ENUG 2021 Day 2: Process Authority: Expediting the Authority Control Task List (Colin Bitter and Yuji Tosaka)

Wednesday, 10/27 4:00 PM

Transcript generated by Zoom Live Transcription

00:00:00.000 --> 00:00:00.000

And now I'd like to introduce column better, and ug to soccer of the College of New Jersey and their presentation process authority expediting the authority control task list, you folks all ready to go.

00:00:00.000 --> 00:00:00.000

Ready to go Thank you can

00:00:00.000 --> 00:00:00.000

start the recording one second.

00:00:00.000 --> 00:00:05.000

Can you see my screen.

00:00:05.000 --> 00:00:08.000

Okay.

00:00:08.000 --> 00:00:16.000

Thank you for attending our presentation today we'll be talking about how you can use Python to assist in your authority working Alma.

00:00:16.000 --> 00:00:25.000

Note that all code can be found in GitHub at GitHub. com slash calendar.

00:00:25.000 --> 00:00:26.000

Starting with introductions.

00:00:26.000 --> 00:00:31.000

My name is Colin Baker and I'm the head of cattle you meditated at the college New Jersey.

00:00:31.000 --> 00:00:48.000

Today I am joined by my colleague, up to Saka cataloguing metadata. Metadata librarian, also from TCNJUG we'll start by talking about authority control at our library, and then I will show how we use Python to process the ACL.

00:00:48.000 --> 00:00:51.000

Take it away.

00:00:51.000 --> 00:01:09.000

Okay, thank you for the introduction. First let us start briefly with some context about the New Jersey and the current authority control practices in general at our library, which was recently named didn't stay library after the previous president was

00:01:09.000 --> 00:01:16.000

served nearly 20 years before retiring, about a few years ago.

00:01:16.000 --> 00:01:34.000

Well, the cause of New Jersey identified us that de Sanjay, in a nutshell shorthand is a mid sized for your public college in us, which is in Central New Jersey a short distance away from the state capital Trenton, and also from Princeton, because he

00:01:34.000 --> 00:01:55.000

is one of the top ranked public colleges in North East, with a focus. We have approximately 7000 undergraduate students in a wide range of disciplines, and also offer a number of small masters and post a call it programs such as business counseling education

00:01:55.000 --> 00:02:14.000

counseling education and nursing public house and get us in library the typical academic library for mid sized institution like peace nj library is the only library on campus, serving the Christ community and holds about half a million titles and its

00:02:14.000 --> 00:02:16.000

physical collections.

00:02:16.000 --> 00:02:28.000

The library also definitely manages over 350,000 electronic titles spread across various joining collections.

00:02:28.000 --> 00:02:47.000

And not surprisingly, the vast majority of Mark records for both the physical and electronic collections that pedalo in one of your department cataloging and meta data services consists of two full time calibers back to our department at college bitter

00:02:47.000 --> 00:02:56.000

and kindly met at a library of me class one part time special formats catalog as opposed to the library.

00:02:56.000 --> 00:03:03.000

And we also have before our professional staff members, the department.

00:03:03.000 --> 00:03:18.000

Well the delivery migrated from Voyager to our men Primo be in back in the summer of 2019 as part of the moose conference during configuration on five state colleges and universities in New Jersey.

00:03:18.000 --> 00:03:37.000

Well, as a side note, this migration to the cloud based system as today's presenters at some of the presenters mentioned proved to be a very timely. One, in retrospect, allowing, which allowed us to work mostly with relative ease and seamlessness during

00:03:37.000 --> 00:03:56.000

the call with 19 pandemic us for authority control practices in general, TCNG has been a nickel library for the last about 15 years or so as part of that we are part of, we have been part of the NATO New Jersey, New York final.

00:03:56.000 --> 00:04:18.000

And we also have very clear reactive NATO program. Nick, a new nickel record contribution was ranked about approximately at the study as in the latest FY 21 and unique was this this test that just has come out today or yesterday.

