Best practice for data/metadata collection and management

Introduction

This document sets out best practice for researchers that handle spatial data collected in the UKOTs in the South Atlantic region. The information provided in the guide aims at helping both researchers and the IMS/GIS data centre manager to produce higher quality research data, ensure their long term use and also their re-use for future science.

The basic concept is to have data that are well organised, documented, preserved and accessible, whose accuracy and validity is controlled at all times. The main results are high quality data, efficient research, saving on time and resources. Researchers themselves benefit greatly from good data management above all when the best practices outlined in this document are taken into consideration before research and a project work start.

There are precise roles and responsibilities of all parties involved: researchers are always responsible for the quality and documentation of their data however, institutions such as the IMS/GIS data centre can provide tools, infrastructures, guidelines to support the UKOTs territories in the SA region to ensure a reliable and efficient data management.

Researchers working in the territories of the South Atlantic region are asked to deliver data that comply with the standards mentioned in this document and check for their overall quality.

The authorities receiving data from the researchers are required to store them in the dedicated GIS server and be responsible for their maintenance, security and accessibility. Data provided by the researchers will not be published for three years in order to give time to the researchers to carry out and complete the data analyses and publication. After the three years, data will become available to others according to the specification in the metadata form.

Use and access restrictions, along with copyrights and intellectual property rights will be defined by this research licence agreement which is signed jointly by the researchers and a designated representative of the UKOTs in the SA region prior the research and a project work starts.

Standards for data collection

Researchers should adopt the following standards when recording their data:

1. DATE as YYYY/MM/DD

2. LAT/LONG (WGS84) in Decimal Degrees (to 5 decimal places)

3. Any blank cell should be avoided. It is preferable state if the value was 0, or not applicable, or unknown.

4. Use meaningful headers for the fields of the attribute table. Do not abuse of shortcuts otherwise interpreting the meaning of each field becomes a guess work. Use upper and lower cases to separate the words, or alternatively the underscore symbol. Do not leave spaces in the names.

5. If frequent updates of the same dataset are planned, please consider versioning each copy, for example “habitat_restoration_2014_02” and the following file is “habitat_restoration_2014_08”. Notice that for the date the format YYYY/MM/DD is always adopted. If the data are presented in a word document, e.g. a report, please think of choosing the style such as “nameofthefile.v01”.

6. Keep always a single copy of the master file, which is the current file, and store the old files in an archive folder. Be careful in deciding the frequency of the versioning: try to define properly what changes determine a “new” file. It is best practice to keep a version control table for recording all the versions and the changes occurred every time.

7. Use meaningful names for the datasets but try to keep them not too long. Once again avoid spaces and utilise underscore or a mix of lower and upper cases to separate the words.

8. Do not add text to a column filled with numeric values. For example, if a column refers to area in square metres it is not necessary to add 23m² as this value is not anymore considered as numeric but text. Instead, add the measure units to the column header.

9. If there is a reference to species names, please consider to have two columns, one for the Latin names AND one for the common names. Avoid using colloquial names, if essential ensure that you have Latin and English names as well.

10. If a list of attributes is used, perhaps create a drop down box so that typos in recording the name of locations or species is minimised. Similarly, try to be systematic when recording your data: if a capital letter is used for the initial of the names please stick to it, as well as if spaces between words were introduced.

11. Avoid recording two names in the same cell. Each record in the table should be unique. A list of names in a cell is acceptable. However, remember that querying the column and retrieving information from the column is not easy.

12. Use codes for each record in the attribute table. Codes must be individual and unique. They can be a mix between letters and numbers, but once again something meaningful, for example, TimeA_site1_recordAA.

13. Add a field for notes and comments as it is likely that during the survey it is hard to assess everything. Hence, writing down (free text) comments helps remember about the specific moment in the survey and it ensures that there are as many usable records as possible in the table. In case abbreviations are used to name the fields of the table, it is recommended to explain their meaning.

Standards for metadata

In the context of data management, metadata play a main role as they are the documentation that explains the origin, purpose, time reference, geographic location, access conditions, terms of use, responsible organisation of a dataset. Usually metadata are used for:

- Resource discovery, since they provide searchable information that helps users to easily find existing data;
- Bibliographic record for citation.

The IMS/GIS data centre has adopted the ISO19115 for compiling metadata. An empty form (.csv file) is available at SAERI webpage in the GIS-data section. Researchers are asked to fill in the metadata form. Once completed the metadata are uploaded into the metadata catalogue online dedicated webpage hosted by SAERI website. As a result, metadata can be viewed with web browsers by enabling field-specific (e.g. keywords or topic category) searching tools.

Researchers should provide metadata with the data to the representative of the UKOTs. Contrary to the data, which can have restrictions, copyrights and confidentiality, metadata are published online as the primary intention is to create a publicly accessible data catalogue which informs on existing data in the whole South Atlantic region.

