has proven to be a powerful format for presentation of complex geospatial content to diverse audiences that are not familiar with geospatial technologies. In September 2008 TerraGo Technologies approached the OGC with a proposal to introduce the GeoPDF encoding specification to the OGC standards process to make it an open standard and it is now published as a "Best Practices" document[54]. In parallel, Adobe introduced its own method for geo-registration into the ISO standards process for PDF.

The preservation challenges[55] that accrue to complex PDF documents will accrue to these documents as well. While the PDF/A specification has been developed to define an archive-friendly version of PDF, some of the more advanced functionality that is put to use in geospatial implementations are not supported by the current PDF/A specification. The history of complex geospatial PDF documents is rather short and risks associated with external dependencies (e.g., fonts) and reliance on specialized software will require close attention by the preservation community.

3.5 Spatial Databases

Spatial databases reach a higher level of complexity than individual data files, as they are capable of storing multiple datasets along with dataset relationships, behaviours,

[51] http://www.adobe.com/devnet/pdf/pdf_reference.html
[52] http://en.wikipedia.org/wiki/GeoPDF
[53] http://www.terragotech.com/
[54] http://portal.opengeospatial.org/files/?artifact_id=33332
[55] See the DPC Technology Watch Report at http://www.dpconline.org/docs/reports/dpctw08-02.pdf for an analysis of PDF for preservation

annotations, and data models, all of which are hosted in a relational database system. Spatial databases have played an increasingly prominent role in data production and management, while dataset-oriented formats are often still used for data distribution.

A variety of commercial database management systems, some using spatial extensions, have the ability to store geospatial data including: Oracle Spatial[56], IBM Informix Spatial DataBlade[57]and Microsoft SQLServer[58]. A prominent open source option is the PostgreSQL-based PostGIS[59] spatial database. These spatial extensions generally allow the user to store raster and vector data by adding spatial data types to the database that supports storing and querying of spatial data. Access to the spatial data in these databases can be directly through the database or, more commonly, through a connection to a desktop or web-based client.

Spatial databases have a number of features in common, including support for:

- Continuous (large geographic extent) datasets
- Large volumes of data (raster and vector)
- Complex data models (spatial data and business models)
- Long transactions, multi-user editing and versioning

These features make the long term preservation of data in spatial databases much more complex as it is often not possible to extract and transfer individual components of this data into other systems without losing some information. Preserving geospatial databases in general is likely to be particularly challenging as all the problems of preserving relational databases[60] are inherited: the need to take snapshots of running databases; storage of snapshots in proprietary database dump formats; complex dump formats; and large, monolithic sizes of snapshots.

3.5.1 ESRI Geodatabases

A prominent spatial database format is the ESRI Geodatabase[61]. The ESRI Geodatabase (often just referred to as Geodatabase) came into use in the late 1990s with the advent of the ArcGIS software environment. The Geodatabase can store a range of data types including geographic features, attribute information, satellite and aerial imagery, surface modelling data, and survey measurements. In addition to storing data, Geodatabases can also model the relationships between data and handle data validation and versioning.

Until recently, there were two forms of the Geodatabase:[62] ArcSDE Geodatabases and Personal Geodatabases. ArcSDE Geodatabases store the data in a relational database management system (RDBMS) and support multiple users; Personal Geodatabases are stored in Microsoft Access and cannot be larger than two gigabytes in size. The requirement of a commercial relational database connection has made transfers of ESRI Geodatabases greater than two gigabytes of size difficult.

---

[56] http://www.oracle.com/technology/products/spatial/htdocs/data_sheet_9i/9iR2_spatial_ds.html
[57] http://www-01.ibm.com/software/data/informix/blades/spatial/
[58] http://www.microsoft.com/sqlserver/2008/en/us/spatial-data.aspx
[59] http://postgis.refractions.net/
[60] Database preservation as such is outside the scope of this report. However there is much research going on in this area, see http://www.dcc.ac.uk/resource/briefing-papers/database-archiving/ for a brief summary of the topic.
[61] http://www.esri.com/software/arcgis/geodatabase/index.html
[62] http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Types_of_geodatabases

In ArcGIS version 9.2 the File Geodatabase was created as a standalone database not requiring a commercial back-end database. All information is stored in a directory of files that can scale up to one terabyte of size, potentially increasing portability and making the format more useful in archival transfers. However, as yet the format specifications of the File Geodatabase have not been made publicly available and there are issues over compatibility between versions[63] making its immediate appeal for preservation problematic.