00:04:18.000 --> 00:04:38.000

As a matter of fact, And for Neko work we have all three professional colorist creating name author of the records for the LC may cause with the file, a set of it control crostini out nice also has been also one of the key priority areas for the next

00:04:38.000 --> 00:05:06.000

slide please. Okay, so what are regular Jane already posting workflows in now at this time library, a local practice is to control for only for Lz authorities, excluding LCMPT, and we only review for and suppressed institution records in our schema.

00:05:06.000 --> 00:05:27.000

And the only two full time covers department head and colonial metadata library work on also a different role plus proceeding. Our well the the part time, special for my scholars also do the embed code works as well as specifically the department head

00:05:27.000 --> 00:05:32.000

calling extract also a differential task with reports every Monday.

00:05:32.000 --> 00:05:51.000

Every Monday morning hours they saved into reality available in one of the department's Google Drive folders, as the case uses Google Apps for Education, and generally the two of us as signed a CTO reports for alternate weeks for checking and pause.

00:05:51.000 --> 00:06:12.000

All series access points on a CTO reports need, meeting new series of directors are handed over to me typing in meta data librarian for review as I am, be the only one I have additional training to create series also at the records in addition to regular

00:06:12.000 --> 00:06:16.000

person their corporate names for your skin AF.

00:06:16.000 --> 00:06:34.000

Well, as far as series are concerned, we also want to note here that we can we do continue to a series by default, by default, which is the PCC practice so we have wheels have been vacation will set up in Alabama to feel that any new vehicle of records

00:06:34.000 --> 00:06:47.000

and traced for 93 years for reviews on a regular basis, although will not be covering that in this presentation. So next slide please.

00:06:47.000 --> 00:06:56.000

Okay, so this is just a partial screenshot of one of the ACTA reports from last year.

00:06:56.000 --> 00:07:00.000

Initially, these spreadsheets used to be saved.

00:07:00.000 --> 00:07:18.000

Used to be saved separately for a while in multiple weekly files by category like pending found no matching before heading found multiple matching and so on and so forth so we had the multiple spreadsheet, Jamie rate for review.

00:07:18.000 --> 00:07:32.000

Each week, but then calling us had since then switch to say being there to gather all together in Kumbaya and save them together in a single, single spreadsheet.

00:07:32.000 --> 00:07:48.000

Each week since he has found that may, that makes posting a CPU reports from our lives easier in subsequent steps which is going to be explaining going over again, is part.

00:07:48.000 --> 00:07:51.000

So as mentioned in the previous slide.

00:07:51.000 --> 00:08:11.000

These ACTA reports assigned to them calendars for alternate weeks for checking with creating contribute new name also direct was in OCLC as needed for a day for work being us being an active member library and we also revise headings in our must need

00:08:11.000 --> 00:08:29.000

it. record by record when only a handful directors need to be revised but also using a normalization rules as needed for making that we have to make any changes in a larger set of records needs to be changed at once.

00:08:29.000 --> 00:08:43.000

And so these anyhow these basic also in the control workflow in Alma should be fairly obvious to any user so will not be getting into it. In this presentation too much.

00:08:43.000 --> 00:08:48.000

So with this brief kind of overview, with these brief remarks.

00:08:48.000 --> 00:09:09.000

Let me turn things over to client to discuss how python programming as be used by him to enable additional processing of these regular ACTA reports against the LC link data service to enhance our internal automated workflow workflow at TZJXUZ.

00:09:09.000 --> 00:09:13.000

Moving on to AC to enhance stop pie.

00:09:13.000 --> 00:09:16.000

So, why should we use Python at all.

00:09:16.000 --> 00:09:30.000

For starters, our library only treats LCGFTLC names and LCS hy vocabularies can be manually eliminated from the ACL TL with an Alma, it is quite time consuming.

00:09:30.000 --> 00:09:46.000

Also, there are timing issues between the LC authority file, and the CZ. For example, I wants to update it in any era and OCLC on Monday, July, 12, but this wasn't updated in Alma until Thursday, July 20 seconds.

