Graduate Center Academic Works: The First Few Projects

Nomenclature:

- **Vendor Name:** bepress
- **Product Name:** Digital Commons
- **Our Instance:** Academic Works / Graduate Center Academic Works
- **URL:** http://works.gc.cuny.edu
- **Generic Term:** Institutional repository (IR)

Project 1: Faculty Works
http://works.gc.cuny.edu/facworks/

- **Upload method:** Individual submissions via Faculty Works submission form (which was customization of the default Digital Commons upload form)
- All submissions go into the central Faculty Works collection, and they are also "auto-collected" into subject-specific Faculty Works collections (e.g., Library Faculty Works, Physics Faculty Works) according to rules I set up.
- Once the fields and submission form are finalized, there's no need for cataloging/metadata expertise. Submitters are prompted to supply basic metadata (e.g., abstract, keywords, and disciplines). Submissions do not require additional cataloging.
- However, there is the question of how much metadata clean-up we should do for submitters. Fix their capitalization? Fix typos and grammatical errors in their abstract? Add keywords/disciplines if they didn't include any?
- Even bigger question: Which is better, self-submission by faculty or submission by us (or student workers) on behalf of faculty? Other IR admins report that faculty can be sloppy in their self-submissions, that it often works better to offer to do it for them. What would work best at CUNY? An open question!
- Everything submitted goes into holding pen for approval. Nothing is publicly viewable until it is approved.
- **Important:** Anyone can create a Digital Commons account and submit to any Digital Commons repository. (This is because many journals are housed in Digital Commons repositories, and of course anyone should be able to submit to a journal.) As a result, anyone anywhere can submit a supposed faculty work to Academic Works. Therefore, every submission **must** be checked to confirm that the submitter really is a faculty member at your campus!

Project 2: Dissertations/Theses 2014-Present
http://works.gc.cuny.edu/etd/

- **Upload method:** XML batch uploads
- As of 2014, all new dissertations and theses are made open access in Academic Works, with an optional embargo of up to two years (extendable upon request).
- We have three degree conferral dates per year, and we aim to upload the dissertations/theses from each within a couple of months of that date.
- Students submit their own dissertations/theses to ProQuest ETD Administrator, as they did in the past. ProQuest sends their files and XML metadata to a Graduate Center server. (Because the metadata — title, abstract, etc. — was provided by the student, it is sometimes sloppy.)
- Our Metadata Librarian massages the XML from ProQuest into the XML format required by Digital Commons. She does this with the help ETD_CON, a tool created for exactly this purpose (see http://journal.code4lib.org/articles/1647 and http://digitalcommons.bepress.com/collaboratory/58/). Note that ETD_CON required adjustments before we could use it.
- A batch upload is a mass upload of XML metadata, not a mass upload of files. In order to associate files with the metadata, the metadata must point to files that already live online somewhere. Therefore, our Digital Services Librarian temporarily puts all the dissertation/thesis files on our web server, so Digital Commons can see and ingest them. (While on our server, these files are theoretically public, but they're not findable, so we're not violating students' embargo requests.) He takes them down once the ingestion is complete, which sometimes takes a few days.
- Our Metadata Librarian then uploads the XML metadata, and Digital Commons ingests both the XML and the files it points to. Once the ingestion is complete (which sometimes takes a few days), our Digital Services Librarian removes the files from our web server.
- **Issue:** ETD_CON does not correctly process foreign accents and symbols, so each item needs to be checked after upload. We're still hoping to figure out a solution.

Project 3: Computer Science Technical Reports
http://works.gc.cuny.edu/cs_tr/

- **Upload method:** Excel batch upload
- Computer Science had their own file server for technical reports, but it died suddenly. The 201 reports (all in either PDF or PS format) survived, but all metadata was lost. They needed help, and IT directed them to us.
- I started by distilling all PS files into PDF format, making it a collection of PDF files only.
- I then pulled the PDFs into Zotero and used its "Retrieve Metadata for PDF" tool, which scrapes information from a PDF, searches Google Scholar for matches, and adds the corresponding metadata to the Zotero record. Zotero identified about 150 of the papers (sometimes perfectly, sometimes requiring some edits). About 50 required manual metadata entry.
- I only cared about these metadata fields, as they were the only ones going into the repository: Technical Report number (e.g., TR-2014002), which was part of each item's filename; author; title; and year.
- Once the metadata was all correct in Zotero, I exported it to a CSV, imported that into Excel, and massaged the spreadsheet into the form that Digital Commons requires. This required deleting/hiding a bunch of columns, reordering columns, and performing a variety of Excel functions, including left truncation, right truncation, concatenation, and transforming formulas into text.
- Our Digital Services Librarian temporarily put all of the PDFs on our web server.
- I added each item's URL to the metadata spreadsheet and uploaded the metadata spreadsheet. (I happened to upload during peak upload season, so it took three days for the batch to be processed.)
- Once the batch was processed, we took the PDFs down from our server.
- IT set up a up redirect from tr.cs.gc.cuny.edu (the old Technical Reports server) to works.gc.cuny.edu/cs_tr/. Specifically, any URL that starts with http://tr.cs.gc.cuny.edu gets redirected to http://works.gc.cuny.edu/cs_tr/.
- **After-the-fact change:** I uploaded the batch with the titles in the form "Title (TR Number)" but Computer Science preferred "TR Number: Title." This change could easily be made through a batch revision: I downloaded the metadata spreadsheet for the collection, manipulated the spreadsheet, and re-uploaded it. As soon as it was processed, the titles appeared in the desired form.

