# How to set up a metadata catalogue.
## The South Atlantic UK Overseas Territories example.

### Introduction

Metadata (data about data) is information that describes the “who, what, where and when” of a dataset. Hence it enables the contextual understanding and long term usability of it.

Metadata is one of the fundamental aspects of data management. A dataset without associated documentation loses 90% of its value. For instance, a department may have a collection of benthic samples stored without metadata. The samples are almost useless because whoever needs them for conducting a research or a project, will need to know as well “who, what, where and when” has been collected.
If the sample cannot be reused, the time, effort and funding that went into their data collection is essentially lost. Then, it may be necessary to recollect the samples, wasting once again time and money and duplicating the effort, and future sponsorships may not be given due to this lack of poor data management and loss of efficiency and money.

The conclusion is that metadata needs to be provided with each dataset. In a data management system it is responsibility of the data manager to ensure that all the datasets are secured and stored with the metadata. It is also a best practice to provide people with the tools for filling in metadata appropriately. Hence, a standard metadata form should be adopted and made available to the wider community as data providers or collectors can be an independent or university affiliated researcher, a NGO, a governmental department, or a single or group of people who are keen and passionate about a subject.

In each territory, in order to get metadata with the data it is essential to promote the importance of documentation (documenting) across the whole community. Some suggestions could be publicising examples of best practice versus bad data management; enforcing metadata submission within research licence agreements; persuading data collectors through the development of user-friendly tools for metadata harvesting and giving free accessibility to these tools.

Here below the practical example of what has been developed and implemented in the UK overseas territories of the South Atlantic region. The procedure of harvesting metadata is straightforward and should be easily applicable to other territories. The system is open and can be enhanced further, thus it is suggested to look at the online page where updates will be posted.

### From an excel file to a database

In 2013 only three islands in the South Atlantic could offer examples of metadata collection: Ascension, Saint Helena and South Georgia. Nevertheless, the metadata records in Ascension and Saint Helena were last updated in 2007 and those of South Georgia were available in xml format and very basic information. Therefore, the situation we faced in the South Atlantic was: two territories without metadata catalogue, two territories with metadata approximately seven years old and new data generated without providing metadata, one territory with metadata with basic and not fully informative content. On this basis and considering other factors such as the general workload of people working on the five UKOTs

and the high turnover of staff, it was decided to focus the collection of metadata on the most recent datasets. The likelihood of having the data collectors still on their job position or able to record detailed information about the data were high.

The initial step was to adopt a standard metadata form which could allow the metadata to be internationally recognised. Documenting data is important but it is less appealing than field works and there is a risk to be stuck in filling in a boring and complicated “paper work” instead of telling relevant details on data collected. As a result, we opted for the ISO 19115 which seemed to be relatively more simple than other metadata form such as that one used by INSPIRE. Additionally, we slightly shortened the ISO 19115 form as we maintained the mandatory and few optional fields.

Initially, the ISO 19115 form was saved as excel file (figure 1) and instructions on how to fill in the requested fields were provided (figure 2).

| | |
|---|---|
| **uniqueResourceID** | FK-FC-22 |
| **title** | Movement and feeding ecology of the Striated |
| **language** | eng |
| **abstract** | Movement and feeding ecology of the Striated Caracara in the Falkland Islands. The study was supported by several grants: The Darwin Initiative, Falklands Conservation, Jim and Jean Macaleer, The Acopian Family, Hawk Mountain Sanctuary. |
| **topicCategory** | biota |
| **keyword** | Falkland Islands, striated Caracara, |
| **temporalExtent** | |
| **datasetReferenceDate** | |
| **lineage** | Research was conducted on the movement and feeding ecology of the Striated Caracara, the world’s most southerly distributed raptor in its Falkland Islands stronghold. Method used: satellite tracking. The effort is linked directly to behavioral studies of a “saturation banded” population of individuals in the northwestern-most part of the Falklands Archipelago. The studies are observational and satellite-tracking work is expected to continue for at least the next 3-5 years. Data are available on request at Movebank, https://www.movebank.org/ |
| **W_long** | |
| **S_lat** | |
| **E_long** | |
| **N_lat** | |
| **spatialReferenceSystem** | WGS84 |
| **responsibleOrganisationN** | Falkland Conservation |
| **contactMailAddress** | CO@conservation.org.fk |
| **responsiblePartyRole** | user, point of contact? owner? |
| **frequencyUpdate** | unknown |
| **limitationsAccess** | Data can be used for educational purposes, but |
| **useConstraints** | Data can be used for educational purposes, but |
| **dataFormat** | |
| **accuracy** | |
| **metadataDate** | 2014/11/22 |
| **metadataPointContact** | imarengo@env.institute.ac.fk |
| **resourceType** | Dataseries |
| | |
| **original title** | |

