Geographic Information Systems Procedures

Version 1.148

Created date:	26 January 2012
Last updated:	25 August 2017
Review Due:	24 August 2018
Authors:	Michael Charno, Jen Mitcham, Tim Evans, Kieron Niven and Jenny O'Brien
Maintained by:	Digital Archivists
Previous version:	Live

1 Purpose of this document

From the 2009 DPC Technology Watch Report

> 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.[1]

The following is therefore a guide to our in-house procedures for preserving the type of datasets we most commonly receive, and include geo-referenced vector and raster datasets.

[1] McGarva, G; Morris S and Janée G (2009) *Technology Watch Report Preserving Geospatial Data.* DPC Technology Watch Series Report 09-01. Digital Preservation Coalition. http://www.dpconline.org/knowledge-base/tech-watch-reports.

2 Formats

Offered format	Accepted	Preservation	Presentation	Notes
Geo-referenced Vector				
ArcInfo Interchange **.e00**	YES	ArcInfo Interchange **.e00**	ESRI Shapefile **.shp** (zipped)	Old versions of ArcGIS still create this and we still accept it. The ESRI E00 interchange data format combines spatial and descriptive information for vectors and rasters in a single ASCII file. It was mainly used to exchange files between different versions of ArcInfo, but can also be read by many other GIS programs.
ESRI Shapefile **.shp**	YES	Geographic Markup Language 3.2 **.gml**	ESRI Shapefile **.shp** (zipped)	See footnote.[2]

[2] A shapefile is actually a collection of files the number and combination depends upon the type of data stored in the file. Shapefile is an openly published format. It stores non topological geometry as part of a set of data files making up a spatial dataset. It must be accompanied by in index file (SHX) and a dBASE file that holds the attributes of the shapes in the shp file. SHP - the file that stores the feature geometry. Required. SHX - the file that stores the index of the feature geometry. Required., DBF - the dBASE file that stores the attribute information of features. Required. SBN, SBX - the files that store the spatial index of the features. Optional. FBN, FBX - the files that store the spatial index of the features for shapefiles that are read-only. Optional. AIN, AIH - the files that store the attribute index of the active fields in a table or a theme's attribute table. Optional. PRJ - the file that stores the coordinate system information. Optional. XML - metadata. Optional.

.lyr	NO	N/A	N/A	A layer file (**.lyr**) is an ESRI file that stores the path to a source dataset and other layer properties, including symbology. It is not data - send us the shapefile instead!
MapInfo Interchange Format **.mif** & **.mid**	YES	Geographic Markup Language 3.2 **.gml**	ESRI Shapefile **.shp** (zipped)	See footnote.[3]
Spatial Data transfer standard **.ddf**	NO	N/A	N/A	See footnote.[4]
Vector product Format **.vpf**	NO	N/A	N/A	See footnote.[5]

[3] MapInfo is a commonly used GIS software package .mif contains the graphics and .mid contains any attribute data and is optional. This is a suitable deposit format as is can be imported into ArcCatalog. MIF and MID is MapInfos standard format, but most other GIS programs can also read it. The format holds three types of information: geometry (geography), attributes and display. The MIF file contains the geometric data whilst the MID file header and attribute data as delimited text. Like the ArcInfo export format this format is ASCII based and open and thus a possible preservation format. Note: We can open these in QGIS

[4] The Spatial Data Transfer Standard (SDTS) was a data exchange format for transferring different databases between dissimilar computing systems, preserving meaning and minimizing the amount of external information needed to describe the data. It can only be used for certain types of feature point, arc and grid data. One coverage would produce many files all with extension .ddf. This can be imported into ArcInfo but support in later versions of ESRI (and other) software is limited. Better to export as shapefile instead.

