Appendices for Research Paper "A Maturity Model for Urban Dataset Metadata"

Appendix 1: Metadata Ontologies and Considerations

The DataCite project (Rueda, Fenner, & Cruse, 2017) seeks to create an interoperable e-infrastructure for research data. It highlights the importance of unique, persistent identifiers in datasets for achieving an interoperable e-infrastructure. "Persistent identifiers allow different platforms to exchange information consistently and unambiguously and provide a reliable way to track citations and reuse." Additionally, there is the adoption of a common set of metadata properties, partitioned into mandatory, recommended, and optional (Table 17). **Error! Reference source not found.**

Table 1. DataCite metadata properties.

Mandatory	Recommended	Optional
Identifier	Subject	Language
Creator	Contributor	Alternate identifier
Title	Date	Size
Publisher	Related identifier	Format
Publication year	Description	Version
Resource type	Geolocation	Rights

Fenner et al. (2019) define a roadmap for data citation. They identify two types of metadata that need to be represented. The first is citation metadata. Table 18 lists the types of citation metadata in the first column and the corresponding properties as found in Dublin Core, Schema.org, DataCite, and DATS (Sansone et al., 2017).

Table 2. Citation metadata.

Citation metadata	Dublin Core	Schema.org	DataCite	DATS
Dataset identifier	identifier	@id	identifier	identifier
Title	title	name	title	title
Creator	creator	author	creator	creator
Data repository or archive	publisher	publisher	publisher	publisher
Publication date	date	datePublished	publication Year	date
Version	not available	version	version	version

Type	type	type	resourceTypeGeneral	type

The second is discovery metadata used to enable the discovery of relevant datasets (Table 19).

Table 3. Discovery metadata.

Discovery metadata	Dublin Core	Schema.org	DataCite	DATS
Description	description	description	description	datatype is a dimension, isAbout Material Material
Keywords	subject	keywords	subject	keywords
License	license	license	rights	license
Related dataset	isPartOf is VersionOf references	isPartOf citation	relatedIdentifier	isPartOf
Related publication	bibliographicCitation	citation	relatedIdentifier	publication

Chapman et al. (2020) state that repositories need to consider data provenance, annotations, quality, granularity of content, data schema, language, and temporal coverage.

Thornton and Shiri (2021) analysed several Canadian open health data repositories regarding the richness of their metadata. As part of their analysis, they used metadata defined in the Dataverse North metadata best practices guide (Cooper et al., 2019) and Data Citation Roadmap (Fenner et al., 2019). The following are the metadata in the Dataverse North guide:

- Title
- Author
- Description
- Subject
- Producer
- Contact name
- Contact affiliation
- Contact email

Gebru et al. (2021) in their "Datasheets for Datasets" proposal defined 56 questions to document the provenance of machine-learning datasets. These questions are divided into seven categories:

1. Motivation: Who created the dataset? For what purpose? Who funded it?
2. Composition: What is the dataset composed of? Size? Completeness?

3. Collection process: How was the data collected? When? Ethical process?
4. Pre-processing/cleaning/labelling: Was any cleaning or labelling performed?
5. Uses: How has the data been used? What can it be used for, or not?
6. Distribution: How and when will the dataset be distributed? Any restrictions?
7. Maintenance: Who supports the dataset? Will it be updated? Will older versions be maintained?

Appendix 2 contains the complete list of questions for each category.

Licensing Metadata

Another important category of metadata are the licences that dictate by whom and how a dataset may be used. To ascertain the metadata required to ascertain the latter, we review licences under which datasets are often published.

The Creative Commons Organization has six types of licence,[1] spanning the continuum from free use of the material for both commercial and non-commercial uses, to limitations on remixing, adapting, and building upon, and for commercial use. Common to all these licences is the requirement to give attribution to the creator of the material.

The Open Knowledge Foundation has three types of licence[2] that focus specifically on data. The licences allow users of the data to:

- Share: To copy, distribute and use the database.
- Create: To produce works from the database.
- Adapt: To modify, transform and build upon the database.

