http://doi.org/10.7207/twgn21-16

Preserving GIS

Data Types Series

Artefactual Systems and the Digital Preservation Coalition

DPC Technology Watch Guidance Note

July 2021

DigitalPreservationCoalition

ISSN: 2048-7916

First published in Great Britain in 2021 by the Digital Preservation Coalition.

The Data Type Guidance Note Series

Each Guidance Note in the Data Types series is designed to provide a primer on the current state of community knowledge about data types commonly encountered by those seeking to preserve digital holdings. Digital preservation is about keeping information findable, usable, and trustworthy over the long-term. The best approach for any repository will vary according to the scope and content of its holdings, available resources, and the expectations of its funders and users. There are however, broadly applicable good practices that have been established as a result of many years of research, practical implementation, and consensus building. These are presented here as a starting point, along with additional resources for further exploration.

This series of Data Type Guidance Notes has been authored by staff at Artefactual Systems in collaboration with the Digital Preservation Coalition. These notes have been developed in conjunction with the UK Nuclear Decommissioning Authority.

Digital preservation is an evolving field and continues to change and develop in response to external drivers and fresh challenges. New formats, standards, and examples of good practice will emerge over time and the information contained within this report will need to be updated. We welcome comments and feedback to: info@dpconline.org.

1 Overview of data type

Geographic Information Systems (GIS) are multi-user software tools that provide the ability to capture and analyse geographic and spatial (geospatial) data. They present this information as multi-faceted, interactive maps. Geospatial data represents information about locations (latitude, longitude, altitude, temporal) within set geographical boundaries (extents). The distinguishing characteristic of a GIS is that it uses a space-time location as the common key to which all other information in the system is cross-indexed. GIS maps are organized into data layers such as streets, restaurants, rivers, trees, or fire hydrants that can be toggled on or off to overlay information on a single map. GIS data is timestamped, enabling the layers to illustrate changes over time.

A GIS may consist of some or all of the following:

- one or more databases for information persistence (see also: Guidance Note on Preserving Databases (DPC, 2021)).
- middleware software for business logic.
- map tiles.
- aerial/satellite photograph tiles.
- vector shape files.
- map graphic assets.
- several graphical user interfaces for data entry and display.

A GIS might depend on an online, third-party platform to render its maps. The two most commonly used are Google Earth (Google Earth, 2021) and OpenStreetMap (OpenStreetMap, 2021). Non-GIS applications also make use of these platforms in a practice referred to as “geo-tagging” to embed maps with location points into any web-accessible software.

GIS are used heavily in applications for the management of property boundaries, urban planning, licensing administration, public utilities, emergency response, meteorology data, natural resource and historic environment management, and more. GIS are also widely used by engineering, transportation, logistics, insurance, and telecommunications businesses.

A helpful introduction to geospatial resources and formats has been produced by the Library of Congress (LoC, 2017).

2 Preservation challenges

The challenges with preserving GIS are summarised below and include issues frequently encountered with complex data types such high volume of files, relationships between files and links with external systems and data.

2.1 Ambiguous scope

GIS data are often used in applications ranging far beyond those initially intended. Their location-based time-series information tends to have high utility in other contexts, which means that GIS often spiral out and spread into many other applications, such as document management systems, computer-aided design (CAD) software, building information modelling (BIM) systems, enterprise resource planning (ERP) systems, shared file directories, mobile apps, or online mapping services. This often means that there are no clear delineations of where GIS starts and ends, which makes appraisal and capture difficult. More information on 3D models and CAD can be found in Guidance Notes on 3D (Artefactual & DPC, 2021a) and CAD (Artefactual & DPC, 2021b).

2.2 Volume and volatility

A GIS is typically large in terms of number of files, and can be large in size where raster image files are heavily represented. Geospatial information is frequently updated, and many data producers overwrite existing data values rather than versioning them. Practitioners use the term 'archiving' to refer to specific application functionality (ESRI, 2021). This may create challenges for capturing GIS data for historical or evidentiary purposes.

2.3 Content copyright

Some GIS content, such as aerial photographs and LiDAR surveys, might be protected under copyright. In this case, use of this content may be limited to the licensees of the source GIS.

