Metadata Application Profile for the Dataset Metadata Model (DATMM MAP)

NLM Working Group
Jeff Beck
Nancy Fallgren
Peter Seibert (Chair)
Alvin Stockdale

Revised:
January 4, 2024

Original:
June 9, 2020

General Introduction

"Big Data presents an exciting opportunity to pursue large-scale analyses over collections of data in order to uncover valuable insights across a myriad of fields and disciplines. Yet, as more and more data is made available, researchers are finding it increasingly difficult to discover and reuse these data. One problem is that ***data are insufficiently described*** *to understand what they are or how they were produced. A second issue is that* ***no single vocabulary provides all key metadata fields*** *required to support basic scientific use cases. A third issue is that* ***data catalogs and data repositories all use different metadata standards****, if they use any standard at all, and this prevents easy search and aggregation of data. Therefore, we need a guide to indicate what are the essential metadata for a dataset description, and the manner in which we can express it."* [14]

Background

The National Library of Medicine (NLM) determined to explore metadata schema and a prototype platform toward potential development of a national catalog of biomedical datasets to facilitate finding, sharing, and re-use of these valuable research resources. The charge specified that the metadata scheme should adhere to the FAIR principles [13]:

1. Findable
2. Accessible
3. Interoperable
4. Re-usable

The Dataset Metadata Modeling Working Group (DATMM WG), charged with this exploratory task, decided that a Web-friendly semantic model based on the W3C's Resource Description Framework (RDF) [20] would best fulfill the FAIR requirements and meet NLM's needs.

Based on its combined experience with metadata model implementation, the DATMM WG's Metadata Modeling subgroup (Metadata SG) agreed that to achieve adoption, the model would need to be simple and brief, while still allowing for meaningful discovery. The biomedical community encompasses a broad range of disciplines, each with its own terminology and research foci; hence, a single metadata model that seeks to provide granular discovery in each of these disciplines is doomed to be unwieldy and of little practical use. Therefore, the Metadata SG chose to follow a model based on a usability theory similar to bibliographic catalogs: users should have sufficient descriptive data to find datasets of potential interest across disciplines and shall be provided with links to the dataset source repository for a more granular determination of pertinence and usability. This resembles the working model adopted by the Digital Public Library of America (DPLA) to "aggregate metadata only, primarily to power a dataset for discovery, passing users back to the source content at our Hub Partner's repository." [11, p.1]

The Metadata SG determined that the focus of its model should be the datasets themselves. All other resources in the model (e.g., agents, contextual documentation, repositories, etc.) exist

only insofar as they have a relationship to a described biomedical dataset. As defining a "dataset" is no simple exercise, it was ultimately agreed that original research datasets should comprise the scope of the model. Therefore, "datasets" for this project do not include standard descriptive terminologies, such as NCBI Taxonomy, MeSH, and SNOMED CT, or sites that gather and display factual information from multiple sources, creating a set of data around a topic.

DATMM Metadata Application Profile (DATMM MAP)

Overview

The Metadata SG reviewed RDF schema with a specific focus on describing datasets e.g., schema.org [19], Data Catalog Vocabulary (DCAT) [1], Dataset Tag Suite (DATS) [25], and the W3C Semantic Web for Health Care and Life Sciences (HCLS) Interest Group Community Profile [11]; however, none of the extant schema entirely met our perceived needs. Nevertheless, in keeping with both the FAIR principle of re-use and the spirit of RDF to promote re-use of existing schema, the Metadata SG was determined not to reinvent the wheel but rather to adopt parts of existing efforts to the extent possible. Therefore, the SG created a metadata application profile (MAP) for high level description of biomedical datasets using parts of existing schema, rather than creating an entirely new RDF schema.

The DATMM MAP is designed to provide high level search results across a broad biomedical landscape, with an expectation that either the dataset itself or more granular description in the source repository will provide sufficient detailed information to assist users with selection for use. The field of information retrieval differentiates between precision (finding only relevant results) and recall (finding all potentially relevant results) in search systems. The DATMM MAP seeks to provide search results that represent broad recall, but not necessarily precision.

