# A New Generic Data Access System for Antarctic Data

**Tim D. Barnes**, Paul M. Breen, Peter J. Kirsch

Physical Sciences Division, British Antarctic Survey,
High Cross, Madingley Road, Cambridge, CB3 0ET, UK

## Abstract

The Physical Sciences Division of the British Antarctic Survey are developing a framework within which multiple, inhomogeneous datasets and associated metadata can be discovered, visualised and accessed. The framework is designed to provide a flexible interface to information pertaining to digital, non-digital and physical (e.g. ice cores) data holdings from multifarious sources. The primary elements of the framework; physical data organisation, database catalogue, and the associated web discovery utilities are described. We aim to provide a system into which bespoke visualisation and discovery mechanisms specific to a dataset, discipline or user group can be straightforwardly incorporated and administered. From the user perspective; the intent is to provide an intuitive interface allowing rapid identification, browsing and download of the data sought. An example case study highlights the flexibility of the system.

## 1.Introduction

The Physical Sciences Division of the British Antarctic Survey (BAS) undertakes studies in the fields of space physics, meteorology, atmospheric chemistry, glaciology, and oceanography. Many datasets are collected within each of the areas. The inhomogeneity between datasets presents particular challenges for the Data Managers who have the responsibility of providing the utilities for the acquisition, curation, discovery, examination and dissemination of such data. The data are of many formats and are required by a broad community of end users including scientists internal to BAS, their collaborators, other research institutes world wide, government departments, educational establishments and the general public. The introduction of a 24/7, limited bandwidth satellite link with Antarctic stations and ships enables the possibility of near real-time return for some data. This improvement in speed of return, combined with the management of many datasets from multiple remote sources increases the need for enhanced data and information management systems capable of cataloguing and discovering disparate data in an efficient, configurable and easily usable manner. This paper aims to highlight some implementations of solutions to the cataloguing, discovery and access issues.

## 2.Prior Processes

Before looking in detail at the cataloguing and discovery aspects of data management it is sensible to consider briefly the broader context and the other processes which precede them. The issues prior to cataloguing can be divided into two areas, acquisition and curation.

### 2.1 Acquisition

Until recently data acquisition was confined to annual return of media (tapes, CD's etc.) via ship. The data could be up to 15 months old before analysis. In 2005 the BASnet satellite link was established allowing 24 hour limited bandwidth (max 256 Kb/s) connectivity allowing

- Incorporation of datasets into real-time applications and forecasting models, giving "added" value to the whole dataset.
- Interactivity between scientist and remote experiments, widening the scope of scientific possibilities.
- Reduced logistics costs for retrieval of data from isolated sites by potentially reducing the number of visits for data download.
- Immediate quality assurance and control.
- Increased volumes of data as instruments can now be operated at higher resolutions and not be restricted by local disk storage limitations.

Scripts in Cambridge remotely access data on a frequency dictated by the nature of the experiment from every few minutes to once a

night. They are designed to be robust enough to cope with failures in the satellite link.

### 2.2 Curation

Once returned the data are subjected to some, if not all of the following processes

- Quality assurance and quality control filtering algorithms.
- Metadata extraction and creation.
- Post processing for format conversions, extraction of specific scientific parameters, plot generation, WWW pages etc.
- Placed on the BAS Storage Area Network (SAN).
- Real-time critical data forwarded to specific end users.
- Electronic catalogue updated with generated metadata. This catalogue is described in detail below.

## 3.Metadata catalogue

For each dataset metadata records are generated. These records may be created in a number of ways:

- Automatically generated from information contained within the data files.
- Automatically generated by post processing algorithms (E.g. QA/QC status flags).
- Automatically generated by virtue of the datafile location in a hierarchical folder structure. This predominantly applies to legacy data holdings where little if no information is held within file headers and or file names.
- Manually.

The metadata may include information regarding the data (source, location, date, time, QA/QC status etc.), auxiliary information (description of media storage type, links to files and or documents describing the experiment, software, log books and manuals etc.) and ancillary information (links to other areas of associated project, responsible scientist, etc.). It should be noted the catalogue is designed to hold information irrespective of the format of the raw data; it is not purely for electronically held data.

The metadata records are held within the tables of an Oracle database. Each table stores a different subset of related metadata, for example the Site table contains information such as latitude, longitude and geographical location of all the data recording sites. Similar tables exist for media types, data storage locations, data classifications and so forth. This database structure is flexible and additional tables can be added with the minimum of inconvenience and time. The data classifications are designed to be compatible with NASA's Global Change Master Directory [1], however any information exported from the database as an XML document could be transformed using XSLT into any appropriate schema as required [2].

A number of web applications have been developed, whereby administrators can add, edit or remove data from the database quickly and easily. Any changes made immediately filter down into the rest of the system.

## 4.Data Discovery

Once curated and logged the data must be made accessible; this is of particular importance to BAS, bound as it is by the Antarctic Treaty to make freely available all data under its ownership collected South of 60ºS [3]. In addition to this, the NERC Data Policy also states that Data Managers must “facilitate access by customers to NERC data holdings” [4].