[5] Vector Product Format (VPF) is a U.S. Department of Defense Standard. The National Imagery and Mapping Agency (NIMA) is using VPF for digital vector products developed at a variety of scales. VPF has also been adopted into an international spatial standard as the Digital Geographic Information Exchange Standard (DIGEST). VPF coverages and tables can be translated into ARC/INFO coverages and INFO tables with the VPFIMPORT command. However, it's best if the depositor can give us the data in shapefile or e00 instead

Geographic Markup Language **.gml**	YES	Geographic Markup Language 3.2 **.gml**	ESRI Shapefile **.shp** (zipped)	See footnote.[6]
Other				
Geodatabases (see notes below)	YES - but see notes	**.csv** and **.shp**	**.csv** and **.shp**	See footnote.[7]
GeoJSON **.geojson**	YES	GeoJSON **.geojson**	GeoJSON **.geojson**	An open standard format designed for representing simple geographical features, along with their non-spatial attributes, based on JavaScript Object Notation (**.json**). Can open and export in QGIS.
ESRI project file **.mxd** or **.apr**	NO	N/A	N/A	See footnote.[8]
Geo-referenced Raster				

[6] A preferred format. It can serve as a modelling language for geographic systems as well as an open interchange format for geographic data. It is an ISO standard (ISO 19136) and is built on a number of other ISO standards collectively known as the 19100 family. GML is defined by the Open Geospatial Consortium. In being an XML based schema and an ISO standard GML is very suitable as a preservation format for Geographical data. See http://www.opengeospatial.org/standards/gml.

[7] Although we can take Geodatabases in their original flavour, due to the geometry element we can't establish a simple migration path (i.e. there's no 1-1 format for these things). Our preferred option at the moment is to have the data exported as CSV and SHP which will maintain the 'database' and 'geo' elements respectively. This isn't perfect, but does the job. Our preference is the depositor to do this themselves.

[8] Project files such as APR or MXD. data present 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.

GeoTIFF Image **.tif** (+ **.rrd**, **.aux**, **.xml**)	YES	Geo-referenced TIF Image **.tif** (+ **.aux, .xml** if present)	Geo-referenced TIF Image zipped **.tif** (+ **.aux**, **.xml** if present) (zipped)	See footnote.[9]
ESRI GRID (ascii) **.asc** or **.grd**	YES	ESRI Grid (ascii) **.asc** or **.grd**	ESRI Grid (ascii) **.asc** or **.grd** and/or Geo-referenced TIF Image zipped **.tif (**& **.aux**, **.xml)** (zipped)	See footnote.[10]
ESRI GRID (binary) **.adf**	YES	ESRI Grid (ascii) **.asc** or **.grd**	ESRI Grid (ascii) **.asc** or **.grd** and/or Geo-referenced TIF Image **zipped .tif** (& **.aux, .xml)** (zipped)	As above - note that we can open this and convert.
ERDAS Imagine files **.img (.rrd**)	YES	Geo-referenced TIF Image **.tif** (& **.aux, .xml**)	Geo-referenced TIF Image **zipped .tif** (& **.aux, .xml**) (zipped)	

[9] Geo-Referenced TIFs are TIF images with optional associated files. One file (TIF or TIFF) contains the coordinates of the top-left corner of the image and the spacing of the pixels in the image (units per pixel). There can also be other related files such as RRD etc. These aren't essential, but can be kept. The GeoTIFF file structure allows both the metadata and the image data to be encoded into the same file. GeoTIFF is a preferred format to TIF World files TFW. if needed coordinates can be located by looking at the layer in QGIS, Right clicking, going to layer properties, Metadata and scrolling to find: 'Layer Spatial Reference System, which gives the projection and the zone number followed by the UTF coordinates in 'Layer Extent' This can then be checked against any metadata provided. Images saved using GeoTIFF require only one file with a TIFF or TIF file extension. True GeoTIFF files will import automatically with correct georegistration.

[10] An ESRI GRD is a raster GIS file format developed by ESRI, which has two formats: A proprietary binary format, also known as an ARC/INFO GRD, ARC GRD and many other variations. See ESRI documentation on the binary version. A non-proprietary ASCII format, also known as an ARC/INFO ASCII GRD. The file extension is .asc, but recent versions of ESRI software also recognize the extension .grd. See notes on the ASCII format. The ASCII format is used as an exchange, or export format, due to the simple and portable ASCII file structure. They can be offered to users as downloads (zipped up within their directories) or if appropriate, GeoTIFFs can be created. GeoTIFFs of course are only a georeferenced TIF image and therefore do not display any values associated with a grid file. The binary version can be converted in ArcCatalog.