There are a number of approaches to exporting content from the ESRI Geodatabase. Feature classes (vector layers) may be extracted as Shapefiles or converted to other formats such as GML for distribution or archiving. Raster datasets may also be extracted from a Geodatabase in a number of formats, including ERDAS Imagine, JPEG and TIFF. Starting with ArcGIS version 9 a new, openly specified XML export option[64] became available for the Geodatabase, making it possible to interchange Geodatabase content with other technical environments, yet it is not clear what support there will be in future versions of ArcGIS for re-importing XML exports created from previous versions of the Geodatabase.

3.6 Dynamic Geospatial Data

Geospatial web services allow end-user applications as well as server applications to make requests for sets of data over the web. Requests might also be made for particular data processes, such as finding a route or locating a street address.

In web service client applications, data is drawn from one or possibly many different sources and presented in map form to the user. These mapping environments take the burden of data acquisition and processing away from the user. While it is typically possible for the user to save service state (e.g., map area or view, zoom level, what data is shown etc.), it is usually not possible to save the state of the data within the service, creating a preservation challenge with regard to capturing such interactions.

3.6.1 Web Map Services (WMS)

The OGC WMS specification was released in 2000 and by virtue of its simplicity gained wide adoption and vendor support. WMS is a lightweight web service at the core of which is the "Get Map" request, which allows the client application to request an image representation of a specific data layer. Requests can be made from individual clients such as desktop GIS software, web browsers, as well as other map servers which might blend data sources from a number of different servers. The Web Map Context specification was developed by the OGC to formalize how a specific grouping of one or more maps from one or more map servers can be described in a portable, platform-independent format. The Styled Layer Descriptor profile of the Web Map Service (SLD) provides a means of specifying the styling of features delivered by a WMS using the Symbology Encoding (SE) language. If preservation of the cartographic representation of a map delivered by a WMS is important then it may be necessary to preserve the associated SLD (if there is one). WMS tiling efforts have come as a response to the experience of Google Maps and other commercial map services, which demonstrated the speed with which static tiled imagery could be

[63] http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Client_and_geodatabase_compatibility
[64] http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Geodatabase_XML

presented in user applications. Efforts have been made to develop a standard approach to provide access to static map tiles and the OGC have produced a candidate Web Map Tiling Service (WMTS) Interface Standard[65].

3.6.2 Web Feature Services (WFS)

Web Feature Services, which handle vector data, stream the actual data in the form of GML. WFS which was first released as a standard in 2002 has not been implemented on as wide of a scale as WMS, partly due to a higher level of complexity. WFS could potentially be used in the future to automate data harvests, perhaps using Transactional Web Feature Service (WFS-T) for making updates to a central archive.

3.6.3 Other OGC Web Services

Many other web services specifications have been released by the OGC, including the Web Coverage Service (WCS), which addresses content such as satellite images, digital aerial photos, digital elevation data, and other phenomena represented by values at each measurement point. OGC members are also specifying a variety of interoperability interfaces and metadata encodings that enable real time integration of sensor webs into the information infrastructure. In general OGC services will pose data persistence challenges related to schema evolution, URI/URN persistence and schema access.

Due to the ephemeral nature of the data in web services, new challenges in maintaining data persistence are also created. It might also be argued that the availability of web services-based access to data has decreased the incentive to replicate data resources to additional locations that might otherwise retain copies of the data. Details of all the OGC specifications can be found on the Open Geospatial Consortium (OGC) website.[66]

3.7 Legal Issues

The legal framework in which geospatial data is made available can cause a considerable amount of uncertainty, and this may have an impact on the ability to preserve and make use of geospatial data in the future. Intellectual property rights in geospatial data are carefully – sometimes aggressively – protected. Most geospatial data originates with an underlying dataset licensed from a third party – either from a mapping agency or through a satellite imagery supplier. This means that many geospatial datasets have an implied dependence on a third party supplier who may take a view on preservation and access. Consequently, archivists and repository managers would be well advised to examine the licences under which data is presented to them. There have been various studies and books[67] written about GIS legal issues including a report produced by the JISC funded GRADE[68] project which considered the licensing issues for sharing and re-using geospatial data within the UK research and education sector.

[65] http://www.opengeospatial.org/standards/requests/54
[66] http://www.opengeospatial.org/standards
[67] For example, George Cho, Geographic Information Science: Mastering the Legal Issues (WileyBlackwell, 2005)
[68] http://edina.ac.uk/projects/grade/gradeDigitalRightsIssues.pdf

3.7.1 UK Legal Landscape

In the UK there are significant legal issues around the access and re-use of geospatial data, particularly data that is produced by, or on behalf of, government agencies and protected by Crown Copyright. As an example from a major data provider, Ordnance Survey[69] data is typically licensed in return for a regular payment, and entitles the user access to the data for a period of time. If the licence is not renewed, the normal disposition is that the data must be deleted once the licence term has expired. If the data is required for preservation purposes then it is important to ensure that the data is covered by an appropriate licence. For instance, the "Plan, Design and Build"[70] licence for OS MasterMap provides the right to archive the data for up to 13 years beyond the licence term but the data can only be used for certain purposes during that period. Ordnance Survey has recently published a new strategy that aims to simplify and improve access to geospatial data, including reforming the licensing framework, although details are not currently available.

