Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY, darkened

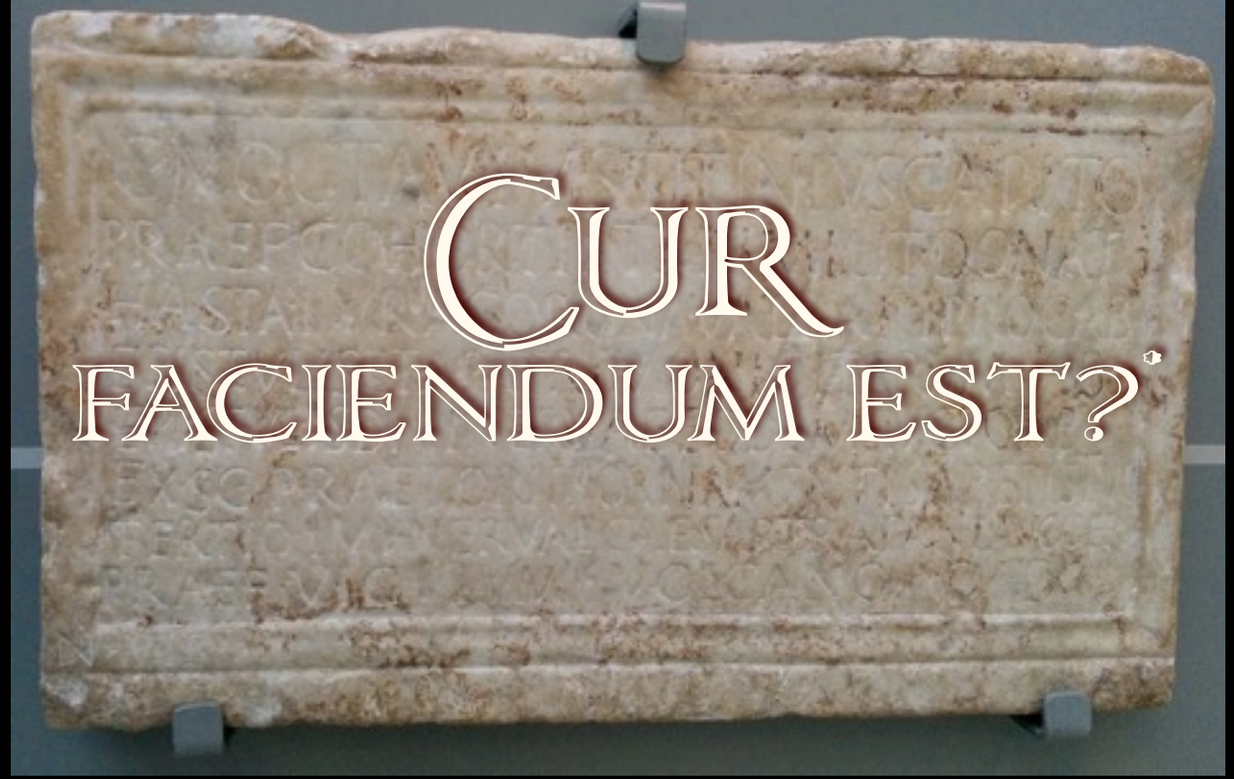
Thanks for that gracious introduction, I very much appreciate it! I'm Dorothea Salo, and I am not nor have I ever been a cataloger. *pause* Though I may be one of the last generation of library-school students required to take a full course in cataloging, I don't know. I just know we at the iSchool are doing a major curriculum revision, and current odds are that we're keeping the core organization-of-information course, just changing it to de-emphasize MARC cataloging in order to include more material relevant to non-MARC environments. Interesting times, not that they're ever not.

So, curriculum revisions, they're NEVER fun, they eat everybody's time for months on end, they ALWAYS cause bureaucratic hassles out the wazoo, for a while you have to deal with two different incompatible curricula and remember which rules apply to which student you're advising, and argh, it's a mess.

So why do we DO THIS to ourselves? Why do we bother, if it's such an awful hassle?

This is my answer. QUIA FACIENDUM EST *CLICK*, because it must be done (just in Latin it sounds cooler). We at the iSchool can't just sit back and do what we've always done because we've always done it that way, not when the world our graduates will need to fit into is way different from what it was when we built the old curriculum. And, I mean, it doesn't mean we did a bad job on the old curriculum, I don't think we did! It's just that the world has changed out from under it.

So we have to change the curriculum. We don't have to enjoy it. We just have to do it. Because it must be done.



Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY

*Why must it be done?

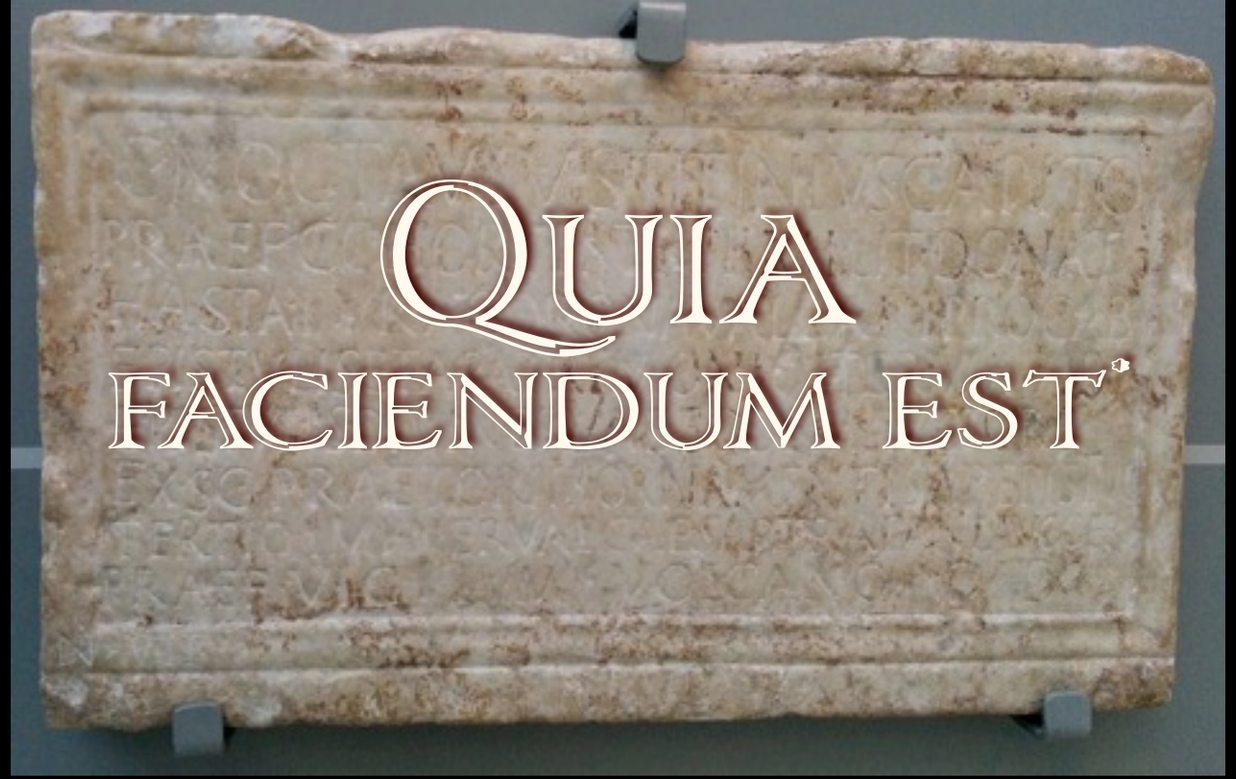
Not coincidentally, that's pretty much how I feel about the move away from MARC. It feels to me like a lot of the library profession has spent half a decade now, at least, on the question *CLICK* "Cur faciendum est?" or "why must it be done?" And even though this question sounds REALLY cool in Latin, I have completely run out of patience with it.

And don't even tell me nobody's asking this still, I straight-up heard it just last May at a conference, it's totally still out there floating in the water.

Alex Krasavtsev, "Empty Bronte Harbour Marina" https://www.flickr.com/photos/krasava/1689792663/ CC-BY, cropped

You like water? Here's some water in a nice harbor, really pretty, love the lighthouse, but the thing that isn't in this picture of a pretty harbor is a ship. Because the "why can't we still use MARC?" ship has sailed, people! It has sailed. I'm not even having the why-do-we-have-to-change discussion today, I honestly don't see the point, that ship has SAILED.

