Module 1: Intro to Metadata

Post Notes

- social bookmarks & folksonomies already feel so last decade

Post

1. What do you think might be the title of this table?
2. "Endangered and Threatened Species by National Park"? I have to admit I'm not finding any obvious criteria for selecting these particular parks and these particular species (many of which can be found in other parks), and I can't tell what the seasons are meant to indicate, either (at first glance they don't seem to correlate with visitor stats, for instance).
3. What do you think might be the headers of this table?
    a. Name
    b. State
    c. Established
    d. Species
    e. Season (?)
4. What other information do you think should also be included?
5. Without knowing the real purpose of the table, it's hard to tell. At the moment I'm inclined to think there's *too much* information—if it's a list of national parks, the species seem superfluous, and if it's a list of species, the years of establishment of the park seem superfluous.

Module 4: Descriptive Metadata

Post

I thought it was amusing that Gartner (p.31) and Pomerantz (p. 6) were clearly both telling the same story about the evolution from Callimachus' *Pinake* to the card catalog—it must be required for anyone who writes a textbook called *Metadata*. Still, though I've run across most of the acronyms and buzzwords in the last couple of modules' reading before, it's good to place them in historical context—not so much Ptolemaic Egypt as the past 20 or 30 years. Having previously worked at the California Digital Library and UC Berkeley, I know some of the characters in these stories personally—I didn't work with Mary Elings directly at Berkeley, but my supervisor certainly did, and I joined CDL about the time Günther Waibel became its executive director; at Berkeley I worked with one of the original developers of the EAD standard. But I still didn't know exactly what order things happened in, or have a sense of just when some of these standards came into being. It's also interesting to read the concerns and predictions of 30 or 20 or even 10 years ago and see how things turned out—Pomerantz (p. 90) notes that the 1995 assumption that web search tools would require descriptive metadata turned out to be false; Elings' and Waibel's prediction that users would come to prefer Google over their local OPAC certainly seems to have come to pass, though not in the structured way they expected.

By the time I was working for UC, these groundbreaking projects were all mature—one of my first projects at Berkeley was to work with a former colleague at CDL on getting the OAI-PMH feed for Berkeley's new digital asset management system talking properly to the metadata harvesting system at the Online Archive of California (successor to MOAC—I'm still not sure when the museums dropped out). There were still a lot of the original CDL staff there, but its technically innovative period was over, and I know more than one person who either left or was pushed due to being unable to adjust from a startup mentality to a staid, buttoned-down service organization.

Module 6: Provenance Metadata

Post

I appreciate the way Bettivia et al. cut through the complex specifics and philosophical abstractions of the PREMIS data model to the nature of provenance, the purpose behind the metadata—even if they do provide two similar but distinct definitions: "Provenance is a description of how something comes to be" (p. 6), and "Provenance is the weaving together of metadata items to tell a cohesive story about an object" (pp. 3-4). It seems important to keep grounding PREMIS in specific tasks digital preservation organizations are trying to accomplish—to treat PREMIS as a tool for getting work done and for making it possible to get work done in the future, rather than as an end in itself; it would be easy to get lost for days in questions of just what to model (do we stop at *Animal Crossing: New Horizons*, or do we need to model the entire *Animal Crossing* family of games? Where are the boundaries of the environment needed to play the game—does it include Nintendo's servers? Does it include the consoles of other players connected to those same servers?), but if what we have is a Switch cartridge in hand and we're trying to document the provenance of the physical object and the hardware and software needed to get the software on it to run, the task starts to seem much more manageable.

References

- Coyle, K. (2016). *FRBR, before and after: A look at our bibliographic models*. ALA Editions.

Response (Carolyn Quimby)

Hi Carolyn,

I groaned a bit at the mention of NFTs—I agree, I think the boom does seem to have busted, and thank God—but I agree that Prelinger has a point about the authenticity of digital objects, and another one about the financial realities of GLAM institutions. I think there are better (cheaper, more environmentally sound) technical solutions to the problem of authentication and digital provenance that don't require buying into the NFT / cryptocurrency / blockchain hype—we already use cheap, reliable digital signatures every time we connect to an HTTPS website. But the promise NFTs and blockchain tech are dangling in front of people is creating something from nothing—taking digital objects, which are fundamentally cheap to create and practically free to reproduce, and somehow making them scarce and expensive. I think it's a fundamentally doomed enterprise, but as long as GLAM institutions are constantly scrambling for cash (and even the big well-off ones seem to end up being led around by their development offices and endowment funds) they're going to be as vulnerable as anyone else to that gold-rush mentality.

