ENHANCEMENTS TO A DIGITAL LIBRARY WEB PORTAL FOR OCEAN AND CLIMATE DATA

Peter Blain [1], Ray Williams [12], Pauline Mak [14], Paola Petrelli [1], Nathan Bindoff [134]

[1] Tasmanian Partnership for Advanced Computing, Private Bag 37, Hobart, Tasmania, 7001, Australia. Email: peter.blain@utas.edu.au

[2] School of Computing and Information Systems, University of Tasmania, Private Bag 87, Hobart, Tasmania, 7001, Australia. Email: r.williams@utas.edu.au

[3] Antarctic Climate and Ecosystems Cooperative Research Centre (ACE CRC), Centenary Building, Grosvenor Crescent , Sandy Bay Tasmania 7005, Australia. Email: enquiries@acecrc.org.au

[4] Australian Research Collaboration Service (ARCS), TPAC, Private Bag 37, Hobart, Tasmania, 7001, Australia. pauline.mak@arcs.org.au

The Tasmanian Partnership for Advanced Computing (TPAC) currently hosts a digital library web portal that provides the marine and climate scientific communities with ready access to a large number of heterogeneous and geographically distributed ocean and climate datasets – some huge in scale. The portal uses the OPeNDAP framework for delivery of files. It employs an associated data harvester that crawls new and existing datasets. Metadata is retrieved by the harvester, and made available for search through the digital library web portal. The portal is being enhanced to cater for larger amounts of data, and to provide an increasingly powerful search capability to the scientific community.

1. GEOSPATIAL EXTENTS

The non-compliance of datasets to a common standard, and the lack of comprehensive metadata to accompany each dataset is a significant problem that hampers data discovery. The strategy in developing the TPAC Digital Library has been to accept all datasets, and to collect all associated metadata. This strategy has enabled the Digital Library to expand quickly and to become a useful facility, but, as it expands further, a lack of critical metadata, particularly information on the geospatial extent of each dataset, is limiting its capacity to guide researchers to the datasets in which they are interested. To help address this problem, the harvester has been modified to retrieve geospatial extents, and the portal has been modified to allow spatial searches.

The modified harvester retrieves spatial information using one of three methods. The first method uses the OGC Web Coverage Service standard. In this case, the harvester issues a WCS “get capabilities” request via HTTP. The server responds with an XML document containing the coverages in the form of latitude/longitude bounding boxes.

The WCS approach works well, but many datasets do not have an associated Web Coverage Server. Where no WCS is available, the geospatial extents are retrieved directly from the dataset's constituent files. This approach is used as a second resort because there is some overhead associated with reading a file in its entirety. NetCDF files typically contain variables that hold the axis values for each dimension. Such variables are called coordinate variables. The latitude coordinate variable, for example, is usually a one-dimensional array that holds the latitudes at each point along the axis. The harvester uses the latitude and longitude coordinate variables to determine the geospatial extents of each file.

The problem with reading files directly is that coordinate variables are not always consistently named. Instead of being named “lat” and “lon”, they might, for example, be named “x” and “y”. The harvester uses NetCDF Markup Language (NcML) to deal with this unpredictability of naming conventions. NcML can, among other things, be used to rename variables. A default

NcML document is auto-generated when a new dataset is configured, and is editable from within the portal's administration area by the digital librarian. The default NcML document follows the CF-convention version 1.4. If there is no WCS, and if the files cannot be read directly (because they require translation) the harvester uses the NetCDF-Java library to read the file using the configured NcML.

For gridded data, the harvester assumes that the first element in each axis array denotes the lower-left corner of the encompassing bounding box – and that the last element denotes the upper-right corner. For non-gridded data (such as ARGO floats) the maximum and minimum *latitudes* are found by reading through the latitude array and finding the largest and smallest values. The calculation of the maximum and minimum *longitudes* is slightly more complicated. If any of the longitudes are greater than 180 degrees, they are translated to the -180 to 180 degree range by subtracting 360. The values are then sorted sequentially. The sorted list is searched to find the largest gap between sequential pairs. The largest gap may also lie between the first and last values in the array – in other words, the gap may cross the 180 degree line. The pair of sequential longitudes on either side of the largest gap represents the maximum and minimum.