Preservation of Ordnance Survey data for the long term is carried out under the "guidance, supervision and coordination"[71] of The National Archives (TNA)[72]. However the UK Legal Deposit Libraries have an agreement[73] with Ordnance Survey whereby they receive an updated snapshot copy every year of detailed mapping, including OS MasterMap. The legal deposit libraries provide a facility[74] whereby users in the libraries can view contemporary and historic versions of OS MasterMap and Land-Line (the precursor dataset to OS MasterMap) going back to 1998 as online mapping and to print out small extracts. However, it is the responsibility of users of OS MasterMap data to maintain their own archives of data (e.g. in GML format) as necessary for future use.

As an example of some of the issues relating to licensing, Ordnance Survey data obtained for educational purposes though the EDINA Digimap[75] service can only be used as for long as the user is an authorised Digimap user. If a user leaves a subscribing institution then the user must delete any data that they have obtained through Digimap. There are also cases (for instance Land-Line data) where the licence for a particular product may not be renewed or the product withdrawn and so that data must be deleted when the licence period ends. This raises issues regarding future access to datasets which may have been used in research or used to derive other datasets which have inherited the same licensing conditions and residual IPR as the source data.

---

[69] http://www.ordnancesurvey.co.uk/oswebsite/

[70] http://www.ordnancesurvey.co.uk/oswebsite/products/ossitemap/pricing.html

[71] Eunice Gill and Jonathon Holmes, The Cartographic Journal, Vol 41 No. 1 pp55-57, June 2004. See also the article in http://www.nationalarchives.gov.uk/documents/winter2005.pdf, pages 16-18

[72] http://www.nationalarchives.gov.uk/

[73] http://www.ordnancesurvey.co.uk/oswebsite/media/news/2008/march/depositlibraries.html

[74] See the British Library website http://www.bl.uk/reshelp/findhelprestype/maps/digitalmapping/ordnancesurvey/osdigitalmaps.html or the National Library of Scotland website at: http://www.nls.uk/collections/maps/subjectinfo/os-mastermap.html for further information.

[75] http://edina.ac.uk/digimap/

3.7.2 US Legal Landscape

In the US, the philosophy is that if the data has been paid for using taxpayers funding then the data should be available without additional cost (except for distribution costs). Works by the US government are not eligible for copyright protection.

While public agency data is typically in the public domain, there are a number of rights-related issues that can complicate preservation. Public Records Law varies from state to state, and even within a single state interpretation may vary widely. Restrictions on commercial use or resale of data can result in restrictions on open secondary redistribution of that data. In general there has been a trend towards more open access to data in recognition of the positive societal benefit that derives from free data access, and the negative burden on local agencies related to mediated or fee-based data request handling. However, since 9/11 some geospatial data resources have been subject to restricted access in accordance with FGDC security guidelines[76].

The situation in other jurisdictions can be quite different. For instance, in Canada, much government spatial data has recently been made freely available through portals such as GeoGratis[77] and GeoBase[78] with very limited restrictions on what can be done with the data.

3.7.3 'Open' Geospatial Data

As a response to the complex licensing issues arising from geospatial data produced by national and local governments, private companies and others, there is a strong and growing movement for more availability, openness and transparency in licensing geospatial data, including making data more accessible and with less restrictive licensing terms. There are several licensing initiatives that have been created including Creative Commons[79] and the Open Data Commons Licences[80] that aim to achieve these goals. These licences let data creators specify less restrictive licensing conditions up to and including putting the work in the 'public domain'. Initiatives such as OpenStreetMap[81] have adopted this approach with user contributed geospatial data currently being licensed under a Creative Commons licence, however this may be changed to the Open Database Licence (ODbL)[82] in the future[83].

3.8 Geospatial Metadata

Metadata plays a central role in the current and future use of geospatial data by making data discoverable through data catalogues and search systems, by providing the means for prospective users to evaluate the data for use, and by allowing data producers to better manage their data holdings and encourage use of the data in the manner in which it was intended. Metadata also provides end users with key information about geographic positioning information including coordinate reference information (such as projection and datum), entity and attribute information, data quality, provenance and rights information that are essential for proper use of the data.