I mean, I'm next year's program planner for the IT division of Special Libraries Association, and I was talking to SLA's technical-services planner Betty Landesmann, some of you probably know her, I was talking to her about a linked-data session, and she rolled her eyes at me and said "can we NOT do another intro to linked data and why it's better than MARC please? I've seen a ton of those and they don't help." Okay. If I've got catalogers yelling at me not to do this, I'm not gonna do it!



Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY, darkened

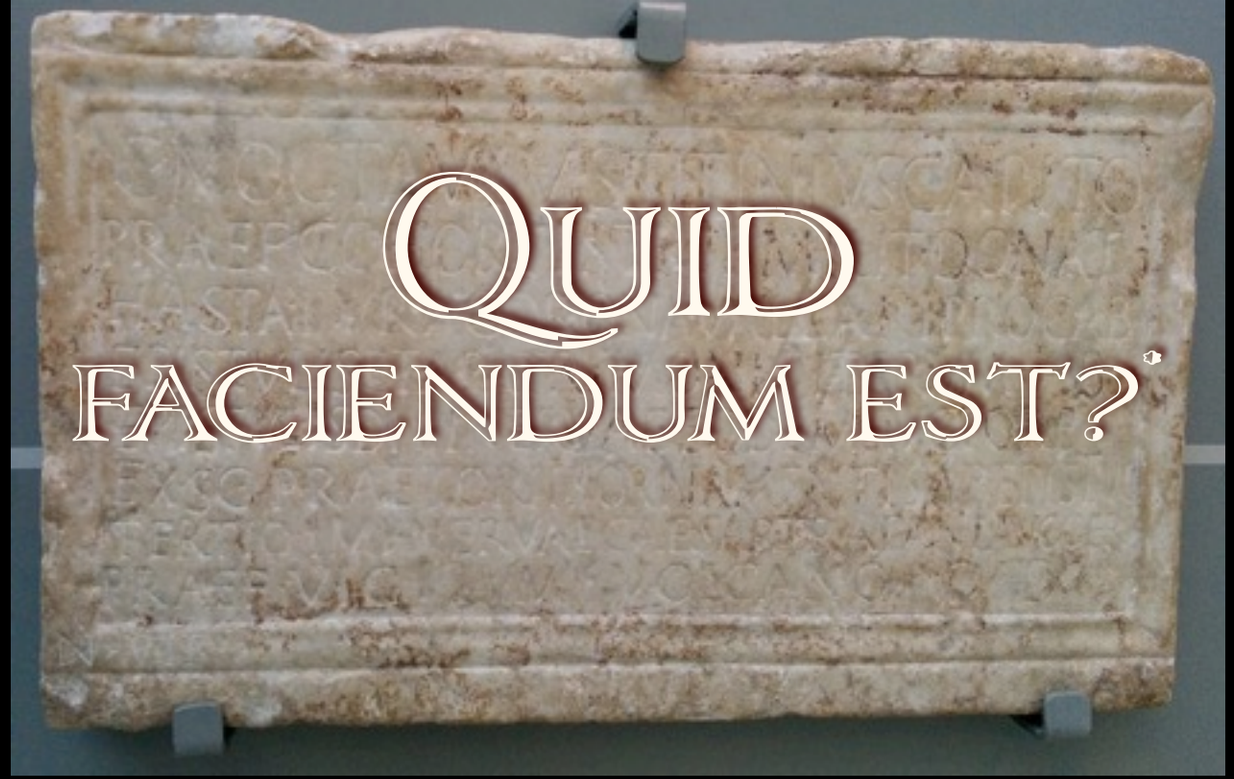
*BECAUSE IT MUST BE DONE

Because really, the answer is exactly the same as it is for our curriculum revisions at the iSchool. QUIA FACIENDUM EST. Because it must be done.

File:Museo_Civico_Archeologico_di_Castelleone_-_St_66634_-_campanello.jpg CC-BY-SA, cropped

I do want to mention, though, because I come at technical services from an XML-and-metadata background rather than a MARC background, that it isn't just MARC cataloging the bell is tolling for here. There's zero chance XML-based metadata practice will stay the way it is today; I already see it changing. I'm not even sure XML will stay ALIVE as a pure metadata format, as opposed to uses like TEI for the digital humanities and EAD for archives, where you're dealing with narrative-type documents intended mostly for human beings.

And, you know, I'm okay with XML's decline as a metadata serialization. I never liked my nice elegant document standard getting worked over by the data engineers anyway — do not even TALK TO ME about XML Schema, argh, it's just horrific. Maybe now I can have XML back. For documents. As it should be.



Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY

*WHAT MUST BE DONE?

No, I'm much more interested in this question. QUID FACIENDUM EST, what actually is it that must be done? What do we have to do to our catalog data and metadata so it works in this world where SO MUCH has changed about how we find information?

I like this question because it's pragmatic, I like it because it's intriguingly complicated, I like it because it's nerdy in all the best ways, I like it because I am an inveterate fiddler-with-things and there's just great huge masses of MARC and XML right there to be fiddled with, and —

And it's another of those questions we have to work on or we just stay stuck, right?

Roberto Venturini, “Roman Mosaic, Alcazar Cordoba” https://www.flickr.com/photos/robven/3069911545/ CC-BY

And I don’t think it’s enough to just say “well, we have to migrate our data from MARC and MODS and METS and the various Cores — Dublin Core, Darwin Core, VRA Core, PBCore and so on — we have to migrate all that to linked data.” That’s skipping all the steps!

Ken & Nyetta, "Unusual 2nd Century AD Mosaic Combining Geometric and Sea Forms
https://www.flickr.com/photos/kjfnjy/14121552821/ CC-BY, cropped

That's like saying "we have to pick up some rocks and turn them into a giant mosaic." Whoa, wait, not enough information! What's our mosaic design gonna be? Where's the mosaic gonna be built? Where do we find the right-colored rocks, and how many rocks of each color do we need, and how do we cut them down if they're too big or not the right shape? How do we glue the rocks down? What if somebody makes a mistake? What if there's an earthquake?

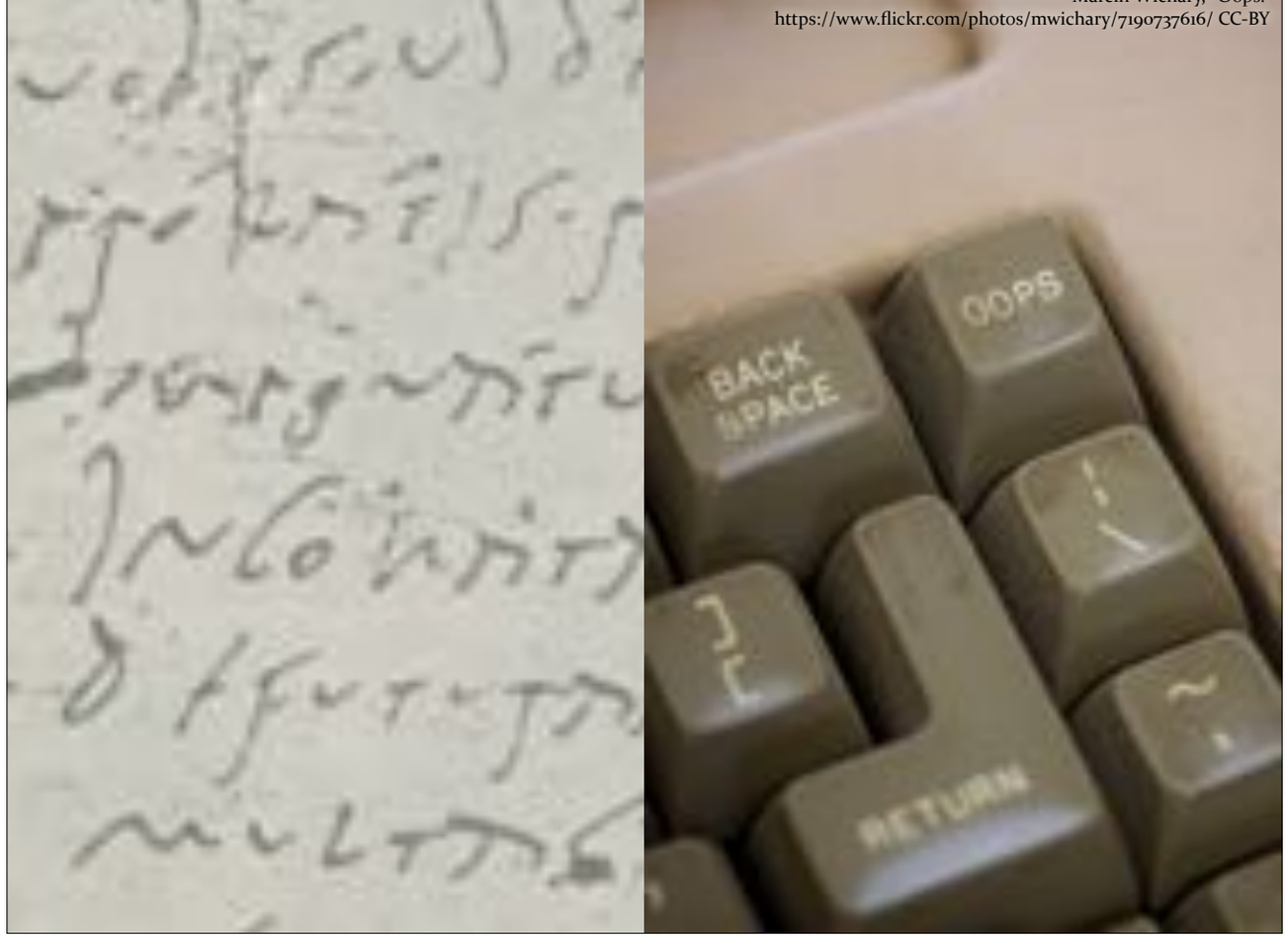
Process. That's what we need here. Some process, right?

