MarcEdit task lists and vendor-supplied metadata: revisiting the supplier-vendor relationship at the University of Leeds

Ceilan Hunter-Green, Metadata and Discovery Coordinator, University of Leeds Libraries

Like all other academic libraries, the University of Leeds needed to hugely increase our electronic provision after the first Covid lockdown in March 2020. Most of the pressure was on providing e-textbooks and e-books, but there were also many modules that suddenly needed access to streaming video. One of these streaming video suppliers was the British Film Institute, who launched a new Institutional Subscription offer during lockdown, but could not provide records. This provided an opportunity for Leeds' Metadata and Discovery team to develop a process of automated record creation using MarcEdit templates and task lists and provide those records to the subscribing community at no increased cost, whilst receiving a discount on our own subscription.

British Film Institute subscription

The BFI streaming video collection is strong in areas that our other institutional suppliers aren't, and their films were particularly needed on Japanese and history of film modules where foundational material is difficult to come by otherwise. There were some challenges, though, as their institutional subscription was new at the time and the streaming video service's metadata format was very different to what we were used to as librarians. They also have a comparatively small collection with around 600 active titles at any given time, which is updated very frequently compared to the more static collections we were used to. Most significantly to us, they didn't offer MARC records, which meant that we spent a lot of time on maintaining local records for their films so that we could link reading lists to individual catalogue pages rather than linking to the BFI Player site for students to repeat their search. But that lack of MARC records became an opportunity for us to offer something back to them as vendors.

For the first year of our subscription we, like all of BFI's other subscribing institutions, received metadata for the streaming video collection's films in the form of a very basic spreadsheet, originally 10 columns and upgraded to 12 columns in February 2022. We used MarcEdit's Delimited Text Translator to create bare-bones records, and then as an Ex Libris library running Alma as our LMS we used Alma's import profiles to run normalization rules on those basic records to get them into our system in semi-decent MARC shape.

The provided metadata covered the film's title, internal BFI ID number, access start and end dates, country of origin, release year, two Genre terms and two Director names, though usually the films just had one director. The remaining metadata that we needed wasn't provided—so it was copied from the BFI Player streaming platform field by field into Alma's metadata editor with each monthly update. The average monthly addition was 22 films, but in practice this meant some months had five additions and some months had fifty, so it was an unpredictable demand on our staffing resource.

Renegotiation and enhanced metadata provision

This time-consuming and inefficient record production continued for a year, until colleagues in the Acquisitions team were preparing to go into negotiations with BFI for a new year's subscription. Previously BFI had asked about the process of us providing the records but it didn't feel practical when the process was so time-consuming. But as we went into that second year of subscription, with more familiarity with the metadata from BFI Player, more experience with MarcEdit and more confidence in the value and usefulness of our records to the community, we thought we would be able to take another look at the process of record creation.

At the University of Leeds, purchasing is handled by a separate Acquisitions and Reading Lists team, and my Metadata and Discovery team were responsible for handling the access and the discoverability of these resources once we had them. Acquisitions colleagues studied the average price for MARC records that we paid to other subscribers and the amount of time we were spending on the process of record creation, and were successful in negotiating a discount. At that point, we just had to develop a workflow to create the records in a much more efficient way than we had been; we needed much more (and better quality) metadata to start with, and the resulting records had to be an even higher standard if they were going to be shared with other institutions.

The team at BFI worked with their systems team to extract all relevant metadata from their streaming platform in the form of a 56-column .csv file, a huge improvement on the previous spreadsheet. We now have metadata for all cast members in separate columns, a unique system identifier from BFI, runtime, rating, and everything else we need which previously had to be copied from the BFI Player site to fill out our records. We also started to receive some additional metadata like original language titles in addition to English titles, which isn't on the public BFI Player site, and a direct URL instead of the general landing page URL we added before.

