GEOGRAPHIC INFORMATION SYSTEM (GIS) PROCEDURES (VERSION 1.92)

DIGITAL ARHIVISTS
ARCHAEOLOGY DATA SERVICE
https://archaeologydataservice.ac.uk/

Geographic Information System (GIS) Procedures (Version 1.92)

Created date:	26 January 2012
Last updated:	05 May 2020
Review Due:	31 March 2021
Authors:	Michael Charno, Jen Mitcham, Tim Evans, Kieron Niven, Jenny O'Brien, Leontien Talboom, Teagan Zoldoske, Olivia Foster, Digital Archivists
Maintained by:	Digital Archivists
Required Action:	
Status:	Live
Location:	https://archaeologydataservice.ac.uk/advice/PolicyDocuments.xhtml

1. Purpose of this document

1.0.1 This documents current ADS procedures for production of dissemination and preservation copies of GIS data. It contains a list of current dissemination and preservation formats and how to migrate files to the required formats. More information on this data type, can be found in the G2GP for GIS https://guides.archaeologydataservice.ac.uk/g2gp/Gis_Toc.

1.0.2 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. Consequently, 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. (From the 2009 DPC Technology Watch Report) 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.

2. Formats[1]

Offered format	Accepted	Preservation	Presentation	Notes
Geo-referenced Vector				
ArcInfo Interchange **.e00**	YES	ArcInfo Interchange **.e00**	ESRI Shapefile **.shp** For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	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 wass mainly used to exchange files between different versions of ArcInfo, but can also be read by many other GIS programs.

[1] Very extensive overview of raster formats here: http://desktop.arcgis.com/en/arcmap/latest/manage-data/raster-and-images/supported-raster-dataset-file-formats.htm.

ESRI Shapefile **.shp**	YES	Geographic Markup Language 3.2 **.gml**[2]	ESRI Shapefile **.shp** For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	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 nontopological geometry as part of a set of data files making up a spatial dataset.[3]
DataBase File **.dbf**	YES	Comma Separated Values **.csv**	DataBase File **.dbf** (+ accompanying files) For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	Occasionally .dbf files (and accompanying xml and cpg files) may be deposited as extra dbf files to an over-arching shp file (which should also have a main .dbf file), these should be exported to csv using Arc Map as well as viewed in QGIS with the main shp file to ensure that all of the data is there that should be.
.lyr	NO			
MapInfo Interchange Format **.mif** & **.mid**	YES	Geographic Markup	ESRI Shapefile **.shp** For groups of files, **.zip**	MapInfo is a commonly used GIS software package .mif contains the graphics and .mid contains any attribute

[2] From February 2016 we are now migrating to version 3.2 (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.

[3] It must be accompanied by in index file (.shx) and a dBASE DBF file that holds the attributes of the shapes in the shp file. Details on the function of each file extension are outlined in the footnotes below. **.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. **.cpg** - ESRI character encoding file. Plain text. Optional.

		Language 3.2 **.gml**[4]	archives of the formats listed above should be used for dissemination.	data and is optional. This is a suitable deposit format as is can be imported into ArcCatalog and qGIS. MIF/MID is MapInfos standard format, but most other GIS programs can also read it.[5]
MapInfo **.tab** + **.map**, .**dat**, **.id**, **.ind**	NO			
Spatial Data transfer standard **.ddf**	NO			
Vector product Format **.vpf**	NO			
Geographic Markup Language **.gml**	YES	Geographic Markup Language 3.2 **.gml**[2]	ESRI Shapefile **.shp** For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	A preferred format. It can serve as a modelling language for geographic systems as well as an open interchange format for geographic data. GML is defined by the Open Geospatial Consortium as being an XML based schema and an ISO standard.[6] GML is very suitable as a preservation format for

[4] From February 2016 we are now migrating to version 3.2 (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.

[5] 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. The file may contain multiple feature types/geometries and may require export to multiple (pont/line/poly) shapefiles.