Definitions of MAPs range from very generic to platform specific (see Appendix B for examples). The Metadata SG follows the generic definition of a MAP as

> a mixing of only those relevant properties from different (and perhaps diverse) standard metadata schemas, combined for the purpose of describing resources (records, documents, etc.) in a particular context. [24]

At the same time, the Metadata SG closely aligns its DATMM MAP with the platform specific definition of the DPLA MAP, described as

> a set of metadata elements, taken from multiple schemas for a particular local use. It is also a semantic metadata model, or an abstract structure that describes the relationships between different types of data about the same thing. This means it is more robust and abstract than a metadata schema like Dublin Core or MODS in that it describes entities and the relationships between them. [11, p.1]

In accordance with these definitions, the DATMM MAP defines only 3 new RDF Classes (datmm:Dataset, datmm:Repository, and datmm:Documentation), while mainly re-using Classes

and properties from other schema. The following table indicates the namespaces used in the DATMM MAP:

Description	Prefix	URI
BIBFRAME 2.0	bf:	http://id.loc.gov/ontologies/bibframe/
Dataset Metadata Model (DATMM)	datmm:	http://id.nlm.nih.gov/datmm/
DCMI Metadata Terms	dct:	http://purl.org/dc/terms/
DCMI Type Vocabulary	dcmitype:	http://purl.org/dc/terms/DCMIType/
Friend-of-a-Friend (FOAF)	foaf:	http://xmlns.com/foaf/0.1/
Provenance Authoring and Versioning Ontology (PAV)	pav:	http://purl.org/pav/
RDF Schema	rdfs:	http://www.w3.org/2000/01/rdf-schema#
RDF Syntax	rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
Schema.org	schema:	http://schema.org/
Simple Knowledge Organization System (SKOS)	skos:	http://www.w3.org/2004/02/skos/core#

As stated in the introductory quote, a major stumbling block to finding, sharing, and re-using datasets is that there may be no extant metadata, whether in an established standard or a local scheme, to describe them. It is the WG's hope that biomedical dataset repositories might adopt the DATMM MAP as a core set of minimum viable descriptive metadata and then extend or augment it locally to meet the more specific descriptive domain of their general biomedical disciplines or specific areas of research. Domain specific extension of the DATMM MAP is a more viable model for providing the descriptive detail that researchers need to make informed decisions about the pertinence of a given dataset for re-use or further study. Repository adoption of the DATMM MAP would also facilitate participation in the NLM Dataset Catalog.

Structure

The DATMM MAP is based on RDF, the semantic framework behind Linked Open Data. RDF is a precise descriptive structure composed of simple statements called "triples". Triples are akin to simple, grammatical statements posed as "Subject predicate Object", e.g., Article1 has author Author1. Each triple must be true and must be complete in and of itself, i.e., it can be understood as a standalone statement without other context.

Conceptually, RDF organizes the Things or Resources in the world into generic Classes, e.g., the Class of Datasets, the Class of People, et al., while each individual in a Class is a specific Instance of that Class. Relating that to triples, Subjects are generally one Instance of a Class, e.g., a specific dataset with a unique identifier. Predicates, generally called "properties" in RDF, are the verbs that relate Subjects to Objects. Objects are property values, i.e., data values which may be a single Instance of a Class or a literal string value (such as a scopenote or a