---

[76] http://www.fgdc.gov/policyandplanning/Access%20Guidelines.pdf
[77] http://geogratis.cgdi.gc.ca/geogratis/en/index.html
[78] http://www.geobase.ca/geobase/en/index.html
[79] http://creativecommons.org/
[80] http://www.opendatacommons.org/licenses/
[81] http://www.openstreetmap.org/
[82] http://www.opendatacommons.org/licenses/odbl/
[83] http://wiki.openstreetmap.org/wiki/Open_Data_License

3.8.1 Metadata Standards

In 2003 the ISO standard: 19115 Geographic Information - Metadata[84], was finalized, providing a new international standard for geospatial metadata. Prior to that, a number of national metadata standards had emerged around the world, providing several years of initial experience as a starting point to inform the development of the international standard. For example, in the United States the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata[85] was released in 1994 (version 2.0 was released in 1998). Federal agencies were mandated to begin using the standard in 1995, and the standard came into wide use by state agencies and commercial data producers as well. Profiles of the standard have also been developed, for example the NBII (National Biological Information Infrastructure) Profile and the ESRI Profile.

In the UK, EDINA[86] have developed and implemented a metadata profile based on the ISO 19115 standard but with extensions to support the needs of the UK academic community, called AGMAP (Academic Geospatial Metadata Application Profile (AGMAP)[87] which is used in the Go-Geo![88] metadata portal. Gigateway[89] is another UK based metadata portal and implements the UK GEMINI[90] metadata standard (based on ISO 19115). It is run by the Association for Geographic Information (AGI) and provides access to UK geospatial metadata. Work is also ongoing to develop an application profile of Dublin Core called the Geospatial Application Profile (GAP)[91] which is focusing on geospatial data.

INSPIRE (Infrastructure for Spatial Information in Europe) is an initiative of the EU that "intends to trigger the creation of a European spatial information infrastructure that delivers to the users integrated spatial information services".[92] One of the first deliverables of the INSPIRE initiative has been the development of regulations and rules[93] regarding the implementation of geospatial metadata to describe relevant datasets. The INSPIRE metadata specifications are based on ISO 19115 and other appropriate ISO standards.

Although INSPIRE does not currently address preservation issues specifically, it has the aim of making environmental data available for applications such as monitoring climate change which by its nature necessitates accessing data that covers a significant period of time.

3.8.2 Metadata Challenges for Archives

Geospatial metadata, either by its presence or its absence, creates numerous archival challenges, if:

- Metadata is not created by the data producer

[84] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020
[85] http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html
[86] http://edina.ac.uk/
[87] http://www.gogeo.ac.uk/Help/AGMAP_Introduction.htm
[88] http://www.gogeo.ac.uk/cgi-bin/index.cgi
[89] http://www.gigateway.org.uk/
[90] http://www.gigateway.org.uk/metadata/standards.html
[91] http://www.ukoln.ac.uk/repositories/digirep/index/Geospatial_Application_Profile
[92] http://inspire.jrc.ec.europa.eu/whyinspire.cfm
[93] http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/metadata/MD_IR_and_ISO_20081219.pdf

- Metadata is not distributed with the data
- The metadata is not concurrent with the data (i.e., the data has been updated but the metadata has not)
- The metadata file does not adhere to a widely supported encoding standard, making automated handling of the metadata difficult
- Different versions of the same metadata record are available from different sources

If metadata is **not** available or has not been created, a recipient archive can attempt to assemble a metadata record. Many elements of metadata records may be auto-extracted by software and metadata templates for different producer agencies or data collections can further help aid the metadata production process. Unfortunately, many portions of a metadata record, including data quality information, lineage information, and detailed explanations of the meaning of attribute information, can only be provided by the data producer.

If metadata **does** exist, the recipient archive will often find it necessary to:

a) Normalize the structure of the metadata to some understood schema,
b) Synchronize the metadata to reflect the current state of the data, or
c) Remediate errors found in the metadata.

3.8.3 Geospatial Metadata vs. Preservation Metadata

Geospatial metadata standards lack some features which would be useful in the archival management of data. Most notably, geospatial metadata standards do not provide a wrapper function that would allow additional technical or administrative metadata elements to be associated with (rather than replace) the data producer-originated metadata. Examples of such metadata elements that archives might wish to associate with data include:

- Archival rights information, either in text form or in a rights expression language, that does not replace any rights statements provided by the data producer in the original metadata record
- Administrative metadata related to the manner of the data acquisition
- Technical metadata related to the actual transfer of the data, including provision of assurances about data integrity
- Metadata related to any transformations carried out by the archive post-acquisition
- The outcomes of any assessments of data validity or any assessments of risk associated with the data

In the digital library community efforts have been made to use a combination of METS[94] (Metadata Encoding and Transfer Standard) and PREMIS[95] (Preservation Metadata: Implementation Strategies) to address the metadata wrapper need, however there is no parallel in the geospatial community to date.