Standards for formatting

All digital information is designed to be interpreted by computer programs and is by nature software dependent. Therefore, in order to make digital data open to all and not dependant on proprietary software, it is best practice to convert and save the final datasets in standard formats that most software are capable to read and that are suitable for data sharing, reuse and preservation.
On this basis, here below a suggestion of standard formats:

- .pdf
- .csv
- .xls
- .txt
- .odt
- .doc
- .tif. .tifw
- .mp4
- .flac
- ESRI shape file
- .dwg
- .tab

Notice that the conversion from a format to another may generate changes in the data itself. Thus, it is best practice to check the data for errors after the conversion. Since researchers are the people that best know their data, they should decide on the format which insures more integrity during the conversion process. In case it is impossible to make any conversion then hand the data in their raw format specifying the type of software/program that is able to read the data.

Standards for quality control and assurance

Quality control of data is an integral part of every job and it takes place at various stages: during data collection, data entry and digitisation, and data checking.
During data collection the researchers must ensure that the data recorded reflects the actual events and observations. Thus, quality control measures during the data collection may include:

1. Calibration of the instruments in order to ensure the precision and the accuracy of the measurement.
2. Taking multiple samples, observations and measurements.
3. Being consistent while capturing data using standardised methods and follow the same instructions.
4. Check with an expert if the data recorded are realistic.

During the data digitisation in a spreadsheet the researchers can avoid errors by setting up validation rules and forms (e.g. forms created in access database); by using drop down menu that refers to controlled vocabularies, code and choice lists; by labelling the variables and record names to avoid confusion.

Data acquire more value if researchers keep their minds open and consider including additional variables and parameters that widen the possible applications of the dataset; in fact data can be reused as they can be useful for other analyses and provide also new avenue for research.

Standards for data security

Data security arrangements need to be proportionate to the nature of the data and the risks involved. Researchers should state the level of data security they want to be applied to their data when they deliver them to the representative of the UKOTs in the South Atlantic region. In fact, data security may be needed to protect intellectual property rights, copyrights, commercial and environmental interests, personal or sensitive information.

To be precise: personal data are those related to a living individual who can be identified from those data. Confidential and sensitive data are data given in confidence or agreed to be kept confidential between two parties due to the sensitivity of their content. These data are not in the public domain and for example can relate to business, income, health, and environment.

The IMS/GIS data centre will ensure data security and will prevent unauthorised access, changes and destruction of data by implementing the following computer systems and files procedures:

1. Controlling the access to the individual file e.g. setting passwords, read-only and write or administrator-only permissions
2. Physically protecting the server dedicated to GIS data and protect it from power cuts by UPS systems
3. Imposing non-disclosure agreements for managers and users of confidential data
4. Not sending personal or confidential data via mail or File Transfer Protocol (FTP), but transmit them as encrypted data
5. The GIS server will be behind a firewall protected network
6. Linking the GIS server only to controlled network such as that one of the UKOTs governments and NGO (Falkland Conservation and Saint Helena National Trust)
7. Having in place back up procedures, which are encrypted, that copy the entire server system (GIS main folders and sub folders)

If researchers have particular issues with the way data security is tackled by the IMS/GIS data centre, then they should state these clearly prior to signing the research license agreement so that the issues can be addressed.

Standards for copyrights and data sharing

Researchers will have full copyright of their data unless there is a contract that assigns copyright differently or there is a written transfer of copyright signed by the copyright owner. In addition, researchers will keep control of their data after delivering them to the representative of the UKOTs (and submitted to the data centre) because the researchers will have set their conditions on data sharing in the metadata form.

By signing the research licence agreement, researchers consent to provide the raw data to the representative of the UKOTs who will save them into the server dedicated to GIS data which is managed by the IMS/GIS data centre.

The IMS/GIS data centre will never transfer data from the GIS server to anybody requesting data, without permission. All the data requests will be dealt directly by the person/organisation responsible for the dataset, named in the metadata record. This person is the only one with the rights of transferring the data to the individual or organisation that asked for the data. The data centre can pass on the data with the permission of the data holder.

If copyright applies to the data, secondary users of the data must obtain copyright clearance from the right holder before data can be reproduced. Data can be entirely or partly copied or distributed for non-commercial, teaching and research purposes without infringing copyrights provided that the owner of the data is acknowledged.

The acknowledgement should give credit to the data source used, the data distributor and the copyright holder. The researcher should specify how data should be acknowledged and cited in the metadata form.

Standards for data anonymisation

Researchers should also consider anonymising personal or commercially and environmentally sensitive data in order to make them more accessible (data sharing) but still secured. Anonymising data can be time consuming and costly, if this is not planned ahead, however if this is part of the data collection strategy it needs not to take time and the data can become part of a wider scientific resource.