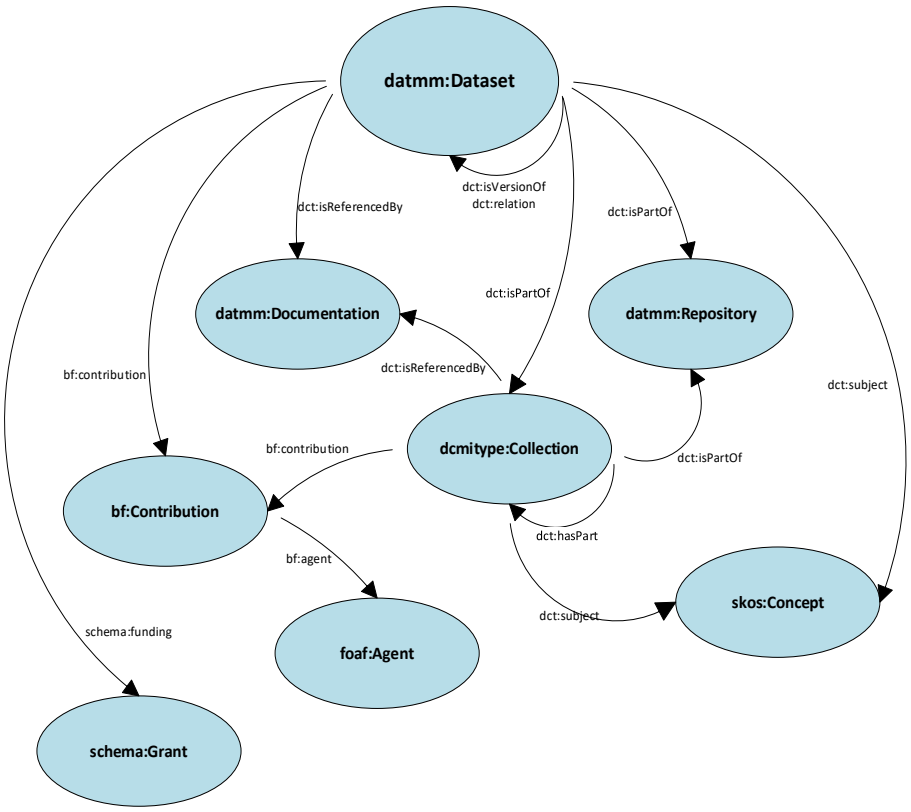
date) or a node for grouping multiple values, such as aggregating discrete components of a subject string.

The focal entity of the DATMM MAP is the Dataset Class. Every other Class in the MAP exists only insofar as it relates to a dataset. Since the DATMM MAP is designed for use in a catalog describing biomedical datasets, it does not seek to describe other dataset catalogs or to provide more than cursory information about source repositories in which the described datasets reside. The DATMM MAP is intended to facilitate finding and using biomedical datasets by providing high level description of the dataset, a link to the dataset, related subjects terms, a link to the home repository, references to contextual documentation, access rights, etc.

The following table defines the top level Classes in the DATMM MAP:

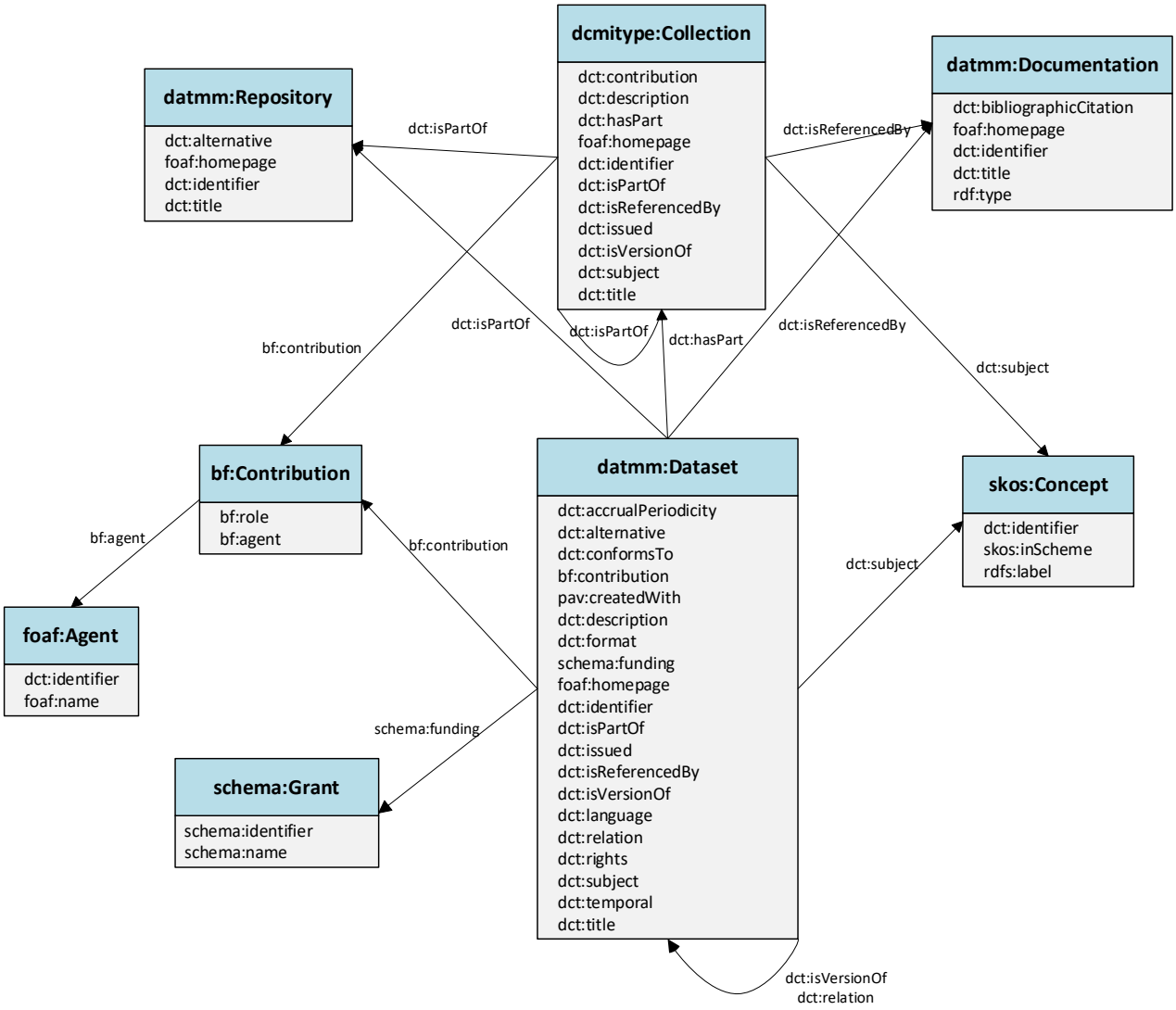
Class Name	Source	Definition
Agent	FOAF	An agent (eg. person, group, software or physical artifact); things that do stuff
Collection	DCMIType	An aggregation of resources.
Concept	SKOS	An idea or notion; a unit of thought
Contribution	BIBFRAME	Agent and role with respect to the resource being described
Dataset	DATMM	A discrete collection of data gathered for use in research
Documentation	DATMM	Contextual documents about the dataset
Grant	Schema.org	A grant, typically financial or otherwise quantifiable, of resources.
Repository	DATMM	A network storage resource from which datasets are accessible

Figure 1. DATMM MAP model of top level Classes and their relationships to each other



Within each Class, the Metadata SG limited properties to data that can be reasonably expected to be known or easily found and recorded, i.e., not onerous to provide. As part of prototype testing and beta implementations, the Metadata SG expects these properties to be refined and/or augmented, while maintaining the underlying premises of simplicity and high level description.

Figure 2. DATMM MAP visualization showing Classes and properties



Serialization and Validation

A human readable spreadsheet rendering of the DATMM MAP is available in Appendix A. The DATMM MAP is available in RDF at JSON-LD Links/DATMM RDF on [DATMM Quick Links - Smartsheetgov.com]

Acknowledgements –

The original DATMM MAP resulted from the work of the National Library of Medicine's DATaset Metadata Modeling (DATMM) Working Group in 2020. Members of that group included Jeff Beck, Philip Chuang, Daniel Davis, Nancy Fallgren, Lisa Federer, Robert Guzman, Karen Nimerick, Peter Seibert (Chair), Alvin Stockdale, Ying Sun, Dianne Babski (Project Sponsor), and Michael Huerta (Project Sponsor).