**Figure 1.** Version of the ISO19115 metadata form adopted by four South Atlantic UKOTs.

| | A | B |
|---|---|---|
| 1 | FIELD | RULE |
| 2 | title | name of the file. Possibly add the year in the title if the data are collected repeatadly every year. Or add year and month in case data are collected seasonally. DO NOT LEAVE SPACE AND USE UNDE |
| 3 | language | eng (english); spa (spanish); fre (french); ita (italian); ger (german) |
| 4 | abstract | it should provide a clear and concise statement that enables the reader to understand the content of the dataset. E.g. "SSSI in Wales classified by habitat type with the limit of each SSSi recorde |
| 5 | topicCategory | please refer to the worksheet called topic_category |
| 6 | keyword | use keywords to idicate the general subject area of the data resource provided. Multiple entries are allowed but separated by COMMA E.g. landcover, fields boundaries TO FIND APPROPRIATE N |
| 7 | temporalExtent | this is the date or date range that identifies the content of the data. It may refer to the period of collection or the date at which the data are current IT MUST BE: YYYY-MM-DD or YYYY/YYYY if a rang |
| 8 | datasetReferenceDate | this is the date that identifies the currency of the publication data. IT MUST BE: YYYY-MM-DD |
| 9 | lineage | it should provide indication of how the data was created. The source of the data, the process used to create the data, the method of updating and quality control process. E.g. |
| 10 | W_long | long in decimal degrees XX,XX with plus or minus according the position to Greenwhich this is obtainable from QGIS |
| 11 | E_long | long in decimal degrees XX,XX with plus or minus according the position to Greenwhich |
| 12 | N_lat | lat in decimal degrees XX,XX with plus or minus according the position to equator this is obtainable from QGIS |
| 13 | S_lat | lat in decimal degrees XX,XX with plus or minus according the position to equator |
| 14 | spatialReferenceSystem | ESPG:XXXX this is obtainable from QGIS. Anyway, each territory can look at the worksheet "spatial reference system" for reference |
| 15 | responsibleOrganisationNan | this informs on who is responsible for the data resource E.g. Ordnance Survey of Great Britain |
| 16 | contactMailAddress | mail address to contact organisation or individual |
| 17 | responsiblePartyRole | role of the responsible organisation (pick one or many from the list in the party role worksheet) |
| 18 | frequencyUpdate | it identifies how often the data resource is updated (pick one from the list in the update worksheet) |
| 19 | limitationsAccess | it identifies any external restriction on ACCESS to the data. Look at the list provided. |
| 20 | useConstraints | it describes restrictions on USAGE of the data. E.g. No conditions apply; not for navigation; to be used for research purposes. Look at the list provided. |
| 21 | dataFormat | this is simply defining the type of data: shape file, raster (tiff, jpeg, ESRI grid or others), excel, csv,pdf etc |
| 22 | accuracy | how accurate are the measurament and data that you took during the survey. It is important to define it so that users can use the data correctly for further analysis or in their reports |
| 23 | metadataDate | date at which the METADATA was created or last updated. USE THE FORM YYYY-MM-DD |
| 24 | metadataPointContact | it indicates the organisation responsible for the creation and the maintenance of the metadata |
| 25 | resourceType | it informs whether the data is: dataset; dataset series (collection of datasets with common specifications); nonGeographicalDataset (pdf, word doc or excel file) |
| 26 | | |
| 27 | | |
| 28 | NOTE: for limitationsAccess and useConstraints if you have other suggestions please let me know. The idea is to make a list of sentences to be used as reference. | |

**Figure 2.** Each item in the metadata form is described so that people entering the metadata are aware of what they need to write in each cell.

Afterwards, it was decided to move from the excel file to a database. The reason was that we needed to manage and documenting information and excel is merely a file for computations. The database was necessary also because the metadata records are going to be related to the data requests sent by people looking for the data. Figure 3 depicts the structure of the table in the database.

| Column Name | Datatype | PK | NN | UQ | BIN | UN | ZF | AI | Default |
|---|---|---|---|---|---|---|---|---|---|
| region | VARCHAR(5) | | | | | | | | NULL |
| organisation | VARCHAR(45) | | | | | | | | NULL |
| id_text | VARCHAR(45) | | | ✓ | | | | | NULL |
| uniqueResourceID | VARCHAR(45) | | | ✓ | | | | | NULL |
| title | VARCHAR(255) | | | | | | | | NULL |
| language | VARCHAR(5) | | | | | | | | NULL |
| abstract | TEXT | | | | | | | | NULL |
| topicCategory | VARCHAR(100) | | | | | | | | NULL |
| keywords | VARCHAR(255) | | | | | | | | NULL |
| temporal_extent | VARCHAR(35) | | | | | | | | NULL |
| datasetReferenceDate | VARCHAR(35) | | | | | | | | NULL |
| lineage | TEXT | | | | | | | | NULL |
| W_long | DOUBLE | | | | | | | | NULL |
| S_lat | DOUBLE | | | | | | | | NULL |
| E_long | DOUBLE | | | | | | | | NULL |
| N_lat | DOUBLE | | | | | | | | NULL |
| spatialReferenceSystem | VARCHAR(35) | | | | | | | | NULL |
| responsibleOrganisationName | TEXT | | | | | | | | NULL |
| contactMailAddress | VARCHAR(80) | | | | | | | | NULL |
| responsiblePartyRole | TEXT | | | | | | | | NULL |
| frequencyUpdate | VARCHAR(45) | | | | | | | | NULL |
| limitationsAccesss | VARCHAR(255) | | | | | | | | NULL |
| useConstraints | VARCHAR(255) | | | | | | | | NULL |
| dataFormat | VARCHAR(45) | | | | | | | | NULL |
| accuracy | VARCHAR(255) | | | | | | | | NULL |
| metadataDate | DATE | | | | | | | | NULL |
| metadataPointContact | VARCHAR(255) | | | | | | | | NULL |
| resourceType | VARCHAR(45) | | | | | | | | NULL |
| originalTitle | VARCHAR(255) | | | | | | | | NULL |
| id | INT(11) | ✓ | ✓ | | | | | ✓ | |
| status | BIT(15) | | | | | | | | NULL |
| X_location | DOUBLE | | | | | | | | NULL |
| Y_location | DOUBLE | | | | | | | | NULL |

**Figure 3.** The metadata table structure and properties as saved in MySQL database. The uniqueResourceID is the fields that will join the metadata table to the table of data requests.

