

## Pathfinder Enrichment Project

**Category I** proposal for "One-Time Investments to Improve Library Operations"

**Proposed by:** Roy Tennant, Library Systems Office, 3-9494, rtennant@library

**Primary Units Affected:** Main Stack/Moffitt and NRLF

**Other Units Affected:** Library Systems.

**Project Description:**

**GOAL**

*To enable library users to make more informed decisions about the books they need.*

**STRATEGY**

Provide online images of the tables of contents and indexes of English language non-fiction books via links from Pathfinder records. Additionally, provide a growing database of the raw text of these items to provide an additional tool for locating items of interest (a searchable database of tables of contents and indexes).

**POSSIBLE BENEFITS**

- \* Library users will increasingly have more information at their fingertips to enable better decisions about which books they need to review, recall, fetch from storage, or check out.
- \* Library staff may see a drop in the numbers of items needing to be re-shelved or fetched from storage. This project should reduce the need to for the user to have to book in hand to discover whether it is useful enough to check out.
- \* A growing database of searchable table of contents and index information will be an increasingly useful method by which Library users can discover useful items. Often individual chapters within non-fiction works would be useful for a particular purpose, but remain undiscovered due to the paucity of information currently available (often just the title and subject headings).

**POSSIBLE DRAWBACKS**

- \* The copyright issues are unclear but also unlikely to be a significant barrier. *Please see Appendix I: Legal Considerations.*

**IMPLEMENTATION**

For information on implementation, please see *Appendix II: Implementation Details*. For a prototype of one aspect of this service, see <http://sunsite.berkeley.edu/PEP/> and search on "mandela". Entry into these pages would also occur from linkages in Pathfinder records.

**Staff Consulted:**

Melissa Smith Levine, Library of Congress

Scott Miller, Main Circulation

Ginny Moon, NRLF

Ralph Moon, Library Systems

**Detailed Budget:**

Scanning Workstations:

6 enhanced computers(faster CPU, more RAM @ \$2,500/each)	\$15,000.00
Additional software (e.g, Photoshop, OCR package, \$500/each)	\$3,000.00
6 Scanners (@ \$800/each)	\$4,800.00
6 Barcode readers (@ \$500/each)	\$3,000.00
6 FTE Student Employees @ \$7/hour (annual figure approx.)	\$81,350.00
Book labels (guesstimate)	\$500.00
Staff time to train staff and manage the project	
Roy Tennant, project manager (unit cost share)	\$0.00
Site managers (Main Stack, NRLF)	?
<b>TOTAL</b>	<b>\$104,653.00</b>

**Timetable:**

Late November - December 1998:	Develop procedures, benchmarks, and documentation
December 1998:	Purchase hardware and software Plan training program Install hardware and software
Late January - February 1999:	Hire and train student staff
February 1999 - end of project:	Production phase
June 1999:	Adjust procedures as needed
December 1999:	Hire and train replacement staff as needed
January 2000:	Project ends
	Project report due, including strategies for continuing the work

**Relation to Category I Criteria:**

There is perhaps no goal more central to our mission than to make it easier for our clientele to find the information they need. With the decades-long project to convert our catalog to digital form now complete, and with the accessibility of that catalog via the Web, we now have a firm foundation upon which to enhance access to our collection.

*Number of Users*

The number of users who would be affected by this project is difficult to predict, but it is likely to be significant. By selecting items that have been checked out at least once, we are likely to be targeting items that are popular. These items will show up more frequently in catalog searches, and users will be more likely to discover that they can browse the table of contents and index online. Publicity may also help alert users to watch for these links.

*Efficiency for User:*

This is a clear win, as it is much more efficient to determine whether an item will be helpful or not *before* taking a trip to the stacks. By providing the key ways in which a user can determine the appropriateness of a non-fiction work, this project has the potential to greatly enhance the efficiency of our user's interaction with The Library.

*Visibility of Service Improvement:*

The benefit of having this additional information about our collection easily available on the Web is clear to everyone. I believe it is one of the most visible and effective means to demonstrate the Library's commitment to its core mission — to improve access to information wherever it is, and certainly to our own print collection. We also demonstrate that the Library can use digital technologies to enhance access to our print collections as well as mount digital collections.

**Relation to Category II Criteria:***Efficiency of Staff:*

By providing methods by which our users can make informed decisions regarding the items they need to review or check out, staff workload should decrease over time. This may be most dramatic in sweeps and storage requests, as users will often pull a number of items off the shelf or request them from storage merely to discover that they are not worth checking out. If they have the opportunity to do this “virtually” rather than actually, the number of items requiring reshelving and returning to storage should decrease.

## **Pathfinder Enrichment Project**

### **Appendix I: Legal Considerations**

#### **The Bad News**

After discussing this with the Counsel for the Library of Congress National Digital Library Program, Melissa Smith Levine, it is clear that this is a legal grey area. There is as yet no case law to determine whether this activity is a violation of copyright. Therefore, we may be at some slight level of risk of being requested by a publisher to "cease and desist". However, Amazon.com makes available the Tables of Contents of many of the titles that they sell.

#### **The Good News**

If we take some proactive steps to address the copyright issue, we are probably unlikely to be threatened with a lawsuit. In reality, given the diversity of material we will be digitizing, the most we are probably looking at would be the necessity of removing selected items published by a publisher that requests we cease and desist.