There is also much work going on in the area of identifying "significant properties" of digital objects which aims to help the development of preservation metadata and to

[94] http://www.loc.gov/standards/mets/
[95] http://www.oclc.org/research/projects/pmwg/

assist in other aspects of digital preservation. Although not dealing specifically with geospatial data many of the studies that have been carried out, including one on Vector Images, are applicable to geospatial data. For more details see the relevant JISC website[96] for studies and related documents.

3.8.4 Metadata Creation

A particular challenge with some pre-ISO geospatial metadata standards created before the arrival of XML has been the absence of standard methods of encoding metadata. The lack of consistent structure to metadata records makes receipt and management of metadata from other sources difficult. To accompany the ISO 19115 geospatial metadata standard a separate XML schema implementation standard, ISO 19139, was finalized in 2007.

Desktop and online tools are available for creating metadata in appropriate standards including: ESRI ArcCatalog which supports FGDC, ISO 19115 and UK Gemini among others; the MetaGenie tool from Gigateway; GeoDoc from Go-Geo! and the Ramona[97] GIS inventory tool in the U.S.

4 Standards Bodies and Working Groups

The following are a selection of international standards bodies and working groups that are addressing the issues of geospatial data preservation.

4.1 Open Geospatial Consortium (OGC)[98]

The OGC is an international industry consortium of companies, government agencies and universities that work together to develop publicly available interface specifications. OGC specifications support interoperable solutions that "geo-enable" the Web, wireless and location-based services, and mainstream information technology. Examples of OGC specifications include Web Mapping Service (WMS), Web Feature Service (WFS), Geography Markup Language (GML), and OGC KML. The OGC has a close relationship with ISO/TC 211, which addresses standardization in the field of digital geographic information, and a subset of OGC standards are now ISO standards. The OGC also works with other international standards bodies such as W3C, OASIS, WfMC, and the IETF.

4.1.1 OGC Data Preservation Working Group[99]

In December 2006 the OGC Data Preservation Working Group was formed "to address technical and institutional challenges posed by data preservation, to interface with other OGC working groups that address technical areas that are affected by the data preservation problem, and to engage in outreach and communication with the preservation and archival information community." A goal of the group is to "create and dialog with the broad spectrum of geospatial community and archival community constituents that have a stake in addressing data preservation issues." To date the work of the group has been focused on identifying points of intersection between data preservation issues and OGC standards efforts, and to introduce temporal data management use cases into OGC discussions.

[96] http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.aspx
[97] http://staging.gisinventory.net/
[98] http://www.opengeospatial.org/
[99] http://www.opengeospatial.org/projects/groups/preservwg

4.2 U.S. Federal Geographic Data Committee (FGDC)[100]

The FGDC is an interagency committee that promotes the coordinated development, use, sharing, and dissemination of geospatial data within the U.S as part of the National Spatial Data Infrastructure (NSDI), a physical, organizational, and virtual network designed to enable the development and sharing of geospatial data. A wide range of stakeholder organizations participate in FGDC activities representing the interests of state and local government, industry, and professional organizations.

4.2.1 FGDC Historical Data Working Group[101]

The FGDC Historical Data Working Group was established to promote and coordinate activities among Federal agencies relating to the historical dimension to geospatial data. The role of the Working Group is to "promote an awareness among Federal agencies of the historical dimension to geospatial data; to facilitate the long-term retention, storage, and accessibility of selected historically valuable geospatial data; and to establish a mechanism for the coordinated development, use, sharing, and dissemination of historically valuable geospatial data which have been financed in whole or part by Federal funds." The Working Group has played a coordinating role in the development of a Historical Collections community within the national Geospatial One Stop portal.

5 Technology and Tools

In addition to the tools and technologies described above, there are a number of facilities that may contribute to effective long term management of geospatial data.

5.1 Digital Globe Tools

Increasingly virtual or digital globe tools such as Google Earth[102], Microsoft Virtual Earth[103], NASA Worldwind[104] and ESRI ArcGIS Explorer[105] are being used for accessing geospatial data in many communities. These tools provide a simple means of visualising, analysing and integrating different datasets based on a 'global' view. The tools allow the user to display their own data, or data from another source overlaid on top of existing base map imagery and vector layers.

Google Earth has recently (Feb 2009) added an easily accessible historic imagery layer[106] and an ability to move through a timeline of available images for an area, although currently imagery is available only over a relatively short period of time for some areas.

Google also offers the possibility to share current and historical imagery with them through the Imagery Partner Program[107]. Other datasets can also be shared including vector and terrain datasets. However, Google maintains discretion over what is included and when, and Google does not provide a data download facility, so it is 'view-only' data.