In the process of moving from the excel file to a structured metadata database, four fields were introduced. These fields would help the identification and retrieval of the datasets prompted by a data request; in fact the new fields refer to codes for the territory, the organisation owner of the data, and the serial number of each dataset. The concatenation of these fields results in the “unique reference ID” which makes each metadata record identifiable from the others. The Xloc and Yloc fields were added in the last iteration of the metadata form and are used to add a marker on the online map which comes with each metadata record.

The following goal was to design a simple and user-friendly interface for entering the metadata and uploading them to the catalogue online hosted at the SAERI webpage. Consequently, two points needed to be considered in order to achieve the goal: define a simple data entry form and look at how to publish the database online.

It is likely that the highest percentage of people with little knowledge in databases show familiarity with Microsoft Access. However, Access is not recognised by by Netfirm, which is SAERI web host service. On the contrary, Netfirm has MySQL as reference database. The advantage of Access is that it can be linked to external data sources, e.g. MySQL, through customised ODBC drives. This opportunity allows having a user-friendly front end form for the metadata entry in Access (figure 4) and a solid data storage and management back end in MySQL (figure 5). In this way the data entered in Access and managed in MySQL can be published online.

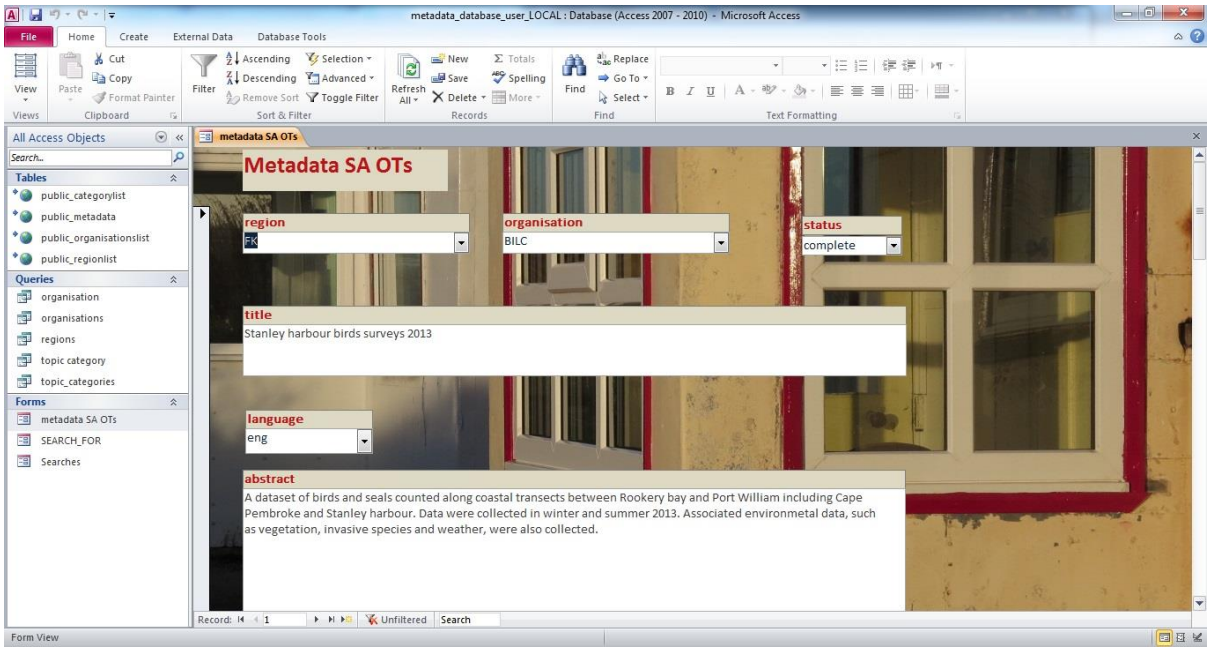



**Figure 4.** Customised user friendly interface to enter metadata into the form in Access.

It is worth mentioning that in Access the interface, which allows the metadata entering, can be customised according to what is the best solution for the people using it.