#### **Steps to Consider**

We should consider taking the following steps to minimize our exposure, which are basically aimed at documenting a record of responsible behavior:

- 1) Make available a clear statement of our intent (to support research and education) in making this material available and the requirement of those using the material to ascertain that their use of it does not violate copyright (see *Attachment 1* for the statement used by the Library of Congress).
- 2) Send a letter to University Counsel describing the project.
- 3) Limit access to the material to the campus or university community.

#### **If All Else Fails**

If The Library is unwilling to take any risk whatsoever, Project management staff could first solicit permission from major publishing houses, and only do items from those companies. For example, one would hope that the major university presses would be amenable to the goals of the project.

## **Pathfinder Enrichment Project**

### **Appendix II: Implementation Details**

One of the goals of the implementation plan is to devise procedures that enable automation of key tasks. By devising a production process that is simple and efficient, I hope to create an infrastructure that is sustainable beyond the term of this project.

#### *Hardware*

Nine scanning workstations will be purchased and installed in Main Circulation, NRLF, and branch libraries that wish to participate.

#### *Staffing*

The project will be managed by Roy Tennant (Digital Library Project Manager), Scott Miller (Main Circulation Supervisor), Ginny Moon (NRLF Manager), and branch managers to be determined.

Approximately 27 students will be hired and trained to scan, edit images, perform OCR, and transfer the files to the server via FTP. The staffing goal is to keep all of the nine workstations in use for about eight hours a day or more.

#### *Workflow*

After books are checked in by circulation staff, project staff will select books appropriate for scanning (non-fiction works with useful tables of contents and indexes that haven't been scanned before). Selected items will be loaded on a truck and taken over to one of the project scanning stations.

Student project staff will scan each page of the table of contents and each index page, save a copy for optical character recognition (OCR, turns the image into indexable text), edit the images using Photoshop according to established practice (resampling, sharpening, etc.), and save them.

The scanned page images will be uploaded to the SunSITE server into a directory for new files. This can happen either as the work progresses or in batch mode at the end of the shift. The page scans previously saved for OCR will be processed into one file of uncorrected OCR per table of contents or index. After finishing with an item, a project label will be attached to the spine of the book so it can easily be passed over when checking returned books for items appropriate to the project.

Each day, a program will run that will collect all the filenames from their upload location, construct the appropriate URL for insertion into the Pathfinder records, and write out the update file. A Pathfinder routine will then update the appropriate Pathfinder records from the information in the file. The new files will then be transferred to the appropriate location on the server for serving on demand. Each evening, an indexing routine will run that will reindex the files of uncorrected OCR.

The two main ways users will be able to access these files will be through using Pathfinder (a link will appear in the record), and by (eventually) searching the database of uncorrected OCR. A Web interface will present this service to the user, and fetch the page images when an item is selected.

#### **Workstation Specifications**

The following applications will be available on each workstation:

GLADIS (to find the GLADIS number of the item)

Adobe Photoshop (to perform scanning and image editing)  
OmniPage Pro or equivalent (to perform OCR)  
FTP (to transfer files to the server)

Most of these applications will need to be open concurrently, and image scanning requires a good deal of RAM, so these workstations should have 256 MB of RAM. A fast CPU (400Mhz or above Pentium II/MMX) will also be required for the efficient use of Photoshop and OCR software.

### **File Naming Schemes**

Image scans will be saved using the following file naming scheme for each file:

GLADIS Number, delimiting character (-), table of contents (T) or index (I)  
identifier, delimiting character (-), page number, delimiting character  
(-), total number of pages, file type (.jpg)

For example:

1234567-T-001-003.jpg

The file naming scheme for OCR'd text files will be the same except for using ".html" as the filename extension. For example, the OCR'd text for all of the three image files that comprise the table of contents for GLADIS item #1234567 would look like this:

1234567-T-001-003.html

These file naming schemes will allow for several automated steps:

\* *Pathfinder Record Updates:* To discover what GLADIS records need to be updated, a program will each evening check the directory for uploads. Each file name will be parsed for the GLADIS record number and whether there are tables of contents pages or index pages or both. A Pathfinder update file will be written out that includes the GLADIS number of the records that require updating and the URL to insert in the 856 field of the Pathfinder MARC record. This URL will point to a Web CGI program and will pass along the filename of the first file.

\* *On-the-Fly Display:* When a user clicks on a Pathfinder link to a table of contents or index, the Web CGI program builds a Web page with navigational controls and the image of the first page. The CGI program knows various things simply from the filename: whether a requested item is a table of contents page, which number of a sequence this page represents, and how many pages there are total. Therefore, appropriate navigational controls (such as "Next Page" ) can be built-in on-the-fly by the program. The program can also use the GLADIS number to provide a direct link to the Pathfinder record.

\* *"Intelligent" Displays:* When a user clicks on a table of contents link in Pathfinder, it is likely they will also be interested in seeing the index pages as well. The Web CGI program can easily check for the existence of such files, and provide a link on the table of contents display for the user to select. This will prevent the user from having to backtrack to Pathfinder to see the index files.

### **Prototype Implementation**

To test the viability of this project, a prototype system has been constructed in about a day's work that can be viewed at:

<http://sunsite.Berkeley.EDU/PEP/>

A search on "mandela" demonstrates how the searchable database of OCR'd text would work. The word appears in the index, but not the table of contents (jump to page 11 to see it). A search on "apartheid" demonstrates retrieval of both the table of contents and the index. In any event, displaying either the index or table of contents will automatically provide a link to the other.