TIFW **.tif** & .**tifw**	YES	TIFW **.tif** & **.tifw**	TIFW **.tif** & **.tifw** (zipped)	See footnote.[11]
JPG World **.jpg** & **.jgw** (**.rrd**, **.aux**, **.xml**)	YES	Geo-referenced TIF Image **.tif** (& **.aux**, **.xml**)	JPG World **.jpg** & **.jgw** (**.rrd**, **.aux**, **.xml**) (zipped)	As above, this has geospatial information stored in the jgw element. Best just to disseminate what was given.
PNG World **.png** & **.pgw** (**.rrd**, **.aux**, **.xml**)	YES	Geo-referenced TIF Image **.tif** (& **.aux**, **.xml**)	PNG World .png **&** .pgw **(.rrd**, **.aux**, **.xml**) (zipped)	As above, this has geospatial information stored in the pgw element. Best just to disseminate what was given.
Keyhole Markup Language **.kml**	YES	**.kml** + others (e.g. **.tif**)	**.kmz**	An XML-based format primarily used for web display (e.g. Google maps / Earth). **.kmz** is a compressed format and can be unzipped to its constituent parts. A **.kmz** file will contain the main **.kml** file along with other images, overlays, or even 3D (COLLADA) files.

Geodatabases

ESRI geodatabases are essentially containers and come in a number of flavours, the two we are likely to deal with are Personal Geodatabases and File Geodatabases. (The other type is ArcSDE Geodatabase). Be aware that "The whole of these databases is often greater than the sum of the parts in that they are capable of storing not only a large set of datasets but also object relationships, behaviors, annotations, tools, and data models that may span or connect the stored datasets" (GeoMAPP, 'Archival Challenges...')

- **Personal Geodatabases (pGDB)** are the original format and use an Access (MDB) database file, this limits its size (2GB per file) and compatibility beyond Windows platforms. Has advantages via its use of Access (e.g. form input) but even ESRI state a preference for fGDBs. These can be exported to fGDBs as well as Shapefiles.
- **File Geodatabases (fGDB)** is a native format for ArcGIS which stores data in various files within a folder structure. Size is limited to 1TB per dataset (through multiple dataset can exist within each geodatabase). Data can optionally be stored in

[11] Not to be confused with the GeoTIFF format (above). This is a different format using .tif files called the TFW format. This format uses two files, a TIF file and a TFW "world" file to provide some georeferencing information. TFW is not the same as GeoTIFF.!!! Adding to the confusion is that some packages will create both a GeoTIFF file as well as a TFW "world" file. The TFW file provided in such cases is not part of the GeoTIFF standard.

a read-only compressed format. fGDB can be exported to Shapefiles and GML. The API for fGDB was made available in 2011.
- Also ArcSDE GDB (see Katsianis)

What are those rrd/aux/xml files?

.xml	ESRI software can (depending on settings) automatically XML files for shapefiles and GRID files. It is created to contain Geospatial metadata such as ISO 19115, but can incorporate other schemas. The default setting for a shapefile is just an ESRI default format (normally tagged something like "esriprof80") which contains non-pertinent information, just local variables such as the folder it was created in/date created and so forth. Nothing which shouldn't already be in the ADS file-metadata requirements. A decision should be made (a lot like with EXIF metadata for rasters) about the quality of this metadata, and preservation strategy planned accordingly.
.rrd	Raster pyramids or "Reduced Resolution Dataset" if you're not into the whole brevity thing. These are created when the dataset is opened in ArcGIS to allow quicker loading times, so are not worth preserving. Can be disseminated.
.aux	Auxiliary file. Can contain Colour map, Histogram or table, Coordinate system,Transformation or Projection information. These should be retained.

3 Documentation / Metadata

Documentation	
Project Level	
Project Title	
Hardware	
Software	
Date of Creation	
Coverage	
Creator	
File Level	
File name	
Data type	
Coordinate System	
Description	
Scale of data capture	

Scale of data storage	
Assessment of data quality	
Method of original data capture	
Purpose of data creation	
Ownership	
Attribute tables	

4 Accessioning checks

- Check files are in accepted formats (for example no .ecw files as described below)
- Check that all files are necessary e.g. if LYR files are submitted alongside a SHP group, we should not accession these as they are just another iteration of the shapefile and could just confuse things in future.
- GIS project and file-level metadata is present
- Open files, do they appear in the right place?[12]
- Check georeferencing is correct
- Check content, especially any third party content (OS, BGS, Seazone etc), check depositor have permission to deposit (this should be in the 'source' metadata). This also includes things like HER data (need permissions), see below for discussion.
- Project files - we can't archive these. If present then we should remove from the SIP, unless specifically requested to keep in order to aid interface design.