2.4 Complex data relationships

For most user interaction with a GIS, multiple files with different formats are brought together in one map-centric view. Managing the data required to recreate these views accurately is complex. Simply preserving GIS database snapshots along with copies of their raster tiles and vector shape files is often not enough to provide adequate access to historical geospatial information. A functioning GIS is needed to orchestrate the relationships and behaviours between these components. However, many organisations responsible for digital preservation do not operate GIS software.

3 File formats

There is no single perfect format for the preservation and future use of GIS. Decisions made on file formats should be dependent on the features and functionality to be preserved and the future use cases to be supported. Note that the table below does not provide an exhaustive list of formats suitable for preservation and access. The most suitable format for preserving the important features and functionality of a file may be the original format that it was created in. The Library of Congress (2020-2021) recommends target preservation formats that maintain the completeness of the data, 'with a preference for preserving the native format and projection of the data.' This includes the use of proprietary file formats of the native version. It is recommended therefore that careful research and analysis is carried out before migrating files to a new format.

File format	Extensions	Brief summary
ArcGIS Pro project file	.aprx	APRX project files contain maps, toolboxes, databases, folders, and styles. They can also contain connections to databases, servers, and folders, and can include multiple maps and layouts in a single project. They are intended as a container to "archive" all the assets associated with a project. However, APRX project files are a proprietary format that requires the ESRI ArcGIS Pro application to render them with their functionality intact.
ASPRS LiDAR Data Exchange Format	.las, .lasd, .laz	Light Detection and Ranging (LiDAR) is an optical remote sensing technology that is used in a wide variety of sectors to visualize the earth's surface and to make high-definition maps. LiDAR data can be imported into applications like ArcGIS to add rich new layers to maps. LiDAR consists of a dense network of three-dimensional point cloud coordinates with elevation values.

		The Lidar Data Exchange Format (LAS) is a binary file format which has replaced proprietary and ASCII-based formats as the industry-standard exchange format for LiDAR data (ASPRS, 2021). Using the LAS file format, compression is applied to LAS files to save significant storage space without information loss. LAS Datasets (LASD) reference a set of LAS files to enable the examination of three-dimensional point cloud properties, visualization of triangulated surfaces, and statistical analysis.
ESRI Geodatabase	.gdb	The ESRI geodatabase is the native data structure for ArcGIS applications (LC, 2020). It is the physical store of geographic information, primarily using a DBMS or file system. ESRI geodatabases are the most ubiquitous examples of spatial databases, database management systems that are optimized for storing and querying data that represent objects in a geometric space. ESRI enables users to create "datasets" to organize and manage geographic information. The three primary dataset types are feature classes, raster datasets, and attribute tables. Users extend the datasets in their geodatabase with more advanced capabilities (by adding topologies, networks, or domain-specific schemas, for example) to model GIS behaviour, maintain data integrity and work with a set of spatial relationships.
ESRI Shapefile	.shp, .dbf, .shx	Shapefile is by far the most ubiquitous geospatial file format. It is the industry standard and recommended as a preservation format for GIS feature data (LC, 2020). A shapefile stores a feature's geometry as a set of vector coordinates. It must include three sub-files to be functional: SHP (feature geometry), SHX (shape index position), and DBF (attribute data).
GeoTIFF	.tif .tiff	GeoTIFF (GeoTIFF, 2021) is an extension of the TIFF raster image format which associates its content to a map projection by using a predefined set of extended TIFF tags that are embedded into the file's header. GeoTIFF has become the industry standard for GIS image files, and is recommended as a preservation format for raster data (LC, 2009). It is supported by all major GIS software applications. Most agencies and services that produce geospatial data provide their imagery in GeoTIFF format (LC, 2009). Cloud Optimized GeoTIFF (COG) is another noteworthy TIFF variant. It is a regular GeoTIFF file that is hosted on an HTTP file server. It uses HTTP GET range requests to retrieve just the specific parts of a GeoTIFF file that are required for a cloud-based workflow (Cloud Optimized GeoTIFF, 2021).

