Evergreen Conference

Implications of Earwyrms for Evergreen Metadata

May 27, 2021

>> KATIE GREENLEAF MARTIN: I went to thank everyone for hanging in with us here it is the last regular session before the close of the Evergreen conference. I want to thank the Evergreen community development initiative and Mobius for being our two wonderful sponsors but I hope everyone that attends to visit the sponsor in the Expo area. I am excited that we are, we have Galen Charlton here for Implications of Earwyrms for Evergreen Metadata. I will drop the link for the captains in the chat and if there are other questions we can put those in the chat as well as ensure Galen will pause for questions. All right.

>> GALEN CHARLTON: Okay. So good afternoon. Good morning and possibly good night to everybody depending on time zones. My name is Galen Charlton the implementation and IT manager at the Equinox Open Library Initiative. I am also an Evergreen developer and committer with a fascination with many things. But including library metadata. So we will start this presentation with a discussion of a completely different area of study musicology.

Most of us are familiar with the concept of earworms I don't think I need to define it but of course it is something that's probably best demonstrated.

I have put up a quote from a piece by Mark Twain. Who may have been the first person to suggest that earworms can be weapon eyes. And if you were to search any library catalog for A Literary Nightmare and search on a topic you will come to a realization. And that is unfortunately earworms are not adequately described by library category standards. The library Congress subject heading for your worms I am afraid you cannot have one. In fact, if you do work around medical subject headings there's not even so much as a heading for involuntary musical imagery.

I must say this is quite disappointing. But of course like I said, sometimes it is best to demonstrate this by example. Or example. Or example. Or example. But fortunately you all can be grateful that I am far too shy saying that to you so I will not be infecting you with any earworms today. And that is my promise.

However, there are in fact things that are worse than earworms. And those are library catalog earwyrms. Is a catalog entry that has been, shall we say, sadly mistyped. This will be the focus of my presentation is the earwyrm, bear the flame of failing to find stuff this will actually relate libraries healthy Evergreen theme of the conference but I will get there in a bit.

By the way I am keeping an eye on the comments and this is meant to be more of a discussion than a lecture from on high. So please feel free to ask questions. And I will get to them when I see them and where it makes sense with the flow. There will in fact be 10 points to the first person who spots it.

When we are talking about metadata quality issues, with a particular focus on actual errors as distinguished from issues of incompleteness, there are four factors to think about when dealing with them.

But I will step back a bit and speak on the distinction that I'm making between an error and an omission.

A cataloging record that has not as many subject headings as you might want it to -- that may not be ideal but it is not necessarily an error. Of course it depends on the level of the cataloging that you need to do or that you can afford to do for your particular circumstances. But if we focus today on your worms a.k.a. errors some things that we need to know is how these your worms get classified and when seen other species actually finding them and counting them. Once you have found them the point of the exercise is not to stare at these errors if federal possible. But to fix them.

And going beyond fixing them where they occur to working to prevent them from multiplying. So if we look at the first question how to classify these your worms here's the beginnings of a list of some issues you can run across. Of course typos and misspellings affect us all. There's a matter of outdated headings for example with the change of illegal aliens to a better heading and perhaps one of the more recent and topical examples and congratulations to Debbie and Sarah for catching the Rick roll first. As far as outdated coding is concerned we have the issue that entropy affects us all but especially MARC records.

Another source of errors or class of errors are simply inconsistencies and disagreements. This can be particularly prevalent in the case of describing, so what is the format of the title anyway? Obviously the physical or electronic format of resource can matter a lot to patrons but if you look at a MARC bib record and the items and volumes that are attached to it, you do not have to work too hard to come up with easily a half-dozen ways of expressing the format of an item incorrectly. That can include this can include the fixed field record and include the new RDA fields. But of course we shouldn't forget about the 300 field is any statement about the physical format doesn't need to necessarily correspond with the fixed field or the 33x's and then the situation gets worse when you look at the item -- you can have code number prefixes and suffixes which makes claims about physical format and you can have shopping locations in certain amount of errors that also make claims about the physical format. So with six different ways to make statements about the physical format that gives you six different ways to be wrong so another example of a type of your worm is bad links, this also ties into the problem with entropy because perfectly good thank in 2005 be a very bad Lincoln 2021 and in fact the probability is that a 16 year old link probably is bad.

