**NOVEMBER 2009**

**TEI PROJECTS AND SMALL LIBRARIES: EXAMINING TEI MARKUP DECISIONS AND PROCEDURES**

**Richard Wisneski**, Head, Bibliographic and Metadata Services. rlw54@case.edu

**Virginia Dressler**, Digital Librarian. vad17@case.edu

**Stephanie Pasadyn**, Technical Services Librarian sap68@case.edu

**Kelvin Smith Library**

**Case Western Reserve University**

**Cleveland, OH**

1. **INTRODUCTION**
* Digitizing and Encoding Kelvin Smith Library’s (KSL) *Books on Cleveland, Ohio and the Western Reserve Digital Text Collection*
	+ Currently, there are 120 texts in Digital Case, in PDF format
	+ Goal is to text-encode approx. 110 of these, and add more in the future
	+ Texts Date from late 19th to early 20th centuries
* Using “Book Viewer” in KSL’s “Digital Case” (institutional digital repository) to Display texts’ PDF, Page Images, and TEI
* Have applied for an NEH Humanities Collections and Reference Resources Grant to fund project
* Will collaborate with neighboring institutions to incorporate into our collection their texts on the history of Cleveland and the Western Reserve

**Why Do This Project?**

* Availability of Texts is Limited
* No other institution has a project akin to this in Northeast Ohio
* Interest in Cleveland and Western Reserve history among historians and scholars. Cleveland…
	+ was the northern terminus of the Ohio and Erie Canal
	+ was a leading U.S. manufacturing center, second only to Detroit in the automotive industry, and a leader in steel production, ship building, and other industrial sectors
	+ had the highest concentrations of some Eastern European heritage populations outside of their home countries during the 19th and early 20th centuries

**Why TEI?**

* To allow researchers to have access to an electronic text that does not require special-purpose software or hardware
* To analyze information – provide a standard text-encoding scheme and metadata language which accommodates searching, retrieval, etc.
* To share information – have a standard format for data interchange in humanities research
* Texts are being encoded in Level 3 (structural)
* To create stand-alone electronic text with hierarchy identified
* Emphasis on divisions within text, tables, lists, notes, front and back matter
1. **CURRENT PROJECT PRACTICES**

See Flowchart and Spreadsheet

**Digitization Process**

* + Transfer content from CDs
	+ Assess quality of images (dpi, original canvas size, overall quality)
	+ Rescan titles if needed
	+ Organize/sort for OCR and image conversion

**Key Points in Assessment**

* Complete, uncorrupted files
* Ascertain image quality as to current practices and standards
* Check for legibility of text for OCR process
* Compare illustrations and photos with original source if needed
* Rescan if needed

**OCR**

* Using OCR software (Abbyy FineReader) to create a text file from the image files
* Time saving options in software
	+ removing hyphens between page breaks
	+ retaining page breaks

**Processing TIFF files for the Book Viewer**

* Bit map to grayscale
* Aware batch conversion to jpeg2000 format
* Batch renaming tool
* Ingestion into Fedora

**Text Clean-Up**

Student Workers, Volunteers do work in OpenOffice and oXygen

* Spell-check
* Insert page breaks and numbers
* Replace images with notation
* Remove hyphenations that are erroneously inserted from OCR
1. **LEARNING TEI**

Learning Text Encoding as Secondary Job Responsibility

* Practical Application
* Internal Documentation
* CaseLearns

**Coding a New Text**

* One on one overview
* Creating master outline
* Coding page by page
* Referring to and updating documentation

**Challenges**

* Human Error
* Evolution of Institutional Practice
* Minimal Time Allotment
* Limited Opportunity for Continuing Education
1. **ISSUES**

**To Be Done**

* Re-Scan some of the books
* Continue to encode
* Hold half- full-day workshops on text encoding to full-time staff
* Create of MODS, MARC-XML, and METS records
* Re-examine “Book Viewer”

**Discussion Questions**

* Ways to expedite text encoding
* Ways to scan texts – outsourcing?
* Funding challenges (outsourcing, scanning, equipment)
* Book viewer – effective? Ineffective?
* Text-Encoding Level – change?
* Learning TEI – in-house classes and documentation, TEI-C documentation. Webinars? Online tutorials? Certificate program?