The digital conversion of the CM Doke Collection of personal letters from MK Gandhi, together with other related materials: A case study by the Unisa Library Digital Resource Centre

Ansie Watkins
Digital Resource Centre
Department of Library Services
University of South Africa

Abstract

The digitisation of the Gandhi Papers was a pilot study conducted by the Unisa Library Digital Resource Centre during the year 2000. This article provides an overview of this project in terms of the practical, theoretical and technological issues involved.

1 INTRODUCTION

The digital conversion of the CM Doke Collection of personal letters from MK Gandhi was the pilot project of the Unisa Library Digital Resource Centre. The purpose of this presentation is to focus on the technical aspects of the project, after giving a brief overview of the contents and origin of the collection to serve as background information. The search engine, XPAT, and developments of the University of Michigan Digital Library Extension Service (UM DLXS) played a prominent role in this project. This service has recently been introduced at Unisa and putting the Doke Collection online offered an excellent experimentation opportunity. The presentation will also focus on various aspects of the planning and implementation of the

project. The goals and benefits of the project, as well as the resources, will be briefly outlined. The roles of the project staff and the implementation will be discussed, with the emphasis on the creation of digital images and electronic text and reference to best practices and standards applied. An overview of the website will be given.

2 BACKGROUND TO THE CM DOKE COLLECTION OF PERSONAL LETTERS FROM MK GANDHI, TOGETHER WITH OTHER RELATED MATERIALS

The collection was donated to Unisa by Dr Eunice van den Aardweg, granddaughter of the Baptist minister and biographer of MK Gandhi, the Reverend Joseph John Doke. A lifelong friendship existed between the Doke and Gandhi families. It had its origin in February 1908 when Doke offered to care for Gandhi after he had been assaulted and left unconscious by Pathan Indians on Von Brandis Square in Johannesburg. Gandhi never forgot the love and kindness of the Doke family. Doke became deeply concerned about the welfare of Indians in South Africa. In 1909 he published the biography, *MK Gandhi, an Indian Patriot in South Africa* with a foreword by Lord Ampthill which was translated into several languages. He also edited the weekly newspaper, *Indian Opinion* which supported the battle for civil rights for the Indian community in South Africa. The collection consisted of the personal correspondence, published materials and photographs. These were all pasted into a scrapbook by Prof CM Doke, Dr van den Aardweg's father and reflect the friendship which spanned generations.

3 OVERVIEW OF THE UNIVERSITY OF MICHIGAN DIGITAL LIBRARY EXTENSION SERVICE

The University of Michigan Digital Library Extension Service offers a suite of tools which played a major role in the process of putting this collection online. This service originated in 1996 when the SGML Server Program (SSP) was established to assist educational and other non-profit institutions to make encoded text collections web-accessible. The UM DLXS replaced the SSP and offers a search engine as well as tools to support it. XPAT is a powerful search engine which is extremely suitable for the highly structured electronic documents typical of an educational digital library and is also used for searching the digital images in the Doke Collection.

3.1 Search engine (XPAT)

This is an SGML aware search engine, based on one previously distributed by Opentext as OT5 (TM) or Pat, as it is also known. UM now has a license to the source code for the distribution and enhancement of its functionality. XPAT provides excellent support for word and phrase searching, indexing of SGML elements, fast retrieval and open systems integration. The University of Michigan Digital Library Production Service (http://www.um-dl.umich.edu/) is continuously adding to its functionality, for example as regards basic XML awareness.

3.2 Support

By becoming a UM DLXS member an institution is entitled to XPAT support. This includes the Class Middleware, On-call support and training.

Resources, known as the class middleware, are developed to support many types of collections found in the digital library, for example page image books with associated OCR and SGML/XML encoded books and journals. The broad classes include the text class (for monograph length books) a class for images and image metadata (which was used for the images in the CM Doke Collection), a finding aid class for archival materials, a class for bibliographic information and a class for encyclopaedic reference works.

On-call support is available to members via telephone, e-mail and postal enquiries. Workshops are offered annually, namely one for programmers and one for managers of digital collections.

3.3 The DLXS Image Class

The Image Class was the module that was explored during this project. It will therefore be discussed in more detail in terms of its purpose and functionality. The Image Class is a set of tools and methods providing web-access to diverse image databases. It provides broad access to a wide range of library, museum and archive collections, visual resources and special collections.

The creation and management of the images is handled by a separate system. This process will be discussed under the section on the creation of the digital images.

The general characteristics of the Image Class are as follows:

- text-based search and retrieval of images

- searching simultaneously across multiple collections
- searching of each collection independently
- useful display and customisation of interfaces
- advanced image display capabilities (e g zooming)
- bookmarkable records and images
- minimal descriptive requirements
- reasonable data submission process for database managers
- simple data transformation process for online deployment
- uses a single data model and shared middleware for all databases in the system
- storage of multiple image solutions in a single file format
- assimilates data from practically any management system
- access restrictions at collection and image levels

Searching behaviour is as follows:

- cross collection or collection specific searching
- simple or Boolean searching
- results viewing as thumbnail, text, slideshow or full record display
- flipping among pages
- comparison of two images from two searches
- digital images are retrieved by the unique record identifiers of the text database (not the filename)

4 DESCRIPTION AND PURPOSE OF THE PROJECT

This project has been identified as a pilot study for the implementation of the Digital Resource Centre (http://www.unisa.ac.za/library/main/resource/etcweb/index.html).