---

[100] http://www.fgdc.gov/
[101] http://www.fgdc.gov/participation/working-groups-subcommittees/hdwg/index_html
[102] http://earth.google.com/
[103] http://www.microsoft.com/virtualearth/
[104] http://worldwind.arc.nasa.gov/
[105] http://www.esri.com/software/arcgis/explorer/index.html
[106] http://google-latlong.blogspot.com/2009/02/new-in-google-earth-50-historical.html
[107] http://maps.google.com/help/maps/imagery/

Other examples of online access to historic mapping include the Rumsey Map Collection[108], a subset of which is available as a layer in Google Earth, and the National Library of Scotland (NLS)[109] Map Library, which has geo-referenced historic maps and allows its data to be displayed overlaid on a variety of mapping backgrounds, including Google Maps and Virtual Earth layers.

Tools have been developed to aid geo-referencing and registration of historic maps with imagery or basemaps such as the tools developed by the Old Maps Online[110] project which provides simple means of registering a scanned image to an existing source. Tools are also available for tiling[111] data so it can be displayed more easily in web mapping applications.

5.2 Geospatial Format Registries and Validation Tools

Format registries support preservation by maintaining knowledge of file formats. Registries under development include PRONOM [112] from the National Archives, the Global Digital Format Registry (GDFR)[113] from Harvard University, and geospatial specific registries such as that being developed by the National Geospatial Digital Archive (NGDA)[114]. Commercial companies and projects directly involved in translating and manipulating data in various formats such as Safe Software or the GDAL/OGR open source project also maintain extensive knowledge bases of geospatial formats.

5.3 ESRI Geodatabase Archiving

Maintenance of archived versions of datasets within an ESRI Geodatabase is a challenge that was addressed in ArcGIS version 9.2 by the Geodatabase Archiving feature. Previously, data change could only be tracked by managing transactional versions of the data, and the history of the data could easily be lost if the versions were deleted or if versioning was disabled. Geodatabase Archiving supports the creation of an historical version that represents the data at a specific moment in time and provides a read-only representation of the Geodatabase. The ArcGIS "History Viewer" tool allows user examination of data at specific points of time, and ArcMap provides the capability to run queries to show how the data has evolved over time.

5.4 Digital Repository Software

Digital repository software such as Fedora[115] and DSpace[116] are increasingly being used for retention and management of some types of geospatial data, and tools such as OpenLayers[117] are being used to construct access services, including mapping functionality, on top of such repositories. For example, the ShareGeo[118] geospatial data sharing facility uses DSpace as the underlying repository with an OpenLayers

[108] http://www.davidrumsey.com/
[109] http://geo.nls.uk/maps/
[110] http://blog.oldmapsonline.org/search/label/about
[111] http://www.maptiler.org/
[112] http://www.nationalarchives.gov.uk/pronom/
[113] http://www.gdfr.info/
[114] http://www.ngda.org/research.php#FR
[115] http://www.fedora-commons.org/
[116] http://www.dspace.org/
[117] http://openlayers.org/
[118] http://edina.ac.uk/projects/sharegeo/index.shtml

map interface for searching and GDAL/OGR for data identification. A major challenge in adapting some types of geospatial data with digital repository environments is that of reconciling the "item" orientation of many repositories with the "collection" orientation of many geospatial data types. The item formation process associated with repository ingest can lead to atomization of large, complex, and interrelated sets of geospatial content unless proper component relationships are built into the repository structure. Data that is item-like in nature (e.g. individual digital maps or datasets, which may themselves be multi-file and multi-format in nature) may fit best in digital repositories, while more complex content might need to be managed in a file system structure or within a spatial database.

6 Conclusions and Recommendations

There is no single best approach to preserving geospatial data. Each of the various types of geospatial data will likely call for a mix of seemingly redundant approaches, each of which is intended to mitigate a different perceived risk to the data in terms of technical failure or loss of content. These are early days for geospatial data preservation and further exploration of each of these approaches is necessary, and a longer history of documented successes and failures in preservation efforts is needed in order to arrive at a set of more mature approaches to preserving geospatial data.