Once calculated (or retrieved), the bounding boxes are written to the database as MySQL spatial polygons. The TPAC portal has been enhanced to allow users to search for files with particular geospatial extents. Currently users can enter coordinate values defining a bounding box. The portal issues a spatial database query and returns files that overlap with the specified bounding box. The geospatial search can be used in conjunction with the previously available attribute search capability. Modifications are currently under-way to allow the spatial coordinates to be selected graphically using the Google Web Toolkit (GWT) map widget. The next section discusses architectural enhancements (such as GWT) in more detail.

2. ARCHITECTURE

The TPAC portal is being restructured based on the Spring framework. Spring is a light-weight open source framework for Java that comprises a number of modules including transaction management, security, and other enterprise-grade features. The core features of Spring are dependency injection and aspect-oriented programming. Dependency injection relieves objects of the need to create (or look up) their dependencies. All objects under dependency injection remain passive in this regard, and rely on the container to provide any dependencies at instantiation. Aspect-oriented programming separates business logic from system services (such as logging and transaction support.) Such functions (that span multiple points of an application) are isolated in special objects called aspects. This keeps Java classes focused on their intended purpose and free from ancillary concerns. Another advantage of Spring is that it emphasises the importance of programming to interfaces. Java interfaces define the methods that a class must implement – but contain no functionality themselves. Spring's emphasis on interfaces, in conjunction with dependency injection, makes it easy to automate unit testing. The result of the migration to Spring will, consequently, be a hardened enterprise application that is easily scalable and maintainable.

The TPAC portal's presentation tier is being re-written using GWT (which integrates well with Spring). GWT is a Google product that allows AJAX interfaces to be developed in Java. AJAX is a technology that facilitates the development of feature-rich web applications that are comparable in look and feel to desktop applications – but which run in a browser. GWT allows AJAX applications to be written in Java and cross-compiled into the necessary Javascript. GWT also takes care of most browser differences – making the development of AJAX applications more straight-forward than they otherwise would be. Once migrated to GWT, the TPAC portal's presentation tier will provide an improved user experience. The interface will be more interactive and responsive. Datasets and files, for example, will be searchable by drawing polygons on a map.

3. HARVESTER PERFORMANCE

Another problem facing the digital library web portal has been the fact that, although the harvester is capable of handling datasets with tens of thousands of files, in some cases datasets can include a million or more files. The harvester was struggling to retrieve the required meta-data in a practicable time frame when confronted with the largest of these datasets. Strategies have been identified and implemented to improve the harvester's performance in these situations. The improvements have achieved a three-fold increase in the speed of the harvester. The performance improvement strategies include the following:

- Multiple SQL *insert* statements have been consolidated into a single query. An insert statement containing one hundred inserts is significantly quicker than issuing one hundred separate insert statements.
- Multiple and repeated SQL *select* statements have been consolidated, and the results cached in memory, eliminating the need for additional database queries.
- The harvester now writes to memory (or heap) tables in place of permanent database tables. Memory tables are much faster than normal database tables, leading to significantly improved performance. The contents of the memory tables are copied to the permanent tables at regular intervals.
- The administration area of the TPAC portal has been modified to allow the digital librarian to block particular URLs. In this way, it is possible to avoid crawling extraneous web pages.
- The administration area of the TPAC portal has been modified to allow the harvester to be configured such that it scans datasets progressively. This means that large datasets can be segmented into more manageable parts, which can be crawled individually, but still stored as a single dataset.

4. SUMMARY

The TPAC digital library and portal is an increasingly powerful e-research tool available to the marine and climate scientific communities. The system is currently undergoing a number of enhancements. The harvester performance has been improved through database query optimisation - the improved speed allows a larger number of datasets to be harvested and made available to the scientific community. The harvester has been further enhanced with the ability to retrieve geospatial extents from dataset files - this improves the portal's search capability by allowing researchers to search for data relevant to the geographic locations in which they are interested. Other enhancements are under development, including migration to the Spring framework, which will improve scalability and improve code quality through automated unit testing. The user interface is being redeveloped using GWT, which will make the portal more interactive and help researchers quickly locate and access the data they require.