Geography Markup Language (GML)	.gml	The Geography Markup Language (GML) is an XML schema defined by the Open Geospatial Consortium (OGC) to express geographical features (GML, 2021). The core GML geometry object types are points, line strings, and polygons. They define locations or extents (regions). GML also includes feature objects to represent physical entities (e.g. buildings and rivers), which may or may not have geometric aspects. GML serves as a modelling language and an open interchange format. It is cross-published as ISO 19136:2007. GML is suitable for preservation because its XML format is compatible with older and disparate GIS. It also enables the integration of multiple forms of geographic information, including, for example, vector objects, coverages, and sensor data. Specific communities of interest (e.g. tourism) have created their own XML application schemas based on GML to describe the object types that their community GIS applications must be able to process and expose.
GeoJSON	.json	GeoJSON is a geospatial data interchange format maintained as an open standard by the Internet Engineering Task Force as RFC 7946 (IETF, 2016). GeoJSON represents geographic features and their nonspatial attributes, supporting the following feature types: point, line string, polygon, and topologies. Since JSON is more compact than XML, it is ideal for transferring spatial data to web and mobile applications. GeoJSON is widely used in the GIS community, and both ArcGIS and QGIS support it. It is considered an acceptable preservation format (LC, 2020-2021).
Google Keyhole Markup Language (KML/KMZ)	.kml .kmz	KML was developed by Keyhole Inc., which was acquired by Google (Wikipedia, 2021). KML is an XML format used predominantly in the Google Earth browser, which has ensured a large user base for the format. KML has been maintained by the Open Geospatial Consortium, Inc. (OGC) since 2008. KML files are often distributed in compressed, zipped containers with a KMZ extension. The contents of a KMZ file are a single root KML document and optional overlays, images, icons, or COLLADA 3D model files. KML can be used to render GML content, but differs significantly from GML in that it is first and foremost a 3D rendering format, not a data exchange format. Data creators should therefore avoid encoding GML content for portrayal with KML as this results in significant and unrecoverable loss of content and context.
PostGIS + PostgreSQL	https://www.postgresql.org/	PostGIS is a free and open-source software application that provides support for geographic objects in the PostgreSQL object-relational database. PostGIS adds geometry, geography, raster, and other geospatial types to the PostgreSQL database, along with dedicated functions, operators, and index enhancements that apply to them. These features allow location queries to be run in SQL. PostGIS follows the Open Geospatial Consortium's Simple Features set of standards which specify a common storage and access model for geographic

		features. Many GIS, including ArcGIS and QGIS, can use PostGIS as their database backend. PostGIS + PostgreSQL is recommended as a preservation format for tabular GIS data.
QGIS project file	.qgs, .qgz	QGS project files store the project's map layers with links to the underlying datasets and other layer properties including the map layer tree, extents, spatial reference system, coordinate reference system, styles, renderers, blend modes, and opacity values. It also contains print layouts, table relations, project macros, plugin settings, and QGIS server settings (QGIS.org Association, 2020). The QGS project file is saved in an XML format that can be edited outside of QGIS. The application default is to compress project files into a QGZ zipped format containing a QGS file and a QGD file. The QGD file is the associated SQLite database of the QGIS project which contains project auxiliary data. The QGZ file can be opened with any ZIP utility.
Adobe Geospatial PDF	.pdf	Geospatial PDF is a set of geospatial extensions to the Portable Document Format (PDF) 1.7 specification that enables the ability to relate a region in the document page to a region in physical space (Adobe, 2021). This is often referred to as "geo-tagging" or "georeferencing". Note that "GeoPDF" is a different format which refers specifically to files produced by TerraGo applications (TerraGo, 2021). A geospatial PDF can contain geometry such as points, lines, and polygons that may represent features such as buildings, roads, or city boundaries. These files can be written using Adobe Acrobat or GIS applications such as ArcGIS and QGIS. While Adobe offers its own extension, the Open Geospatial Consortium's (OGC, 2021) standards should be used when adding geospatial metadata to PDF documents. This is an acceptable access format, but preference for the native file format for preservation is recommended (LC, 2020-2021).
MapInfo TAB	.tab .dat .id .map .ind	This is a proprietary format developed for use with the popular MapInfo GIS software application. (LC, 2011). While Mapinfo files can be opened in the free and open-source QGIS software, it is a closed format without a complete specification, so not recommended for long-term preservation. It may however be used as an access format.

4 Metadata standards

International metadata standards exist for geospatial data:

- *ISO 19115-1:2014 Geographic information — Metadata — Part 1: Fundamentals* articulates the schema for describing geographic information and services (ISO, 2014).
- *ISO 19115-2:2019 Geographic information — Metadata — Part 2: Extensions for acquisition and processing* extends the ISO 19115-1:2014 schema to include acquisition and processing metadata, such as numerical methods and computational procedures used to derive geographic information (ISO, 2019).