Roberto Venturini, “Roman Mosaic, Alcazar Cordoba” https://www.flickr.com/photos/robven/3069911545/ CC-BY

The other reason I don’t think it’s enough to just say “well, we have to migrate our data from what we have to linked data” is that it assumes without proof that linked data is the ultimate destination for it. Which, yeah, it’s the horse to bet on, I’m not saying it isn’t, but I just —

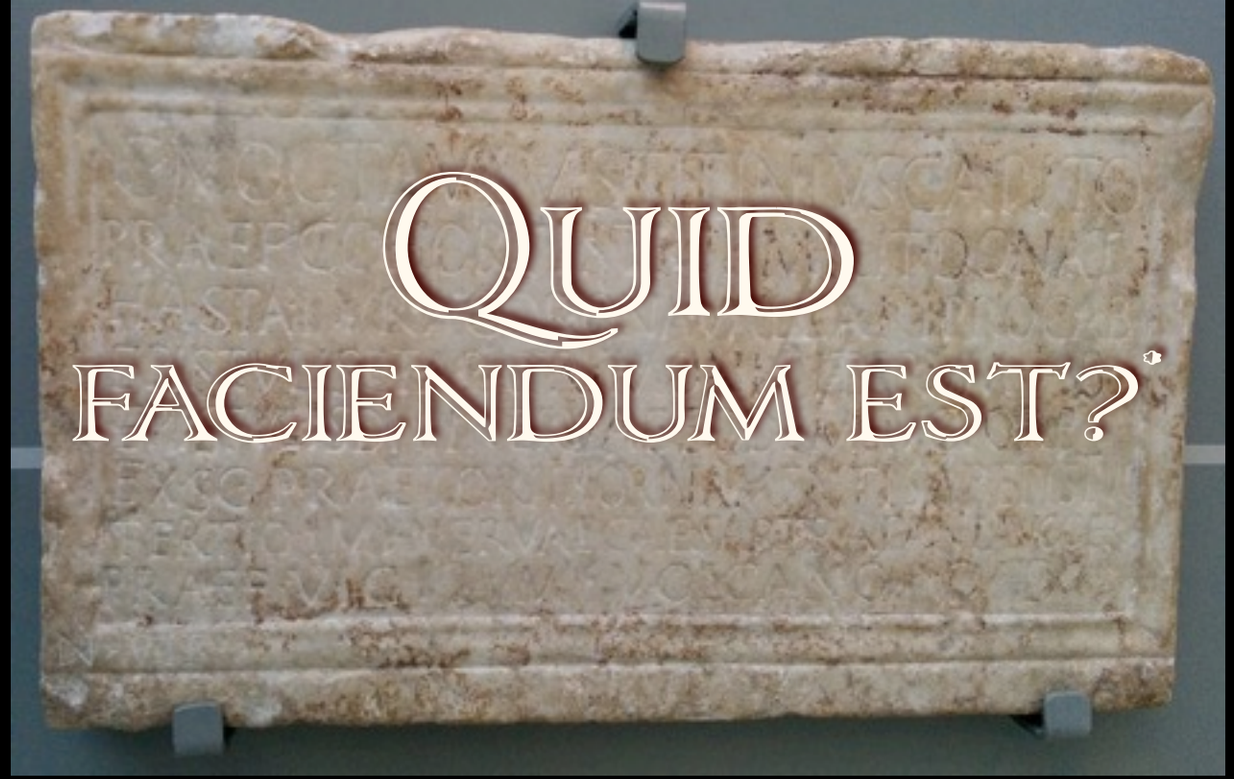
I just think linked data gets used as a stalking horse sometimes, a scapegoat. It’s linked data’s fault we can’t use MARC, it’s linked data’s fault all these changes are happening, if it weren’t for linked data supposedly being the new hotness we could stay the way we are and everything would be fine.

And I don’t actually think that’s true. If linked data didn’t exist — and let me tell you, I have lots of days I’d love to wipe RDF off the face of the earth — if there were no linked data, we’d STILL have to make changes in how we collect and organize our catalog data and our metadata. And we have to make those changes for the same reason we’re changing the LIS curriculum at the iSchool: the world has just plain changed out from under the old ways. And that didn’t happen when the Library of Congress or the British Library announced their linked-data plans. It happened WAY before that.



It happened when paper cards gave way to the Web as the main way patrons interact with library catalogs. (Okay, okay, I cheated, the left-hand picture is actually papyrus, not paper.) But no, that's when it happened. And, I mean, it's not that we didn't notice, of course we did, it's just taken us a while to figure out what we need to DO about it.

Oops. Which, I don't know how to say that in Latin, but yeah. Oops. Think we maybe waited longer than we should have, but you know, water under the bridge now.



Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY

*WHAT MUST BE DONE?

Which leads me right back here. Quid faciendum est. What do we have to DO, now that the work we do has to play nicely with computers? And not just computers, MARC was designed for computers, but NETWORKED computers, computers that can talk to one another. Because the network really does change the game.

Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY

This is where I start with my students when I teach our core organization-of-information course, actually. Ordinatra stulta sunt. *CLICK* COMPUTERS AREN'T REAL BRIGHT. *pause* No, I mean it. Computers are not all that bright, I tell my students, you're way smarter than a computer.

And I do this for a lot of reasons.

One reason is knocking computers off pedestals *pause* not actually literally knocking computers off pedestals, though hey that would be kind of awesome, but you know what I mean, right?

A lot of my students come into the iSchool thinking that computers are small gods, magical and capricious and liable to mess you up, like Diana here's about to do to that antelope. Impossible to understand much less work with, and I'm saying, I have to get them to NOT THINK THAT, because the more they understand about how computers DO work, the better off they are, and the better off we all are.

Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY

*COMPUTERS AREN'T REAL BRIGHT.

But the main reason that the notion that computers aren't real bright is relevant to this talk today is that from our point of view as literate human beings, computers are not-too-bright in some very specific and fairly easy-to-understand ways, and those ways tell us pretty clearly what our catalog data and metadata have to look like if we want computers to work effectively with it.


Deborah Fitchett, “Catalogue cards” https://www.flickr.com/photos/deborahfitchett/2970373235/ CC-BY

And really this is no different from how the shape and size of catalog cards and the standard size of typewritten lettering shaped how the MARC record had to look. The technology you have available—and the card catalog is totally a technology, don’t let anybody tell you it isn’t—the technologies available to you shape how it makes the most sense to do things, because different technologies are good and bad at different things and need different things to function best. That’s basic design theory, read Donald Norman’s Design of Everyday Things and its sequels if you haven’t already, it’s brilliant stuff.

