Importing data to Leganto from Equella

Created By: Deborah Fitchett
Created on: 6/22/2021

Equella is repository software used by some libraries to host course readings in order to manage copyright. A couple of years ago the company that provided it made it open source and moved on to other things. While support can still be accessed, this has possibly accelerated some libraries' interest in also moving to other methods of managing course readings – including Leganto.

Some libraries have done this move on a "clean slate" basis rather than migrating content. But it is possible to do so – at least mostly. The method below will let you migrate:

- bibliographic data
- the files themselves

It *doesn't* include importing past "activation" data (ie dates that copyright readings were made available to students) into Leganto in a machine-readable format.

These instructions are focused on getting the data and files into a format you can import via FTP as part of your initial migration/implementation of Leganto. If you want to import data/files after you've already implemented, you may need to adjust some later steps to work with the Leganto APIs instead. But the steps for extracting the data from Equella should still be the same.

Software required

(Note it's beyond the scope of this overview to go in depth into how this software should be set up and configured.)

- BIRT reporting software can be used to create a report of journal/book portion metadata (including activation details – getting it out of Equella is no problem, Leganto just didn't seem to have a way to import it when we looked). We based this on the report we'd already created to send to our copyright licensing vendor.
- Equella Bulk Importer (note the releases on the right of the page for the actual .exe file if you don't want to compile it from source!) This can be used to download metadata (but not activation info) and the actual files.
- OpenRefine can be used to refine messy data in spreadsheet format, eg using regular expressions.
  - OpenRefine requires you to install Java for it to work.
  - When it runs, it will bring up a command line to Do Computer Things – give it a few moments – then it will load an interface in your browser that will let you upload a CSV
- Microsoft Excel!

Downloading metadata and files

1. Use **BIRT** to create a report containing all the metadata fields you need, including the filename of the item.
    - If you've got a report you've used for copyright reporting, this probably already has most of what you need.
2. Using Equella's **Administration Console** tool, upload your report onto your Equella site.
3. From your **Equella** site's Reporting tab, run your report.
4. Export as Excel
5. **In Excel**, remove any details above the header row, and unmerge anything that's merged.
6. Data > Remove duplicates (with all columns ticked)
    - NB This means a citation will only be listed once per course, but it might be listed under multiple courses.
7. Make a duplicate of the spreadsheet as we're now doing two processes before merging them together.
    - Important: from this point it's best if you don't do any kind of sorting of data; or if you do, immediately undo it. Otherwise when you come to merging the two processes back together you may have to resort to a VLOOKUP and I make no guarantees!

Process 1 – Tidying metadata

In Excel:

1. Add columns to match https://knowledge.exlibrisgroup.com/Leganto/Implementation_and_Migration/Migration_Tools/Reading_List_Migration_Tool_(English)
2. If you want to add course data from an external system (eg to use the course title in the reading list name, or to add instructor details)
    a. Add the external data to another sheet in the same Excel file, making sure that column A contains a unique identifier that also appears in your Equella data – eg a course code.
    b. Back on the sheet containing Equella data, in the column where you need the course data, use a VLOOKUP formula. This is one of my favourite formulae, and the way it works is eg:
    `=VLOOKUP(C2,ExternalData!$1:$10485,2, FALSE)`
        - `C2` is the field in your Equella data that contains the unique ID (eg course code) you want to look up
        - `ExternalData` is the name of the sheet where you've pasted your external data
        - `$1:$10485` selects *all* the rows you want to search through. You can select a smaller selection, but use $ to fix both the rows and columns so the area doesn't shift when you fill down
        - `2` identifies the column in ExternalData that you want to get the information from, except just to confuse you it counts with numbers instead of letters. So 2 will bring back information in column B
        - `FALSE` makes sure you match exactly on your unique identifier. If for some reason you want fuzzy matching, you can use `TRUE`, but most of the time it'll cause more problems than it's

worth.

Fill the formula down the column.

3. Fill down any fixed parameters you need eg:
    a. reading_list_status = BeingPrepared
    b. RLStatus = DRAFT
    c. visibility = DRAFT
    d. citation_status = BeingPrepared
    e. external_system_id = [a blank space, not empty]

In OpenRefine:

This allows you to do a lot of complex data cleaning using regular expressions.

1. Upload a CSV
2. For each thing you want to clean, go to "Edit column > Add column based on this column" and type in `value.replace(/your regular expression/,"your new text")` – for example:
    a. citation_isbn – Remove hyphens with `value.replace(/-/,"")` Then replace every sequence of stuff that isn't a number or X with the delimiter ; (semi-colon followed by space): `value.replace(/[^0-9X]+/,"; ")`
    b. citation_issn – leave the hyphens but delete every other unexpected character with `value.replace(/[^-0-9X]+/,"; ")`
    c. citation_volume and citation_issue – you could theoretically delete everything that's not a number (and this conforms to the data entry rules in Leganto) but note you'll lose information about issues called "Spring", "March-April", etc. Work out the balance between clean / detailed that's appropriate for your organisation
    d. citation_issue – clean up best you can bearing in mind lots of issues are named Spring, May, etc
    e. citation_start_page and citation_end_page – If you currently have only a column with a range of pages (eg "45-54") then you can easily break it on the hyphen eg add one column with `value.replace(/-[0-9]*/,"")` and another with `value.replace(/[0-9]*-/,"")`
    f. citation_chapter – we had a number of items where the citation_title was something like "Chapter 1: Introduction" (Or "Ch. 1" or other variations). So we looked for where "Ch" appeared at the beginning (with optional extra letters then optional full-stop and/or space) followed by a number followed by any other text – and extracted that number: `value.replace(/Ch[a-z]*[. ]*((\d)*).*/,"$1")`
        ◦ If you're confident with this you can then edit citation_title to delete all of that prefix and any punctuation and just include the actual title.
    g. citation_edition – similar to above, where citation_title often *finished* with something like " (2nd Edition)" or " 2nd ed." etc
    h. citation_author – this was messiest for us as lecturers had entered these into Equella in all sorts of formats. But they were delimited by semi-colon so we just split it on the first semi-colon (cf citation_start_page) so anything before was citation_author and anything after was additional_person_name.
3. When you're happy with the data, Export as CSV