The metadata catalogue described above forms the core utility around which applications have been written to search, examine and download data.

### 4.1 Data Search

A user will first wish to ascertain whether the specific data required actually exist. This application needs to be straightforward to use, require minimal user input and link seamlessly with other utilities allowing the data to be visualised and examined. Two user interfaces by which to discover data holdings have been created; the first a traditional text based form, the second an intuitive graphical presentation for which no user input is necessary. Each interface can navigate through the catalogue and lead the user to a useful plot or summary, the data itself, or a location in which the data can be found.

#### 4.1.1 Specific Search

The specific search interface allows the user to search by all of the metadata fields. This is achieved via a series of combo boxes and a single free text field. The input from this search is used to generate an SQL query and is then

executed; with the results being displayed in a PHP generated web page. This interface is designed for more experienced users (perhaps with a prior knowledge of the dataset properties) and administrators as selections can be narrowly targeted.

The results page lists all of the data items returned as a result of the query. It shows what the data are, where they were recorded and when they were recorded. It also provides links to allow the user to get more detail on the data item and to view metadata about a particular recording site. Figure 1 below shows a sample output table.

| Data Label | Data Class | Media Type | Start Date | Start Tim |
|---|---|---|---|---|
| 97022021d to 97041122d | SuperDARN | 8mm DAT | 20-Feb-1997 | 21:00 |
| 97041200d to 97061022d | SuperDARN | 8mm DAT | 12-Apr-1997 | 00:00 |
| 97060200d to 97071922d | SuperDARN | 8mm DAT | 02-Jun-1997 | 00:00 |
| 97072000d to 97080522d, 98012713d to 98021908d | SuperDARN | 8mm DAT | 20-Jul-1997 | 00:00 |
| 97072000d to 97080522d, 98012713d to 98021908d | SuperDARN | 8mm DAT | 27-Jan-1998 | 13:00 |
| 98021908d to 98051822d | SuperDARN | 8mm DAT | 19-Feb-1998 | 08:00 |
| 98051900d to 98073122d | SuperDARN | 8mm DAT | 19-May-1998 | 00:00 |
| 98080100d to 98101822d | SuperDARN | 8mm DAT | 01-Aug-1998 | 00:00 |
| 98101200d to 98123122d | SuperDARN | 8mm DAT | 12-Oct-1998 | 00:00 |
| 99010100d to 99021222d | SuperDARN | 8mm DAT | 01-Jan-1999 | 00:00 |
| Goose Bay SuperDARN Data 1994 | SuperDARN | Electronic Data | 03-Apr-1994 | |

*Figure 1 – Output from the Specific Search*

### 4.1.2 Graphical Search

The graphical search interface offers users a straightforward and intuitive method of accessing information regarding data holdings. It is designed such that a user need not make any input other than mouse clicks.

A main menu presents the data holdings currently available. When the streams of interest have been selected, the data availability is displayed in a form shown in Figure 2. The coverage availability is currently displayed in temporal form, however for some datasets a spatial representation may be more appropriate.

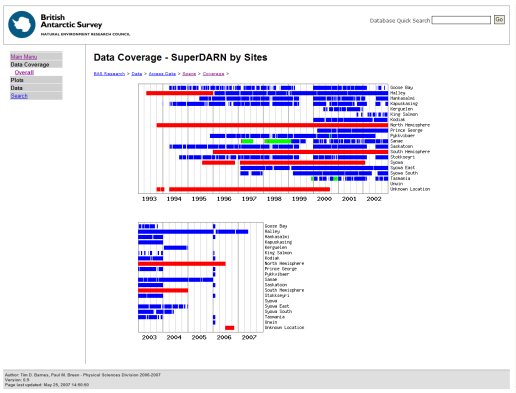


*Figure 2 – Output from the Coverage System*

The coverage of this first page may be several years. The colour bars are indicators of storage media types. A red bar means that the data only exist on hard media (including paper archives), a blue bar represents data that only exist on the SAN and green bars represent data stored on both hard media and the SAN. The coverage bars form a hyperlinked navigation and data mining tool, with the bars themselves being hyperlinks that will take the user either to a more detailed coverage of their dataset, the data itself, or plots associated with the data. The bars are divided into coverage segments and each of these is a hyperlink as well, which if selected allows the user to drill further into the database; the page will refresh the scale, now expanded and in the case of temporal data to show one years data coverage. For instance, clicking on "2006" would show the user the data coverage of their datasets for 2006 only. At that point, further mining can occur in the same way as before. The number of levels through which the coverage utility will drill is dependant upon the data being found but generally most searches can be made to that collected on a specific day. A powerful aspect of this approach other than its simplicity is the immediacy of the result in that if a bar has gaps it is immediately apparent that no data exist, the user need not make any further attempt and can proceed with another query. It is important at all times to maintain maximum expectation for the user and inform at the earliest possible stage of any likelihood of failure of request. The deeper the mining, the more specific and specialised the information relating to the datasets becomes; appropriate contextualised metadata links may appear in the left-hand menu allowing links to relevant resources such as Wiki's, documents and related data.

