Batch Processing Phase 1 Report

Phase I Report. 2CUL TSI Batch Processing Working Group

Leads: Mark Wilson (Columbia -- chair), Gary Branch (Cornell)

Other Members: Columbia: Evelyn Ocken, Gary Bertchume

Methodology -- The Phase I Report is a summary of batch processing workflows and processes in place to support the Technical Services units at Cornell and Columbia University Libraries. The report was built on the foundation of two supporting and accompanying documents compiled to facilitate comparison. The Batch Inventory spreadsheet serves as a single view repository of the multiple batch jobs in place at both institutions. The Batch Processing Outline serves as the structural basis for this summary and follows the five Phase 1 points of the Working Group Charge.

The compiled batch routines extend in places to service needs with Technical Services aspects yet have more of an Access emphasis or function. Likewise, examples of report generation from library systems are also mentioned as data retrieval and data extraction/manipulation are closely related. However, the focus of this report will center on processes associated with traditional Technical Services functions, namely Acquisitions and Cataloging.

Compile an inventory of batch processing staff and expertise at both Columbia and Cornell; compare current job assignments.

Quite likely the organizational and staffing differences as to where batch processing happens at each institution will prove the most significant contrast that permeates through all aspects of the batch processing workflows and policies covered in this report. The essential difference being that these routines are centrally programmed and maintained by high level systems personnel at Columbia and are more dispersed at Cornell, conducted by a variety of staff within technical services units.

Gary Bertchume, Manager, Library Systems, Libraries Information Technology Office (LITO) of the Digital Programs and Technology Services group has 35 years of experience. Gary oversees and has scripted the procedures for MARCIVE record loads, Rapid exports, OTF record deletes, CLIO SOLR ingests, smart barcode projects, global data changes, bibliographic file statistics, OCLC batch projects, cataloging production and change of collection size statistical extracts, system deltas to backup/mirror server and he also manages system data migrations for Columbia University Libraries.

Evelyn Ocken, Senior Systems Analyst/Programmer, Libraries Information Technology Office (LITO) of the Digital Programs and Technology Services group has 20 years of experience. Evelyn oversees and has scripted the procedures for ReCAP accessions, OCLC exports, LTI author records loads and loads, MARC record approval and firm order loads, EDI orders and invoices, Borrow Direct record deletes, Acquisitions and EOFY reports, collections commitments and expenditures, audit reports and ebook, ejournal and eresource MARC record loads.

Jennifer Chong, Associate Systems Engineer, LITO, DPTS, Transferred from User Services Librarian as of 7/13. Jennifer has moved to a vacated position within LITO that initiates and manually executes many of the scripted procedures and reports set in place by Gary and Evelyn. Additionally, Jennifer updates domains in EZ Proxy for e-resources identified as unproxied after batch loading.

The distributed batch processing model of Cornell is facilitated by the creation of LSTools, which acts as a middleware interface permitting LTS staff interacting with the physical materials to also coordinate the bulk importing of the associated records.

Peter Hoyt, Programmer/Analyst, Library Systems, has 15 years of Library programming experience. Peter is the creator and maintainer of the LSTools and he is the creator of many Perl scripts used by Library Technical Services (LTS) staff loading MARC records with LSTools.

Gary Branch, Library Administrator II, is the Batch Processing Supervisor of LTS, with 23 years of experience, including 5 years supervising the Batch unit.

Peter Martinez, Technical Services Assistant TSA V, Batch processing & Metadata Specialist, LTS, 13 years of experience. Peter vets MARC data to be loaded, does Perl scripting, runs Access reports and handles character encoding issues that are problematic to successful data loads.

Natalya Pikulik, TSA IV, Batch Processing & Standing Orders, 7 years of experience. Natalya loads vendor records for designated record sets (e.g. SpringerLink & Books 24x7), runs batch withdrawals and location flips with Strawn tools and configures vendor load specifications.

Nancy Solia, TSA IV, Batch Processing & Image cataloger, 16 years of experience. Nancy loads vendor records for designated record sets (e.g. Safari & Knovel), runs batch withdrawals and location flips with Strawn tools and configures vendor load specifications.

Joe McNamara, TSA IV, Batch Processing & Standing Orders, 23 years in technical services. Joe loads vendor records for designated record sets (e.g. Rand & Memso), runs batch withdrawals and location flips with Strawn tools and configures vendor load specifications. He also handles authority record loads and headings maintenance.

Batch processing at Columbia is managed centrally by Systems personnel scripting processes directly into the Production server. Scheduling of all batch processing (including Cron jobs) is easily coordinated. Large batch processing jobs at Cornell are also handled by Systems personnel but a good amount batch processing is conducted by technical services personnel who have defined batch job responsibilities. With this model, record loads are easier to coordinate with the arrival of material shipments. The locally developed LSTools is used along with the Strawn tools “Location Changer” and “Record Reloader” to load and extract files.

