Section 2. Creating and Using GIS Datasets#

2.8 Why document your data?#

Working with your Geographic Information System on a regular basis as you do, you probably have a pretty good idea about what it contains, the area of the country it covers, and what its major strengths and weaknesses are likely to be. You know, for example, that your data cover the city of York, that period information is only stored to the nearest century, and that the aerial photographic interpretation to the south?west of the city is a bit dubious.

Data offered to a digital archive, however, may potentially be used by researchers from many different parts of the planet, and with widely varied levels of expertise. *They* have no way of knowing anything at all about your data unless you tell them.

In order to make sure that the maximum amount of information is delivered to the user whilst involving you, the depositor, in minimal effort, these Guides suggest a number of procedures to standardise and simplify the documentation process.

Documentation for you

Some form of record about your data ? and about what you've done to it ? is also, of course, undoubtedly useful within your own organisation. Even using data every day, it is still possible to forget about where *some* of it came from, or how the data you currently used were originally compiled from various sources.

This section introduces the issues relevant to both types of documentation, as well as discussing the detail relevant to one or the other.

2.8.1 Levels of Documentation

In documenting something so complicated as a Geographic Information System, it is possible to enter into great detail, and record everything from the data sets comprising the GIS to the sequence in which individual commands were applied to the data in order to produce your current system.

As with most things, there are situations in which great detail is required, and others where a more slimmed?down level of recording might be more appropriate. It is generally up to you as creator, maintainer, and primary user of your data to decide how much documentation is justified, and to select a suitable level for the language of your documentation; is 'cleaned coverage' appropriate, for example, or is your situation such that the more expressive:

Using the Arc/Info `clean` command, tidied up the new pottery layer in order to remove errors introduced whilst digitising from the paper map. The command was
clean pottery # # # poly

is more suitable? The former has implications for decreasing the interpretability of your documentation, whilst the latter has implications for the effort required in producing this level of detail.

In documenting data which are to be made available to others, it is often necessary to describe things more clearly ? and with a greater degree of contextualisation ? than is normally the case for internal use.

2.8.2 'Documentation' versus 'Metadata'

Metadata is discussed in detail in the general section 'Metadata'. Several definitions are offered for metadata, but one which might usefully be given here is that metadata is the means by which your *data* are transformed into *information*, interpretable to and re-usable by those other than yourself. In other words, metadata is a label for the extra details associated with any data set which enable someone else to place them into some form of context. Metadata might include information on the computer format in which the data are stored, the area of the country they relate to, etc.

Metadata in its widest sense may be considered *all* of the documentation conceivably associated with GIS data, but this guide simplifies things somewhat by using the metadata label only to apply to metadata used for *resource discovery*. As such, information suitable for entry into a digital archive catalogue itself, and which can be used to facilitate the discovery of your data by others, can be thought of as metadata, whilst the information you provide which helps people to use your data after they have accessed it may be thought of as ancillary documentation.

So... how much is enough?

Well, it depends... If you are documenting data for your own internal use, you are of course free to use as much or as little of what is recommended here as you like.

If, however, you are preparing data for deposit with a particular archive e.g. the ADS or tDAR, then you will need to comply with certain archive-specific guidelines to enable the ingest and re? use of your data. Where your data are particularly complicated, archives may recommend other ? specialised ? documentation to accompany them, and will hope to enter into dialogue with you at an early stage in order to define this.

2.8.3 Information to be Recorded

It is generally a good idea to start recording information about your data as early as possible, and ideally you should begin recording as soon as you start using or creating the data. If you wait until just before depositing to start creating metadata and documentation, it will be difficult for you to provide some pieces of information at all, and far harder to write most of the rest than it would have been at the time you were actually *doing* it.

Assuming that you choose to record relevant details as you go along, it might be useful for you to start a formal **log book** of some kind. This way, it will be easier to find information later, rather than having to rifle through various old envelopes, scrap paper, and whatever else you scribbled on at the time.

