Newspaper Digitization Project at the Press Institute of Mongolia

International Implementation of Digital Library Software/Platforms
2009 ASIS&T Annual Meeting
Vancouver, November 2009

Krystyna K. Matusiak
Digital Collections Librarian
University of Wisconsin-Milwaukee
Milwaukee, USA

Background

- **1924 –1990** Mongolian People's Republic
  - Single-party system
    - Mongolian People's Revolutionary Party
  - Allied with the Soviet Union
  - Media and the press controlled by the communist government
    - Daily *Unen* (Truth), the equivalent of the Russian *Pravda*
- **1990-1995** Transition Period
  - December 1989 - March 1990
    - Pro-democracy protests
  - Multi-party system
  - Transition to market economy
  - Explosion of media
    - 600 newspapers by the mid 1990s

Mongoliin Zaluuchuud, April 4, 1990

Digitization Project

- First large-scale digitization project in Mongolia
  - Two-year project: 2006-2008
- Initiated by the Press Institute of Mongolia
  - The goal was to digitize Mongolian newspapers in the holdings of the Press Institute
- Supported by a grant from the Endangered Archive Programme of the British Library
  http://www.bl.uk/about/policies/endangeredarch/homepage.html

The Press Institute of Mongolia

- A non-profit organization
  - Located in Ulaan Baatar, Mongolia
  - Established in 1995 under the Free Press project
  - Promotes the development of independent and pluralistic media in Mongolia
    - Publishes a comprehensive annual report on the state of the media in Mongolia
  - Houses a research department and a special library with a collection of newspapers and magazines from 1923 to 1996 http://pressinst.org.mn/

ХЭВЛЭЛИЙН ХҮРЭЭЛЭН

Нүүр хуудас   Танилцуулга   Сургалт   Судалгаа   Холбоо барих   Фото архив   Сэтгүүлч коллеж   English

Newspaper Collection

- A collection of rare Mongolian serials published from 1923 to 1996
  - 80 newspaper and magazine titles acquired from private individuals and several Mongolian media organizations
- Particularly rare newspapers from the transition period 1990-1995
  - Unavailable at other cultural institutions in Mongolia
  - Document the transformation period in Mongolia after the fall of Communism in the early 1990s
  - Used extensively by journalism students and researchers and face the risk of rapid deterioration

Project Goals

- **Preservation**
  - Lack of preservation program
    - Microfilm preservation strategies not available
    - Digitization as a means of creating digital surrogate copies of deteriorating newspapers for preservation purpose
  - Preservation as the project's primary goal
    - Create digital archival copies of rare Mongolian serial publications
    - Safeguard the collection housed at the Press Institute from the risk of physical deterioration and destruction

Project Goals

- **Access – secondary goal**
  - Providing wide and remote access through an online collection
  - Reducing the handling of the fragile print materials
  - Providing optional print version in the PDF format
- Enhancing indexing and searchability of the newspapers
  - Providing full-text keyword search and multiple browse capabilities

Selection

- Coverage
  - The titles selected for the project cover the transition period 1990-1995
  - Document the transition to democracy and the development of the independent press in Mongolia
- Language
  - Mongolian language using the Cyrillic alphabet
- Format
  - Magazines - A4
  - Newspapers - A3 and large A2 newspaper format
- 59 titles
  - 6189 issues
  - 39029 pages

Image Capture

- Original paper copies as a source for digital images during the scanning process
- Scanning conducted in-house using oversize flatbed scanners
  - Each scanned newspaper page treated as a separate image and saved as an archival master in the TIFF format
  - Large newspaper pages in A2 format scanned in two sections and merged in Photoshop
  - Resolution at 400 dpi
    - Determined after preliminary testing
    - Provided more accurate OCR

Digitization Guidelines

- Project guidelines based on digital library standards and best practices
  - Use-neutral approach with the notion of digital master files and derivatives
- The standards established for the scanning process:
  - Resolution: 400 dpi
  - File Format: Tagged Image File Format (TIFF)
  - Compression: None
  - Bit-depth: 8-bit greyscale for black and white newspapers

    24-bit RGB for pages with color images
  - Consistent file naming convention
    - File names consist of a unique newspaper code (two letters) followed by year (two digits), month (two digits), and day (two digits)

    mt950101_01 – page 1 of the *Mongolian Times*, January 1, 1995

Digitization Guidelines

- Digital masters
  - Images created as a direct result of the scanning process
  - Preservation-quality digital copies
  - A source for creating multiple derivative copies
- Two copies of each digital master in the TIFF format
  - The first set of archival TIFF files stored at the Press Institute of Mongolia
  - The second set deposited at the British Library
- Extensive documentation for preservation purpose
  - Recorded on multiple levels
    - Collection
    - Individual publications
    - Publishers
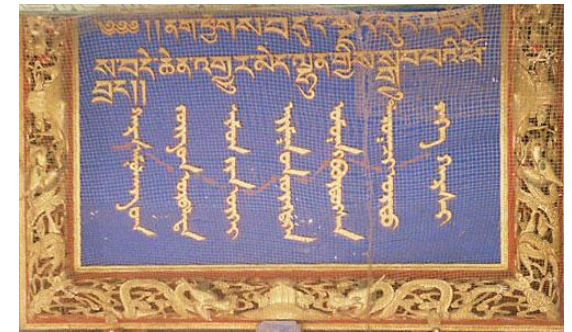 - Deposited at the British Library