Appendix A – DATMM Metadata Application Profile

datmm:Dataset -- A discrete collection of data gathered for use in research

Property	Definition	Constraint	Cardinality	Object Type
dct:accrualPeriodicity	The frequency with which items are added to a collection.	Optional	1:1	URI from standard vocabulary or Literal
dct:alternative	An alternative name for the resource.	Optional	1:N	Literal
dct:conformsTo	An established standard to which the described resource conforms.	Optional	1:N	Literal
bf:contribution	Agent and its role in relation to the resource.	Required	1:N	DATMM Resource/ bf:Contribution
pav:createdWith	The software/tool used by the creator when making the digital resource, for instance a word processor or an annotation tool.	Optional	1:N	Literal
dct:description	An account of the resource.	Required	1:1	Literal
dct:format	The file format, physical medium, or dimensions of the resource.	Optional	1:N	URI from standard vocabulary or Literal
foaf:homepage	A homepage for some thing.	Required	1:1	URL
dct:identifier	An unambiguous reference to the resource within a given context.	Optional	1:N	URI
dct:isPartOf	A related resource in which the described resource is physically or logically included.	Optional	1:N	DATMM Resource/ datmm:Repository AND dct:Collection
dct:issued	Date of formal issuance of the resource.	Optional	1:1	Literal
dct:isReferencedBy	A related resource that references, cites, or otherwise points to the described resource.	Optional	1:N	DATMM Resource/ datmm:Documentation
schema:funding	A Grant that directly or indirectly provide funding or sponsorship for this item.	Optional	1:N	DATMM Resource/ schema:Grant
dct:isVersionOf	A related resource of which the described resource is a version, edition, or adaptation.	Optional	1:N	DATMM Resource/ datmm:Dataset
dct:language	A language of the resource.	Optional	1:N	URI from standard vocabulary or Literal
dct:relation	A related resource.	Optional	1:N	DATMM Resource/ datmm:Dataset
dct:rights	Information about rights held in and over the resource.	Optional	1:N	Literal
dct:subject	A topic of the resource.	Required	1:N	DATMM Resource/ skos:Concept
dct:temporal	Temporal characteristics of the resource./Temporal coverage.	Optional	1:1	Literal
dct:title	A name given to the resource.	Required	1:1	Literal

dcmitype:Collection -- An aggregation of resources

Property	Definition	Constraint	Cardinality	Object Type
bf:contribution	Agent and its role in relation to the resource.	Optional	1:N	DATMM Resource/ bf:Contribution
dct:description	An account of the resource.	Required	1:1	Literal
dct:hasPart	A related resource that is included either physically or logically in the described resource.	Required	1:N	DATMM Resource/ dct:Collection
foaf:homepage	A homepage for some thing.	Required	1:1	URL
dct:identifier	An unambiguous reference to the resource within a given context.	Optional	1:N	URI
dct:isPartOf	A related resource in which the described resource is physically or logically included.	Optional	1:N	DATMM Resource/ datmm:Repository
dct:isReferencedBy	A related resource that references, cites, or otherwise points to the described resource.	Optional	1:N	DATMM Resource/ datmm:Documentation
dct:issued	Date of formal issuance of the resource.	Optional	1:1	Literal
dct:isVersionOf	A related resource of which the described resource is a version, edition, or adaptation.	Optional	1:N	DATMM Resource/ datmm:Collection

dct:subject	A topic of the resource.	Optional	1:N	DATMM Resource/ skos:Concept
dct:title	A name given to the resource.	Required	1:1	Literal

foaf:Agent -- An agent (eg. person, group, software or physical artifact); Things that do stuff

Property	Definition	Constraint	Cardinality	Object Type
dct:identifier	An unambiguous reference to the resource within a given context.	Optional	1:N	URI
foaf:name	The foaf:name of something is a simple textual string.	Required	1:1	Literal

skos:Concept -- An idea or notion; a unit of thought [what the dataset is about]