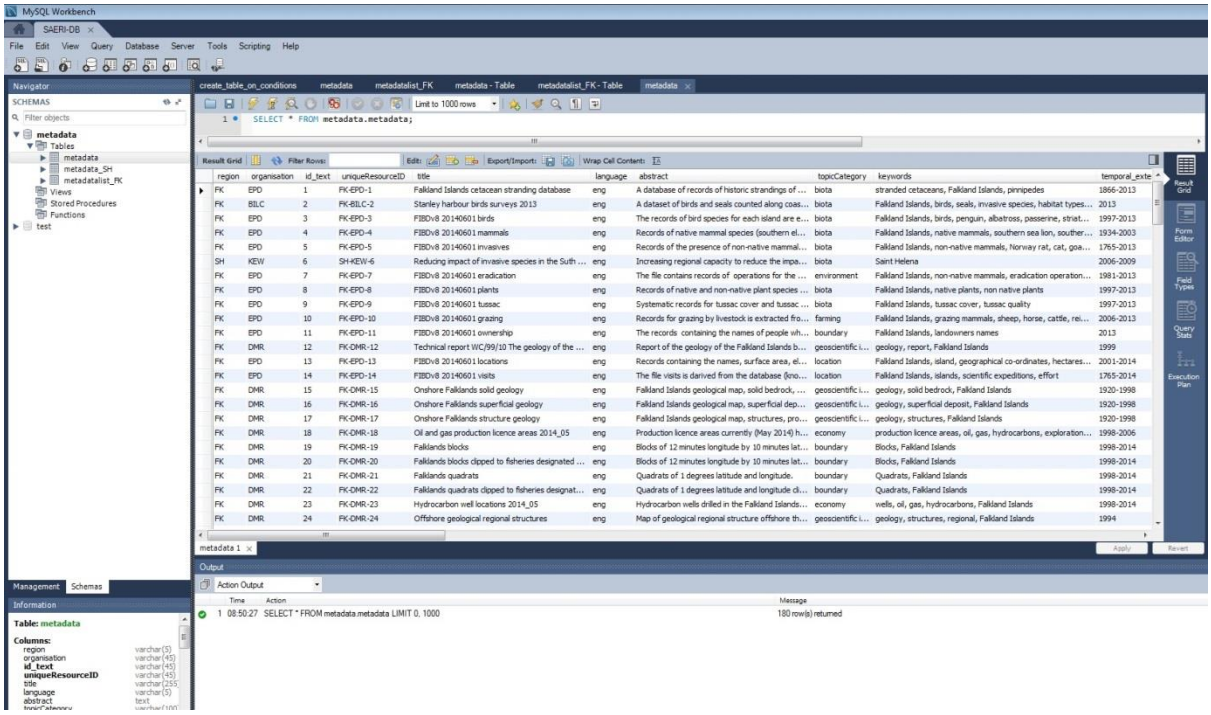



**Figure 5.** Metadata table as stored and managed in MySQL.

The installation of the ODBC drive for MySQL is straightforward. The drive is free and can be downloaded online (http://dev.mysql.com/downloads/connector/odbc/ ). The ODBC drive. In order to work correctly, needs the installation of the connector C++ (http://dev.mysql.com/downloads/connector/cpp/ ). Once both files are built in the system, the ODBC drive appears among the list of available drives (click on Control Panel\All Control Panel Items\Administrative Tools).

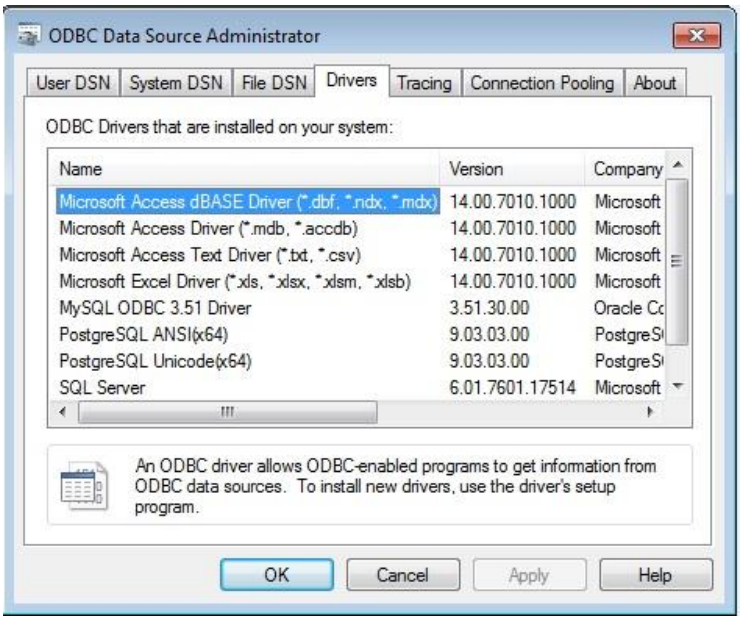



**Figure 6.** The list of ODBC drives available in the local system.

From Access the table stored in MySQL is connected by clicking on external data source, then opting for “link the data source by creating a linked table” (figure 7a), choosing MySQL as machine source (figure 7 b) and then selecting the table that we want to bring into Access (figure 7c).