Similar to the Creative Commons licence, attribution is required (for two of the licences) for any public use of the data and its derivations. In both cases, knowing the creator or owner and the licence is important.

Appendix 2: Datasheets for Datasets Questions

Table 4. Datasheets for Datasets questions.

Category	Question
Motivation	For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
	Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, which company, institution, organisation)?
	Who funded the creation of the dataset?

[1] Creative Commons Licenses: https://creativecommons.org/about/cclicenses/
[2] Open Data Commons Licenses: https://opendatacommons.org/licenses/

Category	Question
Composition	What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instance (for example, movies, users, and ratings; people and interactions between them; nodes and edges)?
	How many instances are there in total (of each type, if appropriate)?
	Does the dataset contain all possible instances, or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, in geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).
	What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features?
	Is there a label or target associated with each instance?
	Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.
	Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
	Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
	Are there any errors, sources of noise, or redundancies in the dataset?
	Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licences, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
	Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?
	Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Category	Question
	Does the dataset identify any subpopulations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
	Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data), from the dataset?
	Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?
Collection process	How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
	What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programmes, software APIs)? How were these mechanisms or procedures validated?
	If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?
	Who was involved in the data collection process (for example, students, crowdworkers, contractors), and how were they compensated (for example, how much were crowdworkers paid)?
	Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
	Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
	If the dataset does not relate to people, you may skip the remaining questions in this section.
	Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
	Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Category	Question
	Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
Pre-processing/ cleaning/ labelling	Was any pre-processing/cleaning/labelling of the data done (for example, discretization or bucketing, tokenisation, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section. Was the "raw" data saved in addition to the pre-processed/cleaned/ labelled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. Is the software that was used to pre-process/clean/label the data available? If so, please provide a link or other access point.
Uses	Has the dataset been used for any tasks already? If so, please provide a description. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. What (other) tasks could the dataset be used for? Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labelled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms? Are there tasks for which the dataset should not be used? If so, please provide a description.
Distribution	Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organisation) on behalf of which the dataset was created? If so, please provide a description.

Category	Question
	How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
	When will the dataset be distributed?
	Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this licence and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
	Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
	Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
Maintenance	Who will be supporting/hosting/maintaining the dataset?
	How can the owner/curator/manager of the dataset be contacted (for example, email address)?
	Is there an erratum? If so, please provide a link or other access point.
	Will the dataset be updated (for example, to correct labelling errors, add new instances, delete instances)? If so, please describe how often, by
	whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub).
	If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits, and explain how they will be enforced.
	Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

Appendix 3: Dataset Metadata Vocabularies

With the growing interest in open data, the importance of vocabularies representing dataset metadata has grown in parallel with the adoption of open portals such as CKAN and Dataverse. This section reviews vocabularies that have been developed to: 1) understand what metadata attributes the vocabularies have chosen to include; and 2) the terms they use for potential reuse in DMCMM.

DCT and VoID

Vocabulary of Interlinked Datasets[3] (Alexander et al., 2011) is one of the early RDF vocabularies for dataset metadata. It identifies Dublin Core Metadata terms to be used for datasets. Table 21 shows the metadata terms. The prefix "dct" denotes the namespace "http://purl.org/dc/terms/".

Table 5. Dublin Core Metadata terms for dataset metadata.

Term	Purpose
dct:title	The name of the dataset.
dct:description	A textual description of the dataset.
dcterms:creator	An entity such as person, organisation or service that is primarily responsible for creating the dataset. The creator should be described as an RDF resource, rather than just providing the name as literal.
dct:publisher	An entity such as a person, organisation or service that is responsible for making the dataset available. The publisher should be described as an RDF resource, rather than just providing the name as a literal.
dct:contributor	An entity such as a person, organisation or service that is responsible for making contributions to the dataset. The contributor should be described as an RDF resource, rather than just providing the name as a literal.
dct:source	A related resource from which the dataset is derived. The source should be described as an RDF resource rather than as literal.
dct:date	A point or period of time associated with an event in the life-cycle of the resource. The value should be formatted and data-typed as an xsd:date.
dct:created	Date of creation of the dataset. The value should be formatted and data-typed as an xsd:date.
dct:issued	Date of formal issuance (e.g., publication) of the dataset. The value should be formatted and data-typed as an xsd:date.
dct:modified	Date on which the dataset was changed. The value should be formatted and data-typed as an xsd:date.