Other national or regional standards may also be relevant. These are often based on the ISO 19115 family of standards but are tailored for specific regions and/or use cases.

- INSPIRE is a European Commission Directive to create a European Union spatial data infrastructure to enable the sharing of environmental spatial information among public sector organizations, facilitate access to spatial data and assist in policy making across boundaries.
- UK GEMINI is the UK geographic metadata standard, providing guidance on how to publish geographic metadata in a way that conforms to UK government guidelines and the relevant ISO standards.

5 Tips for creators

- Creators should document their GIS architecture and processes. For archivists it is critical to understand the context of creation and use of records. However, most GIS architecture and business process information is implicit. If explicit GIS user instructions are proactively preserved and maintained, this will go a long way towards helping archivists make historical copies of the GIS more understandable to future users.
- Wherever possible, document the contents and metadata of specific GIS layers. If the GIS was deployed using formal architecture and business process methodology, then retaining this documentation will be of great assistance in understanding and re-rendering the original GIS interfaces and information. If it is important to future users that changes to GIS content over time can be tracked, and GIS data producers should enable the versioning features offered by GIS platforms. This ensures that historical data is not overwritten.
- Use open or industry standards. Creators should confirm that the tools they are using enable saving or exporting in open or industry standard formats such as Shapefile, GeoJSON, and GeoTIFF. Using open standards will increase the likelihood of keeping GIS data accessible and usable over the long term in current or future applications.

6 Tips for archivists

6.1 General guidance

- The following resources provide guidance on preserving and providing access to GIS:
- Geospatial Multistate Archive and Preservation Partnership's (2011) *Key Findings and Best Practices*.
- Library of Congress' (2020-2021) *Recommended formats statement: GIS, Geospatial and Non-GIS Cartographic*.
- North Carolina Geospatial Data Archiving Project's (2010) *Final Report.*

6.2 Community assistance

Engage with the community for advice and support. Professional communities can provide a wealth of specialist knowledge on the topic and often provide useful publications and ways to get involved:

- The Open Geospatial Consortium is an international consortium with more than 500 members with an interest in making geospatial information and services FAIR – Findable, Accessible, Interoperable and Reusable.
- The Research Data Alliance Geospatial Interest Group is a domain-oriented interest group with a specific interest in data interoperability and quantifying uncertainty in datasets.

6.3 Access

- Consider installing a copy of the free and open-source QGIS software to provide access to GIS data. QGIS can open ESRI Shapefiles, PostGIS, SpatiaLite, Oracle Spatial, MSSQL Spatial, AutoCAD DXF, and many more proprietary and open formats. QGIS supports vector layers as point, line, or polygon features, as well as multiple raster image formats. The software can also georeference image files and interface with web-mapping services, including the Google Geocoding API.
- QGIS will need to be installed on a desktop workstation in the reference room because there is no fully-featured web-accessible version. Using a virtual machine tool like VirtualBox or a container platform like Docker can simplify the deployment of a QGIS instance.

6.4 Acquisition and appraisal

- Perform a business function analysis of the creating organization to determine where and when GIS records should be captured, unless this information is already available in the organization's record retention schedules. Such an analysis should attempt to capture user documentation, business process documentation, and information about the types and frequency of data updates.
- Suggested criteria for appraisal and selection of geospatial data are listed in a report from the National Digital Stewardship Alliance (NDSA, 2013).
- Identify reports or screengrabs which can be captured from the GIS in TIFF or PDF format to supplement the GIS data and documentation.
- To capture GIS data for ingest into a digital repository, use utilities that export data or application programming interfaces (APIs) that make SQL data, vector shapes, and raster tiles available to external applications.
- The Open Geospatial Consortium (OGC) has published a family of API standards to enable uniform programmatic access to geospatial data across applications and services. Data creators can be encouraged to make their systems compatible with them.

6.5 Characterization

Characterization can be useful to identify file formats, extract metadata, identify broken or encrypted content, or check conformance to profiles or standards. Tool support and effectiveness can vary considerably for different file formats.

- Identify file formats with a tool such as DROID, FIDO, or Siegfried that uses the PRONOM file format registry.
- Validation tools for GIS formats are not available at this time.

6.6 Metadata

- Be aware of the different types of metadata required. All GIS contain a common core set of spatial-temporal data about locations. Core metadata may be supplemented by sector-specific data. This wide scope can make it challenging to import, normalize, process, or provide access to GIS metadata in a consistent manner.