Tilemahos Efthimiadis, “Papyrus in Greek regarding tax issues (3rd ca. BC.) ”
https://www.flickr.com/photos/telemax/3210586934/ CC-BY

So the first thing to remember about computers is that text, the ordinary stuff we write for other people to read—the text we literate human beings read and comprehend so fast and easily that we hardly have to THINK about it—it’s all Greek to a computer. (You knew that cliche was showing up at some point, right? I mean, how could I not.) Computers can’t read, they are functionally illiterate. If anybody in this room has, like, a kindergarten-aged child at home? Your kindergartener most likely reads and comprehends text WAY better than a computer can.

504 $aIncludes bibliographies and index.

504 $aIncludes bibliographical references.

504 $aBibliography: p. 238-239.

504 $a"Literature cited": p. 67-68.

504 $a"Spisok izdannikh knig": p. 277.

504 $aSources: p. 125-152.

So, in my head, one step toward coping with illiterate computers is dealing with our addiction to textual notes. By way of example, I took these MARC five-oh-fours straight from the Library of Congress's MARC documentation—thanks for that, by the way, if anybody here is responsible for it, it's super-helpful in my classroom. So, for example, if a patron question we would like our catalogs to answer is "hey, I'm new to this topic, can I get a recent book with a good bibliography please?" these notes being free text means our catalogs can't answer that question.

Because to get to an answer means filtering a list of books by whether they have a bibliography or not, and to do that with MARC notes, a computer has to understand that "bibliography" and "bibliographies" and "bibliographical" and "literature cited" and "sources" and maybe whatever that romanized Russian means, I didn't look it up, the computer has to figure out that they all mean YES, THERE'S SOME KIND OF BIBLIOGRAPHY! And no, a computer can't just look for the existence of a five-oh-four and assume there's a bibliography, because some five-oh-fours don't say anything about bibliographies, just indexes.

Look. Straight-up. The computer is not bright enough to figure this out. It can't read, much less read all the languages we transcribe stuff in, much LESS comprehend what it reads. And that makes a lot of the stuff in our MARC records a lot less useful to patrons than it could be.

☑ Bibliography

☑ Index

Computers get yes or no. That they're real good at. Checkboxes are candy to computers. So for any conceivable criterion we want our patrons to be able to filter their catalog results on, we pretty much gotta quit recording it in text and make it a checkbox. Or, you know, radio buttons if there's more than two options, that works too.

If you get the sense from this that I like MARC fixed fields, you're right! (Though honestly, that festschrift thing? That is just weird.) If a lot more of MARC had been expressed in fixed fields instead of free text, we'd be a LOT better off right now.

504 $aIncludes bibliographies and index.

504 $aIncludes bibliographical references.

504 $aBibliography: p. 238–239.

504 $a"Literature cited": p. 67–68.

504 $a"Spisok izdannikh knig": p. 277.

504 $aSources: p. 125–152.

Now, let's say for a moment that the titanic arguments we in the profession will have to have about when a book can be said to have a bibliography, and what counts as an index, let's say those are over, and we've drawn the best line we can. We still have to deal with this giant horrible mass of free-text notes in our existing catalogs that computers aren't bright enough to understand. FACIENDUM EST, people, it's gotta be done.

And I don't want to get down in the weeds on this—actually I would totally love to because this is exactly the kind of problem I salivate over, but I'm a giant nerd and not everyone is like me in that. So I just want to say—this class of problem can be solved for the great mass of our records without hand-editing, and of course it must be, because there ain't enough staff time in the UNIVERSE to check all those checkboxes.

And it's a thing that must be done, for every characteristic in our notes that we want users to be able to filter or search on.

Quoth the mighty

AACR2R:

"You know, whatevs. We're cool. Just type something."

Another serious and ugly free-text problem we have in our records has to do with places where our content standards don't force us to be consistent about how we record certain information. You know, whatevs. We're cool. Just type something.

And on catalog cards this inconsistency didn't matter, because the information was only ever going to be skimmed by a human being, who doesn't NEED a whole lot of consistency. We're literate humans, we're smart about figuring this stuff out.

Computers are astoundingly literal-minded, however. You can take some text, add a SPACE to it, and to the computer it's suddenly something totally different. Just one space, that to a human isn't even visible!

Quoth the mighty

Dublin Core:

"You know, whatevs. We're cool. Just type something."

And once again, this isn't just a MARC thing. Consistency in Dublin Core metadata? Don't make me laugh. Actually, I'm more likely to cry. It's bad out there, people. I mean, I once got an entire published article out of one poor soul in the institutional repository I was running at the time—poor guy had stuff under EIGHT slightly-different spellings of his name. EIGHT. Just terrible, and yeah, I fixed it as soon I had the screenshot I needed for the article, of course I did, I wasn't just gonna LEAVE it there in that state, but it goes to show.

☑ Dates

☑ Rights statements

And I want to call out two Dublin Core things specifically, noting that you'll find these problems many more places than just Dublin Core. Dates, oh my gosh, dates. Dates are REALLY IMPORTANT to information-seekers, so it's really important that we record them consistently such that a computer can reasonably intelligently filter based on them. We are NOT THERE YET, we're not even CLOSE to there. Dublin Core, MARC, it doesn't matter, people who try to make computers work are tearing out their HAIR about the ways we do dates, and we gotta fix it. Faciendum est, we gotta fix it.

As for rights statements, this is another one that comes from the digital-collections side more than the catalog side. It's important because we gotta be clear about what our users are allowed to do with our digital collections, and to do that, our search engines have to be able to tell which users can do what with which items, and that's all free-text now and it's a total mess. Europeana and the Digital Public Library of America are working on it, thankfully, though I don't envy them that job one bit.

402537 pbk
387406 alk. paper
99260 v # *(e.g., "v. 1", "v. 22", etc.)*
82918 cloth
51125 hbk
42036 electronic bk
41360 acid-free paper
38792 hardcover
28913 set
20358 hardback
19160 ebook
16264 paper
15269 u.s
12770 hd.bd
11793 print
10625 lib. bdg
10520 hc
8772 est
7767 pb
7639 hard

hard / hard adhesive / hard back / hard bd / hard book / hard bound / hard bound book / hard boundhard case / hard casehard copy / hard copy / hard copy set / hard cov / hard cover / hard covers / hard sewn / hard signed / hard backhard backcased / hard bound / hard cover / hard cover acid free / hardb / hardcover / hardback / hardback / hardback book / hardback cover / hardbackcased / hardbd / hardbk / hardbond / hardbook / hardboubd / hardbound / hardboundhardboundtion / hardc / hardcase / hardcopy / hardcopy publication / hardcov / hardcov er / hardcovcer / hardcove / hardcover / hardcover-alk. paper / hardcovercloth / hardcoverflexibound / hardcoverhardcoverwith cd / hardcoverr / hardcovers / hardcoversame / hardcoversame as above / hardcoverset / hardcovertion / hardover / hardcvoer / hardcvr / hardcback / hande / hardocover / hardover / hardpack / hardpaper / hardvocer / hardware / hd / hd bd / hd. bd / hd. bd. in slip case / hd. bd.in sl.cs / hd. bk / hd. cover / hd.bd / hd.bd. in box / hdb / hdbd / hdbk / hdbkb / hdbkhdbk / hdbnd / hdc / hdcvr / hdk / hdp / hdpk / hradback / hsadcover / hvd / hndbk / hrdcover / hrdcvr

Bill Dueber,
"ISBN parenthetical notes: Bad MARC data #1."
http://robotlibrarian.billdueber.com/2011/04/isbn-parenthetical-notes-bad-marc-data-1/

An example of the ugliness of free text that I use in class a lot is from library software developer Bill Dueber, who took a close look at what was after the ISBN in the oh-twenty field in the catalog he was working with. And it's horrific. Just the top twenty responses over there on the left, you can see the inconsistency, and the more you drill down, the worse it gets.

So yeah, our catalogs can't answer the very simple question "yo, this book, print or electronic or both?" At least not based on the oh-twenty, and yeah, I know RDA fixes this and I'm pleased about that. Bottom line, though, a lot of catalog data is hopelessly internally inconsistent.