### 4.1.3 Comparison of Methods

Both methods of data discovery have advantages and disadvantages over the other. In the graphical search, the temporal coverage of a dataset, including its gaps, can be determined immediately. Therefore, it is easy to determine whether the data required exists or not. In the specific search, only the start date and end date of a particular data item are shown meaning the gaps in the data cannot be determined. The specific search allows for more rapid discovery of metadata information via keywords.

The graphical search method is currently biased towards temporal data; however incorporation of spatial coverage is planned. This would include data such as cruise data, where spatial information is more useful than temporal.

## 4.2 Data Examination

Once a user is assured that the data of interest exist it should be possible to seamlessly view those data. Due to the inhomogeneity of the datasets under curation and the differing demands of the user communities accessing the system, utilities have been modularly developed

to allow them to interface with the coverage view described above.

The ability to incorporate specific utilities for each dataset allows us as Data Managers to finely tune our service provision for each dataset, project or even user group. The utility that examines data is developed, in consultation with the Project Investigator who is asked the question "What information from the dataset needs to be displayed, that will satisfy 90% of users, 90% of the time that this, is in fact, the data required?". This question must be asked as the utility is designed for discovery of data, not for directly undertaking science on the data; in addition there are programming resource limitations. The aim is to provide sufficient information to allow the user to be confident that these data will provide useful input into their own specialist utilities. In the Division these data summaries are referred to as "quicklooks". The quicklooks, which may be plots, tables or any suitable summary of the data under investigation, are directly accessible from the coverage utility. A user is able to browse through the quicklooks and when an interesting plot is discovered, the relevant data can be downloaded directly and further analysis can be carried out. An example of the plots page is shown below in Figure 3.

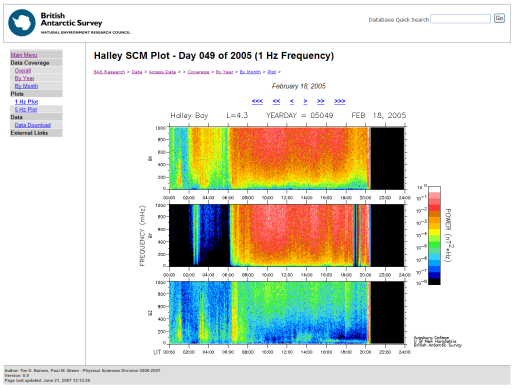


*Figure 3 – Standard plots page showing SCM data*

The style and layout of the quicklooks page varies from dataset to dataset, but they all share some common functionality. Each has a simple navigation system, by which users can go forward or backwards in time by a day, a month or a year. Each also has a link to a data download area. There are also links in the left hand menu that display an alternative plot or plots if they are available, or other information that may be of use. In some cases, instead of one plot being displayed, a series of thumbnails are displayed which represent a whole day's worth of data. A user can select one of these thumbnails to view the image in greater detail. See Figure 4 for an example.


*Figure 4 – Plots page displaying plots as thumbnails*

In some cases, quicklooks can also be a link to other applications. Examples of this include the visualisation of GIS based data in utilities such as Google Earth.

### 4.2.1 Case Study

Recently, the data management group was consulted by the BAS principal investigator concerned with the SuperDARN radar network – a network that studies the ionosphere and upper atmosphere [5].

The scientists outlined their requirements; data are to be returned nightly and placed into folders on the SAN. This data is to be made available to a global user community, with some data (outside the Antarctic Treaty area) having restricted access. Quicklook plots are to be provided at monthly and daily timescales for browsing. Links are to be provided to project specific pages and documentation.

Over a period of a few weeks, and in consultation with the scientists, a utility was created to satisfy the above requirements.

It is intended that a similar process will be carried out for the remaining datasets leading to a situation whereby the large majority of the divisional data holdings can be accessed and managed from within a common framework. There are at present about 100 datasets to develop quicklook or similar applications.

## 4.3 Data Dissemination

The final phase of user interaction is access to the actual data of interest. Again this should be available through both the coverage and quicklook utilities. Currently, the applications take the user directly to the requested raw files, which is presented in the form of a list, where the user can select the files of interest, and download them either as a zip file or as a tarball. However before final release provisions for access and authorization must be in place (not all BAS data holdings are owned by BAS) and utilities to allow conversion into more

appropriate formats for distribution need to be prepared.

## 5.Future Work

Future developments include improvement of the user interface (ability to save searches, a "shopping" basket for data).

The framework will be evolved to improve spatial data access and consideration of model data.

## 6.Summary

The framework outlined above allows users and administrators to access a large collection of disparate datasets and their associated metadata. Through this system, users can quickly browse the database via its two different interfaces, view quicklook graphs and download data directly, all with ease.

## 7.References


[1] NASA's Global Change Master Directory: http://gcmd.gsfc.nasa.gov

[2] W3C XSLT Transformations: http://www.w3.org/TR/xslt

[3] Full text of the Antarctic Treaty: http://www.ats.aq/uploaded/SIGNEDINWASHINGTON.pdf

[4] NERC Data Policy: http://www.nerc.ac.uk/research/sites/data/documents/datahandbook.pdf

[5] SuperDARN home page: http://superdarn.jhuapl.edu/