Selection of Digital Library Software

- Factors considered in the selection process
  - Affordability
  - Language support
  - Required staff skills and technical expertise
    - Availability of software support and training
  - Search and browse capabilities
  - Support of digital library standards
    - Interoperability
  - Usability
- Evaluated DL software packages
  - CONTENTdm
  - Olive software
  - Greenstone
  - Newspaper Management Program
    - Software developed by a local Mongolian company
    - Offered support for the Mongolian language
    - Under development at the time of evaluation

Language Support

- Multilanguage support
  - Capability of displaying non-Latin characters
  - Unicode support
- Mongolian language
  - Traditional Mongolian script
  - Cyrillic alphabet adopted in the 1940s
- Publications selected for the digitization
  - Mongolian language using the Cyrillic alphabet
  - Russian
    - *Novosti Mongolii*

Adherence to DL Standards

- Support metadata schema
  - Dublin Core
- Support controlled vocabulary tools
- Provide access to digital objects in multiple formats
  - Searchable text, image, and PDF for documents
- Provide efficient and effective info access and retrieval
  - Simple and advanced search options
  - Multiple browse pathways
  - Usable and customizable interface
- Ensure interoperability
  - Metadata harvesting
    - Open Archives Protocol for Metadata Harvesting (OAI-PMH)
  - Accommodate future migration of objects and metadata

Open-Source Vs. Proprietary DL Software

- Open-source software
  - Source code and rights in public domain
  - Developed by dedicated developers and a community of users
  - Distributed under Open Source License
    - No initial cost
    - Right to improve and change the software
    - Associated (hidden?) cost
      - Technical expertise, support, and training
      - Maintenance and sustainability
      - Further development
- Proprietary DL Software
  - Source code protected and locked
  - Developed by commercial vendors for profit
  - High cost - license and annual maintenance fee
- Technological competition and innovation
  - Hybrid products

Greenstone

- Greenstone http://www.greenstone.org/
  - Open-source international digital library software
  - Developed at the University of Waikato in New Zealand
  - Supports digital library standards
    - Dublin Core
    - Open Archives Protocol for Metadata Harvesting (OAI-PMH)
  - Offers multilingual support – Unicode Standard
    - Capable of processing and displaying non-Latin characters, including Cyrillic alphabet
    - Capable of building indexes and searching in the Mongolian language
  - Requires training and technical expertise
  - Low score on usability

greenstone digital library software

Collection Building

- Several stages of collection building process:
  - Creating derivative files using ABBYY FineReader software
    - Searchable HTML files through the OCR (Optical Character Recognition) process
    - PDF files
  - Adding descriptive metadata
    - Dublin Core schema
  - Assembling an online collection with Greenstone software
    - Uploading the files to the Web server
    - Building indexes
    - Customizing the interface

Online Collection

Digital Archive of the Mongolian Newspapers 1990-1995
http://www.pressinst.org.mn/cgi-bin/library

МОНГОЛЫН 1990-1995 ОНЫ
СОНИНЫ ЭЛЕКТРОН АРХИВ

DIGITAL ARCHIVE OF THE
MONGOLIAN NEWSPAPERS 1990-1995

- Online collection:
  - Built to support remote and immediate access to copies of the newspapers on the Web
  - Presents customized Greenstone interface in the Mongolian language
    - Lists titles in Cyrillic and Latin alphabet
  - Offers full-text searchability and browse options by year and title
    - Optional PDF versions

Collection Sustainability

- Challenges
  - Grant funded project
    - Grant expired in June 2008
  - Unstable IT infrastructure
  - Limited expertise in Greenstone
    - Staff turnover
    - No access to Greenstone training sessions
- Efforts to ensure sustainability
  - Institutional commitment
  - Detailed project documentation
  - Adherence to DL standards and best practices
  - Integration with other DL developments in Mongolia
    - Mongolian DL portal - Library at the American Center for Mongolian Studies

Questions?

Krystyna K. Matusiak
Digital Collections Librarian
University of Wisconsin-Milwaukee
Milwaukee, USA
kkm@uwm.edu

Myagmar Munkhmandakh
Director
Press Institute of Mongolia
Ulaanbaatar, Mongolia
munkhmandakh@pressinst.org.mn

"Preservation of rare periodical publications in Mongolia." (2008).
Project report available from the Endangered Archive Programme
website at:
http://www.bl.uk/about/policies/endangeredarch/bayarmaaoutcome.html