Examine and evaluate reporting and decision-making structures at both institutions

At the broadest organizational level, Columbia University Libraries branches into three distinct service groups and one administrative group, headed by Associate University Librarians or Vice Presidents reporting to the University Librarian. The Library at Cornell is divided into six organization units, headed by Associate University Librarians under the University Librarian. These structures are detailed in the Batch Processing Outline.
Notable at Columbia is that the batch processing personnel in LITO are in the Digital Programs and Technology Services section. Monographs Processing Services and Monographs Acquisitions are the central technical services operations of the libraries and the beneficiary client of many batch processes. MPS and MAS are in the Bibliographic Services and Collection Development section. Batch processing and technical services therefore operate in different divisions. The East Asian technical services unit reports to the Director of the C.V. Starr Library, reporting to the AUL for Collections and Services, yet another division. Additional batch processing and reporting needs may arise from the technical services units of Barnard College and the Health Sciences Libraries that share the same catalog and fall outside the libraries' organizational structure. The time resources of Columbia's batch processing staff in LITO are highly utilized for multiple reporting and project development efforts across all four divisions of the Libraries.

Despite the organizational separation of batch processing staff and technical services units, communication between these units is both direct and frequent, conducted mostly by email. Technical details are communicated and resolved via direct communication while larger issues are discussed and resolved in operational group meetings that bring together key personnel at all levels from the various divisions.

Cornell batch processing takes place directly in Technical Services and therefore batch processing decision making takes place at lower levels than at Columbia. The Batch Processing Supervisor is a paraprofessional manager who develops and initiates batch jobs. The level of oversight may consist of only a conversation with his library manager. The Supervisor in turn communicates with his staff mostly via informal conversations. Most vendor loads tend to have similar record modifications and can follow established local practices and do not require much deliberation, so new vendors can be implemented routinely and with relative ease.

Large scale batch jobs are managed by Chris Manly or Peter Hoyt, who report to the Associate University Librarian for Information Technology. LTS and Batch Processing report to Xin Li, Associate University Librarian for Central Library Operations.

Both Columbia and Cornell vet records and vendor performance before implementing MARC services. In both institutions these decisions are made in technical services departments, even though the requests for MARC records often come from selectors.

Batch processing activity is organizationally more distant from technical services at Columbia than at Cornell. As a result, more cross divisional communication is necessary to implement MARC record loading from new vendors at Columbia. Cornell technical services units can initiate new MARC services independently without being subject to competing priorities.

Compile an inventory of all policies, practices, and workflows involving automated record loading, export and maintenance activity at both institutions. – including which unit or department is responsible for which activities.

An extensive listing of all batch jobs at Columbia and Cornell exists in the Batch Inventory spreadsheet and there are links to detailed process descriptions in the Outline. One of the fundamental differences in the use of MARC records for approval receipts is that Columbia creates Purchase orders and line items as part of the load whereas Cornell loads only bib and holdings records. Reports of each load arrive as emails to acquisitions staff. Duplicates that are detected via ISBN matching are manually resolved by an MPS supervisor and records are merged. Frequently this requires the deletion and recreation of line items in order to delete the duplicate record. Columbia uses Voyager operators named with the vendor code to display in the history as the creators of the bibliographic and holdings records. Approval items load with the default “sho” locations. Shelf ready approvals load with locations and items, a report of item barcodes is sent to acquisitions staff who scan them as soon as possible to ensure the OPAC reads “in transit to library” rather than “Not checked out”.

The approval records for Columbia’s major vendors (YBP, Casalini, Aux Amateurs and Harrassowitz) are loaded via a scripted cron job, and the records can appear in the catalog before the materials arrive in acquisitions. LTS staff at Cornell is able wait until materials are received to load the records.

Export of Columbia’s records to OCLC takes place every Wednesday for records last edited in the week 10 days before export (it is assumed all edits are by then complete). Cornell sends a nightly export of new cataloging which are triggered by a cataloging statistic in the 948 field. OCLC keys are added as 035 fields from a return report for both libraries. Columbia sends a separate file for East Asian titles as they also load institutional records. Many record sets have a 965 marker that serves to enable easy selection and collection identification and to apply the policy of not uploading certain eresource collections to OCLC. Cornell uses an 899 field for aggregating vendor packages and a 995 field to prevent export. Columbia excludes from export EL 5 preliminary records processed for Precat/Offprecat while Cornell includes records with EL 3, 5 and 8 in its uploads.

Columbia sends a file of EL/5 preliminary records annually to OCLC to match with full level cataloging records. Cornell send files for their EL 3, 5 and 8 records monthly.

Cornell load firm order records from WorldCat Select. Columbia receives firm order catalog MARC records only for YBP and Casalini, and now also records from POOF ordering. GOBI orders will merge and add holdings to existing bibs if they match, if the matched record is for an ebook, the MARC record gets kicked out and is then manually loaded by the Order Unit supervisor.

A monthly extraction of records is sent out for authorities processing selected on criteria in a perl script. Notices are sent to technical services units not to edit records from the specified period as the edits will be lost once the returned bibs replace the place holder record. Cornell does not use an authority vendor and use in house processes to load authority records on a weekly basis and conduct headings maintenance.