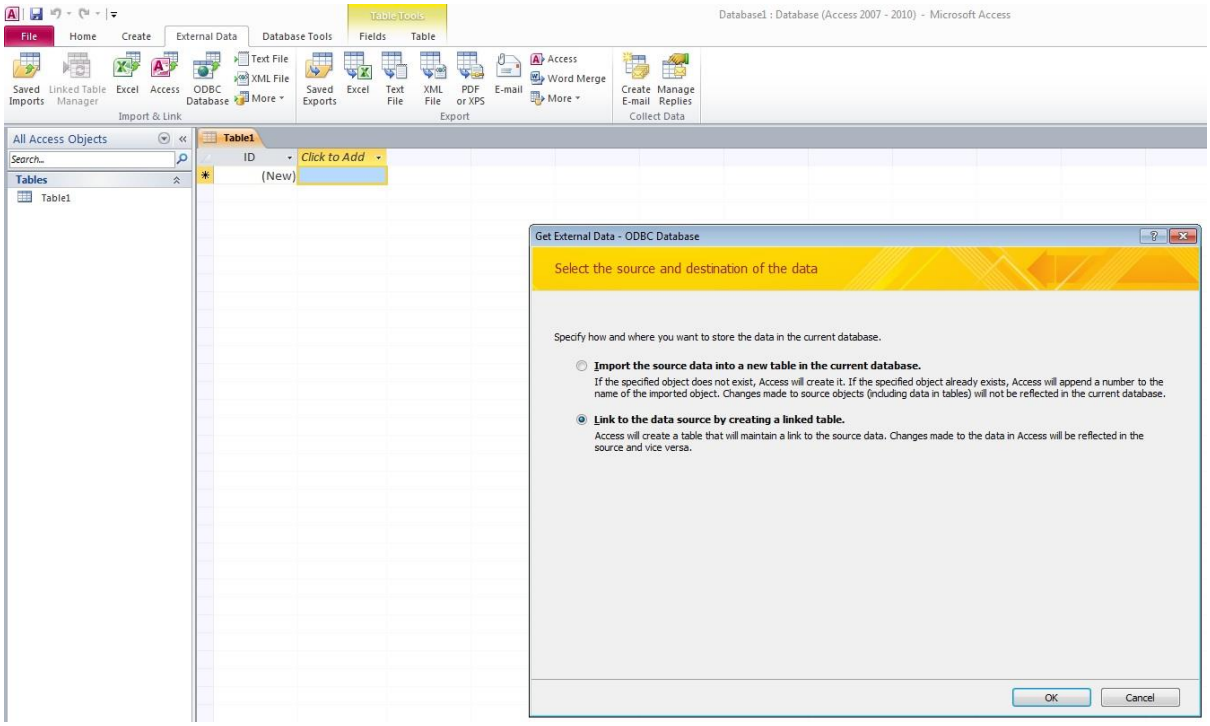



**Figure 7a.** Linking the external data source to Access.

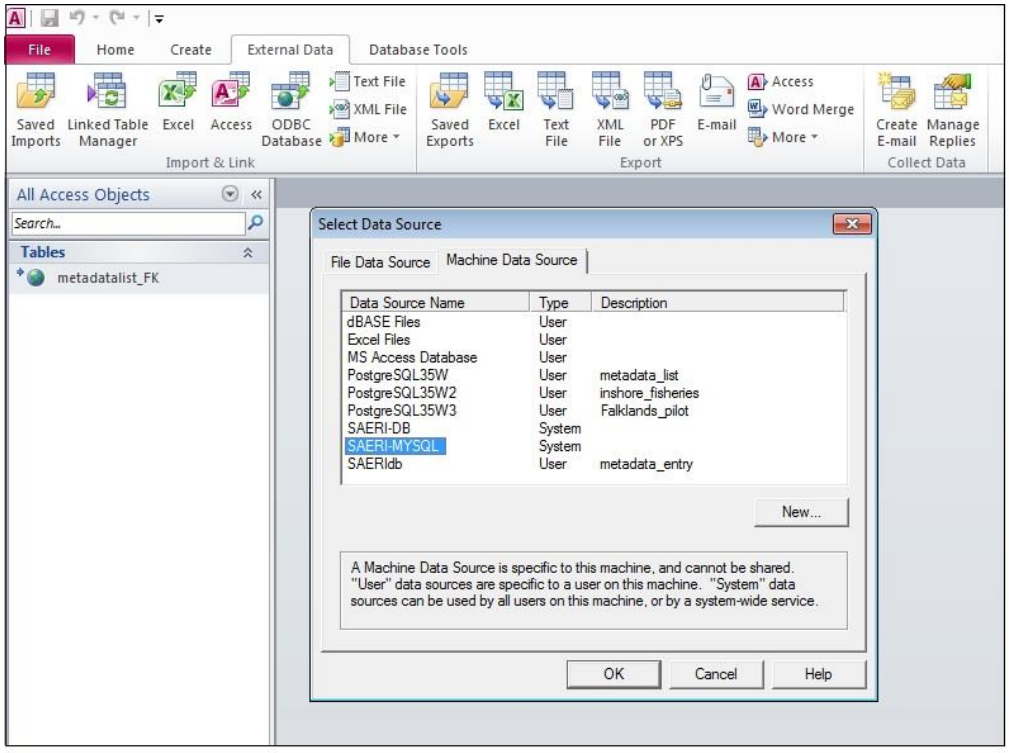



**Figure 7b.** Choosing MySQL as machine data source.

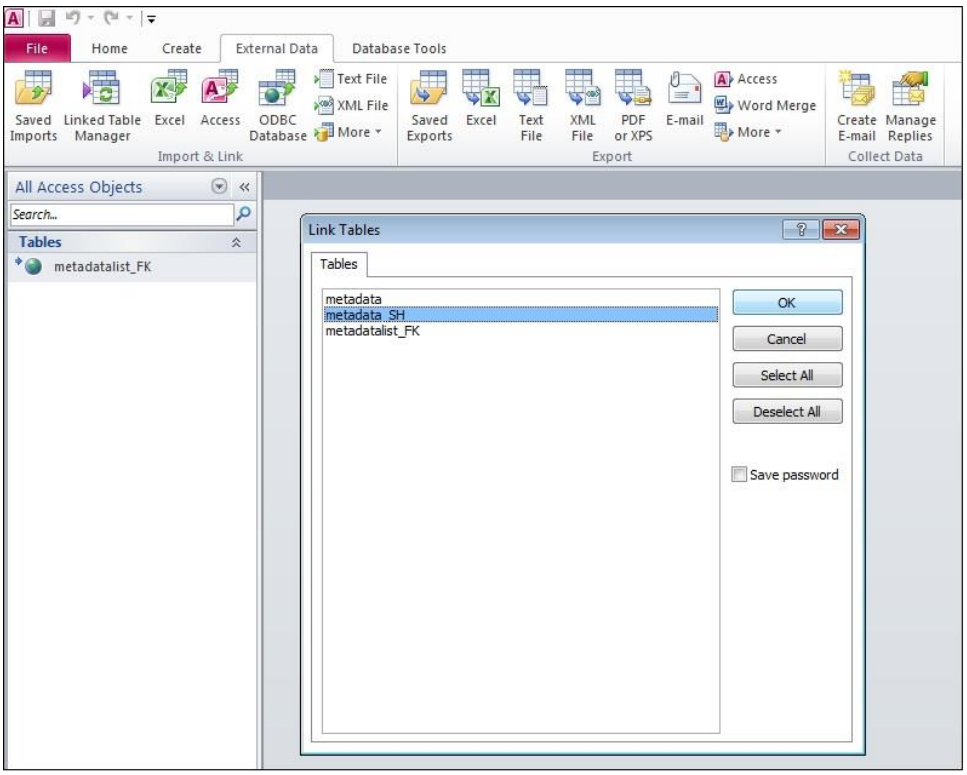



**Figure 7c.** Selecting the table that we want to link to Access

The main result from these series of steps is that now we can manipulate the table which is in MySQL within Access which, as mentioned before, is a database engine familiar to a greater number of people. The data entry form can now be built from this linked metadata table. It is worth bearing in mind that any change of the table in Access is directly recorded in MySQL too since the table is linked and not simply imported.

### Adding metadata from other islands into a central form

Everyone dealing with data is responsible for writing metadata for each dataset that create. The issue of having more than one person, within the same organisation, filling in the metadata form is solved by hosting the metadata table on a server and providing to people the metadata entry form that links to