00:09:46.000 --> 00:09:59.000

As you mentioned, mentioned previously, we extract the AC TL from all my each Monday. So this heading would have unnecessarily made it to the AC TL on our July, 19 Monday report.

00:09:59.000 --> 00:10:05.000

Additionally, preferred term correction will frequently fix lingering headings on the ACL.

00:10:05.000 --> 00:10:07.000

Some other local issues.

00:10:07.000 --> 00:10:11.000

We do not regularly modify or create name titles.

00:10:11.000 --> 00:10:20.000

We also have several areas in the institutional zone which do not receive authority control, for example, minimal records lacking OCLC numbers.

00:10:20.000 --> 00:10:33.000

Further records are imported from goby for on order items each morning and Alma, which are to be catalogues once the materials are received headings for these records don't need review until they are fully catalogs.

00:10:33.000 --> 00:10:45.000

Lastly, the information in the AC TL is not comprehensive, We need OCLC number location and record creation date to display on our reports.

00:10:45.000 --> 00:10:48.000

Our Python script requires three inputs.

00:10:48.000 --> 00:10:55.000

The ACTL in the form of a spreadsheet manually exported from Alma, and to analyses from analytics.

00:10:55.000 --> 00:11:00.000

One analysis contains all fiscal records which recreated or modified in the last month.

00:11:00.000 --> 00:11:05.000

And the other contains electronic records following similar behavior.

00:11:05.000 --> 00:11:12.000

Analytics information can be exploited manually or pulled down using API's.

00:11:12.000 --> 00:11:15.000

Starting with the AZTL.

00:11:15.000 --> 00:11:20.000

If you work with the AC TL you're probably used to seeing the screen.

00:11:20.000 --> 00:11:34.000

At TCNJ, all we do before exporting from Alma is eliminate suppressed records, and records linked to the CZ. These records do not receive authority control at our institution.

00:11:34.000 --> 00:11:50.000

And here is our export spreadsheet with columns, columns report type change date m&s ID, and others know that some institutions will not have been heading before and before heading after, depending on how you have your authority rule set up.

00:11:50.000 --> 00:11:58.000

You might only see a single column for beheading.

00:11:58.000 --> 00:12:02.000

Moving over to analytics, we have two simple analyses.

00:12:02.000 --> 00:12:13.000

One to retrieve records with electronic portfolios and the other for physical inventory, here's the example for electronic records, AC t le.

00:12:13.000 --> 00:12:26.000

This is limited to any record with an electronic portfolio that has been created or modified within the last month.

00:12:26.000 --> 00:12:31.000

And here's our output from analytics named ACTLEXLSX.

00:12:31.000 --> 00:12:49.000

We have columns MMS ID OCLC number are cataloguing statistic field 978 creation date, and electronic collection public name.

00:12:49.000 --> 00:12:51.000

Now on to the actual scripts.

00:12:51.000 --> 00:12:57.000

I won't take you through the entire code in the slides, but I will show several selections.

00:12:57.000 --> 00:13:12.000

Just a reminder, if you want to use this code copy from GitHub, and not the slides for application we use several libraries in Python requests to make calls to id.llc.gov.

00:13:12.000 --> 00:13:22.000

And as a non pie for data manipulation within data frames Catholic for file maintenance warnings to deal with Excel warnings.

00:13:22.000 --> 00:13:26.000

And finally daytime to name our results and file.

00:13:26.000 --> 00:13:35.000

I also use PI installer to compile this script into an executable executable file for ease of use.

00:13:35.000 --> 00:13:44.000

First we import our three spreadsheets into three separate data frames ACTA CTO for the authority control task list.

00:13:44.000 --> 00:13:50.000

And he for the electronic records analysis and Anna p for the physical records.

00:13:50.000 --> 00:13:58.000

One simple beginning part of the script is the string.com contains commands on the lower portion of the slide.