The following *criteria* were applied for the selection of this specific collection for the pilot project:

- Gandhi is a figure of international interest.
- The collection contains unique primary materials which are potentially valuable for research in the fields of history, political science and theological studies.
- The letters are old and brittle and could be preserved by digital conversion.
- The content of the collection represents various types of materials,

namely digital images, electronic texts and finding aids, which is useful for experimentation.

It is important to set clear goals for any digital conversion project. The following *goals* were set for this project:

- to create electronic versions of the documents in the collection in both text and digital image format
- to make the collection accessible via an easy and advanced search and browse web based interface
- to explore digital imaging and search engine technology
- to develop institutional skills and expertise with regard to the electronic preservation and digitisation of archival material
- to serve as a model for other initiatives in South Africa and other African countries

Benefits to the clients, the collection managers and the institution

- Unique archival materials will be accessible to researchers not only in South Africa but also in the rest of the world.
- It will serve as the basis for other similar projects.
- It is an opportunity to gain practical experience and to apply skills obtained during the DLXS workshop.
- Institutional skills and expertise with regard to the electronic preservation and digitisation of archival materials will be developed.
- The use of computer based research tools is a means of enhancing the skills of students, thus preparing them for postgraduate studies and the job market.
- This project has preserved unique and brittle materials and made them accessible.
- Digitisation of this collection will encourage and stimulate the African Renaissance.

5 FUNDING AND RESOURCES

Existing resources were used for the project. The Unisa Foundation granted an amount to be used to set up the infrastructure for the digital resource centre during 1999. The system was to run on a Sun Solaris Unix machine. DLXS XPAT, a system distributed and supported by the University of Michigan, is the information retrieval software that was selected for the Digital Resource Centre. An A3 Epson flatbed scanner was used to create the

digital images of the source materials. Adobe Photoshop was used to convert and enhance master tif images to jpeg for web delivery. Xmetal is the XML editor which is used to encode the electronic text and finding aids. The images are stored in a SQL database from which the data file is generated. The collection is catalogued on the Innopac system, OASIS.

6 THE PROJECT STAFF

A working group, consisting of representatives from the Archives, Digital Resource Centre, Web Development, Department of Computer Services and the relevant staff from the Client Services Division was constituted. The division of duties was as follows:

- The project sponsor was responsible for providing managerial support to the project leader and had to negotiate the funding and resources necessary for the successful implementation of the project.
- The project leader had to coordinate the project, monitor the progress by means of regular meetings and report to the sponsor. This role was fulfilled by the digital resources manager. This person also defined the metadata for the image database and was responsible for processing the data to enable keyword searches.
- The subject librarian for History and Political Science represented the Client Services Division and played an important role in the selection of the sources to be digitised.
- The Archives were well represented, having four staff members on the working group. Their duties were to communicate with the donor and Corporate Marketing and Communication, prepare the materials, do the inventory, type the letters, scan the digital images and contribute to the content of the web page.
- The information was catalogued by the Technical Services Division.
- A programmer from the Department of Computer Services installed the software and set up the environment for the presentation of the digital resources on the Internet. He also attended to the access requirements to the server. The Webmaster of the Library developed the web pages where the collection can be accessed from and the project leader assisted with the search screens.

7 THE PROJECT IMPLEMENTATION SCHEDULE

After the goals had been set a business plan was drawn up and submitted for

managerial approval. The final phase was the implementation of the project. The project proceeded more or less chronologically within the following steps:

7.1 Copyright

The only copyright issue that had to be dealt with in this case, was to get the permission of the donor to publish the collection on the Internet. This was not a problem at all because one of the conditions under which the collection was donated was that the letters should be transcribed in electronic format. The Unisa Library holds the rights to the digital images and electronic text of the source materials.

7.2 Preparation of the materials

An inventory of the materials was compiled and they were divided into correspondence, published materials and photographs. The inventory is accessible from the web site.

7.3 Definition of the metadata

Various types of metadata, for example descriptive, administrative and technical metadata had to be defined and created. The descriptive metadata for the digital images are based on the Dublin Core elements. The metadata for the text transcriptions of the letters were defined according to the Text Encoding Initiative (TEI) Guidelines, which represent the international standard. EAD (encoded archival description) is the international standard for encoding archival finding aids. SGML DTDs are generated from these metadata sets and the encoding is done accordingly. A MARC record was created for the collection in the Library catalogue. These metadata sets all deal with the bibliographic and subject description of the entities. The administrative and technical metadata are also important for the description of digital images and can be defined according to the needs of a specific institution. Examples of these types of metadata elements are the size of an image file, the type of scanner used, the date of the creation of the file, the person who scanned the image, etcetera. There are no limitations on the number of elements or the field lengths or rules of what elements should be included. The rights, and file name, type and size are the technical elements defined for the CM Doke Collection.