MarcEdit transformation – import template

To translate this 56-column spreadsheet into a MARC record, we start with the MarcEdit import template, which brings that mammoth spreadsheet into MarcEdit (Figure 1) through the Delimited Text Translator. It mostly imports one column per field with a few exceptions where data from a single column is added to two fields, such as the Cast column data added both to a concatenated 511 and to individual 700 fields, and the Year of release data added both to the 501 and the 046. Some fields are populated as placeholders, like the 501 for the year of release which will be moved to a 500 later, and the directors are added to a 701 until the addition of a relator term, when they're moved to a 700 as well. Most of the data at this point is pretty blunt.

```
=LDR  00000ngm a2200000la 4500
=008  230831s9999||||xx\|||\|\\\\|\||\\|und||
=024  7\$a9afb773d-ce2e-59dd-b516-b13c84dada79
=046  1\$k1953
=130  0\$aTōkyō Monogatari
=245  00$aTokyo Story.
=300  \\$a136
=500  \\$aJapan.
=501  \\$a1953
=506  1\$a02/09/2025;@.
=511  1\$aChishu Ryu,Chieko Higashiyama,Setsuko Hara,
=520  \\$aA constant fixture in critics' polls, Yasujirō Ozu's most enduring masterpiece, Tokyo Story, is a
beautifully nuanced exploration of filial duty, expectation and regret. From the simple tale of an elderly
husband and wife's visit to Tokyo to see their grown-up children, Ozu draws a compelling contrast
between the measured dignity of age and the hurried insensitivity of a younger generation.
=521  8\$aU
=532  1\$aNo
=540  \\$aNo
=546  \\$aJapanese
=650  \4$aAgeing
=650  \4$aFamily
=653  \6$aDrama
=700  1\$aChieko Higashiyama
=700  1\$aChishu Ryu
=700  1\$aSetsuko Hara
=701  1\$aYasujirō Ozu
=856  40$uhttp://player.bfi.org.uk/subscription/film/watch-tokyo-story-1953-online
```

MarcEdit transformation – task lists

MarcEdit's Task List functions let users join together lists of tasks in order to always run them in the same order and the same way for every set of records. Instead of having to create the tasks from scratch and type out the replacements and regular expressions to use every time, the program will remember the tasks and the order and run them all at once, and it's also possible to share these task lists (like the templates) between users in order for all colleagues to perform the same tasks in the same way across the team. It's incredibly useful!

For our BFI data we run two task lists. This first set of tasks adds the Library of Congress fields for Short or Feature film to a new 655 field depending on the run time in the 300. It's a bit silly that it has to be a separate task list, but when those tasks were integrated into the second list, the second list started to delete the 856 field from every 13th record, so we've given up for the time being.

This first task list has a simple effect but uses a kind of logical puzzle to get there.

First it copies the text of the 300$a into a new 655 if the number in the 300 is over 40, which is the Library of Congress run time threshold for short films. Then it replaces the text of *any* 655 with Feature Film, because that new field is the only 655 in the record at this point. It then adds a new 655 field for Short films to all of the records, adds the necessary $2 to the Feature films 655, then deletes the Short film field (replaces it with nothing) if a Feature films field is also present in the record. It goes around the houses, but gets there in the end.

```
Tasks:
COPY     300      655      false     (\$a)([4-9][0-9]\d{3,})          $1$2      true
REPLACE          (=655)(.{6})(\d*)    $1  \7$aFeature films.      2              0          False
ADD     655     \7$aShort films.$2lcgft          100
REPLACE          =655 \7$aFeature films.    =655 \7$aFeature films.$2lcgft    0              0          False
REPLACE          =655 \7$aShort films.$2lcgft     0         =655 \7$aFeature films.$2lcgft    1          False
REMOVEBLANKLINES
```

<div align="center">Figure 2. Task List 1, 655 for Short and Feature film</div>

Here's what our record looks like after running that first task list (Figure 2). The run time in the 300 field is 136 minutes, so this record needs a Feature Films field. The rest of the record is the same as before, only a that new 655 has been added.

```
LDR  00000ngm a2200000la 4500
=008  230831s9999||||xx\|||\|\\\\||\\\||und||
=024  7\$a9afb773d-ce2e-59dd-b516-b13c84dada79
=046  1\$k1953
=130  0\$aTōkyō Monogatari
=245  00$aTokyo Story.
=300  \\$a136
=500  \\$aJapan.
=501  \\$a1953
=506  1\$a02/09/2025;@.
=511  1\$aChishu Ryu,Chieko Higashiyama,Setsuko Hara,
=520  \\$aA constant fixture in critics' polls, Yasujirō Ozu's most enduring masterpiece, Tokyo Story, is a
beautifully nuanced exploration of filial duty, expectation and regret. From the simple tale of an elderly
husband and wife's visit to Tokyo to see their grown-up children, Ozu draws a compelling contrast
between the measured dignity of age and the hurried insensitivity of a younger generation.
=521  8\$aU
=532  1\$aNo
=540  \\$aNo
=546  \\$aJapanese
=650  \4$aAgeing
=650  \4$aFamily
=653  \6$aDrama
=655  \7$aFeature films.$2lcgft
=700  1\$aChieko Higashiyama
=700  1\$aChishu Ryu
=700  1\$aSetsuko Hara
=701  1\$aYasujirō Ozu
=856  40$uhttp://player.bfi.org.uk/subscription/film/watch-tokyo-story-1953-online
```