Back in Excel

1. **Important!** When importing back to Excel, use Data > Get External Data > From Text and **make sure you select**:
    - Delimited
    - Unicode (UTF-8) (otherwise diacritics will go haywire)
    - Headers
    - Delimiter = comma
    - In the final data preview, select *all* the columns and format them as *Text* (otherwise dates and ISBNs will go haywire)
2. Check all your column names again (as OpenRefine will have created a lot of new columns) and get them back in order.
3. Save early, save often!

Process 2 – Downloading files

With a separate copy of your Equella export:

1. Reduce it to the “Item id” (or may be called “citation_originating_system_id”) column and rename that column “source”
2. Add empty columns with reference to the metadata schema in your Equella (you can browse it via the Administration Console > Metadata Schemas > Edit > Editor; or use BIRT to see which each one contains). As examples, we used the following columns – note the last three were all different places where the full-text attachment might be stored due to our local database architecture – but you’ll want at least the item ID, its version, its title (to confirm you have the right one), and of course the attachment.
    - item_id
    - item_version
    - lom/general/title
    - localData/attachments/attachment/uuid
    - item/copyright/portions/portion/sections/section
    - item/copyright/portions/portion/sections/section/attachment
3. Save this **as a CSV** into an empty directory
4. In the same directory, create another empty directory for the attachment files you’re now going to download.

In EBI:

Connection tab

1. Fill out
    - Institution URL

- Username (you'll need admin access to get files)
- Password

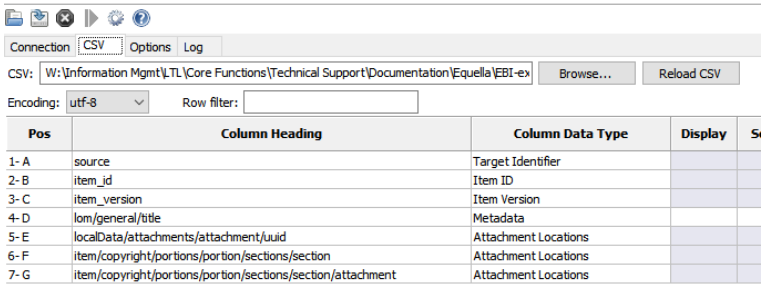
2. Test/Get Collections
3. Select a collection

Options tab

1. Set up a base path pointing to the empty directory you created for attachment files
2. Important: Tick "Export items as CSV" – the "Test Import" button should now change to "Test Export".
3. Filename conflicts: Do not overwrite any files

CSV tab

1. Browse to the CSV file you created at the start of this process and load it.
2. It should pull in your column headings.
3. You need to edit the appropriate column data type for each heading – these are dropdown menus.
   - Source (the only column you have that isn't empty) must be the Target Identifier
   - item_ID = Item ID
   - item_version = Item Version
   - any metadata you want like the title, author, etc = Metadata
   - anwhere you store the attachment file = Attachment Locations



When your screen looks like this you're almost ready, but first:

If you've set up Equella to show users a copyright notice before they can access the file, this is going to just download a file that's named like the file you expect but only contains the copyright notice! So go to:

- Equella > Administration Console > Collection definitions
- Then Edit the collection > Extensions > CAL Licensing Configure
- Untick "Resources with copyright status require agreement?"
- OK
- Save

Now you can either do a Test Export (exports metadata only, not attachments) or go ahead and Export (exports metadata into the CSV and attachments into the files directory).

In File Explorer

1. Some files will have duplicate names – these will be in subfolders of your attachments directory eg 1/documentName
2. In each folder, rename each file with the folder number -> underscore -> documentName eg 1/chapter 1.pdf -> 1_chapter 1.pdf
   - We had only a small number so did this manually. If you've got too many for that, try using a Powershell script (see #4 for an example, and google for syntax)
3. Move all files from these subfolders out into the main directory and delete the empty subfolders.
   - If any names still clash, don't overwrite but rename them – just make sure you also rename them in the CSV sheet as below.
4. Ex Libris won't accept files with spaces in the filenames and we definitely didn't want to edit these manually. So we used Powershell to navigate to the files folder, and ran the command:
   `dir | rename-item -NewName {$_.name -replace " ","_"}`

And back in Excel

1. In the CSV in the filename column, also rename all 1/ -> 1_
2. And also replace all spaces with underscores
3. Now copy this filename column back to the file_name column of the spreadsheet you got from Process 1
4. Spot-check using item ids that the right data has ended up on the right line! If not, take it out again and try using a VLOOKUP similar to when we pulled in extra course data but using the item_id as the unique identifier.

Preparing for upload

You should have:

- An Excel file with metadata including filenames – save as xlsx
- A "files" folder with files

You'll want to double-check:

1. Make sure that for all the mandatory fields, there's something in the field! See especially citation_title, reading_list_code, reading_list_name, RLStatus, section_name, citation_secondary_type. And check course_code too – not mandatory but a missing course_code may confuse.
2. The final column should have a space in every cell.
3. It should be sorted by courses, then reading list codes, and then by section name and/or start date to avoid duplicates

And your Ex Libris implementation team should provide instructions to upload the Excel sheet and files folder via FTP.

Report