Project 4: Dissertations 1965-2013
http://works.gc.cuny.edu/etd_1965-2013/

- **Upload method:** XML batch uploads for 2009-2013 dissertations. TBD for pre-2009 dissertations.
- Approximately 12,000 dissertations, some already digitized but most in microfilm. We paid ProQuest to digitize them all, and they also supplied us with metadata for all the items (XML for 2009-2013 and MARC for 1965-2008).
- Converting MARC to XML or Excel files ready for upload won't be as easy as tweaking XML files with the ETD_CON tool. We haven't started converting MARC files yet, but it will almost certainly involve significant coding by our Metadata Librarian.
- We will process and upload the dissertations one year at a time, moving backwards from 2013. The project is so huge that our Metadata Librarian cannot take the time to dedicate herself entirely to this project. She will integrate this into the rest of her work, and it will take quite a while, perhaps a year.
- **Permission to Post:** Pre-2014 graduates never agreed to inclusion in an open access repository, so how open can we make these works? We're starting by limiting access to CUNY-wide IP space & CUNY proxy — this, we decided, was the online equivalent to availability on the shelf. We will then undertake an outreach campaign to alumni to give them each the option of making their dissertations open access.

Ongoing Project: Encouraging faculty submissions.

Upcoming Project: Creating a "Graduate Student Works" series for each subject and allowing self-submission by students. (Admittedly, allowing student self-submission is not common, and the Graduate Center may be the only CUNY school where this is appropriate. More common is to have curated collections of student works — student journals, award winning papers, etc.)

Upcoming Project: Working with GC Centers/Institutes to identify collections for inclusion.

Goals for 2015:
- Upload all 1965-2013 dissertations
- Hit 500 faculty works (with help of Stephen Flynn's Scopus + SHERPA/RoMEO script, http://tinyurl.com/flynnscript)
- Work toward getting 2-3 collections from GC Centers/Institutes

Skills Required:

- **Comfort with OA concepts:** Tackling these projects requires clarity about "green" open access, authors' rights, pre-print vs. post-print vs. publisher's PDF, etc.
- **Comfort with idiosyncratic software:** The front end of Digital Commons is very user friendly, but the back end is considerably more complicated. Learning the back end takes time, and it's easy to forget the details if you don't use it regularly.
- **Attention to detail:** Anyone who works with the back end and/or deals with batches must be very careful and detail-oriented.
- **Specific technologies:** Batch uploading requires comfort with Excel (intermediate-level skills such as cell concatenation, left truncation, right truncation, etc.) and/or XML. The CS Technical Reports project also required Zotero and Acrobat Distiller.

Time/Staff Required:

- Outreach to faculty can be integrated into the outreach work of subject liaisons.
- At least one person at each library needs to be comfortable with the back end of Digital Commons. It is not necessary or even desirable for all liaisons to be trained on the back end and have admin privileges.
- For those who work with the back end, deal with batches, etc., the work will likely be significant enough to require some release from other duties.
- Student workers or interns could easily perform a lot of the more tedious work.

File Types, Size, Embargoes, Etc.:

- Digital Commons can accept any file type: PDFs, image files, audio files, etc.
- Of course, it's best for each file to be in the appropriate format for its content (e.g., data as an Excel or CSV file rather than a PDF).
- It's best to submit text-based files (articles, chapters, etc.) as PDF files, as Digital Commons automatically creates cover pages for PDFs. It does **not** create cover pages for other file types.
- Digital Commons does not stream content (but you can embed streaming content from sites like YouTube and Vimeo) or serve up anything but files – i.e., no live websites, web apps, etc.
- Each uploaded file can be **up to 2 GB** in size.
- CUNY is currently allotted a total of **30 TB (terabytes) of space**.
- Each file can be open access, or it can be embargoed or subject to access controls:
    - **Embargo:** The file is withheld from the public until the embargo period expires. Digital Commons counts down the embargo and automatically makes the file open access when the embargo expires.
    - **Access control:** The file is available to some specific population. There are several ways of doing this, but the one that will likely make the most sense most of the time is by limiting access to a certain IP address range and possibly also allowing ezproxy access.