Supporting Digital Scholarship: Jump-starting Digital Humanities

Mitsunori Ogihara

Prof. of Computer Science, College of Arts and Sciences
Director of Data Mining, Center for Computational Science
Associate Dean for Digital Library Innovation, College of Arts and Sciences; Otto G. Richter Library

Prof. of Electrical & Computer Engineering, College of Engineering*
Prof. of Music Media and Industry, Frost School of Music*
Prof. in the Center on Aging, Miller School of Medicine*

University or Miami

* Secondary appointments

CNI Spring 2014, St. Louis, MO

Associate Dean for Digital Library Innovation

- Explore possibilities for digital humanities and digital material processing
- Explore opportunities for extramural funding
- Jump start projects
- One-person operation

- Former University Librarian Bill Walker
- Arts & Sciences Dean Leonidas Bachas
- New University Librarian Charles Eckman

Digital Humanities

- Conducts humanities scholarship through aggressive use of digital technologies
  - Digital source material generation
  - Data generation from source
  - Data analysis
  - Presentation and visualization
  - Data curation and archiving
  - Software packaging

Issues 1

- Technical
  - Materials are not digitized
  - Data generation is technically challenging

Issues 2

- Political
  - Needs collaborations among humanists, technologists, and librarians
  - Collaboration is a foreign concept to most humanists
  - Strong skepticisms about the value of digital scholarship
    - Junior faculty may not wish to conduct digital humanities research

Issues 3

- Systematic
  - Digital humanities scholarship is for humanistic research
    - Results must be examined from a humanistic perspective
  - Digital humanities scholarship can be time consuming
    - Any part of the process may require substantial effort and cost

Initial Investigation

- Met with a score or so non-science scholars to explore their interests, in particular, the need for custom software development

Challenges at UM

- A very few humanities faculty conduct or are interested in digital humanities research
- Interesting digital materials exist in the library, but many of them are still in raw formats and require processing

Operational Principle

- Two possible approaches for promoting digital humanities:
  1. Show what other scholars have done at other institutions to drum up interests among the audience
  2. **Work with interested scholars to jump-start projects**

Activities to Date

- Tried to entice humanities scholars for digital projects
- Examined possibilities for generating textual data from the library's digital collection
- Written a number of proposals
  - Two Andrew W. Mellon Foundation Grants
  - Two NEH proposals (one declined, one pending)
  - One NSF pending proposal

Project List

- Digital Humanities Projects
  1. **Text-mining of Horace and Virgil**
  2. **Software development for a digital archive of theater productions**
  3. **Handwritten mediaeval Latin text transcription and computational correction**
  4. Text-mining of the Carlyle letter collection
- Library Projects
  5. **Information extraction from university newspapers**
  6. Mining library patrons resource use

1. TEXT-MINING OF HORACE AND VIRGIL

Text-mining of Horace and Virgil

- Collaborator: Jennifer Ferriss-Hill, Assistant Professor of Classics
- Find clusters of unique words in Horace (BC65 – BC8)'s "Odes" when lexical endings are ignored
- Those words may be under influence of Virgil's "Aeneid"

Method

- Collected files from the Latin Library
- Program
  - Preprocess the html files
  - Dictionary look-up for root-form identification
    - "Whittaker's Words"
  - Identify and cluster unique words
  - Generate html files that show clusters and links among the identified words

- Example: Horace-1-3-Sample.html

Results

- From Jennifer's presentation abstract: "Horace figures the *Aeneid* as a ship transporting Virgil through arduous (literary) hazards into Greek territories, taking him further away from Horace's own satirical and lyric interests.... I show through these results the extent to which Horace in writing Odes 1.3 borrowed from the *Aeneid*'s vocabulary and imagery."

Work to Do

- Improve the program
  - No false negatives, but quite a bit of false positives:
    - Proper nouns
    - Archaic forms
    - Irregular verbs

2. DIGITAL ARCHIVE OF THEATER PRODUCTIONS

Cuban Theater Digital Archive

- An archive of theater productions by Cuban artists, both in Cuba and in the states
  - Archive Director: Lillian Manzur, Assoc. Prof. of Modern Language and Literature
- Videos and images
- Recurrent support from the Andrew W. Mellon Foundation

Goals of the Current Development Efforts

1. Develop a platform on top of Django
2. Add a search engine component using Solr
3. Add Twitter feeds
4. Add tags and comments via Disqus
5. Add publication capability via University of Southern California's "Scalar" project