And sometimes that's material for patrons and sometimes it isn't, but when it IS—this is my call to CONTROL ALL THE THINGS. All of them. Anything useful in a record that isn't actually transcribed off the item needs a controlled vocabulary, or other appropriately-standardized expression if it's something like a date. I cannot with this nonsense, and neither can computers. "Whatevs, just type something" is not okay in twenty-fifteen. Transcribe it or control it, there is no third option. Faciendum est.

Oh, and since I've said the word "transcribe," let me just say, intentionally transcribing typos and other errors in information that's material to a patron's searching and browsing is completely mindboggling to me. We gotta fix that stuff, and stop propagating mistakes. Consider it a service to publishers, as well as our poor patrons.

paul b. toman, “pointing hand of the Constantine Colossus” https://commons.wikimedia.org/wiki/File:Hand-of_Constantine.jpg CC-BY-SA, rotated

I picked on the oh-twenty field for another reason too, having less to do with cataloging practices and more to do with ISBNs. Now, I know I can’t fool y’all the way I fool my org-of-info students with the question “is the ISBN a good book identifier?” We know it’s not, we know lots of books don’t even HAVE ISBNs, and sometimes ISBNs get repeated for different books, and it’s not totally clear what a “book” even is in ISBNland, it’s kind of an edition but not really, it’s kind of a format question but not really, and it’s all very confusing.

Perhaps predictably, it’s confusing to computers too. Computers need to be really super unambiguous when talking about what kind of thing something is — if you and the computer have a different definition of what a “book” is, the computer is going to do random unexpected and unpredictable things from your point of view. And, I mean, the computer is happy to use whatever definition or definitions we’re happy with, the computer doesn’t care… but in spite of FRBR and sometimes, it must be said, BECAUSE of it, we don’t really have clear definitions here that don’t lead us into logical contradictions or bad edge cases.

So that's one thing. We gotta figure out what exactly we're talking about when we say things like "book" and "ebook" and "hardback" and so on, so we can explain the distinctions clearly to the computer… and if this reminds you of Suzanne Briet trying to explain when an antelope is a document and when it isn't, I am RIGHT THERE WITH YOU, it's totally going to be weird and sometimes theoretical like that, and I have to say, this mosaic is now kind of my personal headcanon for Suzanne Briet.

paul b. toman, “pointing hand of the Constantine Colossus” https://commons.wikimedia.org/wiki/File:Hand-of_Constantine.jpg CC-BY-SA, rotated

And then once we know what kinds of things we’re talking about, we have to be able to point clearly and unambiguously at every single example of these things that we—we collectively—have, so that it’s easier to pool the information ABOUT these things that we collectively have. The network can work FOR us if we let it, all of us know more than ANY of us about what we all have, but to let the network work, we have to have a common way to point at things. And for a computer, that means an identifier that (unlike an authority string) never, ever changes. And for a NETWORKED computer, that means an identifier that’s unique not just in your organization—so no, your call number or barcode number won’t work—but it’s got to be unique worldwide, so it absolutely cannot be language-dependent.

Because, you know, we’ve tried collating our records by fuzzy-matching and deduplicating, that’s how metasearch worked. But we pretty much all know that metasearch never worked real great. Computers aren’t bright enough to fuzzy-match well, and catalog data and metadata are sparse enough that they’re not good candidates for that approach to begin with. We’ll still have to use it to assign identifiers to start with, because we don’t have anything better, but it’ll be a long haul for some stuff and some catalogs.

So that means unique identifiers for our stuff that are way more reliable than ISBNs. And if you know the linked data world at all, you know that the scheme that’s been settled on for these is URIs, which mostly look like URLs. And the reasoning there is, we already know how to make URLs globally unique, because they already have to be or the web doesn’t work. That’s all it is. Just keeping the computers from getting confused.

Ken & Nyetta, "Dolfin Mosaic from 4th Century AD Villa at Halicarnassus"
https://www.flickr.com/photos/kjfnjy/6011841788/ CC-BY

And so another problem we have to solve if we're going to take advantage of the network is taking identifiers and quasi-identifiers that we've been using for things, like ISBNs and authority strings, and matching them up with URIs-URLs that have been established for those things. Again, I'm not going down in the weeds here, not even the sea-weeds with these dolphins, but I do want you to know (if you don't already) that the nerd-word for this process is "reconciliation" and it can be partially automated IF you know your source data well, as catalogers generally do.

See, because once you have a URI for something, you can go out to the network and ask a whole bunch of trustworthy sources "hey, what do you know about this thing?" and get back useful answers. To me, that's how what we now think of as copy cataloging is going to work. Ask tiny questions of various reliable sources, get tiny answers, build them up into search indexes and facets and browsing tools and all the other UI chrome we're already familiar with. And it won't have to be done by hand, if you tell a computer "every time I feed you a URI for a book we've bought, ask THIS question of THAT source in THIS way and store the answer THERE" it will happily do that, reliably and consistently, every single time.

And I strongly believe this will be a MUCH better solution to what I think of in capital letters as the Problem Of Vendor Records. You're familiar with this problem, I don't have to elaborate, right? What I'm saying is, the Problem of Vendor Records is nine times out of ten a problem of vendors struggling not only with MARC and AACR2, but with MARC practices that are incredibly inconsistent across libraries. It'll be a lot easier for us AND for vendors if instead our computers ask their computers a lot of tiny questions with tiny answers.

Roberto Venturini, “Roman Mosaic, Alcazar Cordoba” https://www.flickr.com/photos/robven/3069911545/ CC-BY

Tatjana Todorovic, “Mosaic, Aquileia”
https://www.flickr.com/photos/tatto_be/5439676274/ CC-BY, cropped

And that leads to one last thing about the oh-twenty, okay? ISBNs aren’t unique, the field includes inconsistent format information, yeah, yeah, we got that. Here’s my question: what the everliving heck is format information doing in an ISBN field to begin with? Much less information about volumes of a multi-volume series? And if it does have to be there — and I know, I know, I do get why it ended up there — why isn’t it at LEAST in a separate subfield? What’s this nonsense with parentheses?

I’m showing you a mosaic detail here for a reason. You can see all the teensy-tiny individual rock bits here, and you can see how carefully they’re placed, and that none of them actually overlap anywhere. That’s what our catalog data and metadata should look like. No overlaps, nothing jammed together, everything in tiny tiny bits and each tiny bit in its own singular place, very carefully set apart from all the other tiny bits.

That’s called “granularity,” and computers love it. Computers are really really good at building up whole mosaics from tiny granular pieces! What they’re critically BAD at is taking a whole and breaking it into parts. We have to do that for them. And as we saw with the oh-twenty field, we often don’t, or when we do, we do it in ways that the computer finds confusing and inconsistent.

Here’s a relevant factoid to take home with you: Computers CANNOT reliably and consistently split human names and titles in human languages into their component parts. Naming is just too inconsistent across human cultures and languages for that to work. With pre-split names, though, it’s relatively easy to write rules to put the names back together intelligibly.

English 1 — Globe

Object (cartographic ; visual) : unmediated

Scanglobe diameter 30 cm / by Replogle. — Scale 1:42,000,000. — [Chicago, Illinois] : Replogle Globes, [2006?]

1 globe : col., plastic ; 30 cm (diam.)

