The Qatar Heritage Rare Book Project: Digitization and Online Publication of Treasures from the Persian Gulf

Gabrielle V. Michalek

Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA

Abstract -- **In October, 2006 the Qatar Foundation in collaboration with Carnegie Mellon University began a Pilot project to digitize the Qatar Heritage Book Collection and make the collection available to the world. The Pilot project will make available online 300 rare materials and 5000 books from the general collection from the Heritage Library by May, 2007. This paper will describe how the project will address the practical problems surrounding intellectual control of the collection, digitization workflow, OCR and management of the Arabic character set, and the configuration of the information management system for world-wide access to the digital library.**

I. Introduction

In 2004, through the support of the Qatar Foundation for Education, Science and Community Development, Carnegie Mellon University opened a branch campus in Qatar offering undergraduate programs in computer science and business. Dr. Raj Reddy, who helped establish the campus in Qatar, was also with the Million Book Project, a universal digital library that brings together content from United States, India and China. In November, 2005 members of the Carnegie Mellon University Libraries visited Qatar and were introduced to the Heritage Library Collection. It immediately became clear that this magnificent collection should be made available to researchers around the world through the creation of a Digital Library of Qatar. An agreement was reached between the Qatar Foundation and Carnegie Mellon University to undertake a Pilot project to digitize and make available a portion of the Heritage Collection. The Pilot project will illuminate problems on a small scale; however, these problems may become larger when the full-blown project to digitize the rest of the Heritage Book Collection is conducted.

II. The Heritage Book Collection

The Heritage Book Collection, located in Doha, contains approximately 120,000 items; 10,000 of which are extremely rare or unique. The original focus of the collection centered on Western views of the Middle East. The collection consists of books, manuscripts, newspapers, journals and maps. The materials are written in several languages with English and Arabic the most prevalent, but also including German, French, Farsi, and Hebrew.

For the purpose of the project the collection will be divided between rare materials and materials from the general collection. The goal of the Pilot project is to make available 300 rare materials and 5000 materials from the general collection. Gabrielle V. Michalek is responsible for overseeing the digitization of the rare materials as well as cataloging of the entire collection and supervising the information management system. Mr. Kiran Kumar is responsible for the digitization of the general collection. The Pilot project will take place between October, 2006 and May, 2007.

III. Intellectual Control

The curator of the collection, Mr. Mohammed Hammam Fikry, has spent many years creating a FileMaker database of bibliographic information describing the collection. This database has proven to be an effective management tool for local use. The Filmmaker database lays a solid foundation for a bibliographic resource that can provide intellectual control. However, it is usefulness is

limited since it is not a networked database. One of the major challenges of the project is to transfer or recreate intellectual control by creating MARC records and inputting them in the OCLC database. MARC is a reliable and internationally accepted standard for the cataloging of library holdings. This would allow the contents of the library to be surfaced to a world community, providing easier access to users. Creation of the MARC records can be done in one of two ways. The first is to send digital surrogates of the title page to outside cataloguers who would provide a descriptive catalogue of the digital book object. The second method is to copy catalog on-site. Using this method we would be able to catalogue the physical as well as the digital object. The workflow for cataloguing the library collection and entering MARC records into the OCLC database will be resolved during the beginning phases of the Pilot project.

IV. Digitization Equipment and Workflow

The equipment selected to perform the high end scanning for the project will be the i2S DigiBook SupraScan Scanner System. This scanning system will allow us to make high quality digital surrogates without damaging the original materials. Rare materials will be digitized at 300 DPI, 24 bit color, using a lossless JPEG 2000 compression. The general collection will be digitized at 600 DPI, bitonal scan with a CCITT G4 compression.

Staffing for the project will consist of citizens of Qatar, expatriates living in Qatar, and labor brought into Qatar, most likely from India. The workflow will include multiple scanning stations working two shifts.

V. QUALITY Control

Once images are digitized they will be post-processed to de-skew and to remove lines and artifacts. The post-processed files will be OCR'd to create a text file for searching. The images and derivative files will then go through a quality control process where OCR correction may be made. After the process is completed files will be converted to PDF to incorporate the series of scanned images into one file. The files are then exported to the retrieval system for indexing and searching.

VI. OCR and Management of the Arab Character Set

Several languages are represented in the collection of which English and Arabic are the most common. We have excellent OCR technology for English; however, Arabic presents a greater challenge. Arabic is cursive with overlapping letters and diacritic marks making OCR of Arabic a technical challenge. While we acquired commercial software to OCR Arabic we are not satisfied with its rate of accuracy. Another goal of this project is to collaborate with our colleagues, especially the people working on the Bibliotheca Alexandrina, to improve the accuracy of Arabic OCR. By completion of this project, we believe that the quality of Arabic OCR, as well as translation devices for some of the other languages represented in the collection will be vastly improved.

Because the project is centered in the Middle East and the Heritage Book Collection represents one of the finest collection of Middle Eastern culture it was imperative that the end user interface for the project supports both English and Arabic languages. The project will have a custom end user interface that will support searching in multiple languages and provide a multilingual display. The vendor we have chosen for the project already has experience working with the Arabic character set from a pervious project with a group of Jordanian universities.

VII. Information Management System

The information management system that we have selected for the project is a SIRSI/Dynix product called Digital Library System, created by one of their technology partners, PTFS Digital Archive Solutions. PTFS has sold their technology to several agencies within the U.S. government and are experts in the storage and management of large scale databases. The system employs multiple languages and will allow the end user to search the metadata, or the document full text, or both through a web portal. We may also run a parallel system developed by Microsoft but configured and

enhanced by people at Carnegie Mellon University for testing and comparison.

VII. Conclusion

Through the generosity of the Qatar Foundation, Carnegie Mellon University will be able to preserve and make available to the world the splendid collection of the Heritage Library. Application of accepted cataloging standards and library digitization best practices will solve many of the interesting problems presented by digitization of the Qatar Heritage Rare Book Collection. While these standards and practices may not be revolutionary, applying them to an international, multilingual collection will move forward a project of global interest.

Acknowledgements

I would like to thank the Qatar Foundation and its chairperson, Her Highness Sheika Moza Bin Nasser Al Misnad. I would also like to acknowledge the help of the project team, especially Dr. Raj Reddy, Dr. Gloriana St. Clair and Dr. Chuck Thorpe.