<div align="center">Figure 3. The original record with a newly added 655 field for Feature films.</div>

The second, more robust task list is 179 tasks long and handles all of the rest of the transformation from essentially a spreadsheet in MarcEdit form into proper records. It does a few different types of tasks from really basic ones to more nimble ones that cross-reference multiple fields.

Those basic tasks include standard field additions to every record in the file, which aren't conditional on the existence or content of any other fields, and which don't involve altering the order or the content of the field text. So with these tasks we add things like the 006 and 007, the 336, 337, and 338, 347 for the video file format, a 264 \2 for BFI's distribution, a 506 field to say that access is limited to within the UK, and a 588 field to indicate that we've constructed these records from vendor-supplied metadata.

The task list then does some indelicate, brute-force amendments. Library of Congress wants the Country of Production to be present in a 257 field, and those countries should have standardized names according to the Name Authority File source (some of which are not how BFI provides them), so the task list converts them into the correct format. For example BFI might say the country of origin for a particular film was the USSR, but the NAF name is Soviet Union, so this set of tasks (Figure 4) standardizes those in the 257 field.

```
EDITFIELD257USA0United States
EDITFIELD257USSR0Soviet Union
EDITFIELD257Federal Republic of Germany0Germany (West)
EDITFIELD257Russian Federation0Russia (Federation)
EDITFIELD257People's Republic of China0China
EDITFIELD257Republic of Korea0Korea (South)
EDITFIELD257State of Palestine0Palestine
EDITFIELD257Socialist Republic of Vietnam0Vietnam
EDITFIELD257Burkina Faso (Upper Volta)0Burkina Faso
EDITFIELD257Canada (Quebec)0Canada
```

Figure 4. Sample of the tasks to convert BFI country names to LoC format.

Another brute-force task handles the language code in the 008 (Figure 5). When we first received the bulk of BFI's records in May, we pulled out every language represented in their collection regardless of whether the film was active on the streaming platform, and added these all into the task list, cross-referenced with the MARC language codes.

```
REPLACE||und||||alb||0=546  \\$aAlbanian1False
REPLACE||und||||amh||0=546  \\$aAmharic1False
REPLACE||und||||ara||0=546  \\$aArabic1False
REPLACE||und||||arm||0=546  \\$aArmenian1False
REPLACE||und||||bam||0=546  \\$aBambara1False
REPLACE||und||||ben||0=546  \\$aBengali1False
REPLACE||und||||cat||0=546  \\$aCatalan1False
REPLACE||und||||chi||0=546  \\$aChinese1False
REPLACE||und||||chi||0=546  \\$aMandarin1False
REPLACE||und||||chi||0=546  \\$aCantonese1False
REPLACE||und||||cze||0=546  \\$aCzech1False
REPLACE||und||||dan||0=546  \\$aDanish1False
REPLACE||und||||eng||0=546  \\$aEnglish1False
REPLACE||und||||fin||0=546  \\$aFinnish1False
REPLACE||und||||fre||0=546  \\$aFrench1False
REPLACE||und||||ger||0=546  \\$aGerman1False
REPLACE||und||||gre||0=546  \\$aGreek1False
REPLACE||und||||hau||0=546  \\$aHausa1False
REPLACE||und||||heb||0=546  \\$aHebrew1False
REPLACE||und||||hin||0=546  \\$aHindi1False
REPLACE||und||||hun||0=546  \\$aHungarian1False
REPLACE||und||||ice||0=546  \\$aIcelandic1False
REPLACE||und||||alg||0=546  \\$aInnu1False
REPLACE||und||||sla||0=546  \\$aInterslavic1False
REPLACE||und||||gle||0=546  \\$aIrish1False
REPLACE||und||||ita||0=546  \\$aItalian1False
REPLACE||und||||jpn||0=546  \\$aJapanese1False
```