[6] It is an ISO standard (ISO 19136) and is built on a number of other ISO standards collectively known as the 19100 family.

				Geographical data. See OGS descriptions.[7]
Other				
Geodatabases (see notes below)	YES - but see notes	Comma Separated Values **.csv** and ESRI Shapefile **.shp**	. Comma Separated Values **.csv** and ESRI Shapefile **.shp**	Although we can take Geodatabases in their original flavours, 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 Shapefile 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]
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			
Geo-referenced Raster				

[7] http://www.opengeospatial.org/standards/gml.

[8] Alternatively, QGIS has a couple of limited options for file Geodatabases, see here & here which, along with checking the input file in ArcGIS, should allow us to export data if necessary.

GeoTIFF Image **.tif** (+ **.rrd**, **.aux**, **.xml**)	YES	GeoTIFF Image **.tif** (+ **.rrd**, **.aux**, **.xml** if present)	GeoTIFF Image **.tif** (+ **.rrd**, **.aux**, **.xml** if present) For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	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.[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**) For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	An ESRI GRID is a raster GIS file format developed by Esri, which has two formats: A proprietary binary format, also known as an ARC/INFO GRID, ARC GRID and many other variations. See ESRI documentation on the binary version. A non-proprietary ASCII format, also known as an ARC/INFO ASCII GRID. The file extension is .asc, but recent versions of ESRI software also recognize the extension .grd.[10]

[9] 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. Co-ordinates can be located by looking at the file in XnView (GeoTiff section of the EXIF metadata) or by loading the layer into QGIS (Right click on layer, 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 geo-registration.

[10] 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 TIFF image and therefore do not display any values associated with a grid file. The binary version can be converted in ArcCatalog

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**) For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	As above - note that we can open this and convert
ERDAS Imagine files **.img** (**.rrd**, **aux.xml**, **img.xml**)	YES	Geo-referenced TIF Image **.tif** (+ **.aux**, **.xml**)	Geo-referenced TIF Image zipped **.tif** (+ **.aux**, **.xml**) For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	Can be opened and exported in QGIS, EDAL, ArcGIS. Compressed files with exported TIFFs being c.5x larger. Simple metadata can be found in the aux.xml and img.xml files.[11]
TIF World Files **.tif** & **.tfw** or **.tifw**	YES	TIF World Files **.tif** & **.tfw** or **.tifw**	TIF World Files **.tif** & **.tfw** or **.tifw** For groups of files, **.zip** archives of the formats listed above should	Not to be confused with the GeoTIFF format (see above), TIFF World Files use a normal .tif file alongside a .tfw "world" file providing basic georeferencing information. TFW is not the same as GeoTIFF and, adding to the

[11] http://www.loc.gov/preservation/digital/formats/fdd/fdd000420.shtml.

			be used for dissemination.	confusion, 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.[12]
JPG World **.jpg** & **.jgw** (+ **.rrd**, **.aux**, **.xml**)	YES	Geo-referenced TIF Image **.tif** (+ **.aux**, **.xml**)	JPG World **.jpg** & **.jgw** (+ **.rrd**, **.aux**, **.xml**) For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	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**) For groups of files, **.zip** archives of the formats listed above should be used for dissemination.	As above, this has geospatial information stored in the pgw element. Best just to disseminate what was given
Keyhole Markup Language **.kml**	YES	Keyhole Markup Language **.kml** (+ others e.g. .tif)	Keyhole Markup language Zipped **.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

[12] TIFF World Files will generally not include CRS information (e.g. OSGB36) which will need to be set (e.g. in the GIS) in order for them to display correctly (converting them to GeoTIFF allows this to be set).

				other images, overlays, or even 3D (COLLADA) files.
Pyramid files .ovr files (created by ESRI)	NO			
OGC GeoPackage **.gpk**, **.gpkg**	NO			