Additionally, it provides properties for contact information, licensing, dataset domain categories, format, access information, and statistics. Table 22 lists the statistics-related properties. The prefix "void" denotes the namespace "http://rdfs.org/ns/void#".

Table 6. VoID dataset statistics.

Property	Purpose
void:triples	The total number of triples contained in the dataset.
void:entities	The total number of entities that are described in the dataset. To be an entity in a dataset, a resource must have a URI, and the URI must match

[3] Describing Linked Datasets with the VoID Vocabulary: https://www.w3.org/TR/void/

Property	Purpose
	the dataset's void:uriRegexPattern, if any. Authors of VoID files may impose arbitrary additional requirements, for example, they may consider any foaf:Document resources as not being entities.
void:classes	The total number of distinct classes in the dataset. In other words, the number of distinct class URIs occurring as objects of rdf:type triples in the dataset.
void:properties	The total number of distinct properties in the dataset. In other words, the number of distinct property URIs that occur in the predicate position of triples in the dataset.
void:distinctSubjects	The total number of distinct subjects in the dataset. In other words, the number of distinct URIs or blank nodes that occur in the subject position of triples in the dataset.
void:distinctObjects	The total number of distinct objects in the dataset. In other words, the number of distinct URIs, blank nodes, or literals that occur in the object position of triples in the dataset.
void:documents	If the dataset is published as a set of individual documents, such as RDF/XML documents or RDFa-annotated web pages, then this property indicates the total number of such documents. Non-RDF documents, such as web pages in HTML or images, are usually not included in this count. This property is intended for datasets where the total number of triples or entities is hard to determine. void:triples or void:entities should be preferred where practical.