Module 7

Post (activity)

ProvONE elements	**Answer**
Entity	Mini DIY Workbench (output entity)

	(existing) workbench (input entity)
Data	wood x3 hardwood x3 softwood x3 iron nugget x2
User	Emery
Execution	Emery’s Mini DIY Workbench
Program	Mini DIY Workbench recipe
Workflow	Craft mini DIY workbench

(Note: I watched this YouTube how-to on crafting the workbench, and also read some AC:NH crafting information on Nookipedia. It looks like the process is much less interactive than the snowperson example, so I only identified one workflow.)

Post (readings)

I came out of last week’s readings unconvinced that modeling the past history of an object and future actions intended to be taken on an object both as “provenance.” The breakdown of ProvONE that Bettivia et al. give us doesn’t do much to convince me, and neither does their proposal for “subjunctive provenance”, even if it is amusing and even if it does have some grains of truth in it (once you have a system that uses the same data model for past history and future intentions, it would be easy and tempting, but badly mistaken, to have the system simply assume as a shortcut that intentions were carried out according to plan, rather than recording what actually happened). .

The PREMIS data model at least is pitched as a “preservation” model rather than a “provenance” model, so it seems reasonable that it might be general enough to model both past actions and hypothetical actions—if I were more familiar with how “prospective provenance” was used in the real world, I might respond to Bettivia et al. 2023 by arguing that it’s not so much that the conditional and future tenses merit separation (p. 4) as that the mistake is collapsing the two of them into “future,” rather than into “conditional.”

I looked up Zhao et al. 2006, which the ProvONE draft cites for the concept of “prospective provenance”, and feel no better informed—the concept seems to have been pretty well baked in already by that point. That said, there’s obviously a whole history here I’m not familiar with, and I assume the idea of calling all this “provenance” has roots, that somebody raised these semantic objections at some point early in the century and the data provenance community dispensed with them.

Still, though, ProvONE’s whole approach to modeling time seems poorly thought out. When you find yourself encoding semantic relations like *hadPlan* and *wasAssociatedWith*, you should step back and ask yourself some hard questions.

Response (readings) (Nicole Poyer)

Hi Nicole,

I like the carrot cake example, and it fits well both with Bettivia et al.’s IKEA example and with the Animal Crossing snowpeople. What I wonder, though, is what makes it worth taking the time subjunctively (or retro-subjunctively?) document all the possible cakes that could have been made, but weren’t? It’s easy to see the value in documenting the recipe, prospectively, even if calling that “provenance” still doesn’t sit right with me; it’s easy to see the value in documenting how a particular cake was actually made, retrospectively, whether it’s “I left out the nuts, because Philippa’s allergic,” or “I was out of butter and had to sub oil, and it’s kind of… flat, sorry!”. I can even see a lot of value in documenting possible cakes in the conditional, as part of (or variations on) the recipe: “You can add coconut if you want more texture”; “if you can’t find confectioner’s sugar, you can grind white sugar together with corn starch”. But I’m not convinced that we need to include all the cakes not made as the documentation of the particular cake that was.

Module 8: Interoperability

Post notes

Post (MARC-MODS-DC crosswalk)

MARC XML Element	**MODS Element**	**Dublin Core Terms Element**
datafield[@tag="260"]	originInfo	(none)

datafield[@tag="260"]/ subfield[@code="a"]	originInfo/place	(none)
datafield[@tag="260"]/ subfield[@code="b"]	originInfo/publisher	publisher
datafield[@tag="520"]	abstract	abstract
datafield[@tag="245"]/ subfield[@code="c"]	note[@type= "statement of responsibility"]	description

Post (readings)

I like the mathematical formulation of taxonomy mapping in Franz et al. (2016) and Cheng et al. (2017), although I do wonder how much traction these kind of automated reasoning / theorem-proving systems (or results they generate) are getting in the real world—it feels like something of an intellectual throwback to the “expert systems” of the 1970s and 1980s (though presumably much more powerful), and so something of a hard sell when the money and excitement is all in unstructured machine learning. It seems like at least some of the systems in Shvaiko and Euzenat (2013) might be more of that school—although that doesn’t necessarily make them any better at the task, or any easier to apply.

I suppose in general I’d like to know more about how these mappings are applied—it’s certainly *irritating*, from an intellectual perspective, that grass species names are vague, that there is no cross-contextual consensus as to the regional divisions of the United States, or that there are multiple overlapping terms with different granularities for the genre of a visual artwork, but which of these are stopping people from getting what kind of work done? And how are these mapping systems making that work easier?

By contrast, Chan and Zeng (2006, and/or Zeng and Chan 2006) are addressing metadata mapping problems more like the ones I’ve run into in my professional life, and it’s interesting to get a view of the landscape (pace *Aria* and *Iconclass*, as described in Shvaiko and Euzenat) as it stood in 2006.