2.1 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 it's 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 newer native format for ArcGIS which stores data in various files within a folder structure. Size is limited to 1TB per dataset (though multiple dataset can exist within each geodatabase). Data can optionally be stored in a read-only compressed format. fGDB can be exported to Shapefiles and GML. The API for fGDB was made available in 2011.
- Information on exporting Geodatabases with QGIS / GDAL can be found here: https://gis.ucla.edu/node/53 & https://gdal.org/drivers/vector/openfilegdb.html
- Also ArcSDE GDB (see Katsianis)

3. Documentation / Metadata

3.0.1 Alongside the standard metadata for files, the following additional documentation is required for any GIS dataset. The current metadata template is available from the Guidelines for Depositors.[13]

Element	Description

[13] https://archaeologydataservice.ac.uk/advice/guidelinesForDepositors.xhtml.

Scale of data capture	The scale of data capture (digitizing, remote sensing, GPS, etc.), or creation.
Scale of data storage	This is the scale at which the data is stored.
Assessment of data quality	Here you can provide an assessment or statement about the quality and accuracy of the data used and/or created within the GIS.
Method of data capture	How the data used within the GIS was created or captured.
Purpose of data creation	The reason why the data was collected, and the GIS created.
Coordinate grid system	The coordinate grid system utilised within the GIS. This enables you to create a map that accurately shows distances, areas, or directions.
Data type	Geographical features are expressed as vectors, or geometrical shapes.
Source	This is the source of the data used in the creation of the GIS.
Hardware/Operating System	The hardware and operating system used.
Table attributes	Code/description.
Supporting Documentation	
Supporting documentation	Any supporting documentation should be enclosed separately. This would typically include any codes, abbreviations or terminology utilised within the GIS
Copyright	Transcripts of interviews can be important documentation particularly in clarifying those involved in recordings and allowing specific individuals to be identified.

3.0.2 This table is derived from the G2GP

https://guides.archaeologydataservice.ac.uk/g2gp/Gis_Toc.

4. Accessioning checks

4.1 Checks

- Do 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? (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)
- Check geo-referencing 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.
- Copyright/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.
- 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-sampled.
- 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.
- In rare cases multiple shapefiles may be deposited that make up one 'thing' (a plan, with different layers in a single file). These shapefiles are treated as a single object, but a relationship needs to be established
- Note that for our purposes auxiliary files (.xml etc.) are part of the overall object. This is good, as all files with the same name will be computer-matched as one object, so all of the elements (shp, shx, dbf, sbn, sbx, fbn, fbx, ain, aih, xml for the original, gml, xsd/gfs for preservation and the zip dissemination will be grouped together as one object). As long as the filename for all representations is the same (not counting the file extension), then the computer-match facility in the CMS interface should match the objects automatically

4.2 Significant properties

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

4.3 File-naming

4.3.1 Where possible files should retain the same name as the original. On occasion (and normally for dissemination), it may be necessary to create different versions of the same file. In these cases a logical naming strategy should be used, and should be accompanied by explanation in the Processes section of the CMS.

4.3.2 In cases where zip files are created for dissemination (in accordance with the 'formats' above) then follow standard ADS procedure for differentiating content. For example:

original_shapefile_name_shp.zip
original_geotiff_name_tif.zip

4.3.3 All files and metadata should be placed in the appropriate location as outlined below.

5 How to convert files

Starting Format	Procedure	End Format	Checks
ESRI Shapefile **.shp**	**QGIS** After some testing, it is recommended that CATS use QGIS 2.8+ (Wien, Madeira), 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 displayed in footnote, being careful to retain the original layer projection (sometimes can default to WGS84 or similar) (e.g. Ensure CRS: Layer CRS (EPSG:54004, World_Mercator),Datasource Options: Format: GML3.2 and XSISchema:External) . 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)	Geographic Markup Language 3.2 **.gml**[2]	
ESRI Shapefile **.shp** (BATCH)[14]	**OsGEO4W** Open the OsGEO4W Shell or Windows command prompt and type in:	Geographic Markup	

[14] If an archive has a large number of Shapefiles the conversion of these files can be done as a batch process. To make this work GDAL should be installed on your computer. The easiest way of doing this is by downloading OSgeo4W. This will make sure that GDAL is up to date on your computer, as GDAL version 1.9.0 or higher is needed for this to work. If you are unsure about your version of GDAL you can open the command prompt and type in: ' gdalinfo –version.