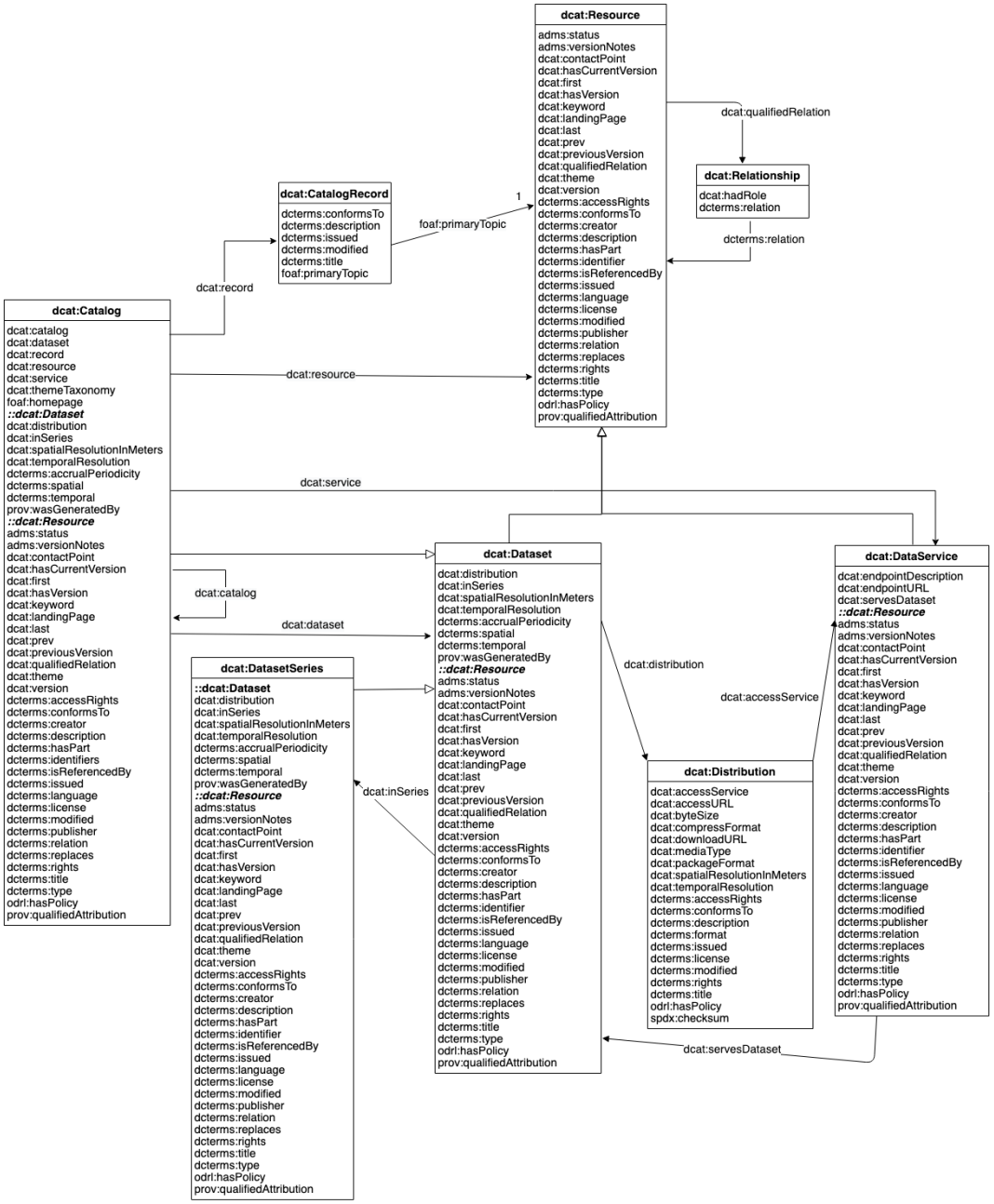


Figure 1. DCAT3 classes and properties (Albertoni et al., 2023).

DCAT

The Data Catalog Vocabulary (DCAT)[4] is a W3C RDF-based vocabulary that enables interoperability among data catalogues published on the Web. The vocabulary defines a set of metadata terms for describing data catalogues and datasets. Figure 9 depicts the classes and properties in Version 3 (working draft) of DCAT. The dcat:Catalog class is used to define a web-accessible catalogue composed of dcat:Resources of which dcat:Dataset and dcat:DataService are subclasses. A rich set of properties are provided to describe both. The prefix "dcat" denotes the namespace "http://www.w3.org/ns/dcat#".

[4] Data Catalog Vocabulary (DCAT) - Version 3: https://www.w3.org/TR/vocab-dcat-3/

DCAT-AP

The DCAT-AP (Van Nuffelen, 2022) provides a standard for describing dataset metadata that is published by data portals across Europe. It identifies a required set of DCAT classes and properties, categorising them as mandatory, recommended, or optional, which can be interpreted as a three-level maturity model. Tables 23 to 25 define the properties of the three categories, respectively.

Table 7. DCAT-AP mandatory dataset properties.

Property	URI	Range	Usage note	Card
Description	dct:description	rdfs:Literal	This property contains a free-text account of the dataset. This property can be repeated for parallel language versions of the description.	1.. n
Title	dct:title	rdfs:Literal	This property contains a name given to the dataset. This property can be repeated for parallel language versions of the name.	1..n

Table 8. DCAT-AP recommended dataset properties.

Property	URI	Range	Usage note	Card
contact point	dcat:contactPoint	vcard:Kind	This property contains contact information that can be used for sending comments about the dataset.	0..n
dataset distribution	dcat:distribution	dcat: Distribution	This property links the dataset to an available distribution.	0..n
keyword/ tag	dcat:keyword	rdfs: Literal	This property contains a keyword or tag describing the dataset.	0..n
publisher	dct:publisher	foaf:Agent	This property refers to an entity (organisation) responsible for making the dataset available.	0..1
spatial/ geographical coverage	dct:spatial	dct:Location	This property refers to a geographic region that is covered by the dataset.	0..n
temporal coverage	dct:temporal	dct:PeriodOfTime	This property refers to a temporal period that the dataset covers.	0..n
theme/category	dcat:theme, subproperty of dct:subject	skos:Concept	This property refers to a category of the dataset. A sataset may be associated with multiple themes.	0..n

Table 9. DCAT-AP Optional Dataset Properties.