00:13:58.000 --> 00:14:02.000

Here we are limiting to the LC vocabularies we care about.

00:14:02.000 --> 00:14:13.000

This will will make things like LCMPTLCSH Kids GND fast and others,

00:14:13.000 --> 00:14:18.000

the headings from the ACTL, the treatment, prior to manipulation Python.

00:14:18.000 --> 00:14:22.000

We need to strip trailing commas, using our strip.

00:14:22.000 --> 00:14:30.000

We also need to strip trailing periods, unless they follow an initial since we are going to be querying id.llc.gov.

00:14:30.000 --> 00:14:35.000

Most authorities will not end with a period, unless it is following an initial.

00:14:35.000 --> 00:14:42.000

In the case of James Andrew 1858 to 1930 day period is removed.

00:14:42.000 --> 00:14:52.000

But in the case of James Andrew be. It has retained our expression highlighted on the slide uses a negative look behind.

00:14:52.000 --> 00:14:57.000

If it finds a space, followed by an uppercase letter than the period is retained.

00:14:57.000 --> 00:15:04.000

All other periods will be removed.

00:15:04.000 --> 00:15:09.000

Trailing semi colons also need to be removed from beheading using our strip.

00:15:09.000 --> 00:15:20.000

Lastly, series headings and xx fields are frequently deposited in the ATL with a subfield x eight digit isn these needs to be removed as well.

00:15:20.000 --> 00:15:32.000

Our expression looks for space semi colon space four digits hyphen four digits, and E string matching that pattern will be removed.

00:15:32.000 --> 00:15:38.000

In order to effectively address the timing problem between the LC authority file and the authorities in the CZ.

00:15:38.000 --> 00:15:45.000

We want to compare each PIP heading from the AC TL against id.llc.gov.

00:15:45.000 --> 00:15:56.000

Here we combine a URL with the heading from the ACL at the bottom of the slide, you can see the example URL which will be used to query id.llc.gov.

00:15:56.000 --> 00:16:13.000

This is a combination of it that LCR gov slash authority slash name slash label and Bruckner Anton 1824 to 1896.

00:16:13.000 --> 00:16:23.000

In this section of code, we query ideas, llc.gov by using some lambda functions we retrieve two pieces of information from LC.

00:16:23.000 --> 00:16:29.000

The status code of the page, and the actual HTML content.

00:16:29.000 --> 00:16:37.000

The status code can be any number of values. For example, 200 for successful call, or 444 and error.

00:16:37.000 --> 00:16:44.000

For the HTML content we decode the page using UTM eight.

00:16:44.000 --> 00:16:50.000

Here's what the UTM a decoded HTML looks like coming from back from LC.

00:16:50.000 --> 00:17:04.000

We want to extract the authorized heading source from Mark authority field 100 highlighted on the slide.

00:17:04.000 --> 00:17:13.000

We can use expressions to get the authorized heading here our expression reads that following the HTML tag title.

00:17:13.000 --> 00:17:18.000

We want to extract everything leading up to space hyphen space, LLC.

00:17:18.000 --> 00:17:29.000

Here we get the preferred name before, Amazon Bruckner.

00:17:29.000 --> 00:17:33.000

This might be a good place to stop and take stock of our data frames.

00:17:33.000 --> 00:17:41.000

We have a CCL consisting of our original data from the task list and return data from it.

00:17:41.000 --> 00:17:52.000

Then we still have our two analyses, and he and Ana p for electronic and physical records respectively.

00:17:52.000 --> 00:17:58.000

Now, we want to combine all three of the data frames.

00:17:58.000 --> 00:18:12.000

First, we can find the two analytics data frames merging outer on MMS ID OCLC number, local program oh one, that was our 978 fields cataloging statistic and creation date.

00:18:12.000 --> 00:18:23.000

Further down in the script we will join the combined analytics data frames with the, with the ACL.

00:18:23.000 --> 00:18:28.000

Now we have our results and data frame, and can work on local customizations.