7.4 Creation and management of digital images

This process consists basically of three steps, namely creating and preparing the images, creating the descriptive data (indexing) and linking the image to the descriptive data

7.4.1 Digital image production

Digitisation is the process of analog to digital conversion. The digital image is a series of picture elements (little squares called pixels) which represent a picture. Images are measured in terms of resolution (the number of dots or pixels that represent the image – dots per inch(dpi)), pixel bit depth (defines the number of shades and colour). Digital masters of the source materials were created in tif format at 600 dpi. Derivatives in jpeg format were processed from these masters for web delivery. They are presented as large images and thumbnails. The naming convention relates to the item numbers in the inventory. All the documents in the collection including the envelopes were scanned.

7.4.2 Loading images and linking them to the descriptive data

After the images had been scanned, the jpeg files were transferred to the Unix Server. The files have to be organised in a specific directory structure to ensure that the images can be retrieved by the system. The images of each collection are organised in its own directory with the large and thumbnail files in distinctive subdirectories.

They were indexed according to the defined structure which is based on the descriptive and technical metadata and consists of elements such as description, type, format, date, creator, filename and rights. A datafile was exported from this database and also transferred to the Unix Server.

7.4.3 Integrating the images with the descriptive data

There were now two sets of files, namely the image files and the descriptive data that had to be integrated to enable keyword searching and retrieval of the images. The datafile was translated into SGML and indexed by means of the XPAT software. HTML templates were used to create an image class search interface from which simple and Boolean searches can be executed and results can be viewed.

7.5 Creation of full text, including markup

Creating encoded text is the most expensive model of electronic text publishing. It is, however, the most flexible and functional for online presentation. The letters were typed from the originals by means of a wordprocessor and saved as ASCII files. These attributes were encoded in SGML (Standard Generalised Markup Language) and converted to XML. This is done by inserting tags in the text to identify certain regions and selections of texts to enable fielded searching. There are several SGML tags that can be used. The Text Encoding Initiative (TEI) Guidelines refer to a subset of these tags that is internationally used for publishing scholarly electronic text. The software tools that were used for the encoding were an intelligent text editor, namely Notetab Lite and an XML editor, namely Xmetal.

The *reasons* for the transcription of the letters were as follows:

- Transcription had been requested by the donor.
- Gandhi's handwriting is difficult to read.
- Page images are not full text searchable. They have to be translated to machine readable ASCII. Retyping or keying from the original was the only way to do it in this case, because the source materials were mostly handwritten letters.
- It was an important part of the pilot study to explore the practices of creating and encoding searchable electronic text.

These SGML/XML files are processed and indexed by means of the DLXS Text Class.

7.6 Overview of the website

http://www.unisa.ac.za/library/main/resource/etcweb/digtexts.html

The collection is not only accessible from the OASIS library catalogue, but also from the Unisa Library home page. It provides detailed background information on the history of the Gandhi/Doke relationship with biographies and photographs of the relevant family members. There is also a section on the arrangement of the materials in the inventory which is also available online. The inventory offers links to the transcribed letters. The link to the image collection leads to the simple and Boolean search interfaces. There are references to books on MK Gandhi and CM Doke and finally a link to the OASIS catalogue.

7.7 Launch, advertising and user evaluation.

The Doke Collection was formally handed over to Unisa during a social function on 14 November 2000. Dr Van den Aardweg and other members of the Doke family, the Principal of Unisa, Prof AM Melck, Unisa staff from the Departments of Corporate Communication, Library Services and Computer Services who were involved in the project, as well as members of the Student Representative Council were present. The press were also invited and an article was published in the Pretoria News the next day. The event was advertised via e-mail and Unisa staff and clients were invited to access the website and send remarks to the project leader. The web site and search interfaces are at present being updated and improved according to recommendations received in this regard.

8 CONCLUSION

The digital conversion of the CM Doke Collection is the first model using the DLXS solution in South Africa. Although there are many other ways to put electronic documents online, this was selected by Unisa because it is cost-effective and has been tailored to the specific requirements of the academic environment. Support is also readily available. Useful lessons were learned from this exercise. Digitisation is expensive and a risk to an organisation. Creating electronic texts and digital images is in many cases regarded as a ''prestige activity'' until people have to do the actual work when they have to start dealing with IT problems and routine tasks. It is therefore advisable to select a small and manageable, but preferably interesting, collection for your pilot study. That will ensure that your staff remain enthusiastic and are able to achieve results before they lose interest. It is advisable to stick to international standards and to follow the ''open systems'' approach. Scan digital images for preservation of the highest quality that can be afforded. This will help your collection outlive the current technology and save your organisation money in the long term. Try to avoid using a collection with copyright problems for a start. South African institutions can benefit a lot from collaboration, but we still have a long way to go. It would be worth exploring these possibilities, however, because we have a wealth of inaccessible cultural collections at some institutions that have few resources and on the other hand other institutions are able to offer expertise. Bringing these two poles together appears impossible at this stage. We might consider this as a challenge for our electronic future.