Property	URI	Range	Usage note	Card
access rights	dct:accessRights	dct:RightsStatement	This property refers to information that indicates whether the dataset is open data, has access restrictions, or is not public. A controlled vocabulary with three members (:public, :restricted, :non-public) will be created and maintained by the Publications Office of the EU.	0..1
creator	dct:creator	foaf:Agent	This property refers to the entity primarily responsible for producing the dataset.	0..1
conforms to	dct:conformsTo	dct:Standard	This property refers to an implementing rule or other specification.	0..n
documentation	foaf:page	foaf: Document	This property refers to a page or document about this dataset.	0..n
frequency	dct:accrualPeriodicity	dct:Frequency	This property refers to the frequency at which the dataset is updated.	0..1
has version	dct:hasVersion	dcat:Dataset	This property refers to a related dataset that is a version, edition, or adaptation of the described dataset.	0..n
identifier	dct:identifier	rdfs:Literal	This property contains the main identifier for the dataset, e.g., the URI or other unique identifier in the context of the catalogue.	0..n
is referenced by	dct:isReferencedBy	rdfs:Resource	This property provides a link to a description of a relationship with another resource.	0..n
is version of	dct:isVersionOf	dcat:Dataset	This property refers to a related dataset of which the described dataset is a version, edition, or adaptation.	0..n
landing page	dcat:landingPage	foaf:Document	This property refers to a web page that provides access to the dataset, its distributions, and/or additional information. It is intended to point to a landing page at the original data	0..n

Property	URI	Range	Usage note	Card
			provider, not to a page on a site of a third party, such as an aggregator.	
language	dct:language	dct:LinguisticSystem	This property refers to a language of the dataset. This property can be repeated if there are multiple languages in the dataset.	0..n
other identifier	adms:identifier	adms:Identifier	This property refers to a secondary identifier of the dataset, such as MAST/ADS[15], DataCite [16], DOI17, EZID18, or W3ID19.	0..n
provenance	dct:provenance	dct:ProvenanceStatement	This property contains a statement about the lineage of a dataset.	0..n
qualified attribution	prov:qualifiedAttribution	prov:Attribution	This property refers to a link to an agent having some form of responsibility for the resource.	0..n
qualified relation	dcat:qualifiedRelation	dcat: Relationship	This property is about a related resource, such as a publication, that references, cites, or otherwise points to the dataset.	0..n
related resource	dct:relation	rdfs: Resource	This property refers to a related resource.	0..n
release date	dct:issued	rdfs: Literal typed as xsd:date or xsd:dateTime	This property contains the date of formal issuance (e.g., publication) of the dataset.	0..1
sample	adms:sample	dcat:Distribution	This property refers to a sample distribution of the dataset.	0..n
source	dct:source	dcat:Dataset	This property refers to a related dataset from which the described dataset is derived.	0..n
spatial resolution	dcat:spatialResolutionInMeters	xsd:decimal	This property refers to the minimum spatial separation resolvable in a dataset, measured in metres.	0..n
temporal resolution	dcat:temporalResolution	xsd:duration	This property refers to the minimum time period resolvable in the dataset.	0..n
Type	dct:type	skos:Concept	This property refers to the type of the dataset. A controlled	0..1

Property	URI	Range	Usage note	Card
			vocabulary for the values has not been established.	
update/ modification date	dct:modified	rdfs:Literal typed as xsd:date or xsd:dateTime	This property contains the most recent date on which the dataset was changed or modified.	0..1
version	owl:versionInfo	rdfs:Literal	This property contains a version number or other version designation of the dataset.	0..1
version notes	adms:versionNotes	rdfs:Literal	This property contains a description of the differences between this version and a previous version of the dataset. This property can be repeated for parallel language versions of the version notes.	0..n
was generated by	prov:wasGeneratedBy	prov:Activity	This property refers to an activity that generated, or provides the business context for, the creation of the dataset.	0..n