I found it interesting in Chan and Zeng how often federated search—searching multiple heterogeneous collections simultaneously (pt. 4.3)—came up as a motivating case for metadata mapping. I feel like federated search is something we used to hear about a lot, but nobody's doing it now, or at least not doing it in that form—i.e., actually sending out multiple search requests to multiple databases and merging the results. This is hard to do well both in terms of performance (users are no longer willing to wait for multiple queries that might take different amounts of time to all return results) and in terms of relevance (it's pretty much impossible to compare the relevance rankings of results from different databases). In practice, most systems I've worked on or encountered are aggregation systems of the sort Zeng and Chan discuss in their second paper (pt. 3.3), where records in a small set of known schemas are harvested from multiple collections and crosswalked to some compromise schema, with links back to the original records.

Worth pointing out for this class is that even using the same schema, say MARC (as discussed in Chan and Zeng, pt. 3), doesn't eliminate metadata mapping as an issue—at Berkeley we had a set of spreadsheets, and then (thanks to my colleague Yuchai Zhou) a Ruby code library, for mapping MARC records from the library's ILS (first Millennium, then Alma) to its TIND digital asset management system, moving some data from one field to another, eliminating some subfields, collapsing others. (You can see some of the mapping rules in CSV format here.) Similarly, anyone who's working on consolidating multiple catalogs into a consortial ILS will tell you how much work is involved in reconciling different libraries' MARC record standards and practices. There are still a lot of semantic ambiguities and system-to-system or institution-to-institution implementation differences even in a very well-known, well-established schema like MARC.

Response (Nicole Poyer)

Hi Nicole,

I like the observation that when one institution uses multiple schemas "they can't expect there to *not* be problems." In my experience there's usually *some* consideration given to interoperability in these projects, but usually only to the extent that there's a known need to import/export metadata to interact with some other system. Mostly a schema gets picked because it's the best, or at least the easiest, for some particular task, and then the problems come later when the metadata created for that task gets taken out of that narrow context into a wider one.

It's also a lot easier to do a one-way crosswalk—"we're going to export these MARC records as METS", or "we're going to extract Dublin Core from these EADs"—than it is to really have two-way interoperability between systems using different schemas. The data loss problem is real and I think it's kind of inevitable—once you convert MARC to

Dublin Core, you can convert that Dublin Core back to *a* MARC record, but you can't convert it back to the original MARC record, because Dublin Core, by design, just isn't rich enough to capture all the complexity MARC can. The best solutions, I think, are the ones that keep the original records around, either embedding them or linking to them—but while doing that doesn't lose information, it also leads to a lot of complexity when trying to *use* that information, whether as a human or as a software system.

No matter what approach you take, it's going to be work. As Dorothea Salo of the University of Wisconsin iSchool puts it in her (great) presentation "🐂💩 people believe about digitization and digital preservation," "there is no magic metadata fairy"!

Module 9: XSD

Module 10: Domain-Specific Metadata

Post

The Lee et al. pieces were interesting, partly because there's some overlap with / analogy to my topic (tabletop role-playing game systems), but also because of the way they detailed their process. I thought the user personas were interesting, and in general liked the aspects that looked at the purpose behind the schema, the ways it might be used. There's a tension between trying to design a Platonic ideal schema that captures everything meaningful about a topic in the most accurate and specific possible way, and designing a schema useful for a particular task, which might not be as comprehensive but might be more workable and more flexible. Some of the difficulties they encountered, like distinguishing developer/publisher/distributor (2013, p. 236) or capturing credits consistently (2015, p. 2619) also might be avoided by taking a less specific, less structured approach, like Dublin Core's creator and contributor elements.

Hu and Downie's work is illustrative here—their mood terms include incommensurable qualities like "literate", "autumnal", "hypnotic", and "street smart", which an ideal system might want to decompose into specifically sonic or specifically lyrical elements, but the gestalt better captures how music experienced by listeners—not necessarily as a set of separable components. I did wonder about the quality of the data that Hu and Downie were working with, and to what extent their statistical methods and cross-testing against different data sources would mitigate it. My personal experience corroborates Daehl—most music metadata is terrible. (Apple seems to be on a longterm project to gradually replace all the album art in my iTunes collection with the cover of the *Fifty Shades of Gray* soundtrack, which I don't even own.)

Daehl of course also points to the drawbacks of taking metadata mostly developed for one purpose (allowing users to select and play or buy music) and applying it to another (distribution of streaming revenue). It doesn't help in this case that the people / companies developing and maintaining the metadata, the streaming companies and distributors, have no incentive to get it right—and in fact have an incentive to get it wrong, insofar as money left on the floor because the people who should be getting it can't be identified is money they can pick up.

