L/ECA/DISD/AVLIN/14

8 May, 2003

UNITED NATIONS
ECONOMIC AND SOCIAL COUNCIL

Original: English

ECONOMIC COMMISSION FOR AFRICA

Workshop on Technical Aspects of Building Digital Libraries and Electronic Information Networks

Addis Ababa, Ethiopia
10 – 11, May 2003

University of Iowa Libraries – University of Ghana, Balme Library Nationalist movement digitalization project technical details [1]

1

[1] Prepared by Prof. Anaba Alemna, University Librarian, University of Ghana

UNIVERSITY OF IOWA LIBRARIES – UNIVERSITY OF GHANA, BALME LIBRARY
NATIONALIST MOVEMENT DIGITIZATION PROJECT
TECHNICAL DETAILS

This document describes some of the technical details for the production scanning in support of the joint project between the University of Iowa Libraries and Balme Library at the University of Ghana to digitize selected material on Ghana's nationalist movement papers of the 1940's and 1950's found in the Africana Collections at Balme Library.

There are two desired outcomes of the digitization project. First, we will be converting the materials from print to electronic format in order to preserve the content. Many of the documents are quite brittle and cannot be handled repeatedly without damage. The newspapers from this time period are especially deteriorating because of the quality of newsprint. Second, by making the materials available electronically, we will be able to increase access beyond the room housing Balme Library's Africana Rare Book Collection. The electronic documents will be made available on the Internet, with searchable text. One copy will be on a server at the University of Ghana, primarily for on-campus use. A second copy will be available from the University of Iowa Libraries' Scholarly Digital Resources Center, which hosts the Center for Electronic Resources in African Studies (CERAS). This copy is intended to provide access to the rest of the world, in order to conserve precious bandwidth at Legon.

Dr. Afeworki Paulos, former International Studies and Political Science Bibliographer at the University of Iowa Libraries, and Mr. S.K. Attah, Assistant Librarian for the Africana Rare Book Collection at the University of Ghana's Balme Library, selected the materials to be scanned. These materials fall into three categories: print documents (generally fewer than 50 pages each), print newspapers, and microfilm newspapers. The print documents will be scanned first and will serve as the primary material in support of the grant. Print newspapers will be converted during phase two, with as much accomplished during the grant period as possible. Because the grant does not provide funding for microfilm equipment, we will work to secure additional funding for the third phase.

Nearly all materials are plain text or line drawings, including figures, tables, and graphs. These will be scanned on high-end scanners as bitonal (black-and-white) images at 600 dots per inch. Pages containing continuous-tone grayscale material (such as pictures) will be scanned as 8-bit images (for 256 shades of gray) at 300 dots per inch. Pages containing color of interest (excluding, for example, pages printed simply on colored paper) will be scanned as 32-bit images (for millions of colors) at 300 dots per inch.

For archival purposes, the documents will be saved as TIFF images and stored on CD-R. Three copies of each CD will be created – one copy will remain in Ghana and the other two will be shipped to Iowa. Once the archival TIFF images are received at the University of Iowa, each document will be processed in four ways. First, a minimal item-level bibliographic record will be created to provide metadata for search and retrieval. Second, optical character recognition (OCR) will be used on the document, in order to provide the capability of full-text searching. Finally, each document will be converted to a PDF file for easy download and printing over the World Wide Web. A unique digital identifier will be used to link the various

permutations of the document. This identifier will appear in the metadata and be used in the directory and file names of the PDF versions of each document.

Because the primary initial focus of the digitization project is for preservation of the materials, quality control over the archival TIFF images, as well as the derived PDF file, will be accomplished by 100% inspection. Due, however, to budgetary and time constraints, a lower threshold will be set for quality control over the cataloging records and the OCR at this initial stage. Cataloging records will be verified to ensure that the digital identifiers remain unique, but the records will generally be skeletal records. No attempt will be made to catalog publication information and physical description, nor will subject headings be assigned.

