U. S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
5200 Auth Road Room 207
Camp Springs, MD 20746

Technical Note

A Lightweight Institutional Repository[†].

Robert Grumbine [‡]
Environmental Modeling Center
Marine Modeling and Analysis Branch

January 8, 2009

THIS IS AN UNREVIEWED MANUSCRIPT, PRIMARILY INTENDED FOR INFORMAL EXCHANGE OF INFORMATION AMONG NCEP STAFF MEMBERS

[†] MMAB Contribution No. 271.

[‡] e-mail: Robert.Grumbine@NOAA.gov

This page is intentionally left blank.

1 Introduction

Institutional Repositories can be major efforts, their goal being to preserve and share the intellectual contributions of the members of an insitution. This may include not merely copies of journal articles, but be fully searchable repositories of journal articles, tech notes, maps, data sets, videos, recordings, and on through all the sorts of material which might be developed at an insitution. Two examples are at Tufts University [2008], and the NOAA Aquaculture Program [2008]. Our concern in the Marine Modeling and Analysis Branch is much more modest. Namely, we wanted to have a better system for our list of branch contributions, all of which are papers of one sort or another, and for our concern content searching is not necessary.

Hence I developed a lightweight insitutional repository. It is capable of accepting publication information, a copy of the document, and automatically loads this to an appropriate location on the branch's web server. It also updates the running listing of branch contributions (at 266 when this effort started), and displays the bibliographic information as entered. There are features still lacking, a better treatment of the author list, for instance. But as this is a lightweight effort, this is not a major problem. At some later date, a heavy duty institutional repository will be brought in to place in NOAA [Campbell, 2008, personal communication], and the information from this lightweight repository can be handed off to the more capable system.

2 Architecture of the Repository

2.1 At the level of individual papers

There are two main things to be done at the level of the individual document – permit their upload (while collecting appropriate information to display the document properly) and to display the information about the paper to interested users. Since the information required for both is strongly structured, I am using XML as the basic data storage scheme. Since XML is a text-based format, this also makes it easy for people to read and edit the data files manually if needed. See Ray [2003] for more details on XML.

The upload page is at http://polar.ncep.noaa.gov/develop/repository/ This location, being under 'develop', has limited access so as to prevent the general public from making contributions to the repository. It is a simple form, the only novelty relative to typical forms being a file upload portion (the 'browse ...' button) which lets branch members search their own system for their document source. This can be a tar file of the paper, source codes, documentation and such rather than just a pdf of a document. The index.html file used is available as part of the tar file with this document in the repository.

The uploading of documents via a browse button requires a suitable script. The Perl script used, repository.pl, was derived from a demonstration script by Doyle [2008].

An author, then, goes to the repository upload page, enters the information to the form – including title, authors, additional bibliographic information such as journal and pages, year of publication, keywords, the status (draft, in press, published), and the abstract. When finished, this is translated to XML form by the script, the document is uploaded to the server and placed in the appropriate location along with the XML index, and the document is assigned a branch contribution number by the software. Users need not track what contribution number is most recent. Any element of information may be omitted. It can then be manually edited in to the XML file later.

Since the data are given in XML format, the display can be reworked separately by appropriately modifying the XSL file – the stylesheet for display [Fitzgerald, 2004]. The initial stylesheet is very plain. The current XSL is also available as part of the tar file with this document.

2.2 Constructing Collected Indices

Beyond providing the new facility of uploading new documents and providing detailed information about both old and new, we also want to retain something like the present format of a summary listing of branch publications. A simple method to accomplish this is to extract from each new submission the summary information, append it to a summary file, and run another XML style sheet to construct the html summary. This is done at the end of the original upload script, which includes invoking a script, update.sh, to manage the construction of the final index html and shtml files and move files to their final location. It also sends an email notification to the branch chief and webmaster that a new paper has been uploaded. The processing requires Xalan to be installed on the system. Update.sh is also in the tar file.

A useful feature of the XML/XSLT management of information is that we can change our minds about how the summary should be done independantly of the more detailed listings that are at the paper level. And conversely – we can also update legacy information at the level of the individual papers, an important matter as we start with a legacy of 266 documents, and propagate the information to the summary (if needed). The summary will always be pointing to the current best information on the paper.

2.3 Practicalities

There are eight files which support constructing and maintaining the repository. The locations of most on a system are arbitrary, aside from repository.pl which must be in the cgi-bin directory. Their names, purpose, location on polar currently is:

index.html	develop/repository	web page interface for authors to upload papers
repository.pl	cgi-bin/	perl script to manage document upload and web site updates
doccount	cgi-bin/	current document count
doccount2	cgi-bin/	temporary new document count
paper.xsl	mmab/images/	describe how detailed paper information should be displayed
summary.xsl	mmab/papers/	describe how summary information should be displayed
index.shtml	tmp/	reference page for the display of detailed paper information in NWS corporate web look
update.sh	tmp/	construct the new summary index to all papers and notify branch chief and webmaster of the new document

All paths given are relative to the root of the web server. All papers are in directories mmab/papers/tnNNN, where NNN is the contribution number.

A separate practicality is that we do have 266 documents from prior to the establishment of the repository. The two main steps are to, first, make links to the documents for those which already exist (about two dozen) on the system, and second, to get copies of existing documents added to the appropriate tn directory. A further refinement will be to add keyword and abstract information for documents. All this can be done incrementally, by the authors or by a knowledgeable third party.

For users establishing a repository elsewhere, this paper's tar file includes all the above files. They can then modify the files to suit their own situation.

3 Conclusions

The lightweight institutional repository described here achieved its main local goals – it provides a simple, automatic method to update the branch's list of documents, and it did not take significant time to write. Further, it is easy to modify and to revise old information. It is also sufficiently simple that it should be easy for sites with similarly lightweight interests to use to establish their own repositories.

4 References

Campbell, Michelle, Librarian, Betty Peterson Reading Room, NOAA Science Center, 2008.

Doyle, Matt Uploading Files Using CGI and Perl Last Accessed 14 May 2008 http://www.sitepoint.com/article/uploading-files-cgi-perl/

Fitzgerald, Michael, *Learning XSLT*, O'Reilly and Associates, 352 pp., 2004.

NOAA Aquaculture Program Last Accessed 26 September 2008 http://www8.nos.noaa.gov/aquaculture/publist.aspx

Ray, Erik T., *Learning XML, 2nd ed.*, O'Reilly, 400 pp., 2003.

Tufts http://dl.tufts.edu Last Accessed 21 November 2008