- Software development is being provided by the Center for Computational Science's Software Engineering Group

Current Status

- Most of the goals completed
- Presentations
- Publications of the project

- CTDA

3. HANDWRITTEN LATIN MANUSCRIPT RECOGNITION

Goal

- Develop tools for generating candidate transcriptions of Carolingian Latin manuscripts through OCR of handwritten texts
- Eventually … develop a database of images and transcriptions of some Carolingian manuscripts

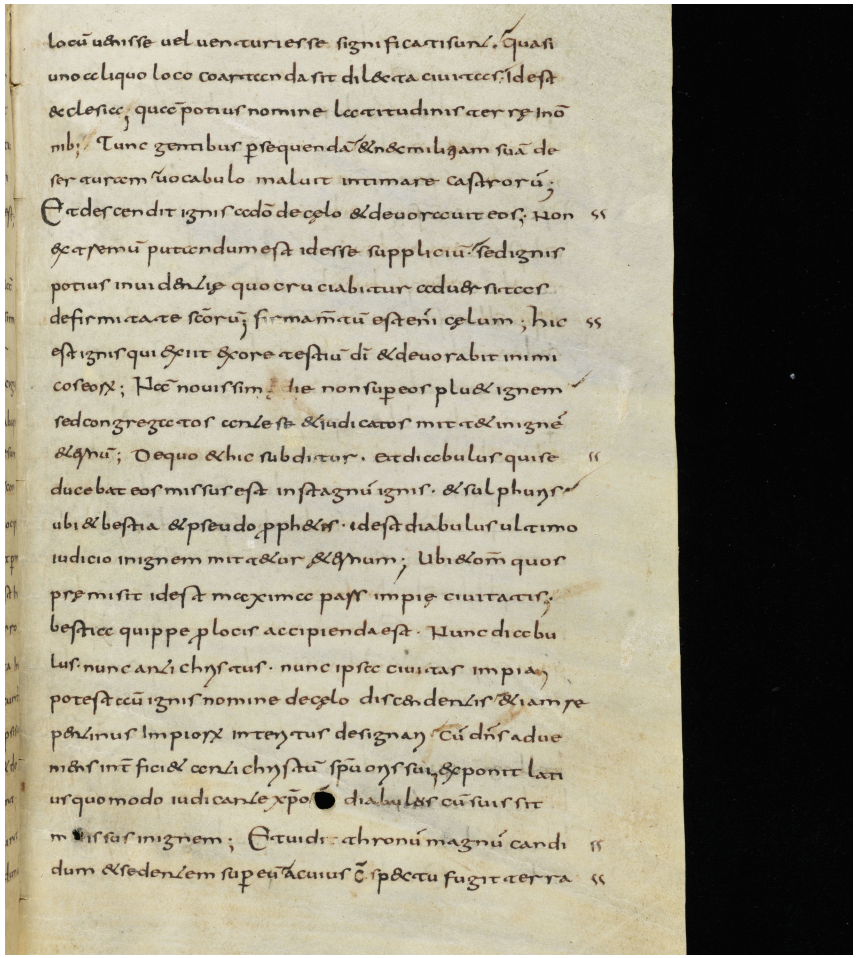
Collaborator/Sponsor

- Collaborator
  - Wilson Shearin, Assist. Prof. of Classics, U. Miami
  - Dimitris Papamichail, Assist. Prof. of Computer Science, The College of New Jersey
  - Gavriil Tsechpenakis, Assist. Prof. of Computer Science, IUPUI

- Pilot funding from the Andrew W. Mellon Foundation

Approach

- Examine two off-the-shelf OCRs
  - ABBYY Fine Reader … commercial
  - Tessaract … open-source
- Provide extensive training
- Develop methods for error correction
  - Develop confusion matrices
  - Use Ngrams and tokenization