Duplication of course is another area of concern especially in consortial catalogs. Incompleteness in some cases can be an error if one part of a record expects a field that is not actually present or if you have an absence of a fixed field. So thank you Janet and Elaine for mentioning examples of how the 07 has changed coding over the years so Laserdiscs are indeed different things. By the way is a channel 1 called technology connections which I recommend if you're interested in the history of the videodiscs. It's quite interesting.

And then there's the most difficult category of your worms to consider. And I confess to making a mistake on this slide. I put it down as a failure of cataloging records to conform to the real world. But of course, that is backwards. Your real problem is the failure of the real world conform to the metadata.

The last question of your worm is one that for the most part I am going to blithely ignore. Because ultimately, if you need to check against the real world you of course are best advised to have the item in hand.

Almost programmer and not a cataloger so I'm not dealing with items in hand so the focus of this presentation will be on things that we can deal with via a combination of human effort and software. So given incomplete categorization of catalog errors -- by the way, I am sure there is some discussion in the library literature about this and to my knowledge there is not a controlled level category for catalog, cataloging errors.

This is an opportunity for us to create one. But once you have the notion of what sorts of issues you are looking for then comes the question of okay the how do you find these errors? And then of course there's also a level of, do you care? In any given case. So one of the possible ways of doing it is simply using the catalog. So serendipity and libraries is nothing to do with users finding things they need but didn't know that they wanted. It is really about promoting systems to find mistakes. Here's an example of doing Evergreen search for your worms because this is a small database this feature in Evergreen 37 -- it is sadly thinking that the word your worms with an out is misspelled. But this is feedback whether coming from tech services, colleagues and library departments or patrons is of course an excellent way to find your worms either blatant such as this one or cases of results showing up either unexpectedly or not showing up unexpectedly so course one thing about serendipitous your worm it's not exactly efficient probably at there's a trope that all patrons believe that librarians do nothing read books all day. You could imagine a trope were catalogers do nothing but search the catalog all day and that's also wrong. If course that's where things like research logs can come into play if you preferably after anonymizing keeping an eye on the searches that patrons are actually doing, particularly the ones that result in zero hits, that cannot be more systematically using the catalog to locate your worms.

It is a little more efficient because if you are basing it on actual searches, conducted by real-life patrons, that's a signal that the searches most likely matter to the patrons and if metadata errors are getting in the way of the results they should give it a higher priority for fixing.

Another way of finding your worms is to do it in the MARC record editor. Depending on your workflow obviously you have opportunities to do a view particularly in sporting records from OCLC or z3950 sources on one on one basis. You may remember from a catalog working group meeting a couple months ago, I shut off an example of using the PERL tool called MARC::Lint to critique MARC records. This is an example of what it can look like in practice. This is something that I intend to finish and get into Evergreen 3.8 so that it's a starting point for put record validation. Beyond authority headings in Evergreen. So this gives you potentially something to work from. You could look at a record in the links button and fix things or disagree with it and it is important to acknowledge two things.

All library catalogs are local. Secondly, library catalogs are global.

The locality means that for your particular patient population some of the cataloging roles they no matter may be, need to be abandoned or banned it all catalogs are global or at least most are global -- village graphic databases as well as the consideration but even if you do not have the Z 39 50 terminal for your Evergreen system -- any library in the world can always copy a catalog from you.

So MARC::Lint is one way of dealing with finding errors, it is a tool that is coming. Of course there is also the traditional authority validation feature in Evergreen were of course authority records can reduce your worms. Typo in the title might not matter too much a typo in an author's name could at minimum be kind of a personal insult to that author. So authority records of course can help avoid your worms and outsource the blame for typos to the library of Congress. So the importance of that perhaps cannot be overstated. What I showed you in scratching the surface for the retail record by record your worm discovery but of course all of us are dealing with catalogs that have more than 10 titles in them.

So ultimately there will be I think much more of a need for wholesale earworm hunting and to give a preview of a later section of my talk, I think this is an area where Evergreen could benefit from additional enhancements. But of course if we are looking at wholesale your worms a key factor or tool will of course be your reporting engine. What I have put here is an example of an SQL report. And this actually should work more or less as in an Evergreen database. So this is saying, let's look at the search format is extracted from the MARC record and compare it to the shopping location, what the shopping location says.

If you look at this example, you come to the inevitable conclusion that in one sense the concerto data set used for Evergreen testing is absolutely terrible in part is intentional but not necessarily the greatest of things. Look at the set of books and you see book records being in locations or adult VHS, are you looking at zebras or horses? Zebra would be an example of multiformat records. So maybe these are legitimately books that happen to have DVDs. But of course, the more likely guess is that these are actually horses in terms of okay groups -- those six, different ways of expressing physical format have struck.

