Scanning Key Content of Text Based Material at Point of Accessioning or Cataloging: Opportunities for Adding Value, Creating Efficiencies, and Expanding Horizons

Proposal:
Librarians, library users and the general public are increasingly exposed to the power of technology to generate data, provide access to that data, and to manipulate and adapt that data. Interactive links allow seamless navigation through the world of information. Such capabilities are radically changing the expectations of us all.

Meanwhile our library catalogs remain largely flat collections of (mostly hand-keyed) descriptive, surrogate data. This project seeks to increase the value of our catalog to users and improve the efficiency of cataloging by putting scanners on the desks of staff who can initiate, complement and supplement the excellent work being done in Imaging Services. By scanning key content early in the process, images can be made available to patrons, and OCR text can facilitate patron searching and support more efficient cataloging.

Machine manipulation of data for improved discovery and "crowd sourcing" user contributions are already established means contributing to serious scholarship. For instance, LigerCat http://ligercat.ubio.org/ generates MESH headings via word clouds generated from clusters of articles; Zooniverse https://www.zooniverse.org organizes and coordinates thousands of citizen scientists to transcribe data about archeology digs, weather logs, galaxies etc.; Project Gutenberg has produced thousands of titles keyed by interested volunteers.

Scope:
The project will focus primarily on 19th and early 20th century ephemeral text-based pamphlets selected from three collections: Divinity, Houghton, and Schlesinger. The choice of material is strategic for a number of reasons: page layouts are fairly predictable and should be legible for computer applications, most require original cataloging, some have interesting cover images, and all are part of Harvard's hidden collections. Most importantly, it is expected that scholarly demand for such materials will increase as we move into the 21st century.

Project Goals:
This proposal seeks to expand capabilities within our current infrastructure, challenge existing methodologies, speed up the process of making our collections available to our users and lay groundwork for a much more dynamic and flexible future system. We therefore would:

1. Explore possibility of mounting images alongside or within the cataloging records in HOLLIS, in the manner of the Google Books icons or bookplate images. For an example of images as they might appear, see the Catalogue of the Sixteenth Century Editions of the Empoli Public Library "Renato Fucini" (http://www.comune.empoli.fi.it/biblioteca/CATALOGO/indici/indicieng.html).

2. Establish procedures to digitize key components (such as the title, table of contents, index, and perhaps key graphics) of these materials. Determine possibility and feasibility of utilizing scans for cataloging efficiency in the following ways:

- Parsing the resulting OCR text and using it to search Aleph, OCLC, Google Books, or even the NSTC to find available cataloging copy
- Using OCR or TXT files of the images to create provisional (or minimal) catalog records for immediate keyword access. The University of Oklahoma recently completed a twenty-three month project that resulted in 8,300 title pages being scanned and run through a program that created corresponding author and title accessible catalog records

3. Identify a stable environment in which to store the scanned images and related metadata for future usability. This would involve identifying the most suitable program-neutral formats for saving this data, e.g. employing URIs, JavaScript, etc. Optimize the likelihood that resulting images and metadata can be used with emergent technologies, such as the DPLA, linked data developments and future online catalogs using RDA and other "FRBRized" models

Methodologies/Workflows:
An imagined basic workflow would go something like this:

- Preliminary scan (cover, title page, TOC, index, colophon, inscriptions) of item is done at time of accessioning. Simplest would be scan linked to barcode perhaps on a removable flag (consider piggy back barcodes) so barcode can be used later.OCR the scan and parse to a search against OCLC, or generate a stub record
- Modify it for display as a bibliographic record in HOLLIS
- Attach the scan in the OPAC or create a link to the image which has been saved in an alternative location (DRS?)

An enhanced workflow might include:

- Making the record available to the public for social tagging (possible models might be the University of Pennsylvia's Penntags program - http://tags.library.upenn.edu/ or the Australia newspaper project - http://www.nla.gov.au/ndp/get_involved/; and Drupal's SOPAC technology)
- If determined as appropriate, add to digitization queue if it meets parameters (i.e. date, extent, fragility, etc)
- If not a candidate for digitization, then put in cataloging queue. Cataloger would call up scan and build a bibliographic record either by keying text, or "grabbing/scanning" the OCRed text and marking it up for appropriate fields. Additional elements of interest such as inscriptions could be scanned and added to the record.
- Consider linking to additional relevant information such as reviews, citations etc.

Project Staffing:
Acknowledging the current transitional climate of library staffing, initial coordinators/catalogers would include:

- Nell Carlson of the Divinity School Library
- Debbie Funkhouser of the Schlesinger Library
- Karen Nipps from Houghton Library.

In developing this proposal, we have had a preliminary conversation with Todd Bachmann from Imaging Services and Scott Wicks who expressed enthusiasm for this project as a building block toward a very exciting future. If funded, we would have further consultations with Imaging Services regarding specifications of scanners etc. We would also speak with staff in Scan and Deliver and ILL in order to benefit from workflows already in place.

This project will require that the participants prioritize work on the predetermined backlogs over other projects for approximately a day a week. The added time required for scanning itself will be minimal. Other Harvard staff involved in the technical aspects of digitization and metadata creation will be consulted as needed; these might include faculty, Berkman staff, and OIS. Time will be needed for project planning and execution, including regular meetings to analyze procedures and data sets.

Equipment:
Scanners will be needed, but the first part of the planning process will be deciding what technology best suits our purposes. Lower-end equipment might suffice for determining feasibility of proposed workflows.

Metrics and evaluation to inform future projects:
This project will also provide data to assist decision making for future projects. We seek to identify relative benefits of various approaches (for instance, scan image only for better identification by users; scan for OCR text for manipulation into searches or MARC records; scan entire item vs. just title page) for various kinds of materials (English vs. foreign language; font types or type quality and legibility; uniformity of layout across the selected items; size/format or other physical characteristics; age; total length/extent of item). To this end summary reports will include data for each material category (to be determined at outset) for:

- Total number of items scanned, pages scanned (and average pages per item)
- Average time per item/scan and associated work naming and depositing file
- Number/average of scans requiring human manipulation (e.g. correcting OCR, lightening or darkening for legibility) and time spent on these activities
- Number/average of scans resulting in successful searches in Aleph and OCLC
- Number/average of scans used to populate MARC records

Other questions this project will inform:

- If an item is found to have an existing record in Aleph, is there benefit in depositing the image?
- Assuming that full-text access is an ultimate goal, what is the threshold where it would be more cost efficient to scan the entire item vs. select page scans?
- What is the minimal acceptable scanning quality required to accomplish the various objectives (image only, OCR, searching using resulting text, text into MARC record)?
- At what point is scanning more feasibly performed as part of the Imaging Services project workflow rather than at cataloger desktops?

Users will be consulted to determine the usefulness of the added data to their work. At the conclusion of the project, a report will be presented analyzing the data gathered and determining if deploying similar efforts more widely would be useful and is practicable. The project might be discussed at a future Catlaloging Discussion Group meeting. The report could be submitted to an appropriate journal, such as *Cataloging and Classification Quarterly*.

Requested funding:
3 scanners with varying capabilities at @ $5,000-$10,000 a piece (such as http://www.thecrowleycompany.com/scanning-equipment/product-type/book-scanners-manual.html#productTypes)
$6,000 -- Use of technical staff at $85.00 an hour – ten days (one day a week)
$1296 – Student assistant help to support cataloging at Divinity, freeing time for project work (3 hours/wk x 32 wks at 13.50/hour)

Estimated project time:
10 months, including:

- 3 months weekly meetings for developing the project and useful tools and procedures
- 5 months of project work and related tweaking
- 2 months for analyzing work