Turning the Page on Digital Content

Presenters: Donald Moses (University of Prince Edward Island) and David Wilcox (DiscoveryGarden, Inc.)
Track: FUG Islandora Track.

Abstract

This FUG Islandora Track presentation will focus on the treatment of paged content within the Islandora framework. Participants will gain an understanding of the content models used, approaches to metadata for different formats, the preparation of content for upload, the ingest and derivative generation process, and the various display methods for paged content in Islandora. Participants will review production sites that include paged content.

Treatment of Paged Content in Islandora

Within the Islandora context there have been a number of ways paged content has been handled. One of the simplest approaches to paged content is to store it as a PDF document. This continues to be a viable option in Islandora and is supported by a number of solution packs including the PDF Solution Pack. A more sophisticated and preservation friendly approach to paged content is to treat pages as individual digital objects that are related to a parent object. The Book Solution Pack uses this model and it can be repurposed to satisfy many paged content models including journal content by modifying the metadata required. For instance, the Newspaper Solution Pack shares much of the same code as the Book Solution pack, but metadata requirements and the viewer are different.

PDFs

The PDF Solution Pack is probably the most straightforward approach to paged content and has been adopted by many repositories. The PDF Solution Pack provides a default MODS metadata form, provides an ingest method for PDFs, indexes the descriptive metadata and the full text of PDF, and provides a download option or a PDF viewer for displaying the content. Using collection objects in combination with appropriate metadata you could build a collections of that represent books, journals or types of paged content.

Books and Journals

The Book Solution Pack provides a number of enhancements over the PDF. While the PDF SP is a fairly basic content model, the Book SP includes a pair of content models: one that describes the parent book object (islandora:bookCModel) and is primarily metadata and another that describes the pages of the book (islandora:pageCModel). The TIFF file that represents the

page goes through a number of processes:

- the image passes through Tesseract and a plaintext OCR datastream is generated along with encoded forms of the OCR that facilitate search and term highlighting.
- image derivatives are created including JP2000, JPEG, and thumbnail images

Once processing is completed a PDF can be created that represents the entire work and is stored with the book object.

Newspapers

The Newspaper Solution Pack follows the same model as the Book SP. The primary differences are the metadata used to describe the objects, the ingest methods used, the potential collection hierarchy that might be employed (Newspaper, Volume, Issue, Page), the large format of newspaper images work best with the zoomable Islandora SeaDragon Viewer. As well methods are provided within the interface to move from page to page, issue to issue, to download the page in various formats, or to create and download a clipping of the page.

Islandora Paged Content Case Studies

The following sites will be used to demonstrate various approaches:

- The Island Magazine
  - http://vre2.upei.ca/islandmag
- PEI Legislative Documents Online
  - http://peildo.ca/
- Prince Edward Island Magazine
  - http://vre2.upei.ca/peimagazine/
- The Charlottetown Guardian
  - http://newspapers.vre.upei.ca/