CUL Metadata Application Profiles

This is an overview of metadata application profiles and related documentation used within the Cornell University Libraries, generated originally from a MWG Working Session in 2016/2017. The goal is to get a high-level overview of metadata profiles and needs at CUL, which we hope will lead us to see collaboration opportunities. We hope to have another session in the fall of 2017 to see where this overview stands and update/expand it.

What are Metadata Application Profiles (MAPs)?

Wikipedia gives a decent stub on Metadata Application Profiles: https://en.wikipedia.org/wiki/Application_profile Metadata Application Profiles are metadata specifications attached (sometimes loosely, sometimes tightly) to a particular application or metadata service - whether it is a datastore, repository, management system, discovery indexing layer, or other. It helps communicate expectations of the metadata being ingested, processed, managed and exposed by that particular application or service. MAPs are the documentation that connect metadata implementations to shared community models and standards, as well as document where implementers need to diverge from community standards. This makes it easier for outsiders to understand and work with metadata from or headed to your application or system.

Metadata Application Profiles can touch on descriptive, technical, administrative, structural or other (or a mix of all of the above) metadata. They can rely heavily on community standards, but good MAPs don't just copy a community standard over. This is due to the fact that, in implementation, there are points your data needs to diverge or to further specify a datapoint's usage. These MAPs can also be made machine-actionable, with the enabled action being validation of your data against a profile, guiding the creation of data that follows your profile, transforming your data to/from another profile or standard. A popular metadata application profile example is the Digital Public Library of America Profile: https://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf which has a machine-actionable representation here: https://github.com/dpla/dpla_map.

What you'll most often see in cultural heritage institutions' metadata management are MAPs represented via spreadsheets or some kind of text documentation with tables. Depending on the data representation (RDF, XML, CSV, JSON, Arvo, (X)HTML, MARC, ...) you are ultimately working with, you might have some technology available for specifying machine-actionable MAPs. A common example is a XSD file which clarifies or validates that metadata is following the MAP specified for your namespace or XML dataset. Another example is the emergence of RDF shapes technologies (ShACL, ShEx, RML) or application-focused RDF-based object management libraries (ActiveTriples / ActiveFedora) for both checking the state of metadata against a specified MAP or converting to/from that MAP.

CUL-Used or Important Community-Used MAPS

CUL MAPs

> ⓘ FYI: These are provided not to guide policy or implementation decisions, but rather to share metadata efforts across CUL for the purpose of increasing the level of common understanding of metadata work and helping highlight collaboration, metadata infrastructure, or cross-pollination opportunities.

Repository or Unit	Data Representation	Resource Type(s) Described	Human-Readable MAP	Machine-Actionable MAP
Digital Portal Hydra PCDM	RDF	Collections, Cultural Heritage Objects, Digital Surrogates, Filesets, Files	(static version from last September that's easier to share than the version currently being used): https://docs.google.com/spreadsheets/d/1z-dddDvijPIa84dml76YTenSJ_gMnZZ0zR3JnHSZVPc/edit?usp=sharing	being built
eCommons	XML, some CSV to XML	Scholarly Output Objects, Digital Surrogates, Collections	https://cornell.box.com/s/7eweceb0dxhbgq3ja5yko69ednbeefb4	being built
SharedShelf	JSON	Visual Resource Objects, Files, Collections, Agents, Events	https://github.com/cul-it/sharedshelf-metadata (being built, probably not public)	being built
ETDs Uber Records	MySQL Data, XML for one representation	ETDs, ETD Agents, ETD Subjects, ETD Binaries	MySQL: https://docs.google.com/spreadsheets/d/1L_06uqiDL5AeBSMm5aDC2BkA-GNABr2_E8iGCaiKyPY/edit#gid=0 Example 1 one MAP fork: eCommons XML: ETD Metadata Profile	being built
Kheel (Bepress & SharedShelf)	CSV to XML	Cultural Heritage Objects, Digital Surrogates, Collections	https://docs.google.com/spreadsheets/d/1X50d8-pnOZ35mU-52mMTVWfWGAgCxoWBW4M4Ih-dlho/edit?usp=sharing	n/a

CULAR (F3)	XML largely?	Preservation Collections, Binaries	Data Model (not really a profile): CULAR Data Model Some parts are here: https://github.com/cmh2166/CULAR	n/a
Embedded Digitization Lab Binaries Metadata	Embedded data in header of digital assets	Primarily technical metadata about the Digital Asset/Binary, as well as administrative metadata for the original analog resource or collection.	https://gist.github.com/cmh2166/84e30b81227e2e5b47f0f51d71e8d9db (in process)	n/a (would be profiles loaded into the various tools used for digitization and binary management)
Vitro Authorities Pilot	RDF	Agents, Topics, Places, Events, Authorities	https://github.com/cul-it/lts-vitro-pilot/wiki (on hold)	n/a
LD4* Ontology (BF2)	RDF	Bibliographic Resources (Work, Instance, Items), Agents, Subjects, Contributions, ...	https://docs.google.com/spreadsheets/d/14ccalbnpr8qhh0O7M43n6vv3xbyE1ydJdJjnQaVDht4/edit?usp=sharing (largely unstarted) http://bibframe.org/bibliomata/profile-edit/#/profile/list (LC's BF Profiles can be seen in their Profile Editor) TODO: add link to Sinopia Profile Editor when it is available	https://github.com/lcnetdev/verso/tree/5a444f7cdd203fbf25563098e18f84302bdd2e6a/data/profiles (LC's BF Editor profiles in json config files) https://github.com/LD4P/HipHop/tree/master/application-profiles (Cornell Hip Hop LPs application profile in SHACL) https://github.com/LD4P/arm/tree/master/application_profiles/raremat_monograph/shacl (ARM extension application profile in SHACL)
KMODDL		1) three-dimensional models designed for demonstration and/or teaching purposes; 2) stereolithography files for creating 3-D replicas of the models; 3) still, moving, and interactive images of the models; 4) tutorials aiding in the use of KMODDL materials; and 5) related textual resources including books and articles.	http://wayback.archive-it.org/2566/20180418122322/http://kmoddl.library.cornell.edu/aboutmeta2.php	no
... (add yours here)				