Relief shown by shading and spot heights. Depths shown by shading and soundings. — "Scanglobe" is a trademark. — Mounted on spindle crowned by a plastic clockface, in a plastic meridian half circle, on plastic base. — Globe lights up from inside by means of electrical cord with switch button and interior bulb

Copyright: Stan Brakhage. — "A hand-painted work whose shapes are scratched on black leader filled with varieties of color; the resultant shapes tend to suggest insect-like movements"–Canyon cinema catalog, no. 8 (2000) online

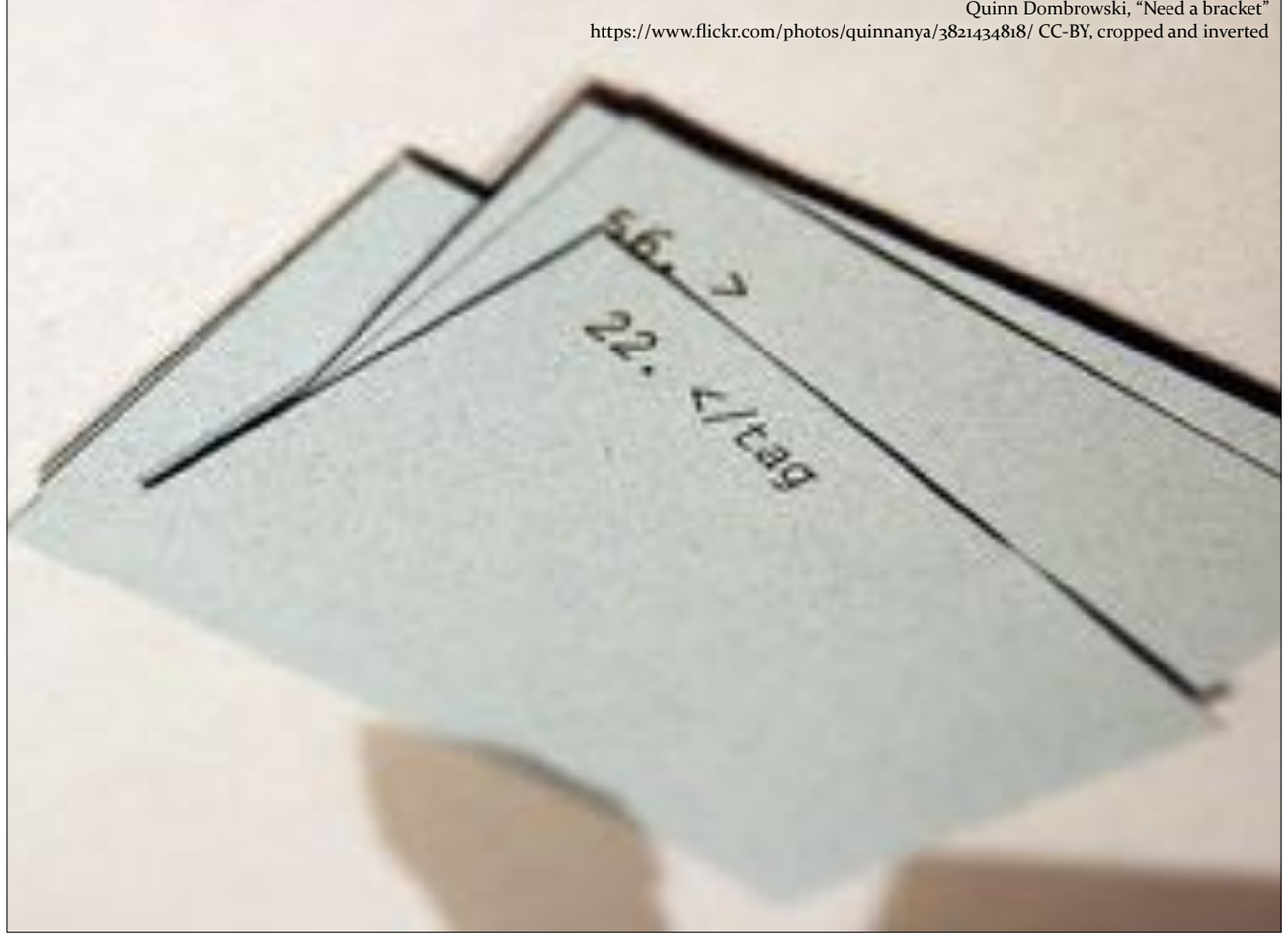
So I grabbed a couple of examples from IFLA's ISBD supplement—and again, thank you for this, if anybody's here from IFLA, this is a classroom life-saver—just to show the difficulties. Try to think like a computer for a second. There's a WHOLE LOT of punctuation all up in here, and it's not at all obvious (even to me, and I'm a human being) what's being set off by it or what it means, or even if it means anything at all!

I mean, look at the physical description of the globe there. Tell me what a period, a dot, means there (1 space globe space colon space col dot comma space plastic space semicolon space thirty cm space parenthesis diam dot parenthesis). Oh, okay, as a human being, I can figure out the dots there are calling out abbreviations. Now, can I just tell the computer to assume a dot always means an abbreviation? Of course I can't! A dot doesn't mean that in the other areas! And can anybody tell me why in the last area of the globe description, everything except the final sentence has a period at the end? It's enough to make a computer programmer cry into her beer.

And I lopped the area labels off for legibility, but I just want to point out, we have two competing sets of delimiters happening here—the areas, that are set off with whitespace, and what's in each individual area, which is funky punctuation city. And when you add that to MARC, we've got a whole 'nother set of delimiters in the form of fields, subfields, and indicators. I respect my colleagues who teach cataloging! I could never do it! Because I cut my teeth on XML, where delimiters are totally cut-and-dried and straightforward, such that I find this mishmash completely bewildering!

And the bottom example there, anybody see an error? *pause* I mean, I think it's an error, it's hard to tell, but that very bottom line, if that's supposed to be an em dash before the word Canyon, it's not. To a human being, no big deal. To a computer, VERY big deal. Just goes to show, when we can't even be sure our documentation is right…

Quinn Dombrowski, “Need a bracket”
https://www.flickr.com/photos/quinnanya/3821434818/ CC-BY, cropped and inverted



So anybody who works with XML is twitching now because of the missing close-angle-bracket on the card here. Sorry about that, but it illustrates an important point about delimiters.

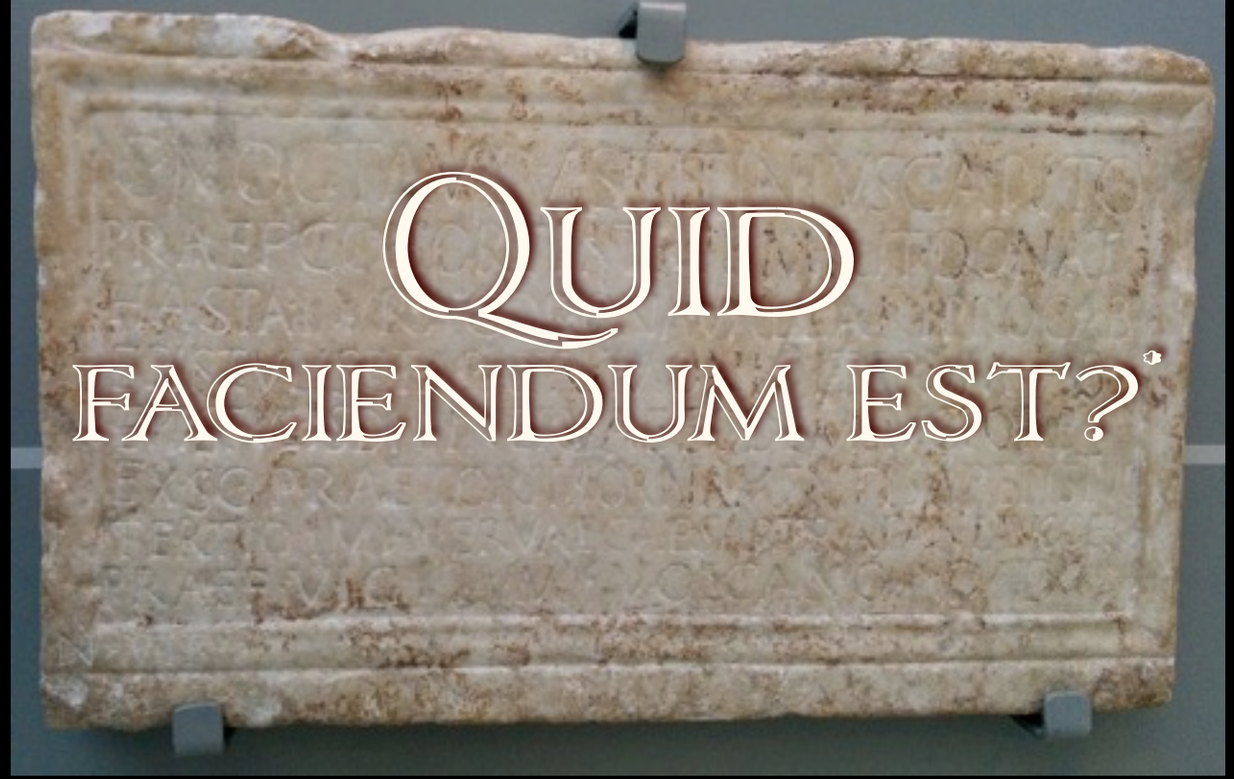
A lot of my students find learning HTML and XML frustrating, because they’ve never had to be one hundred percent consistent about delimiters before. So they make tiny little mistakes like leaving off an angle bracket, and they haven’t learned to SCAN for those mistakes yet, and they don’t understand what the validator that DOES notice problems like that is trying to tell them, and it’s really frustrating for them.