00:18:28.000 --> 00:18:32.000

First, we want to get rid of anything that hasn't been catalogs.

00:18:32.000 --> 00:18:41.000

Here we will drop records born in 2021, missing the 978 as this indicates it hasn't yet been catalogs.

00:18:41.000 --> 00:18:48.000

We limit to 2021 only since we have all the records which never received a 978, but we're in fact catalogs.

00:18:48.000 --> 00:18:58.000

We can remove these uncatalogued resources using a simple drop function.

00:18:58.000 --> 00:19:07.000

At TCNJ, we do not maintain local authorities, any type of authority work is done within the LLC authority file.

00:19:07.000 --> 00:19:15.000

However, we do have dozens of local headings which frequently make it to the AC TL since they are unauthorized.

00:19:15.000 --> 00:19:20.000

For example, we record faculty author collection in 710.

00:19:20.000 --> 00:19:27.000

Then we also use undocumented immigrants, instead of the LCS ah illegal aliens, and 650.

00:19:27.000 --> 00:19:30.000

We maintain a list of these in a spreadsheet.

00:19:30.000 --> 00:19:42.000

All of these headings are eliminated from the report using another drop function.

00:19:42.000 --> 00:19:49.000

As mentioned previously, there are several types of records which will never receive authority control at TCNJ.

00:19:49.000 --> 00:20:06.000

Two examples are Naxos minimal Records, which lack OCLC numbers, and to our entire ICP SR collection Naxos can be eliminated using a drop function for ICP SR we can easily limit the data frame to those collections locations which do not contain string

00:20:06.000 --> 00:20:14.000

ICP Sr.

00:20:14.000 --> 00:20:23.000

Finally, we want to compare those returned values from it that llc.gov against the headings in the ACL.

00:20:23.000 --> 00:20:33.000

If any value is returned for an xo, then it will be eliminated from the report, since these will be fixed via preferred term correction.

00:20:33.000 --> 00:20:38.000

All other fields require a one to one match with LC.

00:20:38.000 --> 00:20:58.000

This is because some headings might have a tag number change, in which case PTC is useless, at least that President, for example, x 102 x three or four treaties, or x 502 xO for fictitious entities.

00:20:58.000 --> 00:21:14.000

One strange thing I had to do in order to get the encoding schemes to match up, was to encode both the AC TL big headings, and the id.llc.gov values in ASCII, and then decode the UTA, as you know, the top of the slide.

00:21:14.000 --> 00:21:18.000

Once everything is decoded in UCF a, we can make our matches.

00:21:18.000 --> 00:21:30.000

Let's look at each line of code here in line one, we create a new column in the data frame Ddf to called match.

00:21:30.000 --> 00:21:50.000

This will determine if the beheading equals the return value from LC in wine to the data frame df two is limited, limited to only non matches that is any one to one match between RACTL and LC will be eliminated from the report in lines three and four,

00:21:50.000 --> 00:21:58.000

we eliminate RXO fields, 100 607 hundred, which had any value returned from LC.

00:21:58.000 --> 00:22:04.000

We do this, do this using the status code from earlier in the request portion of our scripts.

00:22:04.000 --> 00:22:04.000

200 singles that something was retrieved from LC.

00:22:04.000 --> 00:22:25.000

200 singles that something was retrieved from LC. This would account for any Alma headings which contained values from a 400 fields in LC authority records, theoretically be should be corrected via preferred term correction.

00:22:25.000 --> 00:22:35.000

Lastly, after we have carried out all local customizations, the final data frame will be exported as a spreadsheet, named for today's date.

00:22:35.000 --> 00:22:42.000

This is then reconciled by catalog and medicine Data Services Department members.

00:22:42.000 --> 00:22:56.000

And here's our final product, which includes information from the ACL analytics and the Library of Congress.

00:22:56.000 --> 00:23:04.000

And let's look at a quick example and then we'll go to questions.