Within this log book, it is normal to record such general details as the software you are using, the versions thereof, and the type of computer and operating system (e.g. Windows PC, Mac, Sun workstation, etc.) you are running it on. As time passes, people discover problems with earlier versions of software, and if someone finds out that *SuperGIS* version 23.7 displaced all green lines on maps by 3mm, then it is undoubtedly useful for you to be able to look back through the log book and find that all your maps displaying public rights of way were created three years ago using *SuperGIS* 23.7. Knowing there is a problem, you can do something about retrospectively fixing it with adequate documentation.

Sources of Data

Information about where the data you use are acquired from is one of the most important things you can record whilst constructing and using a GIS.

Data are acquired from numerous sources, auch as mapping agencies (e.g. Ordnance Survey, USGS), local authorities, special interest groups, etc., and are gathered and displayed at a wide variety of ? often different ? scales or resolutions.

Each of these sources are of value for a different set of purposes, and each brings with it a different set of problems; data acquired at 1:50,000 scale, for example, may be ideally suited for plotting maps of artefact distributions, but wholly improper for recording the layout of individual excavation trenches (1 centimetre on a 1:50,000 map, after all, is equivalent to 50,000 centimetres, or 500 metres, on the ground).

In order to aid the user in deciding how best to incorporate your data within their own work, it is desirable to provide them with information such as the scale or resolution of the original survey, scale or resolution at which that survey was digitised into the computer, assumed errors from the data capture process (often expressed as a Root Mean Square, or RMS, error on printed maps), and the method by which the data were originally acquired (although both ultimately plotted at a scale of 1:100, a user will presumably be interested to know that one topographic data set was constructed by survey with measuring tapes and dumpy level, whilst the other is the result of a detailed survey by state of the art Total Station Theodolite).

Ownership of data is also an important attribute to record about any data set, and may well prove quite complex. Data owned by the Ordnance Survey, for example, might be used by North Yorkshire County Council to derive a new data set, 'owned' by the County Council. This, in turn, is used by York Archaeological Trust to derive a new data set, now 'owned' by them. Although little, if any, of the original Ordnance Survey resource may survive in this latest incarnation of the data, Ordnance Survey in reality continue to hold intellectual property rights which should be recognised and which may well affect the ease with which, for example, York Archaeological Trust could later *legally* sell 'their' data to Yorkshire Water.

Complicated *data trails* such as this are extremely common with digital data, and it makes life easier for everyone if the evolution of every data set is tracked through every reincarnation.

In short, then, a *non?exhaustive* list of the information you might wish to record during your everyday creation, collection, and use of data includes:

- Computer hardware used
- Computer software used
- Date the data were captured/purchased/whatever
- Who did the work
- Data source ('bought from Ordnance Survey', etc.)
- Scale/resolution of data capture
- Scale/resolution at which data are currently stored
- Root Mean Square error or other assessments of data quality
- Purpose of data set creation, where known
- Method of original data capture (Total Station Survey, etc.)
- Purpose for which *you* acquired the data (might differ from the previous information where the data were *created* by someone else for one purpose, and bought from them by you for another)
- Complete history of data ownership/rights.

Processes applied

As well as recording information such as that suggested above, most of which will probably only need recording once when you start work with a data set, it is also extremely valuable to log the manner in which data are manipulated and modified. Not only does this allow you to keep track of ? and back? track from, if necessary ? changes you make to the data, but it also allows you and others to work out how data you lifted from your local Sites & Monuments Record, for example, and incorporated into your own GIS differs from those same records still residing in the SMR. How many records have you enhanced? For how many have you had to re?enter the grid references, as you discovered that those provided by the SMR actually placed sites in the North Sea?

