InTouch GIS Services

How to Capture Data and Ensure Quality Standards

To ensure information integrity, capture standards should be implemented for all geographic features digitised, derived and produced within all GIS software. Using the functionality, feature definitions and good practice associated with using ArcGIS software will help to maintain and comply with these standards

The Information capture standards outlined below relate to the following key areas:

- Database / attribute accuracy
- Attribute Tables
- Information structure standards
- Information geometry standards – topology accuracy
- Spatial reference accuracy

The standards below are drawn from best practice and should ensure compatibility with clients and other organisations systems. There may well be specific standards that clients request or particular issues relating to clients activities that necessitate additional or more rigorous working practices to be applied. For smaller 'one off' projects, some of the standards and working practices may seem a little 'over the top' but it is worth sticking to them as they are will save time and confusion at a later date, either if the project scope expands or other staff need to pick up on previous work at a later date.

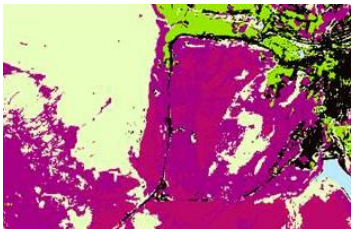
Database / attribute accuracy

Attributes refer to the database information associated with spatial features and assist in identifying the features being drawn, the way they are displayed and the analysis required. Assignment of attribute values needs to be complete, accurate and descriptive in order for the database to be meaningful. The following principles are required as a minimum to ensure data accuracy, fit for purpose and future-proofing of the dataset:

- ***Attribute accuracy*** - All attribute fields should have a valid entry and be in the correct format. Spelling errors should be checked to avoid inconsistencies with symbolisation and should be identical to that of the original source data, especially if these fields link to external data sources.
- ***Attribute completion*** – For text fields, there should be no null values unless intended. If values are not known, it is best to write 'Not Known' or 'N/A'. Where fields are numeric, avoid using a zero '0' value for missing data. This can be miss-interpreted as an actual value (eg. Recording temperature values, missing data values recorded as zero would be interpreted as being very cold!).
- ***Field type*** - Define field type according to intended use, eg, numeric data type for statistical data, text fields for alpha-numeric classifications.
- ***Field properties*** - Field properties should note the level of precision that is required, ie, length of text field, choice of integer or double for number fields. For integer data, make sure you use appropriate 'short' or 'long' integer types. Short integer types record values up to 32,767 and so cannot be used to store 6 figure co-ordinates. If using Long Integers to store co-ordinates, these can only be recorded to the maximum of 1m resolution (for detailed site projects, precision to nearest 10cm (0.1m) or to the nearest 1mm (0.001m) may be required. Be aware that specifying text fields of excessive character length can increase substantially the overall file size for the feature layer particularly if large numbers of records are to be created. Only specify the character length of text fields to contain the largest expected text string (eg. If defining a text field to hold a postcode of 8 characters do not specify a field of 50 characters in length.)

- ***Field names -*** Field names should be short and descriptive, include an underscore if separating two words, and be no more than 10 characters long as most GIS software will truncate longer field names. For example, use 'LA_Bound' rather than 'Local Authority Boundary.' Field names can be explained in more detail in the metadata within the attribute elements. Using Geodatabase files allows the use of longer field names but it is good practice to stick to the 10 character limit where possible to aid data transfer.
- ***Field content*** - Field content should be filled in using 'Sentence case' for memo fields and long description fields and 'Title Case' for filling in coded attribute fields. Caps should not be used unless they are used for codes, acronyms or alphanumeric references. This helps the readability of the attribute information.
- ***Field characters*** - Only use text and numeric characters; avoid using symbols and other conventions that may be misinterpreted, e.g., "! @ # % { } | / \ ~". Commas should not be used in fields. This will create problems when attributes are saved as CSV files for database transfer; commas will force attribute information from one column into a different column, causing inaccuracies. Separations should either be represented as a space, underscore or a hyphen.
- ***Date fields*** - Format of the date field should follow the UK standard: DD-MM-YYYY.
- ***Redundant fields*** - Conversion of data layers from one format to another will often add redundant fields or rename common fields with additional characters to avoid duplication (eg, #, _-.). It is important to clean up the database by redefining these fields or deleting redundant fields, to avoid confusion and reduce file sizes.
- ***Attribute Classification Schemes*** - If using attribute classifications, it is recommended that the character length of these classes is short; they should be brief, concise keyword / code. It is recommended that the maximum length does not exceed 20 characters and / or digits. If classifications are entered via manual data entry, it may be better to enter as a unique classification code, then use a look-up table to populate the text description of the classification. This avoids typos, and increase speed of data entry as only a single value needs to be entered.
- ***Boolean values*** - Eliminate the use of Boolean values 0 and 1 as attribute codes.
- ***Label attributes*** - Labels should be based on attributes in existing fields. Avoid using separate annotation layers for label text as this does not often transfer accurately between different GIS software. For example, if labelling tube stations with their name you would make sure that the dataset has a field called Station_Nm.
- ***Incorporate a unique ID*** - Where possible create a unique ID in your dataset to uniquely identify each feature. If using ArcGIS, do not use the object ID or FID column as a unique ID as it changes as you edit and update the dataset. For example, if each of your features relates to a document, for example borehole locations and their log reports, then each borehole needs a unique reference number or code as an identifier.

