



DIGITAL
LIBRARY
OF THE

Caribbean

SECTION 10

Web Presence

In this Section

- Digital Library Management Systems (DLMS)
- Central dLOC Interface
 - FTP
 - OCR
- Distributed Collections



Web Presence

Digital Library Management Systems (DLMS)

- DLMS allows users to view the on-line resources and allows the library to manage their digital resources.
- The basic functionality required of a DLMS is...
 - Store and provide access to the digitized images of a resource
 - Allow search and retrieval of digitized resources
- Examples of freely-available, open-source, DLMS's include:
 - Fedora
 - Fedora is jointly developed by Cornell University and the University of Virginia Library. Funding comes from the Andrew W. Mellon Foundation and the National Science Foundation.
 - <http://www.fedora.info/>
 - Greenstone
 - Produced by the New Zealand Digital Library Project at the University of Waikato, and developed and distributed in cooperation with UNESCO and the Human Info NGO.
 - <http://www.greenstone.org>
 - DSpace
 - Developed initially as Open Repository software, but increasingly being used as a DLMS.
 - Created jointly by the Massachusetts Institute of Technology (MIT) and Hewlett-Packard (HP).
 - <http://www.dspace.org/>

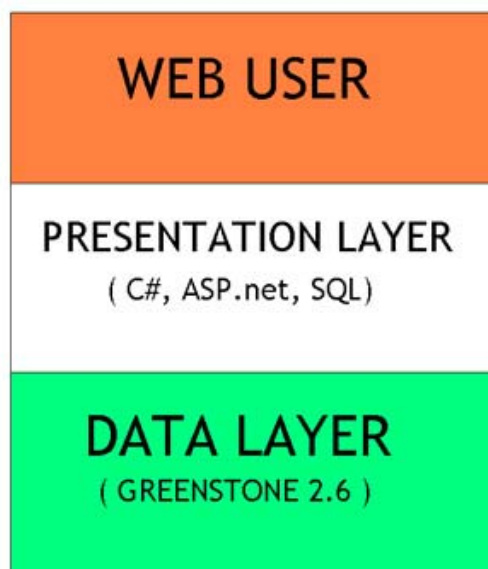
Central dLOC Interface

The Digital Library of the Caribbean project includes a central DLMS for serving all of the contributed resources under a single look-and-feel. This DLMS is hosted by the University of Florida Libraries.

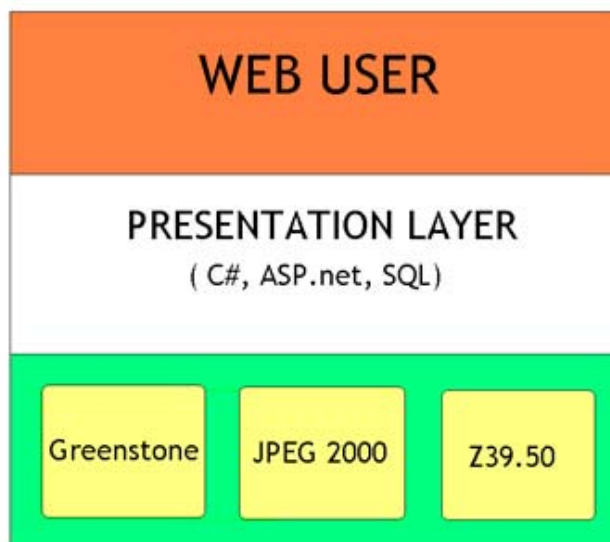
Greenstone's Digital Library System was chosen as the metadata storage, retrieval, and search engine for dLOC. Greenstone has two main components, the metadata portion and the display portion. While the metadata and indexing portion is strong, we felt that the display portion did not provide some of the functionalities we required. As a result, we chose to utilize only the metadata portion of Greenstone. All the bibliographic data ultimately resides in Greenstone 2 running under Linux.

We chose to build a multi-tier architecture with a custom presentation layer. Greenstone forms the foundation of dLOC. A presentation layer provides access to the web user. Work began on creating this layer in C#, utilizing ASP.net. The presentation layer will read all

bibliographic data from Greenstone and interact with Greenstone in real-time to perform searches. The Greenstone server will continue to serve both the data and the image. However, the user will interact with the presentation layer outside of Greenstone.



There are several advantages to this architecture, besides total control over look and feel. This provides for platform independence. Greenstone could be removed from the data layer and a variety of other digital library management systems could be used. Using this architecture will also allow us to store session state and develop user portfolios, should we decide to do so in the future. This architecture can read data from a variety of sources besides Greenstone, and allow the data and images to appear under the same interface. This provides for a continuous look and feel for the users regardless of the source of the images and data.