Figure 5. Sample of the tasks to populate 008 language, conditional

The task list does not have to do anything clever with the 008 positions because the record data is consistent, so the tasks can search for any instance of ||und|| and replace it with the correct language code depending on the text of the 546. This would also be possible to do in Excel before importing the data, but this method limits the number of manual manipulations we have to do with each new spreadsheet.

Another task copies this new 008 language code into a 041 language field (Figure 6). The final language task in the screenshot on the bottom right amends the 546 text to add 'In [language] with English subtitles', since all of the streaming video collection is accessible in English. If the language of the film is English, making the field data 'In English with English subtitles,' the field is deleted.

Figure 6. Task to populate language field with subtitle information.

Here's another look at how the task list manoeuvres data out of and into the 008 field (Figure 7). The first task populates the 008 position 18 from the 300 field, but the 008 will only recognise that value if it's a three-digit runtime. So there are two more tasks to add initial 0s if the runtime is only one or two digits.

```
SUBFIELD_EDIT300a008|18||103|1
EDITFIELD008(\d{2})(\|)20$118:3
EDITFIELD008(\d{1})(\|)(\|)200$118:3
```

Figure 7. Three tasks to add the run time from the 300 to the 008, then add leading 0s to

A final task in Figure 8 reverses the order of the names in all 700s to Last, First, relator term. This is not an exact science, as many Chinese and Korean names are already in the Last First order in the spreadsheet and the 511, but it takes less time to make those incorrect at this stage and then correct them manually when we validate the headings. There is potential here to add another task to only reverse the order of the names if the language codes KOR or CHI are not present in the 008, but that would need extensive testing.

Figure 8. Task to reverse the name order in all 700 fields after the first space.

After running the second task list (Figure 9), you can see our record has effectively doubled as compared to Figure 3. Those accessibility fields are now analysable and also discoverable because they're present correctly in the 341, 347, 532, and also the 655. The relator terms have been added and the 008 has been updated with runtime, year of distribution, year of release, country of distribution, language, form of item, and type of visual material. The 700s have been reversed to Last name First name. And we've also added our local information into the 040 for when this record is shared with other institutions.

```
=LDR  00000ngm a2200000la 4500
=006  m\\\\\o\\c\/\\\\\\
=007  vz#uzazuu
=007  cr#cnunnn|uu|u
=008  230831p20231953enk136\/\\\\\\o\\\v|jpn||
=024  7\$a9afb773d-ce2e-59dd-b516-b13c84dada79$2bfi
=040  \\$aUkLeU$beng$erda$cUkLeU
=041  \\$ajpn$jeng
=046  1\$k1953$2edtf
=130  0\$aTōkyō monogatari (Motion picture)
=245  10$aTokyo Story.
=257  \\$aJapan$2naf
=264  \2$a[England] :$bBFI,$c[2023]
=300  \\$a1 online resource (136 mins.)
=336  \\$atwo-dimensional moving image$btdi$2rdacontent
=337  \\$acomputer$bc$2rdamedia
=338  \\$aonline resource$bcr$2rdacarrier
=347  \\$avideo file$2rdaft
=500  \\$aCountry of origin: Japan.
=500  \\$aOriginally released as a motion picture in 1953.
=506  \\$aBFI allows unrestricted access within the UK.
=506  1\$aAccess ends on 02/09/2025.
=511  1\$aChishu Ryu, Chieko Higashiyama, Setsuko Hara.
=520  \\$aA constant fixture in critics' polls, Yasujirō Ozu's most enduring masterpiece, Tokyo Story, is a beautifully nuanced exploration of filial
duty, expectation and regret. From the simple tale of an elderly husband and wife's visit to Tokyo to see their grown-up children, Ozu draws a
compelling contrast between the measured dignity of age and the hurried insensitivity of a younger generation.
=521  8\$aBBFC certificate: U
=546  \\$aIn Japanese with English subtitles.
=588  \\$aDescription based on vendor-supplied metadata.
=650  \4$aAgeing$vDrama.
=650  \4$aFamily$vDrama.
=653  \6$aDrama
=655  \4$aStreaming video.
=655  \7$aFeature films.$2lcgft
=655  \7$aFiction films.$2lcgft
=700  1\$aHigashiyama, Chieko,$eactor.
=700  1\$aRyu, Chishu,$eactor.
=700  1\$aHara, Setsuko,$eactor.
=700  1\$aOzu, Yasujirō,$edirector.
=856  40$uhttp://player.bfi.org.uk/subscription/film/watch-tokyo-story-1953-online
```