Attribute Classification categories



For clarity and simplicity it is important to choose attribute coding schemes that are well defined and in common use. General best practice principles when encoding and categorising features in the attribute table should reflect relevant best practice, policy themes and / or national data classification categories. For example; national land use database classification criteria.

Information structure standards

Most GIS support relational data structures. It is recommended that where necessary and possible, databases are structured in a logical, relational, and hierarchical or network manner. This will help the organisation and usability of data. If not possible, then related data tables should be submitted with the geographic dataset and the link field is clearly defined in the metadata.

Actual data model structures required for individual projects cannot be defined here as each project will have its own requirements relating to data sources, required functionality and clients own IT structures. However, users are directed to implement best practice and follow national ISO standards wherever possible. For example, if recording address information in the GI layer, these should be structured on the British Standard BS7666.

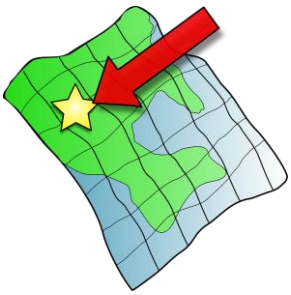
Information geometry standards – topology accuracy

Topology refers to how features are drawn in relation to one another and how they share geometric characteristics. Spatial queries and analysis that involve adjacency, connectivity and containment are often affected by geometric inaccuracies. To avoid such errors, it is important that spatial data is topologically clean and free of error. Although some GIS software allows you to define topology rules as part of setting up database integrity, these are intended as data capture/digitising and editing guidelines irrespective of the software application being used.

To ensure best practice in geometric accuracy, the user should:

- Understand the nature and use of the dataset being created and therefore be sensible in the selection of topology rules to apply; not all apply in all cases.
- Avoid sloppy information capture at all times by using appropriate snapping tolerances and tracing tools; avoid overshoots, slivers, open polygons / undershoots, unsnapped nodes, intersecting lines
- Report positional and geometric accuracy in the metadata. This is important as this defines whether the data is fit for purpose and will highlight to the end user whether to trust the use of the data layer for spatial analysis.

Spatial reference accuracy



Accuracy in spatial reference depends on:

- Coordinate Reference System (CRS);
- Map scale;

These are dependent on how the information has been captured (digitised from base maps or collected using Global Positioning Systems (GPS)) or transformed from its raw format or converted from one file format into another.

Cartographic scaling accuracy

Cartographic accuracy refers to the vertical and horizontal positioning of the data being captured.
It is important that this is captured in the metadata of the dataset to ensure fit for purpose.

- Features should be traced from appropriate base mapping as required (eg.1:1,250 OS MasterMap, 1:10,000 / 1:25,000 base mapping data as precisely as possible to ensure positional accuracy and matching features.
- Consider the capture scale of your data. This is to avoid misinterpretation of the data when someone assumes that the data is more accurate than it actually is. It is also important to consider whether datasets captured at different scales are likely to be used together and overlaid on one map.
- Storage precision must match the capture scale. The database field datatype must allow for the appropriate number of digits to be stored appropriate to the scale of the data being captured. For example, if storing OS grid co-ordinates with 6 digits for the easting and northings to a nominal accuracy of 1m, then a “Long Integer” data type can be used (a “Short Integer” data type would only allow storage of numbers up to 32,767). If sub-metre accuracy is required, then a “floating point” or “Double” data type needs to be used to allow digits to be stored beyond the decimal place (eg. Easting =456765.45 Northing = 345223.87)
- The mixing of digital map information of widely divergent scales into a common database should be avoided, as the positional accuracy of the aggregate database would be considered to be no better than that of the smallest scale. If such mixing should be necessary, however, documentation to that effect should be included in the quality and accuracy report of the metadata.
- Do not confuse the difference between the scale of base data and screen zoom scale when tracing features to create an information layer; if digitising a road network from 1:25,000 paper maps then the accuracy of the network is at best 1:25,000; irrespective of the screen zoom scale you use to digitise the features.
- It should not be assumed in conversion between data formats (ie Shapefile to MapInfo), that cartographic accuracy will be preserved. Careful management of tolerances and precision of the work environment will greatly improve the results of such a conversion. The actual conversion process (ie, which software was used) should be documented in the metadata within the Source Element.
- If possible, point locations should be imported into a database using exact known x and y coordinates rather than by digitising. Where co-ordinates are not know to sufficient precision (eg. only to the nearest 100m), digitising points from detailed mapping based on known site location position is likely to provide more accurate positioning. Batch calculation of X and Y co-ordinates to the nearest 1m can then be achieved for all digitised points.