the same table. Each record in the table should have also an “auto increment” field, hence if two people are entering a record at the same time will have two different unique IDs.

Nevertheless, there is another issue to be taken into account and tackled: what if metadata records from other territories in the South Atlantic need to be gathered and joined to a central metadata table? In this case, the solution is to set up in each territory the same system: a front end in Access and a back end in MySQL, where the metadata table (called child) used in each territory has got exactly the same properties of the central table (called parent) that stores all

the metadata for the South Atlantic UKOTs (figure 8). The database repository of Ascension, Saint Helena and Tristan da Cunha's metadata tables is in the Falkland Islands.

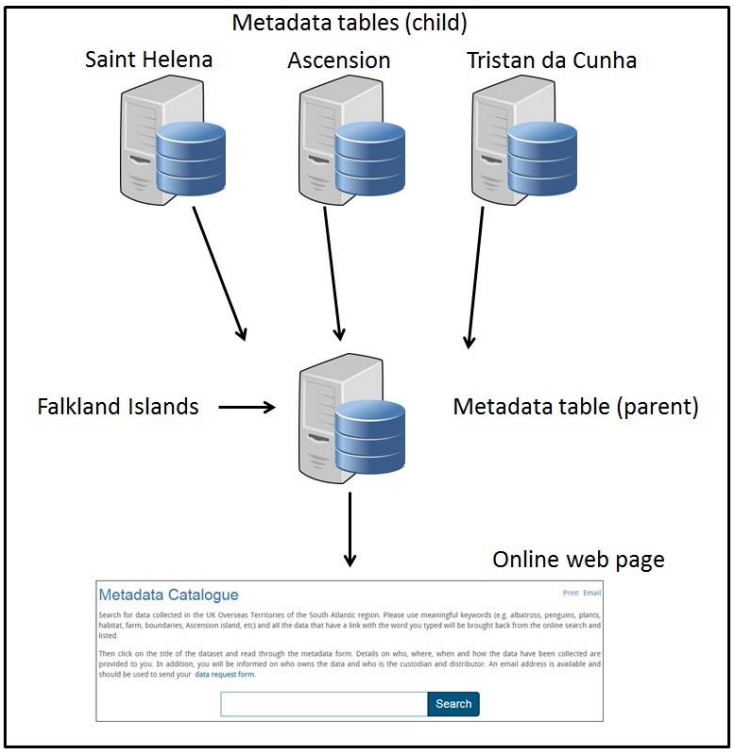



**Figure 8.** Work flow which describes how metadata tables from the territories are sent to the central metadata repository in the Falkland Islands where are joined together and then published online in to the metadata catalogue.