Figure 9. Example record after both task lists have been run.

After this task list process is complete—which takes about ten seconds—we do some spot checking to make sure it all appears as it should, and then use MarcEdit's inbuilt tools to validate the 700 and 130 headings, and correct most of them manually (since the BFI-provided names and original film titles, like the country names, are not in Library of Congress authority format). This manual validation process still takes the longest. We leave the subject headings as we're given them from BFI, but indicate that they're local headings (650 \4, as in Figure 9 above). We then create a much shorter and less detailed Delete file which matches on the BFI unique identifier in the 024 to delete films whose access has expired, and finally we send those two files off to BFI to send to their subscribers.

Development and looking forward

All of that development was a huge amount of work to finish before we started providing the records to other institutions at the end of May 2023. But there have also been a few improvements since then as we continue to develop our relationship with BFI.

First, our major pain point was identifying films whose access had been extended. They wouldn't be obvious in the spreadsheet since we add new films based on the access start date, but sometimes older films who had already expired would be extended without the start date being edited. In response to this need, BFI have added an additional column to indicate which access dates have been edited in the past month, and this has made the process of identifying those extensions much more efficient.

Another enhancement we've made is to start adding Library of Congress URIs to our validated 700 entries, thanks to MarcEdit again - there's an option to do this when validating the headings.

At the moment, we receive this spreadsheet of titles in the last week of the month, and have a few working days to turn around the files before they are then emailed out to other institutions. Many suppliers host files like these for institutions to download rather than emailing them out, so there is potential there for streamlining the supplying process. We're also looking for more clarity on how many other subscribing institutions use Ex Libris Alma for their LMS like we do, because we have an opportunity to use Alma's Community Zone to share records, but that is only helpful with a critical mass of Alma users.

And finally BFI offer a package of freely available material in addition to the subscription films, and some subscribing institutions have expressed interest in getting records for those films as well. We're currently in talks with BFI to understand the turnover and demand for those Free titles and may be able to offer this in future.

Impact

The biggest impact on our team locally is relief on our staffing resource. A process that used to take at least a day, and potentially up to four days with a large import, is now maximum a few hours, including the time it takes to validate the Library of Congress authority headings. It's also been an incredibly useful exercise for us, both in terms of a huge stretch project for our team's understanding of MarcEdit and regular expressions, and also in developing our knowledge of streaming video cataloguing standards. We've been able to offer value to our fellow UK HE institutions - at no price increase - who no longer have to create their own records in the painstaking way that we were. Plus, of course, the discount for our own subscription is a bonus.

Most importantly it's had a positive impact for students and other catalogue users, now that the records are provided faster and to a higher standard - especially for things like the accessibility fields, which are much more user-friendly now.

From the BFI side it's also been a useful partnership. We've gotten some really encouraging feedback from them, including from Simone Pyne who is BFI's Senior Business Development Manager: "Our partnership with the University of Leeds has helped us to deliver a much-requested resource by our BFI player subscribing institutions. I have come to learn how crucial MARC records are in aiding discoverability, which is of the utmost importance to us, as our aim is for students and staff to use their BFI player subscriptions to engage with the cultural value of film and support their studies. We didn't have the expertise to create these records in-house, and the insight of the UoL team has been beyond valuable."

We're really pleased to be ambassadors for library metadata standards and for MARC records, and of course also thrilled to have the feedback that the relationship is mutually beneficial.

This project has been useful for us in the Metadata team, but it's also expanded agency within our supplier-vendor relationships for the wider Collections team, and has provided value to our vendor BFI. We're looking forward to future potential applications of similar task lists and future development of these records, whilst celebrating the work that has got us to this point.