This is something that if you have the resources and staffing to do systematic record quality improvement in your Evergreen database this report of looking for inconsistencies is definitely something I'd recommend.

This can be done at several levels. For example, we could be looking at combinations of location and Circ modifier.

In general, doing wholesale earworm hunting via reports is a way to express some creativity.

Again, harkening to future portion of this presentation, this is a question and challenge that I would like to throw out to the Evergreen community. Sending the cataloging working group -- all the examples of reports that you've created that are in the area of data quality -- I would be really interested to see what people are up to.

Thank you, Katie, that is an excellent example of somebody doing this report already.

For other cases of earworm discovery, Evergreen already provides building tools. So for example, the link checker is something that has been available at Evergreen since one of the early releases. This is a perfect example of something that is good to run periodically.

Of course, to update URLs where possible. Now if you are taking a really close look at the slides, at this slide, you may have a pretty good idea of a bug that I have been working on them recently. That is, one of the ones around dealing with people trying to slip JavaScript into MARC records.

Of course, another thing about wholesale earworm discovery is that sometimes you can take your retail discovery and extend it. So what you are looking at on this slide is it possible to take the MARC::Lint detection and enter it into something you could use in your database to make a query? So in other words, this is a way of running lint against your entire database and deciding what do you actually care about or prioritize?

So, this is also something that will probably be more of something that might show up in the Equinox Recreation Tools but this is a resource available to you to find these things in batch.

So, if you have located your worms, what are various ways of dealing with them? Of course, in the simple case of -- just a typo and something you've catalogs, the remediation is to just go in and fix the record already.

But of course, much of the time you will have a herd or flock or whatever collective noun you choose of earworms. With respect to batch operations, there's a number of approaches. One would be simply saying, okay, let's say your workflow is based on Library of Congress or OCLC, you could potentially address issues by periodically refreshing batches of records from them.

Thank you, Carissa, the official collective term for a group of earwyrms is officially an "ugh" of earwyrms --excellent. Another remediation is to fix the records and re-import them. This would be a good way to make use of tools such as MARC edit. Evergreen does have batch editing the functionality but it is of course limited as compared to what MARC edit can do and other command line and server tools can do. so this is something that you do not have to be stuck within the Evergreen software ecosystem. You can make use of existing tools. Outside of Evergreen.

Of course, Evergreen does have a MARC batch edit feature that can do some things. And there are various command line and server tools -- and I am seeing in the chat but it sounds like -- Blake and Llewellyn have been working on tools presumably on the server side to help do cataloging reports and cataloging error reports and do cleanup. And also something that is present in the Equinox migration tools in cases or classes of earwyrms.

The MARC batch edit tool is an example of something that can do certain limited operations. And here I will place my feet into the hands of the computer. And actually do a little bit of a live demonstration.

Let's bump up the font size. So, one of the things about the Concerto data set, it has a lot of records coming directly from the library of Congress, which has field for workflow that do not matter a bit for anybody else, so the 906 field is an example of those.

We can take a search for records that have 906 fields. Select all of them. Let's go at the basket to a bucket.

Once we have created this bucket, I will double check that it is in good shape. It is. And from here we can go ahead and do a patch update. In this case to say get rid of the 906 field. And for anybody who is looking for an excuse to write a launchpad ticket based on my presentation, the label should probably be changed, this has been used as part of the MARC batch edit interface.

Merge rule doesn't really make sense as it is here.

In any event, this gives us the ability to watch progress bar moved from right to left and then we do the catalog search. And in principle, we can see that we've deleted this offending field. So it is important to emphasize that there is a lot more I wish that Evergreen MARC batch edit would do. But having field that should not be there or during certain kinds of string changes, this can be an effective way of dealing with certain types of errors.

Let us go back to the slide. I will celebrate the fact that the live demo actually worked. So let's go ahead and move on. The question of dealing with earwyrms is really an economic and management question in a sense. The best earwyrm is one that did not exist in the first place. So of course, there is a question of finding the best possible sources for copy cataloging, ideally without paying exorbitant fees for it.

And of course, imported records, particularly from resource vendors etc., do have a reputation for not necessarily being the best overall. So this can be an example were succeeding in convincing a record vendor to do better can sometimes be the best way of controlling earworm. Of course ongoing data quality needs to be reported on. Authority control, record enhancement services can help, and of course, training.