Geospatial data is valuable and faces similar risks and vulnerabilities as other types of data. While some of these risks can be offset by the adoption and adaptation of generic best practice for preservation, and while geospatial data need to be incorporated into the mainstream of digital preservation planning, there are specific actions that need to be considered:

1) Formats:

- Vector data
    - Retain in their original format
    - AND, if proprietary or not widely supported, migrate into widely supported (and openly documented) format
- Raster data
    - Retain in their original format
    - AND, if proprietary or not widely supported, migrate into widely supported (and openly documented) format and compression scheme
    - If possible, retain pre-processed and processed data
- Spatial databases
    - Manage forward in time in active spatial database
    - AND replicate snapshots of spatial database
    - AND extract individual datasets (e.g. feature classes) into stable format
- Dynamic Data and Web Services
    - Take snapshot copies of data and service state and save locally

2) Metadata:

- Maintain technical and administrative metadata in addition to geospatial metadata
- Implement ISO descriptive keywords

- Implement regionally-appropriate profile of ISO 19115 as encoded per ISO 19139
- Retain original metadata AND synchronize/remediate/normalize if feasible

3) Systems:

- Keep archival data in live access systems
- Provide access to superseded datasets
- Avoid 'atomization' of data in digital repository systems
- Capture data as well as representations deemed of value
- Maintain independence of data from specific storage/repository environment

4) Legal:

- Secure archival rights and rights for access to older data
- Develop appropriate rights mechanisms so that future users of the data can be presented with suitable background information

5) Community Actions:

- Develop and promote the business case for preserving geospatial data
- Work with the data producer community to cultivate best practices for frequency of capture of key data layers
- The archival and preservation community needs to engage with existing spatial data infrastructure (SDI) efforts. SDI, in its varying forms, provides an organizational and technical framework for geospatial data access and is instrumental in the development of data sharing networks, the cultivation of metadata, and the implementation of content standards, all of which can prove beneficial to preservation efforts

7 Glossary of Acronyms

Acronym	**Meaning**
API	Application Programming Interface
COU	Change Only Update
DCC	Digital Curation Centre
DEM	Digital Elevation Model
DPC	Digital Preservation Coalition
ESRI	Environmental Systems Research Institute
FGDC	Federal Geographic Data Committee
FME	Feature Manipulation Engine
GDAL	Geospatial Data Abstraction Library
GIS	Geographic Information System
GML	Geography Markup Language
INSPIRE	Infrastructure for Spatial Information in Europe
ISO	International Organization for Standardization
KML	Keyhole Markup Language
METS	Metadata Encoding and Transmission Standard
NCGDAP	North Carolina Geospatial Data Archiving Project

NDIIPP	National Digital Information Infrastructure and Preservation Program
NGDA	National Geospatial Digital Archive
NTF	National Transfer Format
OGC	Open Geospatial Consortium
OS	Ordnance Survey (GB)
PDF	Portable Document Format
PREMIS	Preservation Metadata Implementation Strategies
SDI	Spatial Data Infrastructure
SDTS	Spatial Data Transfer Standard
TOID	Topographic Identifier
USGS	United States Geological Survey
WCS	Web Coverage Service
WFS	Web Feature Service
WMS	Web Map Service

8 Selected References and Resources

The following are a selection of useful references and resources:

The AHDS (Arts and Humanities Data Service) produced a series of handbooks in its Repository Policies and Procedures section. These Preservation Handbooks identify significant properties of various data types and provides information on how best to preserve them. A full list is available at: http://ahds.ac.uk/preservation/ahds-preservation-documents.htm including one on Geographical Information Systems written by Jo Clarke and Jenny Mitcham (2005) at: http://ahds.ac.uk/preservation/gis-preservation-handbook.pdf

The ADS (Archaeology Data Service) has produced a series of "Guides to Good Practice". Specifically there is one devoted to GIS called: GIS Guide to Good Practice, with contributions by Mark Gillings, Peter Halls, Gary Lock, Paul Miller, Greg Phillips, Nick Ryan, David Wheatley, and Alicia Wise (1998);
http://ads.ahds.ac.uk/project/goodguides/gis/index.html
A list of other guides in the series (including ones on CAD data) can be found at:
http://ads.ahds.ac.uk/project/goodguides/g2gp.html

General preservation and curation resources, including a briefing paper on geospatial data are available on the Digital Curation Centre (DCC) website under Resources:
http://www.dcc.ac.uk/resource/

The Digital Preservation Coalition has produced various other Technology Watch Reports which may be relevant, particularly ones on PDF/A and JPEG2000:
http://www.dpconline.org/graphics/reports/index.html#techwatch

EDINA – Digimap service and various other projects on preservation, repositories, metadata and geospatial interoperability:
http://edina.ac.uk/

Go-Geo provides a range of geospatial data resources including links to standards, books, case studies and metadata:

http://www.gogeo.ac.uk/

The North Carolina Geographic Information Coordinating Council (GICC) has produced a report by the Archival and Long Term Access Ad Hoc Committee which makes some recommendations on best practices for preserving geospatial data: http://www.ncgicc.com/Portals/3/documents/Archival_LongTermAccess_FINAL11_08_GICC.pdf

The Sand Report titled "Long-Term Spatial Data Preservation and Archiving: What are the Issues?", Denise R. Bleakly (2002): http://www.prod.sandia.gov/cgi-bin/techlib/access-control.pl/2002/020107.pdf

Preserving Access to Digital Information (PADI) – GIS, National library of Australia: Provides links to articles, projects and case studies dealing with GIS preservation and access.
http://www.nla.gov.au/padi/topics/432.html

UK Data Archive (UKDA) provides a range of resources and guidance on managing and sharing data, including geospatial data: http://www.data-archive.ac.uk/sharing/sharing.asp

8.1 Current Activities and Projects

The following are a selection of activities and projects that have looked at issues of geospatial data preservation or are currently involved in projects related to it.