Property	Definition	Constraint	Cardinality	Object Type
dct:identifier	An unambiguous reference to the resource within a given context.	Optional	1:1	URI
skos:inScheme	[Concept] is in scheme [aka scheme name]	Required	1:1	Literal
rdfs:label	Used to provide a human-readable version of a resource's name.	Required	1:1	Literal

bf:Contribution -- Agent and role with respect to the resource being described

Property	Definition	Constraint	Cardinality	Object Type
bf:agent	Entity associated with a resource or element of description, such as the name of the entity responsible for the content or of the publication, printing, distribution, issue, release or production of a resource.	Required	1:1	DATMM Resource/ foaf:Agent
bf:role	Function provided by a contributor, e.g., author, illustrator, etc.	Optional	1:N	URI or Literal

datmm:Documentation - Contextual documentation about the Dataset, e.g., articles, research study, grant application, clinical trial, research instruments (CDEs), et al. It is assumed that any referenced Documentation will be publicly available in some format.

Property	Definition	Constraint	Cardinality	Object Type
rdf:type	rdf:type is used to state that a resource is an instance of a class.	Optional	1:1	Resource/ subClass from Mesh Publication Characteristics or Literal
dct:bibliographicCitation	A bibliographic reference for the resource.	Optional	1:1	Literal
foaf:homepage	A homepage for some thing.	Optional	1:1	URL
dct:identifier	An unambiguous reference to the resource within a given context.	Optional	1:N	URI
dct:title	A name given to the resource.	Optional	1:1	Literal

schema:Grant -- A grant, typically financial or otherwise quantifiable, of resources.

Property	Definition	Constraint	Cardinality	Object Type
schema:identifier	The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc.	Optional	1:1	Literal
schema:name	The name of the item.	Optional	1:N	Literal

datmm:Repository - A network storage resource from which datasets are accessible

Property	Definition	Constraint	Cardinality	Object Type
dct:alternative	An alternative name for the resource.	Optional	1:N	Literal
foaf:homepage	A homepage for some thing.	Required	1:1	URL

dct:identifier	An unambiguous reference to the resource within a given context.	Optional	1:1	URI
dct:title	A name given to the resource.	Required	1:1	Literal

Appendix B: Some Metadata Application Profile Definitions

A DCAP [Dublin Core Application Profile] defines metadata records which meet specific application needs while providing semantic interoperability with other applications on the basis of globally defined vocabularies and models. [7]

We can think of a metadata application profile as a mixing of only those relevant properties from different (and perhaps diverse) standard metadata schemas, combined for the purpose of describing resources (records, documents, etc.) in a particular context. [24]

DPLA's MAP is an application profile, or a set of metadata elements, taken from multiple schemas for a particular local use. It is also a semantic metadata model, or an abstract structure that describes the relationships between different types of data about the same thing. This means it is more robust and abstract than a metadata schema like Dublin Core or MODS in that it describes entities and the relationships between them. [11]

In the information sciences, an application profile consists of a set of metadata elements, policies, and guidelines defined for a particular application.
The elements may come from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata from several element sets - including locally defined sets. [3]

A metadata application profile (MAP) is a set of recorded decisions about a shared data target for a given community. MAPs declare what models are employed (what types of entities will be described and how they relate to each other), what controlled vocabularies are used, the cardinality of fields/properties (what fields are required and which fields have a cap on the number of times they can be used), data types for string values, and guiding text/scope notes for consistent use of fields/properties. A MAP may be a multipart specification, with human-readable and machine-readable aspects, sometimes in a single file, sometimes in multiple files (e.g., a human-readable file that may include input rules, a machine-readable vocabulary, and a validation schema). [18]

Bibliography

[1] Albertoni, R., Browning, D., Cox, S., Beltran, A. G., Perego, A., & Winstanley, P. (Eds.). (2020, February 4). Data Catalog Vocabulary (DCAT) - Version 2. Retrieved May 29, 2020, from https://www.w3.org/TR/vocab-dcat-2/

[2] Allemang, D., & Hendler, J. (2011). *Semantic web for the working ontologist: modeling in RDFS and OWL* (2nd ed.). Amsterdam: Morgan Kaufmann Publishers/Elsevier.