So what I tell them is, SUCK IT UP AND DEAL. *pause* Okay, no, I’m not actually that evil about it. I’m pretty careful to point out the kinds of errors that beginners usually make, and I tell them that everybody makes those errors, even really skilled and experienced people, and it’s okay, the whole point of validators is to help us get it right.

But fundamentally, yeah, they have to learn to deal. And they don’t like that necessarily, but too bad, FACIENDUM EST, it must be done. Don’t confuse the computer, folks! Reliable and consistent delimiter use is how we avoid confusing the computer. Delimit in just one way, delimit clearly, delimit unambiguously.

And even our XML-based metadata standards aren’t necessarily doing that, much less our MARC-based catalogs! MODS has composite fields and weird ISBD punctuation.

And, you know, I have to share this thing that broke my heart—I was digging into the history of MARC and ISBD internationally, and it turns out that the Germans were totally bent on killing ISBD punctuation out of MARC and relying on MARC delimiters only, which from the point of view of twenty-fifteen totally would have turned out to be the right decision, but English-speaking MARC went the ISBD direction instead, and just argh, twenty-twenty hindsight. Argh, it’s one more mess we now have to clean up.



Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY

*WHAT MUST BE DONE?

So. What is it we have to do now?

Faciendum est:

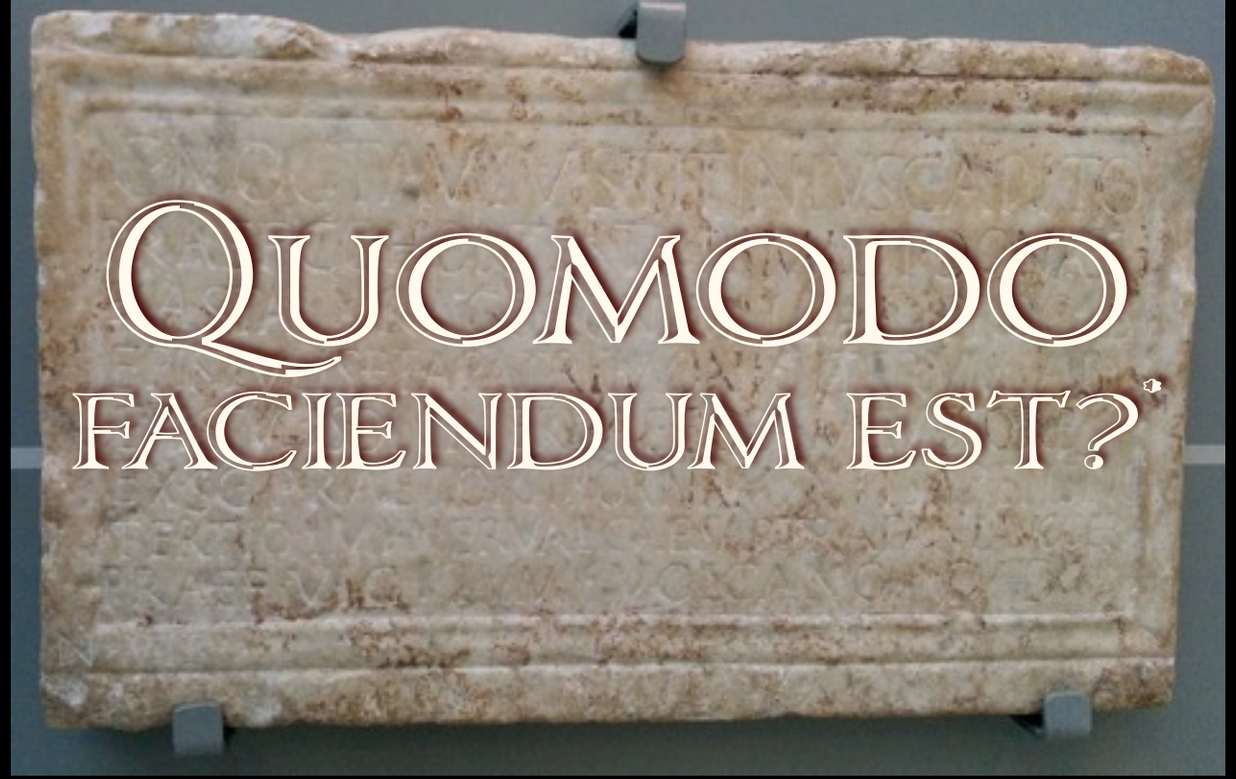
- Free text (other than transcribed) must be controlled.
- The same data must be expressed the same way every time.
- Our data must be broken down to be granular.
- We must fix multiple delimiters and inadequate delimiters.
- Anywhere we can identify something or someone unambiguously and unchangingly, we must.

We gotta get a handle on our free-text issues. When we're saying the same thing, we need to say it the same WAY every time. We need to atomize our data, make it as granular as it can be. The delimiter thing, sometimes we have too many and sometimes we don't have enough, and we need a happy medium. And when we can identify something as well as labeling it, we should, because identifiers make computers happy and useful.

Roberto Venturini, “Roman Mosaic, Alcazar Cordoba” https://www.flickr.com/photos/robven/3069911545/ CC-BY

And just to reiterate, we don’t have to do these things Because Linked Data. We have to do these things Because Databases, and Because Search Engines, and Because Faceted Browsing, and Because Internet, and Because Web. Basically, Because Not-Too-Bright Computers.

Now, no secret, cards on the table, if we DO do these things, we’ll be one heck of a lot closer to having linked data. So it’s not like I’m ignoring that, I’m just saying, linked data is not all this is about.



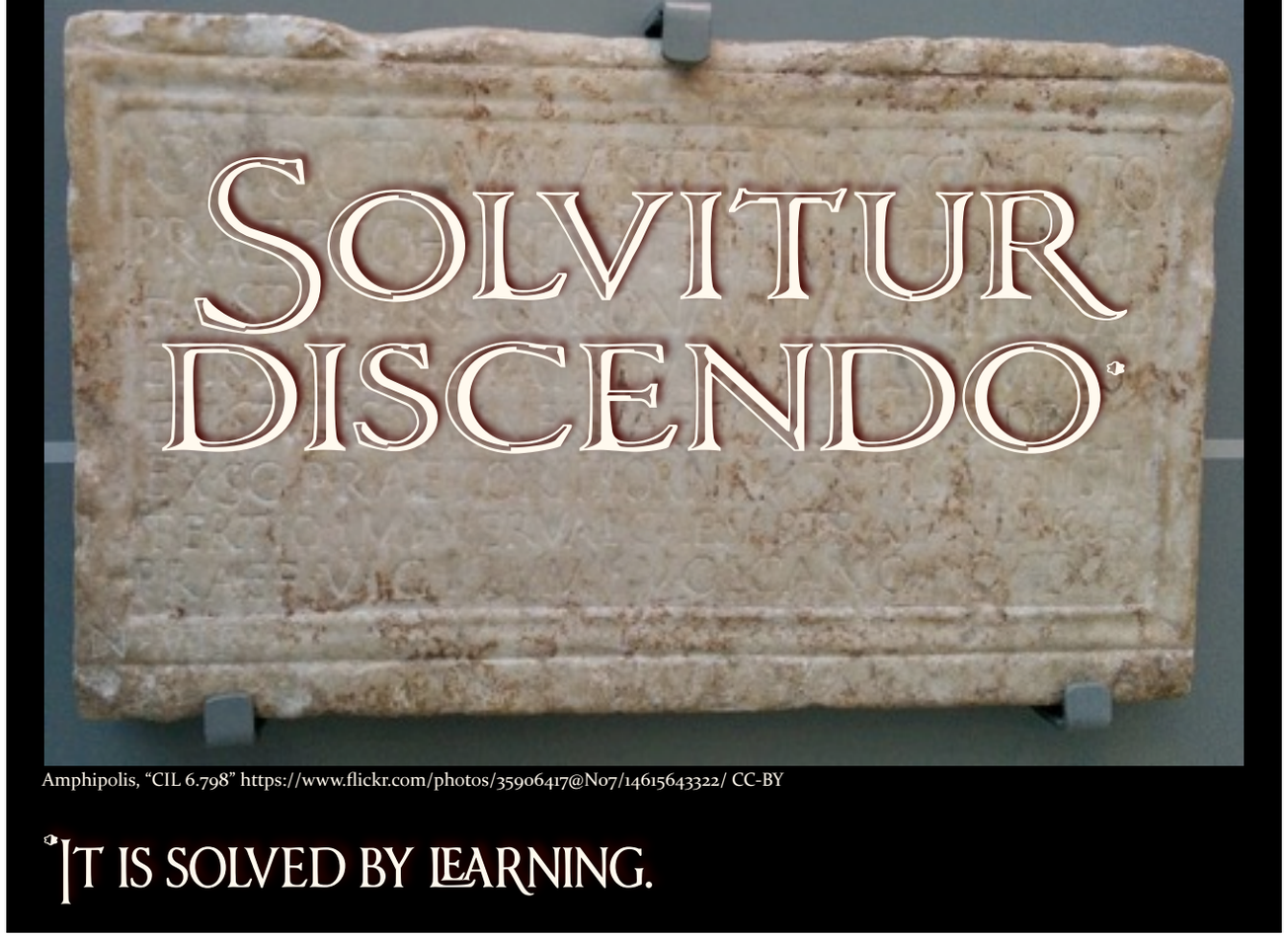
Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY

*How must this be done?

Well, that's great. Plenty of work for catalogers, right? *CLICK* *pause* How do we actually do it, exactly?

“By hand, one record at a time,” is not the answer here. It can’t be. I mean, yes, there will still be weird outliers that we end up fixing by hand, there always are, I could tell you stories—but we need to throw computers at fixing the easier problems and limit our handwork to those weird outliers. We have too much data to do it any other way.

I’m not sure that doing it organization by organization is the way either. In my head, that means a lot of the same problems are getting solved redundantly in parallel. That costs too much and takes too long.



Amphipolis, "CIL 6.798" https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY

But the truth is, we collectively don't yet have the know-how we need to collaborate on this. Most computer programmers don't have the MARC knowledge, and many catalogers don't know how to work with records en masse. So THIS problem? It is solved by learning. Catalogers, developers, we all have some learning to do. And, I mean, I would say that, right? I'm a teacher. But I'm a teacher because I BELIEVE THIS.

Any managers in the room, supervisors? GIVE YOUR PEOPLE TIME AND SPACE TO LEARN, do not make me yell at you about this. Straight up, your cataloging backlog or the new tech thing your developers are working on is far less important than your strategic preparation for what's barrelling down the pipeline at you, okay? By all means hold your people accountable for actually learning as opposed to complaining about learning—again, I could tell you stories, but you can probably tell me stories too—but LET THEM LEARN. Help them learn. Learn yourself, it won't kill you and might make you stronger.

Discendum est:

- MARCEdit
- OpenRefine (try LODRefine)
- Regular expressions (try regexone.com)
- SQL (if you have to extract data from a relational database)
- XSLT (if you have a lot of XML around)
- Catmandu/Fix (don't start here, Here There Be Yaks)

Here are some tools I think are well worth adding to your toolkit if they're not there already, because they're designed to fix stuff in lots of records at once rather than one-record-at-a-time. I've ordered them by the order in which I would recommend that a cataloger learn them, and the last half you may not even need, I mention them because there are situations where they're genuinely going to be useful.

Roberto Venturini, “Roman Mosaic, Alcazar Cordoba” https://www.flickr.com/photos/robven/3069911545/ CC-BY

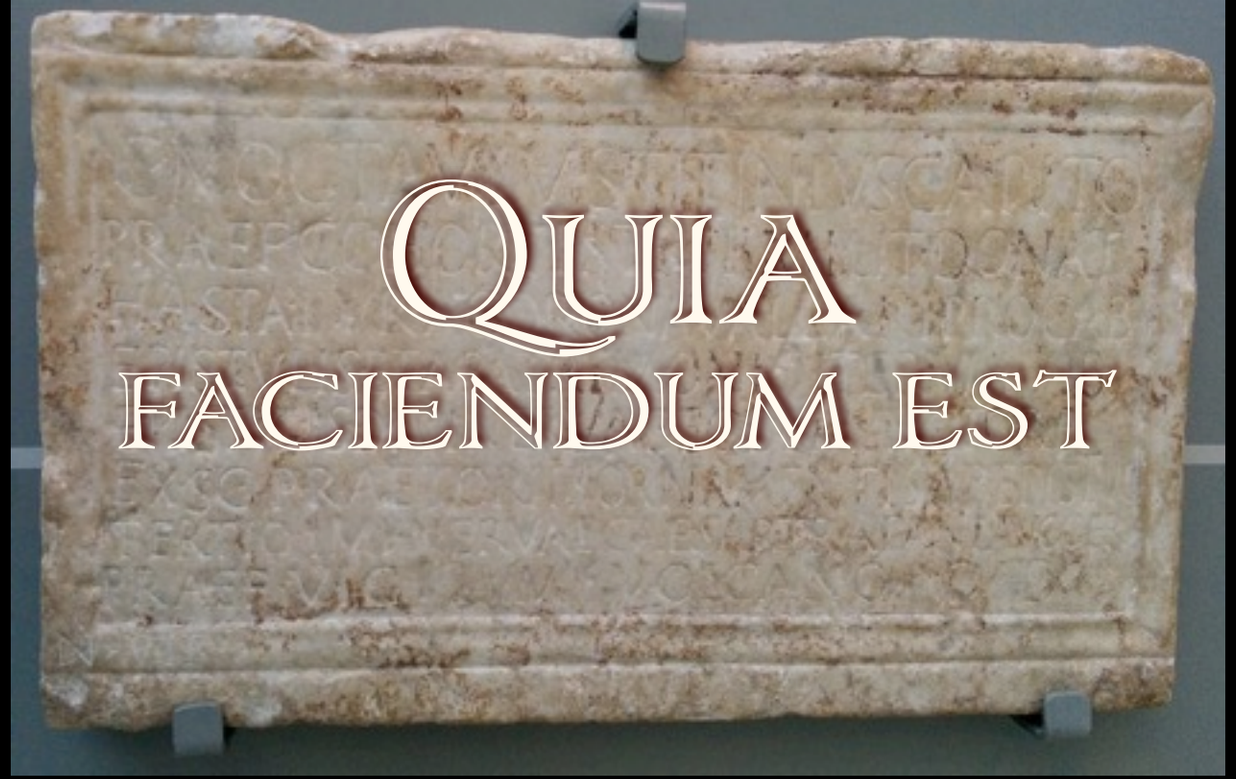
How should you learn, and what should you do with what you learn? In my head, those questions are intimately entwined. In some circles I’m known for the phrase “beating things with rocks until they work.” And yeah, if you use mosaic rocks this probably won’t work, but you get the idea. I think plain old mucking around and breaking things and fixing them is the best way to learn new tools, myself—it’s certainly how I do it, and it’s the method I have my students use. Pick something to do with the tool, then do it, and if a few rocks get beaten on along the way, it’s all good.

And what should you do exactly? Well, try cleaning up your catalog data! Go fix your oh-twenties with MARCEdit, see if you can at LEAST make those format notes more consistent. Export your data into something OpenRefine can read—it’s doable, I’ve talked to catalogers who did it—and see if you can cluster your five-oh-fours such that you can figure out in the majority of cases if there’s a bibliography, if there’s an index. Try fixing your dates, there are all kinds of fun and interesting problems you’ll run into and have to come up with some way to solve.

You know your catalogs way better than I do, you know where the worst problems are. You also know where the most IMPORTANT problems to fix are, which I think is totally crucial knowledge here. So learn the tools by fixing the problems you already know are there.

Gnari spectandi

- Christina Harlow (@cm_harlow, christinaharlow.com)
- Karen Coyle (@karencoyle, kcoyle.blogspot.com)
- Diane Hillman (managemetadata.com)
- Owen Stephens (@ostephens, ostephens.com)
- OCLC Research (especially Thom Hickey)



Amphipolis, “CIL 6.798” https://www.flickr.com/photos/35906417@N07/14615643322/ CC-BY, darkened

We’ve got a lot of work to do, and it’s gotta be done, so let’s jump in and do it. Thanks very much, and I’m happy to take questions.