In MySQL it is possible to export the child metadata table in a format called .sql. Then, the exported file is sent to the person who manages the parent table and it will be imported and added to parent table. This task is delicate and the risk of making fatal mistake is high. Hence it is important that the data manager keeps the files sent from the other territories tidy and archives the old version so that only the new metadata records are added to the parent table.

Researchers and people from external organisations and institutions are requested to fill in metadata form as well and they can use the template (excel file) available online at http://www.south-atlantic-research.org/metadata-catalouge. The template is sent to the data manager of each territory who will be responsible for checking that the form is accurately filled in. If so, the metadata record will be added to the list of metadata table of the territory (figure 9).

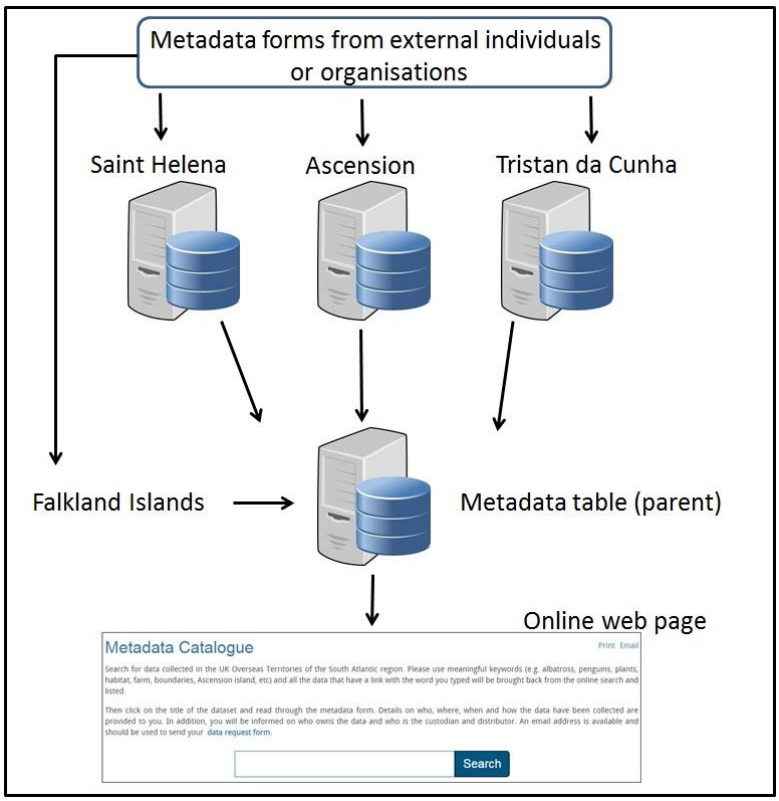



**Figure 9.** Work flow which describes how the metadata forms completed by researchers or organisations external to the territories are sent to each territory. Here are added to the local metadata tables which will be sent to the central metadata repository in the Falkland Islands where are joined together and then published online in to the metadata catalogue.

### Publishing the metadata online

As mentioned in the introduction, harvesting metadata is important as it provides longevity to the datasets, however accessing the metadata is the key to discover the existent datasets. In order to connect people with the metadata records, a catalogue online has been built with the support of the IT team of the Falkland Islands Government (figure 10). The way the catalogue is presented is not a novelty as there are several examples on the internet. Nevertheless, for the UKOTs of the South Atlantic is a great achievement.