	1) ogr2ogr -f GML output.gml input.shp -dsco FORMAT=GML3.2 2) output.gml = name of the output file 3) input.shp = the Shapefile that needs to be converted. An example would be: ogr2ogr -f GML Arch_Brick.gml Arch_Brick.shp -dsco FORMAT=GML3.2[15]	Language 3.2 **.gml**[2]	
ArcInfo Interchange **.e00**	**ArcCatalog** 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. NB: 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 Shapefile **.shp**	
JPG World **.jpg** & **.jgw** (+ **.rrd**, **.aux**, **.xml**)	**ArcCatalog** 1) Open ArcCatalog and navigate to the folder containing the files. 2) Select and right-click file: select Export > Raster to different Format.	Geo-referenced TIF Image **.tif** (+ **.aux**, **.xml**)	

[15] More information on the different options for ogr2ogr can be found on this page - http://www.gdal.org/ogr2ogr.html. Also more information and options for the GML driver can be found on this page - http://www.gdal.org/drv_gml.html.

	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.		

6 Storage

6.1 Storing data

6.1.1 Data should be stored in appropriately named folders, as described in the ADS Repository Operations manual.[16] Any directory structure from the SIP should be retained in the AIP. In some cases editing/restructuring may be necessary, but such restructuring should be recorded in the Processes section of the CMS.

6.1.2 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 loose information. As discussed above, look at the contents on any .xml file associated with rasters and vetors 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.

```
/original
        /{original_structure}
                mygis.shp
                mygis.shx
                mygis.dbf
                mygis.sbn
                mygis.sbx
                mygis.shp.xm
```

6.2 Storing metadata

6.2.1 File metadata should be stored in an appropriate archival format with the preservation/dissemination files in a "documentation" folder within the requisite folder, for example:

```
/preservation
        /{original_structure}
                mygeotif.tif
                mygeotif.aux
                mygeotif.xml
        /documentation
                myraster_metadata.csv

/preservation
        /{original_structure}
                mygis.gml
```

[16] https://archaeologydataservice.ac.uk/advice/PolicyDocuments.xhtml#RepOp.

```
                        mygis.gfs
                        /documentation
                                mygis_metadata.csv
```

6.2.2 For dissemination, where metadata is supplied by depositor then this should be presented with the dissemination data. For example:

```
        /dissemination
                /{original_structure}
                        /mygis.zip
                                mygis.shp
                                mygis.shx
                                mygis.dbf
                                mygis.sbn
                                mygis.sbx
                                mygis.shp.xml
                        /documentation
                                mygis_metadata.csv
```

7. Creating and linking objects in the OMS tables

7.0.1 See Match Objects Overview for general overview {internal access only} see also CMS-OMS TableStructure for MOS data requirements {internal access only}

8. Tech watch / things to note

9. Archival notes

10. References

- Technology Watch Report 09-01: Preserving Geospatial Data by Guy McGarva, Steve Morris and Greg Janée 2009
- Info on Geo-TIFF: http://www.gisdevelopment.net/technology/ip/mi03117pf.htm.
- http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?topicname=types_of_geodatabases.
- http://guides.archaeologydataservice.ac.uk/g2gp/CS_ACE-AUTH-Katsianis (Sections 4 & 5).
- Overview of Multipatches: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//00q8000000mv000000.
- DPC Technology Watch Report Preserving Geospatial Data. Similar conclusion in S.6.

- Archival Challenges Associated with the Esri Personal Geodatabase and File Geodatabase Formats. GeoMAPP.
- http://www.geomapp.net/docs/GIS_Archival_Processing_Process_v1.0_final_20111102.pdf. p12
- Opening GDBs in qGIS. http://gis.stackexchange.com/questions/26285/file-geodatabase-gdb-support-in-qgis.