The Archaeology Data Service (ADS): http://ads.ahds.ac.uk/
The ADS is funded by the UK Arts and Humanities Research Council (AHRC) and provides support for research, learning and teaching in the field of archaeology. It provides preservation and access services for a broad range of archaeological data, including geospatial data and provides a service to archive and preserve user data. The ADS has responsibility for "promoting standards and guidelines for best practice in the creation, description, preservation and use of archaeological information"[119] and has produced a number of publications and policies for dealing with geospatial data, including best practice guides and case studies.

CIESIN Managing and Preserving Geospatial Electronic Records (GER): http://www.ciesin.columbia.edu/ger/
The Managing and Preserving Geospatial Electronic Records (GER) project was conducted by the Center for International Earth Science Information Network (CIESIN) of Columbia University. The GER project resulted in publication of the Data Model for Managing and Preserving Geospatial Electronic Records[120], which provides recommendations with regard to retention of metadata and related information to support the management and preservation of geospatial data records.

GeoMAPP: http://www.geomapp.net/
The Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) is a partnership between the state geospatial agencies and state archives of North Carolina,

[119] http://ads.ahds.ac.uk/project/general.html
[120] http://www.ciesin.columbia.edu/ger/DataModelV1_20050620.pdf

Kentucky and Utah. In addition to directly addressing the selection, appraisal and preservation of at-risk geospatial data, GeoMAPP is directly focused on engagement of state archives within the spatial data infrastructure of each respective state. The project will involve a demonstration of content transfer between states.

Maine GeoArchives: http://www.maine.gov/sos/arc/geoarchives.html
The Maine GeoArchives project was a joint effort between the State Archives and responsible state government agencies to formalize a process for designating a select set of state agency (GIS) records as archival and to develop an archives system prototype. The GeoArchives directly addressed the issue of managing data layers stored within spatial databases.

National Archives and Records Administration (NARA) Guidance on Electronic Geospatial Records: http://www.archives.gov/records-mgmt/initiatives/digital-geospatial-data-records.html
Geospatial data records are a priority electronic records format identified by NARA and partner agencies as part of the Electronic Records Management (ERM) initiative.

The National Archives/NDAD: http://www.nationalarchives.gov.uk/default.htm
The National Archives (TNA) are responsible for preserving and providing access to information and datasets of all kinds from UK government departments, including selected geospatial datasets. NDAD is the National Digital Archive of Datasets and is part of the UK National Archives and "preserves and provides online access to archived digital datasets and documents"[121]. They have carried out a range of projects looking at preserving geospatial datasets and produce a number of advice and guidance notes[122].

NGDA: http://www.ngda.org/
The National Geospatial Digital Archive (NGDA) is focused on the problem of long-term (100+ year) preservation of geospatial data on a national scale. The project is researching long-lived preservation architectures and approaches that transcend individual repositories and storage systems. The project has developed an operational archive and format registry founded on logical and physical data models that unify the representations of file-based geospatial data and data semantics as well as many reports including "An Investigation into Metadata for Long-Lived Geospatial Data Formats": http://www.ngda.org/reports/InvestigateGeoDataFinal_v2.pdf (2008) looking at metadata requirements for long-term preservation of digital information.

NCGDAP: http://www.lib.ncsu.edu/ncgdap/
The North Carolina Geospatial Data Archiving Project (NCGDAP) is focused on preservation of state and local agency digital geospatial data. The project is being carried out as a component of the NC OneMap[123] initiative, which is focused on cultivating seamless access to state, federal, and local data covering the state. While NCGDAP includes a data acquisition and repository development component, development of the archive is intended to serve as a catalyst for engaging elements of

[121] http://www.ndad.nationalarchives.gov.uk/
[122] http://www.nationalarchives.gov.uk/preservation/advice/digital.htm
[123] http://www.nconemap.com/

spatial data infrastructure in the data archiving issue. The project has produced several publications and a detailed interim report from the project[124].

All hyperlinks were accessed on 4th May 2009

[124] http://www.lib.ncsu.edu/ncgdap/documents/NCGDAP_InterimReport_June2008.pdf