Third party data

This can be tricky subject, especially derived data such as BGS Boreholes, HER point data, NMP mapping and so on. In each case we should ensure that the source is specified in the metadata (for example "Staffordshire HER Event data") and permissions given for deposition with the ADS. A good rule of thumb if in doubt is to raise the issue with the depositor and get them to clarify. Most HERS are fine (as their data is present online anyway) with data being deposited as part of a wider project; although certain OS datasets are becoming available under Open Access agreements, we need to err on the side of caution and make sure any such data is accompanied by copyright clearance. A digital copy of such permissions (emails, scans of letters etc) should be stored in the */admin/project_metadata/* folder.

Significant properties

- Coordinate reference system information
- Geometry (e.g. point, polygon, line)
- Attribute fields
- For rasters - source elevation model, bit-type, colour map, pixel type

Strictly speaking, we do not preserve colour in shapefiles. This tailoring of data is stored in the project file (see above) and not in the digital object itself. If the depositor requests that colour/styling of original data should be recorded then this should be supplied as documentation in the form of a document or image. This documentation can then be stored with the data.

[12] Can use UK Grid Reference Finder or the following page to check any coordinates given and check against metadata) http://www.movable-type.co.uk/scripts/latlong-utm-mgrs.html

5 How to convert files

From February 2016 we are now migrating to version 3.2.1 (i.e. the ISO). 3.2 includes the projection (and extents) in the GML. This is quite important, as it means that these are completely self-contained.

In addition, most dissemination files are subsequently zipped for ease of download, but also to reduce size of larger raster datasets.

ESRI shapefile >> GML

After some testing, **it is recommended that CATS use QGIS 2.8 (Wien)**, which has the capacity to select which version of GML to save to.

1. Load the file into QGIS, right click and 'save as'... Then use the following settings, being careful to retain the original layer projection (sometimes can default to WGS84 or similar).
2. A GML and GFS file should be created. GFS appears to be a file QGIS creates after reading GML, it is a schema file created after parsing GML, if there is no XSD present. Although these aren't strictly necessary for re-use, worth keeping with the GML (see below).

ArcInfo Interchange Format >> ESRI shapefile

1. Open ArcCatalog and navigate to the folder containing the files
2. Navigate to Tools => Customize => Toolbars tab.
3. Check the box for the ArcView 8.x Tools toolbar and click Close.
4. Dock the Conversion Tools toolbar.
5. Click the Conversion Tools drop-down menu.
6. Select 'Import from Interchange File'.
7. For the Input file, navigate to the directory location of the E00 file to be imported, and select the E00 file.
8. Specify a name and location for the output dataset.

More than one file can be imported in a batch process. Click the Batch button. To add additional E00 files for processing, click the Add Row button, you'll still have to specify output/name for each file but is quite quick.

ESRI GRID (binary) >> ESRI GRID (ascii)

1. Open ArcCatalog and navigate to the folder containing the files
2. Click the Conversion Tools drop-down menu.
3. From Raster => Raster to ASCII
4. Select the Raster file you want to convert and select the path and filename for the exported ASC file
5. You will then be presented with a mini screen with the following options:

Convert 1 bit data to 8 bit (optional): Choose whether the input 1-bit raster dataset will be converted to an 8-bit raster dataset. In this conversion the value 1 in the input raster dataset will be changed to 255 in the output raster dataset. This is useful when importing a 1-bit raster dataset to ArcSDE. One-bit raster dataset have 8-bit pyramid layers when stored in a file system, but in ArcSDE, 1-bit raster datasets can only have 1-bit pyramid layers, which makes the display unpleasant. By converting the data to 8-bit in ArcSDE, the pyramid layers are built as 8-bit instead of 1-bit, resulting in a proper raster dataset in the display.

- Unchecked—No conversion will be done. This is the default.
- Checked—The input raster will be converted.

Colormap to RGB (optional): If the input raster dataset has a colormap, the output raster dataset can be converted to a three-band output raster dataset. This is useful when mosaicing rasters with different colormaps.

- Unchecked—No conversion will occur. This is the default.
- Checked—The input dataset will be converted.

Pixel Type (optional): Determines the bit depth of the output raster dataset. If left unspecified, the output bit depth will be the same as the input.