7 References

Artefactual & DPC (2021a) *Preserving 3D.* Available at: http://doi.org/10.7207/twgn21-14

Artefactual & DPC (2021b) *Preserving CAD.* Available at: http://doi.org/10.7207/twgn21-15

ASPRS (2021) LASer (LAS) *File Format Exchange Activities.* Available at: https://web.archive.org/web/20200922151120/https://www.asprs.org/divisions-committees/lidar-division/laser-las-file-format-exchange-activities

ESRI (2021) *What is archiving?* Available at: https://web.archive.org/web/20210202215457/https://pro.arcgis.com/en/pro-app/latest/help/data/geodatabases/overview/what-is-archiving-.htm

Geospatial Multistate Archive and Preservation Partnership (2011) *Key Findings and Best Practices.* Available at: https://web.archive.org/web/20200502204535/https://files.nc.gov/ncdit/documents/files/GeoMAPP-Best-Practices-2011-12-31.pdf

GISGeography (2020) *The Ultimate List of GIS Formats and Geospatial File Extensions.* Available at: https://gisgeography.com/gis-formats/

Internet Engineering Task Force (2016) *The GeoJSON Format.* Available at: https://web.archive.org/web/20210104100025/https://tools.ietf.org/html/rfc7946

ISO (2019) *ISO 19115-2:2019 Geographic information — Metadata — Part 2: Extensions for acquisition and processing.* Available at: https://web.archive.org/web/20201206022231/https://www.iso.org/standard/67039.html

ISO (2014) *ISO 19115-1:2014 Geographic information — Metadata — Part 1: Fundamentals*. Available at: https://web.archive.org/web/20210320110509/https://www.iso.org/standard/53798.html

Library of Congress (2020-2021). *Recommended formats statement: GIS, Geospatial and Non-GIS Cartographic*. Available at: https://web.archive.org/web/20210320174757/https://www.loc.gov/preservation/resources/rfs/geo-carto.html

Library of Congress (2020) *ESRI Shapefile.* Available at: https://web.archive.org/web/20201117143240/https://www.loc.gov/preservation/digital/formats/fdd/fdd000280.shtml

Library of Congress (2017) *Introduction to Geospatial Resources and Formats.* Available at: https://www.loc.gov/preservation/digital/formats/content/gis_intro.shtml

Library of Congress (2011) *MapInfo Dataset format.* Available at: https://web.archive.org/web/20200223233237/https://www.loc.gov/preservation/digital/formats/fdd/fdd000300.shtml

Library of Congress (2009) *GeoTIFF, Revision 1.0.* Available at: https://web.archive.org/web/20201030114515/https://www.loc.gov/preservation/digital/formats/fdd/fdd000279.shtml

National Digital Stewardship Alliance (2013) *Issues in the Appraisal and Selection of Geospatial Data.* Available at: https://web.archive.org/web/20170102034153/https://www.digitalpreservation.gov/documents/NDSA_AppraisalSelection_report_final102413.pdf?loclr=blogsig

North Carolina Geographic Information Coordinating Council, Archival and Long Term Access Ad Hoc Committee (2008) *Final Report*. Available at: https://web.archive.org/web/20200706080529/https://files.nc.gov/ncdit/documents/files/GICC-Archival-Long-Term-Access-Report-11-08.pdf

North Carolina Geospatial Data Archiving Project (2010) *Final Report.* Available at: https://web.archive.org/web/20130130005923/https://www.digitalpreservation.gov/partners/documents/ncgdap_final_report.pdf

PostgreSQL Global Development Group (2021) *PostgreSQL: The World's Most Advanced Open Source Relational Database.* Available at: https://web.archive.org/web/20210201221045/https://www.postgresql.org/

QGIS.org Association (2020) *Working with Project Files.* Available at: https://web.archive.org/web/20201129005713/https://docs.qgis.org/3.16/en/docs/user_manual/introduction/project_files.html

TerraGo (2021) *TerraGo.* Available at: https://web.archive.org/web/20201127085818/https://terragotech.com/

Wikipedia (2021) *Keyhole Markup Language.* Available at: https://web.archive.org/web/20210308184851/https://en.wikipedia.org/wiki/Keyhole_Markup_Language

Wikipedia (2020) *Geographic Information System.* Available at: https://web.archive.org/web/20201205232727/https://en.wikipedia.org/wiki/Geographic_information_system