Identify dependencies and limitations inherent in working with other functional areas at Columbia and Cornell, both within and beyond technical services and library systems.

There are multiple inter-dependencies at both institutions listed in bullet points in the Outline. At Columbia, units from all four administrative units depend on the LITO staff either for batch processing, financial and voyager reporting or system development. As a result, there are competing priorities for the time and expertise of LITO personnel. LITO in turn depends on staff in CERM and technical services (including East Asian) to identify any errors in batch processing that are not detected by the load routines. Selectors depend on technical services and LITO to implement desired MARC record services, and technical services and LITO depend on vendors to maintain reliable service performance. HSL depends on timely loading of eresource records to avoid ordering duplicate content. Monographs Recon Projects, the ReCAP coordinator, Collection Development and Access Services likewise rely on LITO support. Limitations mostly consist of EOFY demands and vendor’s ability to meet requirements for MARC services.
The Batch Processing unit at Cornell is highly dependent on Library Systems for scripting and support when tools do not operate as expected. The unit is also dependent on other units to identify problems with the potential for batch correction. Cataloging is depended on for quality control of MARC data. Other units depend on the Batch Processing unit for their functions. Examples are: The ordering unit for ITSO/POOF firm orders processing, Monographs Receiving and Documents for loading of approval MARC records, Database Quality for batch maintenance, reports, and withdrawals, Cataloging and Metadata Services for reports, Annex Library for record cleanup of accessioned volumes, Cornell Weill for Prepared MARC records for shared Springer packages, and Law Library’s Technical Services for batch record load specifications.

If possible, establish baseline productivity numbers for activities and projects at each institution to allow for future assessment of potential changes and development associated with 2CUL TSI

Both CULs use Voyager’s bulk import to create approval MARC records with purchase orders. The main difference is related to different workflows. While Columbia loads almost all approval titles, almost 23,000, with purchase order line items included, Cornell does not. Cornell only creates purchase order line items through bulk import for the vendors, Harrassowitz and Casalini. For the remaining of the 17,400 approval volumes loaded, most of them were loaded without line items. Cornell also loads records for various LC acquisition programs.

While we both use POOF to load records and create purchase orders for firm ordered titles, we differ both in number of line items loaded for firm orders. Columbia uses POOF and YBP’s GOBI for their source while Cornell uses both POOF and WorldCat Selection. This yields a significant difference in the number of loaded records with Columbia producing around 9,000 firm order line items while Cornell produced 14,000 firm orders through bulk loading.

Columbia and Cornell have several Voyager extract and export projects to OCLC. While Cornell does daily exports via a nightly cron job of new cataloging to OCLC, Columbia does weekly extracts. Cornell totaled approximately 250 thousand versus Columbia’s 170 thousand last fiscal year. Columbia also sends preliminary records to OCLC for materials in their circulating backlog, encoding level 5, once a year for another almost 17 thousand and loaded improved records for 8,900. Cornell uses this process, locally called Batchmatch, to send records with encoding levels, 3, 5, and 8 to OCLC. For a total extract of 37 thousand records with updated records for 4150. 2CUL institutions both send extracts for OCLC institutional records to be created. Columbia for their East Asia program sent 19,000 records to OCLC last fiscal year. Cornell sends pre-1900 imprints and bibliographic records for the Law Library which totaled 222,000. Cornell also sends updates to its own manuscript and archival collections totaling another 7900 records as well as exporting all cataloged holdings and changes made to the holdings totaling over 1 and ¾ million holding records exported last fiscal year.

Serial Solutions is the primary source for MARC data for both institutions e-resources. While Columbia discourages getting metadata from other sources, they do upon occasion get their metadata directly from the vendor. Cornell while it prefers Serial Solutions, if the metadata is superior elsewhere will go to that source whether it’s OCLC collection sets or from the vendor. Last year Columbia batch processed over 80,000 e-journal and 440,000, e-book records while Cornell handled 14,000 e-journals and 270,000 e-book records.

Cornell Batch Processing also does many batch maintenance projects, including batch single volume monograph withdrawals, adding 830 fields to those with only 490 indicator 0, many location flips as library collections were moved, removing unneeded 9xx fields, and fixing various MARC errors.

Detailed statistical information is available both in the Batch Processing Outline and the Batch Inventory.

Recommendations regarding a work plan and critical issues to explore in Phase 2 of the group’s assignment

To investigate and learn what batch functionalities and record load features are built in to ALMA and see how easily existing record load workflows at Cornell and Columbia can be expected to migrate.

Leaving established workflows in place, jointly experiment implementing MARC services with new vendors where practicable.

Explore the use of WorldShare Metadata Collection Manager to identify record harvesting for HathiTrust and other record sets.

To make a joint data map/dictionary of terms and MARC fields used in conjunction with record identifications for extracts between the two CULs.

Work with other 2CUL TSI groups in areas of functional overlap to coordinate phase II efforts.