Response (Erin)

Hi Erin,

I didn't focus on the "The Algorithm Sees All, The Algorithm Knows You Better Than You Know Yourself" aspect of the Netflix pieces but you're absolutely right, the subtext there, "this is objective, because we have *daaaaaaata"* is pure 🐴💩. Your choices are shaped by what they put in front of you, and what they put in front of you is shaped by what they've put in front of you in the past and by their own opaque internal incentives to put different things in different places, whether it's pushing their own content to justify the money they're spending on it, or pushing you away from watching content they pay for by the stream, or not having the UI development budget to build an interface that could efficiently show you more than like six rows of twelve movies each, or not having the metadata budget to slice movies up into more than six broad genres. So the Algorithm is eating itself, basically. And I don't know if Netflix knows that, and just doesn't care, or if they're high on their own supply. (My guess is it's a mix of both.)

I can imagine a streaming service that would help me reach my aspirational goals of catching up with Kurosawa's non-samurai work or watching every woman-directed documentary on the National Film Registry or whatever. Unfortunately, it's harder to imagine a streaming service that would actually have an economic incentive to do it. 😒

Module 12: Metadata Quality

Post

I’m not sufficiently familiar with the social science data cleaning practices mentioned by Rawson & Muñoz (2016) to judge whether their methods really are sufficiently well known to pass without comment, but I share the humanities scholar’s suspicion of any procedure that presents processed data without great clarity about exactly how it was processed. Of course there are fields like astronomy, or high-energy particle physics, or some kinds of medical imaging, where it would simply be impractical to store the volumes of raw data the instruments produce, let alone present them, but there the data reduction mechanisms are relatively standardized and relatively well understood.

For the humanities, sheer data volume seems much less likely to be an issue, especially for text data. I agree with Rawson & Muñoz that it’s not right to look at “clean” textual research data as simply a purer form of the original, that it needs to be considered a work product, part of the analysis and generally suitable only for the purpose of that analysis. And it’s not just my humanities background that has me saying that, it’s also dealing professionally with some of the metadata quality problems discussed by Beall (2005) and by Kelly et al. (2005).

In my experience, dealing with these problems properly—especially dealing with them in bulk, in an automated way—always takes time, care, and expertise; and doing so destructively is something that shouldn’t be attempted without many checks and balances. It’s all too easy to imagine the lone humanities scholar losing data—or worse, fabricating data—unknowingly and by accident, in the course of “cleaning”. In which case keeping the original data, and documenting the cleaning process, are both critical for reproducibility.

Response (Carolyn Quimby)

Hi Carolyn,

I like the connection you make between these two articles—in both the biomedical and humanities situations, the researcher who’s head-down in the data probably feels that whatever they’re not telling the reader goes without saying, whether it’s missing geospatial metadata or just how much variation was lost in collapsing all the permutations of “potatoes au gratin.” Of course what they leave out is often something they don’t see as relevant to their particular research question, but leaving it out makes it hard to reproduce their work or assess its validity, as well as making it hard to build on their results in ways they may not have expected. As Schriml et al. note, data publication has become a big deal over the past few years, with funders in many fields starting to insist on data management plans and publication in trusted repositories (this was a big piece of what I worked on at the California Digital Library), but I think we're still only at the very early stages of figuring out what it really means to do a good job of it—the data equivalent of the days when "scientific publication" meant writing a letter to Robert Boyle to tell him "Guess what, I just saw the weirdest cow fetus."

Project

Project Reflection

I appreciated having to think through designing a schema and incorporating elements from other existing schemas, rather than just throwing something together to make a system work. Probably the hardest thing for me was keeping the project scope reasonable, reminding myself that this was a student project that I was building for a semester, not something that was supposed to be globally useful and last indefinitely. If I wanted to do it right, not only would I want to do a much more detailed analysis of all the sample game systems I looked at, and take a closer look at things like schemas for describing video game rules and play (and maybe other analogous domains I didn't get to at all—sports metadata, maybe?), I'd want to identify more user populations and do a lot of workshopping and iteration. (I'd also probably take the time to write a program—or find one somebody else has written—to read a documented schema and generate a data dictionary, because doing it by hand was *just* this side of so annoying I was ready to automate it!)

Response (Alexander McQuade)

Hi Alexander,

Nice work! I'm not up on 5th Edition yet, but I like that you have monsters and PCs covered by the same schema. Really, I think RPG stat blocks are already metadata, just not (usually) in a machine-readable format), so this feels very natural—I imagine the hardest thing was knowing where to stop, as you could easily incorporate everything in this screenshot and more. On the other hand, there's a good argument for just including the basics that your intended users (say, DM trying to pick monsters) would need when skimming, and then link out to the details, especially when you have characters (esp.

monsters and NPCs) that have appeared multiple times in multiple editions with different stats according to different rulesets.

Also, just a random thought, but I was browsing the Getty Art and Architecture Thesaurus the other day because reasons, and it occurred to me that the object genres hierarchy really reads a lot like a D&D treasure table. So if you wanted to have some fun with it that's something you could link out to.