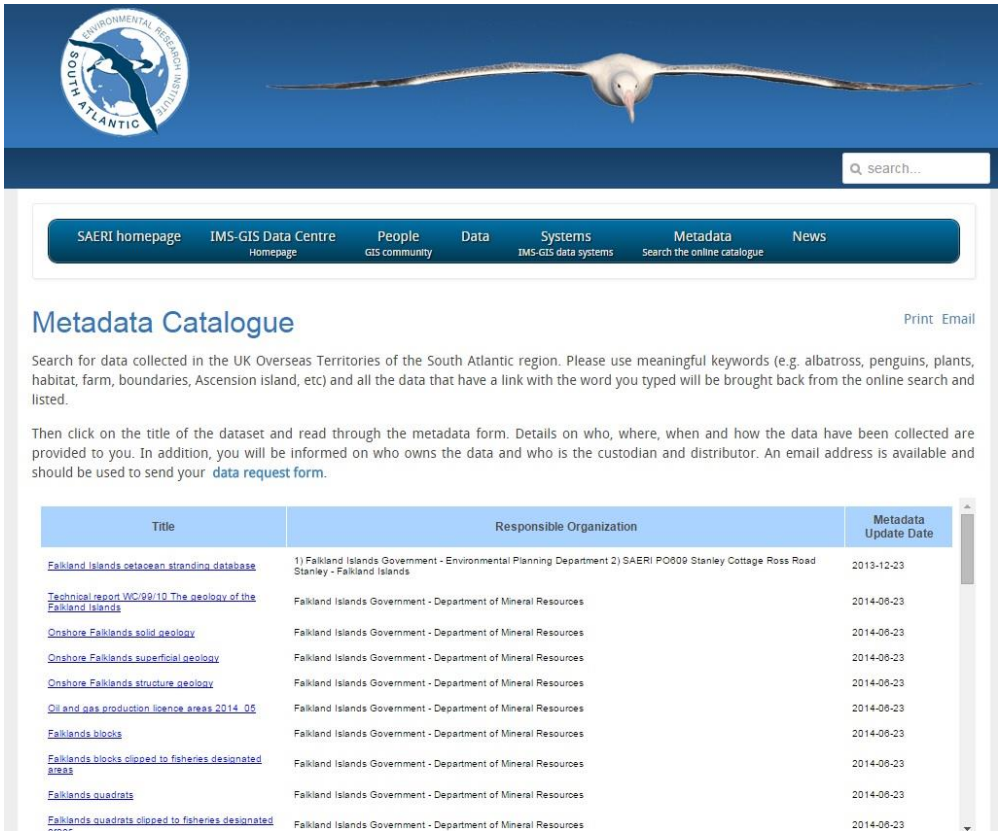
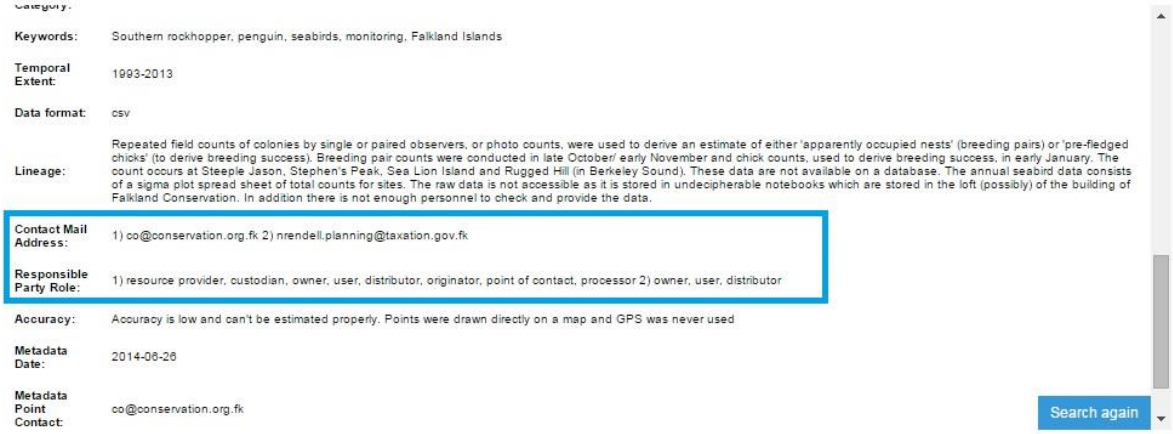



**Figure 10.** Metadata catalogue online main page showing some results from the search online of keywords.

The strength of the catalogue is its simplicity of navigation and management; in fact metadata records are searched either by keywords or by downloading the full list per territory. The information published online come directly from the database built in MySQL.

As said, the choice of MySQL was not random but taken on the consideration that Netfirm, the web host of SAERI, uses MySQL as database of reference. The management of the updates of the metadata records consists in applying the “truncate” function first and then uploading the new version of the metadata table. It is worth recalling that the table uploaded online is the addition of all the metadata tables received from the various territories.

The metadata catalogue online is a useful and helpful resource as any data user can be aware of what has been already collected and what is missing ion terms of data within each territory (figure 11).



| | |
|---|---|
| Keywords: | Southern rockhopper, penguin, seabirds, monitoring, Falkland Islands |
| Temporal Extent: | 1993-2013 |
| Data format: | csv |
| Lineage: | Repeated field counts of colonies by single or paired observers, or photo counts, were used to derive an estimate of either 'apparently occupied nests' (breeding pairs) or 'pre-fledged chicks' (to derive breeding success). Breeding pair counts were conducted in late October/ early November and chick counts, used to derive breeding success, in early January. The count occurs at Steeple Jason, Stephen's Peak, Sea Lion Island and Rugged Hill (in Berkeley Sound). These data are not available on a database. The annual seabird data consists of a sigma plot spread sheet of total counts for sites. The raw data is not accessible as it is stored in undecipherable notebooks which are stored in the loft (possibly) of the building of Falkland Conservation. In addition there is not enough personnel to check and provide the data. |
| Contact Mail Address: | 1) co@conservation.org.fk 2) nrendell.planning@taxation.gov.fk |
| Responsible Party Role: | 1) resource provider, custodian, owner, user, distributor, originator, point of contact, processor 2) owner, user, distributor |
| Accuracy: | Accuracy is low and can't be estimated properly. Points were drawn directly on a map and GPS was never used |
| Metadata Date: | 2014-06-26 |
| Metadata Point Contact: | co@conservation.org.fk |



**Figure 11.** Detail of the type of information available on line from a metadata record.

Overall the catalogue online helps the local administrations to invest money in studies that can bring innovative outcomes; it facilitates partnerships and collaborations among data collectors and users as the data sources and their roles are stated, as well as the restriction of use and access to the data. Ultimately, the catalogue is advantageous for the private sector too because it provides them with the most up-to-date and valuable information on data ownership, availability and accessibility.