It has been both heartening and interesting to see over the past several years and increase in the number of Evergreen consortia that are doing a formal training and certification programs for cataloging for their members.

And of course another aspect is making sure you have local standards certifying and you know what sorts of earwyrms you care about and don't care about.

But of course, there are...I mentioned economics. And that leads to a hidden Dragon. The name of that Dragon is prioritization.

Since of course, nobody has the staff or budget to obtain a perfect catalog, and of course, perfect catalog is a process and if our future goal, but not something that you will actually achieve, entropy if nothing else will make fools of us all.

But of course, one of the things about discovery is sometimes --it can be often the case -- it is better to have an incomplete or even bad record than no record at all.

So there can sometimes be a trade-off that yes, you may really need to load the crappy set of records because that may be the only way to get it into Evergreen and there may or may not be approaches that you search the vendors native database and Evergreen simultaneously.

Another factor for prioritization is what actually matters for functionality as the patrons and your library colleagues actually use these systems. A typo in a content note may not matter as much as going through and making sure that the record material format is consistently coded particularly in areas where patrons are actively holding a request. The particular rules and attribute definitions if you don't index of field at all it might not matter if it is correct.

That said, there can be a subtle trap year. The fixed field that you have ignored for the past 15 years could suddenly become super important with future Evergreen functionality or future ILS functionality. Of course, when the likes of Aqua browser etc. showed up in the marketplace and started using fixed fields more so than some of the second-generation ILS's, that did point out to a lot of library is completely ignoring the fixed fields was false economy.

But that is where there can be hard decisions that you make -- whether to go for short-term gains versus longer-term gains that may or may not ever come to fruition.

The current state of earwyrm, reporting, batch update, record overlay MARC edit special scripts scattered across the Evergreen community. New content we can do better or, so I think one of the key things is getting better tools for the wholesale case of the pile of gold which you need to separate the pyrite.

I want to first mention one of the vital tasks of maintaining useful catalog, whether that catalog is also your discovery system and whether it is something you are doing by yourself or some sort of cooperative mode, these are legitimate questions but ultimately, unless your library literally contains only 10 bucks some sort of discovery mechanism is required and I don't think it is something that can be outsourced by libraries to the likes of Google or the publishers.

I think developing metadata goals, particularly old ones, will help promote Evergreen's health.

This is where I think some of the discussion should happen. What are some enhancements to Evergreen that would help improve data quality. Does Evergreen need to have tools to directly support metadata workflow? There are a lot of bits and pieces that Evergreen handles ranging from import tools to reports to the MARC Editor but does there need, do they need to be tools that explicitly say, okay, let's encapsulate say the copy cataloging record review process?

Do we need some sort of change set mechanism so that we can confidently do and then undo a batch record operations, something to reduce the friction of it and have the ability to rollback if you make a mistake in the operation. The workflow front, is there a notion that maybe Evergreen should let you set up a metadata project as a mechanism that you can track and Evergreen to manage re-labeling or analysis. There are also quarterly things like being able to re-index more quickly would be good.

Should there be standard cataloging quality reporting available with Evergreen either directly in Evergreen or via one of the several banks of canned reports? Probably should be.

I think that is where the Evergreen cataloger community including the cataloging working group has a lot to share and a lot of updates to make. The final tool something on speculation. Would we want something -- let's periodically grab the latest and greatest from the Library of Congress? Can we automate the correction of some of our earworms? Do I have all the answers to that? No. I have some ideas but this I think is very much a community project to come up with ideas and ultimately code tools to have Evergreen be able to do more earwyrm remediation directly.

Of course we are not an island. We shouldn't pretend there aren't other systems and surfaces that aren't providing ideas to learn from.

So thank you for listening. I hope the taxonomy of earworms is useful to you. I am totally putting out a T-shirt idea for an "Ugh for Earwyrms"

the discussion to have I think to have among the Evergreen cataloging community is how to we get the Dragon off the catalogs? Thank you very much.

>> KATIE GREENLEAF MARTIN: [Laughter] Thank you, Galen, the Dragon and the cat, okay.[Laughter]. I am together I promise. I think we all have to head over to the closing session so thank you, Galen, so much that was highly enjoyable and useful to all of us I am sure. I will metaphorically shoo everyone in the direction of the closing session. Thank you everyone for playing along for the Evergreen conference and appreciate you being here. And see you all next year.