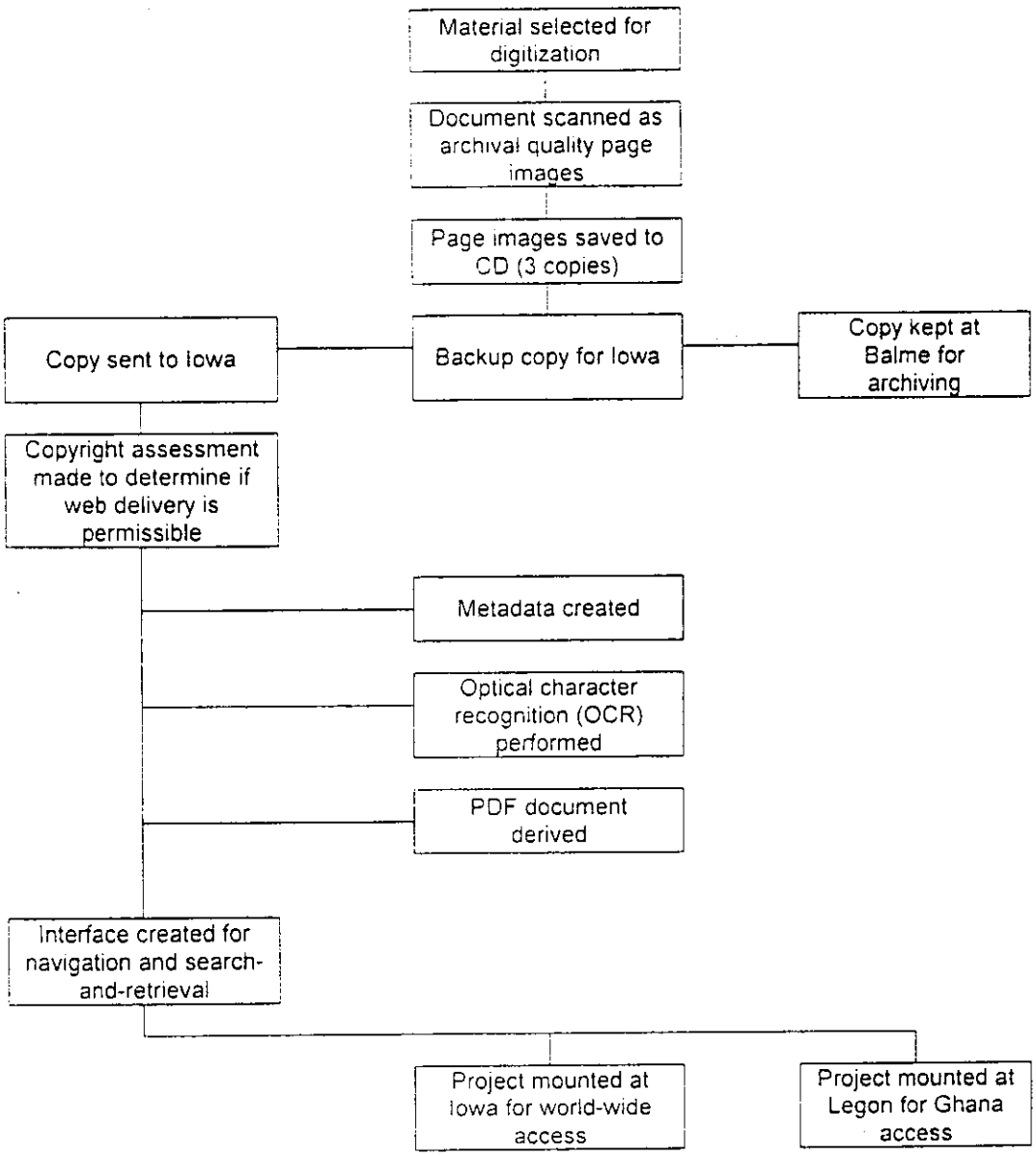
The OCR text will also be subject only to minimal quality control. Given the nature of the documents – both the paper quality and the text font and size – we expect the OCR text to have a success rate in the 80%-90% range. We expect this will have minimal impact on search and retrieval, since the search interface will retrieve whole documents and not just number of occurrences of the search term. In other words, as long as the word being searched occurs in the document more than once, the probability that at least one occurrence of the word is spelled correctly is extremely high. Further, the OCR text will never be displayed to the user. Rather, it will sit only behind the scenes in order to provide the capability for full-text searching.

We are therefore taking a pragmatic approach to quality control given the budgetary and time constraints of the project. We will have high quality assurance on the source material, and an acceptable but lower standard on the text derived from the scanned images, but implemented in such a way as to minimize the impact on the user. This method provides us with the capability of securing additional funding at a later time, in order to do fuller item-level cataloging, achieve a minimal success rate of 99.5% on the full text, and add structural markup to the text using XML. All of these steps can be done as a next step after the immediate project is completed.

A fully developed search and retrieval interface will be designed, modeled after the University of Iowa Libraries' Traveling Culture collection, and will employ a similar page navigation interface. (Traveling Culture can be viewed on the Internet at http://sdrc.lib.uiowa.edu/traveling-culture.)

The flowchart on the following page shows the process at a glance.

Production Scanning Flowchart for Nationalist Movement Digitization Project



ps 28/05/2002