Current Status

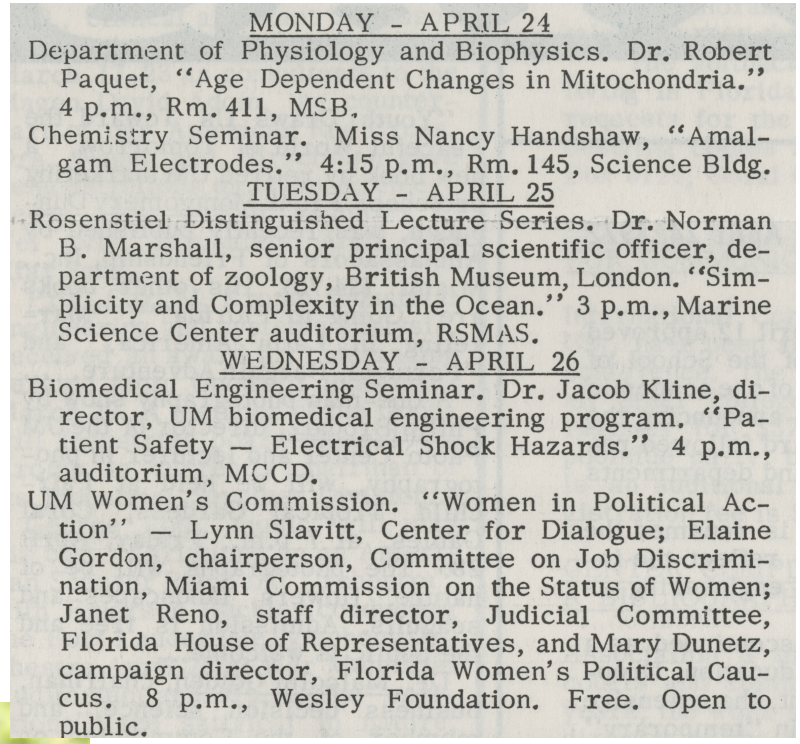
- 60% accurate at the character level
  - Extremely difficult to train a system for hand-written characters
- 63% with computational correction
  - The low accuracy before correction makes it very difficult to correct errors
- We may need to design a preprocessing tool for improving OCR accuracy

4. CALENDAR EVENT EXTRACTION

Goal

- University of Miami's faculty/staff newpaper "Veritas"
- Extract event information from "Upcoming Events" section.

- Collaborator: Koichi Tasa, University Archivist



MONDAY - APRIL 24

Department of Physiology and Biophysics. Dr. Robert Paquet, "Age Dependent Changes in Mitochondria." 4 p.m., Rm. 411, MSB.

Chemistry Seminar. Miss Nancy Handshaw, "Amalgam Electrodes." 4:15 p.m., Rm. 145, Science Bldg.

TUESDAY - APRIL 25

Rosenstiel Distinguished Lecture Series. Dr. Norman B. Marshall, senior principal scientific officer, department of zoology, British Museum, London. "Simplicity and Complexity in the Ocean." 3 p.m., Marine Science Center auditorium, RSMAS.

WEDNESDAY - APRIL 26

Biomedical Engineering Seminar. Dr. Jacob Kline, director, UM biomedical engineering program. "Patient Safety — Electrical Shock Hazards." 4 p.m., auditorium, MCCD.

UM Women's Commission. "Women in Political Action" — Lynn Slavitt, Center for Dialogue; Elaine Gordon, chairperson, Committee on Job Discrimination, Miami Commission on the Status of Women; Janet Reno, staff director, Judicial Committee, Florida House of Representatives, and Mary Dunetz, campaign director, Florida Women's Political Caucus. 8 p.m., Wesley Foundation. Free. Open to public.

Process for Calendar Generation

- Scan the prints (TIFF)
- TIFF to low-resolution PDF
- Assemble multiple pages into a single PDF file
- Run an OCR (ABBYY 11)
- Run an event extraction code to generate an event data file
- Hand-edit the event data files for correction
- Convert the event data file to an iCalendar file

- Sample

Things To Do

- Add geo-coding of location
- Add search-component
- Correct spelling errors

IN RETROSPECT

Lessons Learned

- There is competing time pressures on pre-tenure faculty
- Not being a humanist the whole experience is interesting but time consuming
  - Wound up spending more than 1/3 of time
  - Productivity in the main field declined
- Nicer to work with a team of technologists than to work alone

How Much Was DH Promoted?

- Those scholars I worked with, and those who are around them, are appreciative of the collaboration and enjoyed seeing what technologies could do to explore humanistic questions
- Beyond those people may have skepticisms about the operation and sense of danger having non-humanists poking a nose into humanities research

Future

- Need more propaganda for digital humanities
- Need more resource for promotion
- An inter-school task force is about to be established to consider the future of digital humanities is about to be organized
  - Inventory of technologies, interests, and expertise at the university
  - Having conversations with groups outside the university
  - Recommendations for the future

Thank you so much for listening!