Community MAPs

Institution	Data Representation	Resource Type(s)	Human-Readable MAP	Machine-Actionable MAP
Digital Public Library of America	RDF	Cultural Heritage Objects, Digital Representations, Aggregations	https://pro.dp.la/hubs/metadata-application-profile	https://github.com/dpla/dpla_map (ActiveTriples representation of the MAP, which can guide Ruby applications in creating resources that follow the DPLA MAP.
Europeana	RDF	Cultural Heritage Objects, Digital Representations, Aggregations... and more (works based off of CIDOC-CRM	More generic metadata documentation: http://pro.europeana.eu/page/edm-documentation	?
Sufia / Scholarsphere	RDF, Ruby Objects		See here for a static copy: https://docs.google.com/spreadsheets/d/1FL15HSy0d_Mb6I3r7vjH8yEMkkRCJJkiqOt42Khl8ss/edit?usp=sharing	
NYPL Digital Collections	RDF		See here for a static copy: https://docs.google.com/spreadsheets/d/1FL15HSy0d_Mb6I3r7vjH8yEMkkRCJJkiqOt42Khl8ss/edit?usp=sharing	

Creating a MAP - Guidelines and a Generic Template

First Steps

A MAP should document and specify the expectations of metadata in an application, service, system, or other. This means the first step to creating a MAP is to understand what it is your metadata is attempting to describe or capture at a conceptually level. Here are some questions to help guide creating your MAP:

1. **What are you describing with this metadata**? To what level of conceptual difference do you need to go for your MAP to be accurate and complete? An example: eCommons is an Institutional Repository, which means it tends to manage scholarly output objects. Capturing what type of scholarly output is a required metadata field, but doesn't change (at the moment) the system's management of a scholarly output object's metadata record (i.e., a Presentation eCommons metadata record will not diverge in structure or requirements from an Article eCommons record - the same fields are in play, albeit some will make more sense for different objects). Once you decide this diverges **enough**, you can create forked (but related) MAPs, with a view for system interoperability. This is more determined by needs than any hard and fast rules - i.e., its an art, not a science.
2. **What do you intend to do with this metadata**?

a. **Share with or generate from other systems**? Then you need to make sure shared concepts/fields are captured in both MAPs. Standards can be very helpful here.
b. **Enable some sort of discovery, lookup, resource management, or other functionality**? Clarify this functionality's expectations in the MAPs, particularly the fields that support that work.
c. **Use within a particular system**? Make sure your building your MAP to the abilities of that system, and clarifying limitations where they apply.

3. **How will this metadata be generated, managed, and exposed**? By whom or what processes?
   a. Generation can lead to understanding data expectations and sources, as well as any meta-metadata you might want to capture.
   b. Management helps clarify expectations and understanding of the metadata within the system, as well as guide any enhancement or update work you do to that existing metadata.
   c. Metadata exposure in the format "of record" of that system is **vital**. This is how you can assess metadata against your profile. This is how you can leverage the MAP in play. This is how you can perform analysis to understand what your MAP covers (or doesn't).

Generic Template

This is a generic MAP template that may be helpful. All the columns aren't needed for all cases, but those with stars beside them are strongly recommended for your MAP. You can make a copy of this template from this Google Spreadsheet.

Field*	Schema Mapping*	Domain	Expected Value or Range*	Definition	Obligation*	Usage Notes	Source	Remediation Notes	Exposure or Other Representation Element
This is the name of the field for ease of referring to it in documentation, communication, etc.	This field should represent where the field is mapped to in the metadata records in the application, system, or service. It should also be the field mapped to a shared standard, namespace, or specification. This helps clarify the understanding of what the field means, as well as facilitates mapping that field to other metadata and MAPs.	The expect resource or object type this metadata field is asserted against. Required if you're not going to split MAPs but need to specify fields that apply to only certain types of described resources.	This is the expected metadata value for this field. Here you can specify data types (string, integer, datetime, etc.), specify value sources (controlled vocabularies, authorities, free-text entry, other), and any other specifications around the expected metadata (is identifier for an Agent resource; is a date following EDTF; is a Cornell email address; etc.)	This is the definition for the field. This is helpful if it doesn't map (or doesn't map entirely) to a Namespaced Mapping or a Standard.	Indicates if the field is required and repeatable in the format: {number of times expected, number of times can be present} i.e. {0,1} == Not required (can appear 0 times), Not repeatable (can only appear at most 1 time) {1,1} == Required (must appear 1 time), Not repeatable (can only appear at most 1 time) {0,n} == Not required, repeatable (can appear at most n times) {1,n} == Required, repeatable. You can also break this out into "Required?" and "Repeatable" columns with TRUE/FALSE values if easier.	Any notes on using this field in the metadata generated.	The source of the field - transformed from existing data in XYZ store or format, user-entered in the application, pulled from a shared database, etc.	Any notes on cleanup, normalization, or enhancement of this field.	This is the element or place this field gets captured in alternate exposure points, i.e., here is the element this field is presented as in the application's OAI-PMH stream. Here is the element the field appears in a backup copy of the metadata. Here is the field as the it appears in Solr. Etc.