Data needed to drive the presentation layer are placed in a Microsoft SQL database. This database mainly stores display information. The appearance of collections depends partly on data stored in the database. This data tells the presentation layer where to look for the stylesheets and banners. It also contains the information about the hierarchy of collections. The bridge between the presentation layer and the greenstone collection(s) is stored in this database.

The database also stores basic information to assist with the display of items from Greenstone. This includes the watermarks (or icons) on the left navigation bar, downloads, and the table of contents.

This interface can be seen at the dLOC website (<http://dloc.uvi.edu>).


FTP using Go dLOC!

- Any resource that will be loaded into the central server will need to be FTPd.
- Included in the software toolkit is a FTP Client, named *Go dLOC!*.
- Prior to FTPing the package, *Go dLOC!* performs several other vital functions.
 - The created metadata is validated against several on-line XML schemes
 - JPEG derivatives are renamed for mounting on the web
 - 00001.QC2.jpg is renamed 00001.jpg
 - The information about each JPEG derivative is added to the metadata file, along with checksum information.
 - Any PDF files in the same folder are added to the metadata as downloads. These will appear on the navigation bar of the final item on the web. (see sample navigation bar below)





SEARCH
Search This Collection
Search All Collections
Last Search Results
VIEW
Full Citation
Page Images
DOWNLOAD
PDF (2 MB)
HELP
Using This Site
Contact Us
TABLE OF CONTENTS
Front Cover
Title Page
Foreword
Notes
The steps of the Haitian...
Quick Meringue
Slow Meringue
Back Cover



- If the digital resource folder is in the 'C:\dLOC\Complete' folder, this single file can be FTP'd by selecting the 'FTP this item' link from the single item form.
- *Go dLOC!* can also be run by selecting the icon in the bottom left corner of the main tracking form. When run in this mode, this will allow you to FTP every package pending in the C:\DLOC\Complete folder.
- Finally, if bandwidth is limited in your institution, you can choose to have the application run as a scheduled task from the dLOC Workstation. Instructions on setting this up can be found in Section 3: Software Toolkit Overview.
- Once the item is FTP'd to the server, the digital resource is moved to the C:\dLOC\Archive folder.
- Submittal of your resource, via FTP, triggers the following chain of events.



DIGITAL LIBRARY OF THE CARIBBEAN (dLOC)

- Within 24 hours, your resources is loaded *as is*.
- Technicians are notified that a new dLOC resource has been made available.
- Some name and spatial authority lists are applied to the resource.
- If master TIFFs were included in the resource, and the resource includes text, OCR will be performed on the master TIFFs.
- Once OCR is complete, the item is reloaded, over-laying the original.

OCR

There is one dLOC workstation dedicated to optical-character recognition (OCR): the extraction of textual content from image files. This is the step that allows a digital collection to be full text searchable.

The University of Florida has been using PrimeRecognition's PrimeOCR software for the past three years with great success. This product is actually six OCR engines from four vendors bundled together, governed by a voting engine, that typically yields better than 99% accuracy with little tweaking. It does automatic image enhancement and image zoning, if so configured.

Languages include Danish, English (US or UK), Spanish, Dutch, French, and others. Input file types include TIFF, PDF, color and grayscale images, and others. Output file types include plain text, PDF, and others.

These plain-text files become part of the digital package on the way to the dedicated dLOC server. There, they are indexed by Greenstone and these packages can then be searched through the web.

Distributed Collections

- An alternative to submitting resources to the central DLMS does exist. If an institution prefers to host their resources locally, the central server can harvest the locally-hosted metadata and direct users to that site.
- It is recommended that, at a minimum, a copy of the resource is also submitted, in full, to the central server.
 - This allows the item to be converted to text and be fully text searchable.
 - Additionally, the digital masters can be saved in the central dLOC archive, increasing preservation and reducing risk of loss.
 - The metadata created with the dLOC toolkit increases the ways that a user can locate a resource as well.
 - A user can be directed to the local resource from the central search interface.



DIGITAL LIBRARY OF THE CARIBBEAN (dLOC)

- The dLOC technical team can also work with that institution to convert the dLOC metadata to the form needed for their local DLMS.
- If a copy can not be loaded to the central server, an OAI server will need to be installed with the institution's DLMS.
 - OAI-PMH is a standard for the sharing of metadata across multiple servers.
 - Basically, this exposes your local metadata for harvesting by a central server. Then, the server stores this data, and any searches are applied against this data. If there is a hit on one of your resources, the user will be directed to the resource on your local institution.
 - The dLOC technical team will assist the local institution with establishing the OAI server, as much as possible.