The sorts of information you may wish to consider logging for these purposes include:

- The date of any change/modification
- The reason for any change/modification
- The record numbers affected by the change
- Relationships to other resources; where, for example, you derive a new GIS data set by passing a mathematical filter or some other modification through an existing data set, you may wish to record the relationship formally between the original data and the new set

Where you edit an existing data set to correct spelling in text fields, or some similar operation, it makes more sense to simply record this as 'Corrected spelling throughout data set' and give the numbers of those records altered if relevant, rather than to list every single correction made to every single record. For processes such as converting an elevation matrix to a Triangulated Irregular Network (TIN) or an equally drastic data set-wide modification, it is worth recording the parameters you used in undertaking this process so that you ? and others ? may repeat or undo it in the future.

2.8.4 Dublin Core Metadata#

Much of the information recommended for you to record is most useful when it comes to actually *using* your data, and as such will probably only be downloaded by a potential user at the same time as they access your data. Certain pieces of information, though, are key in aiding the user in *finding* your data in the first place, and it is these that are explored in this section.

Many organisations around the world advocate the use of the *Dublin Core* for recording the information that helps potential users to find - and simply evaluate - your data. This information is known as 'resource discovery metadata'; information about your data (the 'resource') that helps people discover it.

Through more than three years of international development, the Dublin Core has evolved to become a series of fifteen broad categories, or elements. Each of these elements is *optional*, may be *repeated* as many times as required, and may be *refined* through the use of a developing set of sub?elements. The use of the Dublin Core within archives such as the Archaeology Data Service and tDAR is discussed in the general Project Metadata section.

...and how to create it

With complex collections of computer files such as those present in most archaeological GIS, it is extremely difficult to draw up simple rules defining what you should create metadata records *for* (the GIS as a whole, every 'layer', every original data source, etc.). As a basic guide, it is sensible for you to create one record describing the GIS as a whole, plus one subsidiary record for each major resource 'type' stored in the system. If, for example, you created a GIS for a specific region which recorded Neolithic burial monuments and Roman settlement patterns (well ? you *might*!), it would seem sensible to create one record for the whole, one for the Neolithic part, and one for the Roman part, giving three records of which one (the whole GIS) is the 'parent' and two are 'children'. If in doubt, contact the digital archive for advice.

An important factor in ensuring that a user can sensibly compare records you create with those provided by others is the use of relevant standardised terminologies and modes of expression. The Dublin Core system allows users to identify a 'SCHEME' which controls the terms stored in any one occurrence of a Dublin Core element. Thus, a user could identify the *Thesaurus of Monument Types* (RCHME 1995) as a SCHEME for the Dublin Core Subject element, and describe their resource using terms drawn from this thesaurus. As all of the Dublin Core elements are repeatable, the user could then ? if they wanted to ? repeat the Subject element and define the Getty's *Art & Architecture Thesaurus* as their SCHEME. Importantly, each use of a Dublin Core element should only include terms drawn from *one* SCHEME. Where absolutely necessary, it is possible to enter information as 'free text', not qualified by any SCHEME, but such use of free text makes it far harder for users to search across resources meaningfully, and should thus be avoided.

2.8.5 Ancillary Documentation: What to Supply and Why

Possibly the single most important piece of information you can provide above and beyond the Dublin Core catalogue entries discussed above is an idea of your data model.

This model enables potential users to discover relatively quickly what sorts of information your GIS will probably hold, and allows them to work out how the whole thing is tied together.

For a typical archaeological GIS, the information that might usefully be represented in a data model includes:

- a list of field names (and definitions) for your database
  e.g. **Address**: The postal address of the archaeological intervention being described.
- a diagram depicting the relationships between database tables (similar to Figure 3), if relevant
- a list of map/coverage/'layer' names (and definitions)
  e.g. **modernyork**: The modern streetplan for the study area, extracted from Ordnance Survey 1:1,250 scale digital mapping.

Other than the data model itself, much of the information this section advocates for entry into your project log book can usefully be passed on to the digital archive in digital form, as it is equally useful to *others* trying to make use of your data as it was to you.

Previous | Next | Contents