[3] Application profile. (2019, October 2). In *Wikipedia*. Retrieved June 5, 2020, from https://en.wikipedia.org/wiki/Application_profile

[4] Brickley, D., & Miller, L. (2014, January 14). FOAF Vocabulary Specification 0.99. Retrieved May 29, 2020, from http://xmlns.com/foaf/spec/

[5] CEDAR. (n.d.). Retrieved May 29, 2020, from https://metadatacenter.org/

[6] Ciccarese, P., & Soiland-Reyes, S. (2014, August 28). PAV - Provenance, Authoring and Versioning. Retrieved May 29, 2020, from https://pav-ontology.github.io/pav/

[7] Coyle, K., & Baker, T. (2009, May 18). Guidelines for Dublin Core™ Application Profiles. Retrieved May 29, 2020, from https://www.dublincore.org/specifications/dublin-core/profile-guidelines/

[8] Coyle, K. (2018, June 14). Webinar: Introduction to Metadata Application Profiles. Retrieved May 29, 2020, from https://www.dublincore.org/webinars/2018/introduction_to_metadata_application_profiles/

[9] Devey, M., & Côté, M.-C. (2006). The Development and Use of Metadata Application Profiles. *The Serials Librarian*, *51*(2), 103–115. doi: 10.1300/j123v51n02_08

[10] DPLA Metadata Working Group (2017, December 7). Digital Public Library of America Metadata Application Profile, version 5.0. Retrieved May 29, 2020, from https://drive.google.com/file/d/1fJEWhnYy5Ch7_ef_-V48-FAViA72OieG/view

[11] DPLA Metadata Working Group (2017, December 7). Introduction to the DPLA Metadata Application Profile, version 5.0. Retrieved June 5, 2020, from https://drive.google.com/file/d/1kMxXgFrGwu3i7LBLFOj6VZRQFuQzqkHk/view

[12] Dublin Core™ Metadata Initiative. (n.d.). Retrieved May 29, 2020, from https://dublincore.org/

[13] FAIR Principles. (2019, April 03). Retrieved June 05, 2020, from https://www.go-fair.org/fair-principles/

[14] Gray, A. J. G., Baran, J., Marshall, M. S., & Dumontier, M. (Eds.). (2015, May 14). Dataset Descriptions: HCLS Community Profile. Retrieved May 29, 2020, from https://www.w3.org/TR/hcls-dataset/

[15] Hitzler, P., Krötzsch, M., Parsia, B., & Rudolph, S. (Eds.). (2012, December 11). OWL 2 Web Ontology Language Primer (Second Edition). Retrieved June 05, 2020, from https://www.w3.org/TR/2012/REC-owl2-primer-20121211/

[16] Library of Congress. (n.d.). BIBFRAME - Bibliographic Framework Initiative. Retrieved May 29, 2020, from https://www.loc.gov/bibframe/

[17] Miles, A., & Bechhofer, S. (2008, August 20). SKOS Simple Knowledge Organization System RDF Schema. Retrieved May 29, 2020, from https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html

[18] PCC Task Group on Metadata Application Profiles. (2019, April 30). Retrieved May 29, 2020, from https://www.loc.gov/aba/pcc/taskgroup/Metadata-Application-Profiles.html

[19] schema.org. (n.d.). Retrieved May 29, 2020, from https://schema.org/

[20] Schreiber, G., & Raimond, Y. (Eds.). (2014, June 24). RDF 1.1 Primer. Retrieved May 29, 2020, from https://www.w3.org/TR/rdf11-primer/

[21] Tennis, J. T. (2015). Metadata Application Profiles. In *Encyclopedia of Archival Science*. Retrieved from https://www.researchgate.net/publication/326786022_Metadata_Application_Profiles

[22] WG3- DATS Model - Metadata Specifications. (n.d.). Retrieved May 29, 2020, from https://github.com/biocaddie/WG3-MetadataSpecifications