00:23:04.000 --> 00:23:15.000

So I could sit here and we could watch the script run but I know that can be very exciting for anyone for everyone. So, just to show you an example. This is the authority list before.

00:23:15.000 --> 00:23:25.000

So if you work with the ACTA all you're probably pretty used to seeing this. This is a pretty typical weekly report happens to be maybe a little bit larger than some of them.

00:23:25.000 --> 00:23:30.000

This has 7235 headings.

00:23:30.000 --> 00:23:37.000

So you can see we've got all sorts of stuff here LC names fast, LCMPT, it's kind of all over the place.

00:23:37.000 --> 00:23:40.000

And then after running the scripts which would take a few minutes.

00:23:40.000 --> 00:23:43.000

This would be the results and output.

00:23:43.000 --> 00:23:52.000

So we're getting added benefit of like OCLC number catalog statistics of creation date LC returned values, if they're there.

00:23:52.000 --> 00:23:57.000

And of course collection or location.

00:23:57.000 --> 00:24:03.000

And that's it, thanks so much we can go to questions if there are any.

00:24:03.000 --> 00:24:14.000

Well we haven't had any in the chat at this time but hopefully we'll have some from the floor.

00:24:14.000 --> 00:24:17.000

Here we go. Here's one.

00:24:17.000 --> 00:24:27.000

I'm assuming that you know Python pretty well to make this but how much of this Did you find from other places and how much did you do yourself. Also how much local customization went into this.

00:24:27.000 --> 00:24:31.000

That's from Dominique thank somebody.

00:24:31.000 --> 00:24:45.000

Yeah, so the entire script I wrote myself, you know, using a lot of documentation online, especially for, you know, all of the individual libraries that I mentioned in the beginning of the presentation.

00:24:45.000 --> 00:24:52.000

So, each one of those libraries is going to have documentation that you can go and use.

00:24:52.000 --> 00:25:07.000

Beyond that I use a lot of Stack Exchange, as well. So when I ran into problems where documentation didn't answer my question you know Stack Exchange is a great kind of second line of defense.

00:25:07.000 --> 00:25:09.000

Local customization.

00:25:09.000 --> 00:25:22.000

The script is very much tailored to our particular collection so if someone was going to take this down and implemented in for their local environment they would have to do a lot of changes for example all of our local headings they would want to change

00:25:22.000 --> 00:25:23.000

those.

00:25:23.000 --> 00:25:31.000

And of course all of our things that we like to, to focus on, you'd want to change for your particular institution.

00:25:31.000 --> 00:25:38.000

We have a comment from Mary Beth that records can't wait to try this. You may be reaching out to you.

00:25:38.000 --> 00:25:42.000

Sounds good. I'm more than happy to share America.

00:25:42.000 --> 00:25:56.000

And any more questions from the floor. Feel free if you have a question you can have you unmute and just ask.

00:25:56.000 --> 00:26:11.000

No. Okay, I'm not seeing any more questions not hearing anymore. Well then.

00:26:11.000 --> 00:26:15.000

own institutions. Our pleasure thank you karen.

00:26:15.000 --> 00:26:29.000

Okay, thank you, everyone. I'm going to go ahead and end the recording and thank you all for attending there but when will the recordings be available, they will be posted all of them will be posted after the end of the conference so Anthony is in charge

00:26:29.000 --> 00:26:33.000

of that but I think that's probably going to be next week.

00:26:33.000 --> 00:26:49.000

Yeah, it takes a few days, hours for zoom to process the video and make it available and that they'll take a little bit of time for us to get them already, and post them on the web, will try to notify everybody.

00:26:49.000 --> 00:26:50.000

Once they're ready.

00:26:50.000 --> 00:27:02.000

We also like to close them with presentations were asking the presenters to share their slides or PDFs of their slides so that usually takes a little bit as well.

00:27:02.000 --> 00:27:03.000

Right.

00:27:03.000 --> 00:27:23.000

Right. I knew it wouldn't be instant.