JPG World >> Georeferenced TIF

1. Open ArcCatalog and navigate to the folder containing the files
2. Select and right-click file: select Export > Raster to different Format
3. You will then be presented with a mini screen, choose TIFF as export option and follow same guidelines for raster settings as outlined above.

ERDAS Imagine files >> Georeferenced TIF

1. Open ArcCatalog and navigate to the folder containing the files
2. Select and right-click file: select Export > Raster to different Format
3. You will then be presented with a mini screen, choose TIFF as export option and follow same guidelines for raster settings as outlined above.

Mapinfo interchange >> Shapefile

1. Open ArcCatalog and navigate to the folder containing the files
2. Click the Conversion Tools drop-down menu.
3. Select 'MIF to Shapefile'
4. Specify output folder and filename.
5. **Note 1**: Mapinfo to Shapefile conversions will export the different Geometry types as individual shapefiles (e.g. point, rectangle, polygon).
6. **Note 2**: The transfer script will fall over if input or output data are located in a path with folder names longer than 8 characters (I've got no idea why!). See ESRI pages for more details.[13]

Converting Geodatabases

It is recommended that GDBs be broken down into their constituent parts for archiving with any relationships, etc. documented separately.[14] Data loss will occur during this process so it is important that, where possible, the depositor carries out the deconstruction and documentation of files. ArcSDE and pGDB files should first be migrated to fGDB.

Possible migration issues: - GML cannot handle multipatch feature information / complex geometry (i.e. a 3D object stored as a single row the database (inc. texture, color, transparency, and geometric information). - breaking down GDBs to components (e.g. SHP, TIFF, etc.) loses relationships between components. All data structure has to be documented separately.

GeoMAPP have undertaken test migrations in 'Archival Challenges...'.

Possible options

- Geodatabase XML - unclear as to how useful this is outside of Esri software.[15]

File-naming

[13] http://support.esri.com/en/knowledgebase/techarticles/detail/26823

[14] this is discussed in more detail in S5 of the *Guides to Good Practice* http://guides.archaeologydataservice.ac.uk/g2gp/CS_ACE-AUTH-Katsianis

[15] http://desktop.arcgis.com/en/arcmap/10.3/manage-data/gdb-architecture/geodatabase-xml.htm

Where possible files should retain the same name as the original. In the rare cases where zip files may have the same name, then follow standard ADS procedure for differentiating content. For example:

- original_shapefile_name_shp.zip
- original_geotiff_name_tif.zip

6 Post-migration checking

- For raster images open the file in a GIS - check the locational attributes have not changed, and that the image has not been truncated or down-scaled.
- For vector conversions - it's good practice to convert a selection of your GML files (say 10%) back to Shapefile (you can do this in QGIS) and importing into a GIS, compare geometry and location and attribute fields with the original.

Storage

Data should be stored in appropriately named folders, as described in the ADS Repository Operations.[16] Any directory structure from the SIP should try to retained in the AIP. In some cases editing/restructuring may be required, any restructuring must be recorded in the Process table in the CMS.

If your original file has, or you conversion has created, an AUX or XSD (or GFS) file. Keep these with the main file (such as ASC, or GML). not to do so would be to lose information. As discussed above, look at the contents on any .xml file associated with rasters and vectors and make a judgement over it's value. As ever, a good rule of thumb is to keep anything your not sure about (they are only ever about 1Kb). Again, store these with the main file.

Preservation Geotiff should look like this:

```
/preservation
        /{original_structure}
                mygeotiff.tif
                mygeotiff.aux
                mygeotiff.xml
                /documentation
                        myraster_metadata.docx
```

Preservation GML should look like this:

```
/preservation
        /{original_structure}
                myfile1.gml
                myfile1.gfs (if present)
                /documentation
                        myfile_metadata.docx
```

Dissemination shapefile should look like this:

```
/dissemination
        /{original_structure}
                myshapefile.zip
                        myshapefile.shp
                        myshapefile.shx
                        myshapefile.dbf
```

[16] http://archaeologydataservice.ac.uk/advice/RepositoryOperations.xhtml

/documentation
myshapefile_metadata.pdf

Dissemination raster should look like this:

/dissemination
/{original_structure}
mygeotif.zip/
mygeotiff.tif
mygeotiff.aux
mygeotiff.xml
/documentation
mygeotiff_metadata.csv