Schema.org

Schema.org contains several classes and properties relevant to documenting datasets. Google provides a guide[5] for developers to enable dataset discovery. It distinguishes between required Schema.org properties[6] (a sample of the Schema.org class definitions are given in Figure 10):

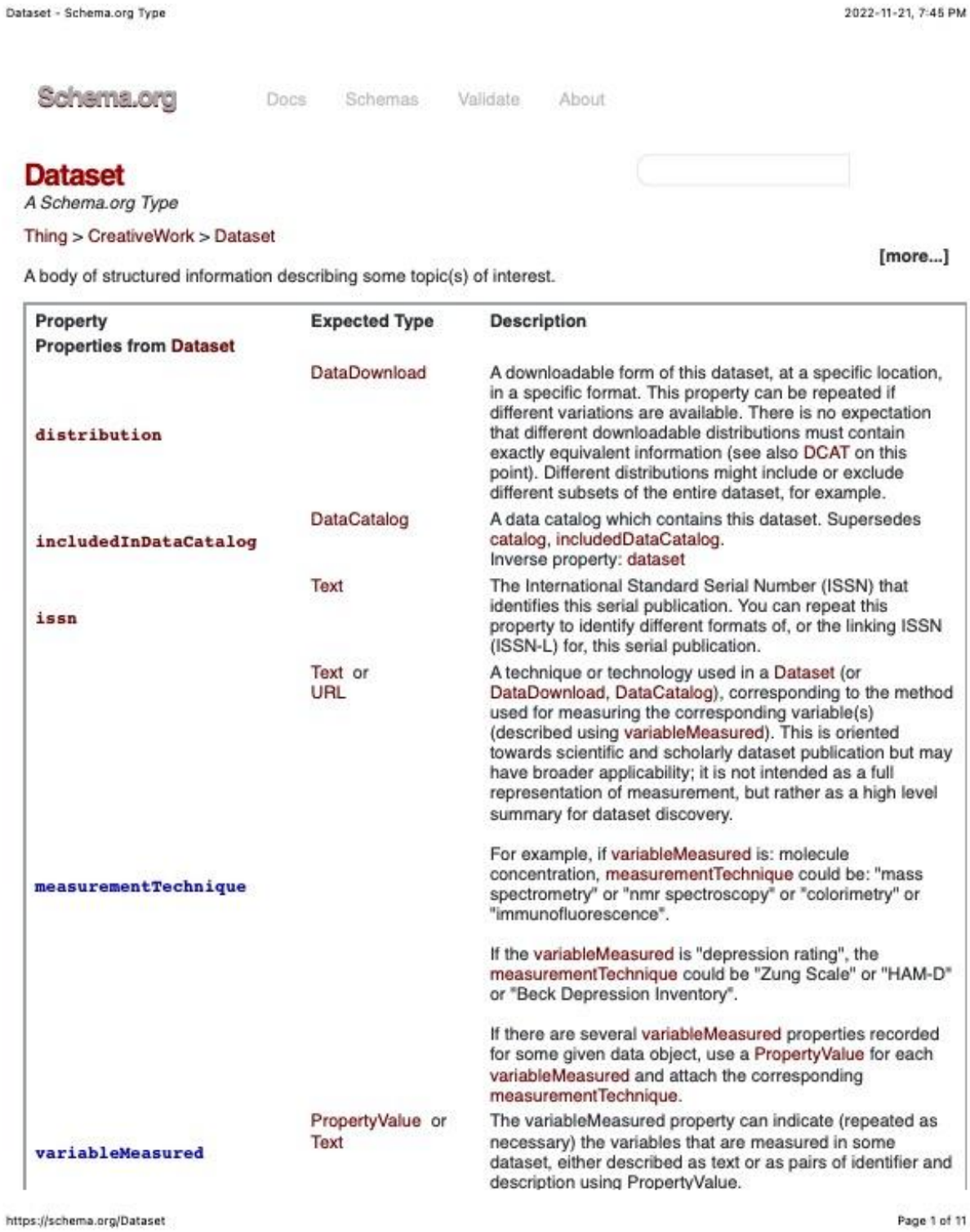
- name – A descriptive name of a dataset (e.g., "Snow depth in northern hemisphere");
- description – A short summary describing a dataset;

and recommended Schema.org properties:

- url – Location of a page describing the dataset;
- sameAs – Other URLs that can be used to access the dataset page. A link to a page that provides more information about the same dataset, usually in a different repository;
- version – The version number or identifier for this dataset (text or numeric);
- isAccessibleForFree – Boolean (true|false) specifying if the dataset is accessible for free;
- keywords – Keywords summarising the dataset;
- identifier – An identifier for the dataset, such as a DOI (text, URL, or PropertyValue);

[5] Dataset (Dataset, DataCatalog, DataDownload) structured data: https://developers.google.com/search/docs/appearance/structured-data/dataset
[6] Describing a Dataset: https://github.com/ESIPFed/science-on-schema.org/blob/master/guides/Dataset.md

- variableMeasured – What does the dataset measure? (e.g., temperature, pressure).



Dataset - Schema.org Type 2022-11-21, 7:45 PM

Schema.org Docs Schemas Validate About

Dataset

A Schema.org Type

Thing > CreativeWork > Dataset

[more...]

A body of structured information describing some topic(s) of interest.

Property	Expected Type	Description
Properties from Dataset		
distribution	DataDownload	A downloadable form of this dataset, at a specific location, in a specific format. This property can be repeated if different variations are available. There is no expectation that different downloadable distributions must contain exactly equivalent information (see also DCAT on this point). Different distributions might include or exclude different subsets of the entire dataset, for example.
includedInDataCatalog	DataCatalog	A data catalog which contains this dataset. Supersedes catalog, includedDataCatalog. Inverse property: dataset
issn	Text	The International Standard Serial Number (ISSN) that identifies this serial publication. You can repeat this property to identify different formats of, or the linking ISSN (ISSN-L) for, this serial publication.
measurementTechnique	Text or URL	A technique or technology used in a Dataset (or DataDownload, DataCatalog), corresponding to the method used for measuring the corresponding variable(s) (described using variableMeasured). This is oriented towards scientific and scholarly dataset publication but may have broader applicability; it is not intended as a full representation of measurement, but rather as a high level summary for dataset discovery. For example, if variableMeasured is: molecule concentration, measurementTechnique could be: "mass spectrometry" or "nmr spectroscopy" or "colorimetry" or "immunofluorescence". If the variableMeasured is "depression rating", the measurementTechnique could be "Zung Scale" or "HAM-D" or "Beck Depression Inventory". If there are several variableMeasured properties recorded for some given data object, use a PropertyValue for each variableMeasured and attach the corresponding measurementTechnique.
variableMeasured	PropertyValue or Text	The variableMeasured property can indicate (repeated as necessary) the variables that are measured in some dataset, either described as text or as pairs of identifier and description using PropertyValue.

https://schema.org/Dataset Page 1 of 11

Figure 2. Sample of Schema.org dataset class definitions.

DQV

Data Quality Vocabulary[7] is an extension of DCAT that focuses on "the quality of the data, how frequently it is updated, whether it accepts user corrections, persistence commitments, etc." (Albertoni & Isaac, 2016). The following lists the core classes. The prefix "dqv" denotes the namespace "http://www.w3.org/ns/dqv#".

- dqv:QualityAnnotation represents feedback and quality certificates given about the dataset or its distribution.
- dct:Standard represents a standard the dataset or its distribution conforms to.
- dqv:QualityPolicy represents a policy or agreement that is chiefly governed by data quality concerns.
- dqv:QualityMeasurement represents a metric value providing quantitative or qualitative information about the dataset or distribution.
- prov:Entity represents an entity involved in the provenance of the dataset or distribution.

Figure 11 depicts the core classes and properties of DQV.

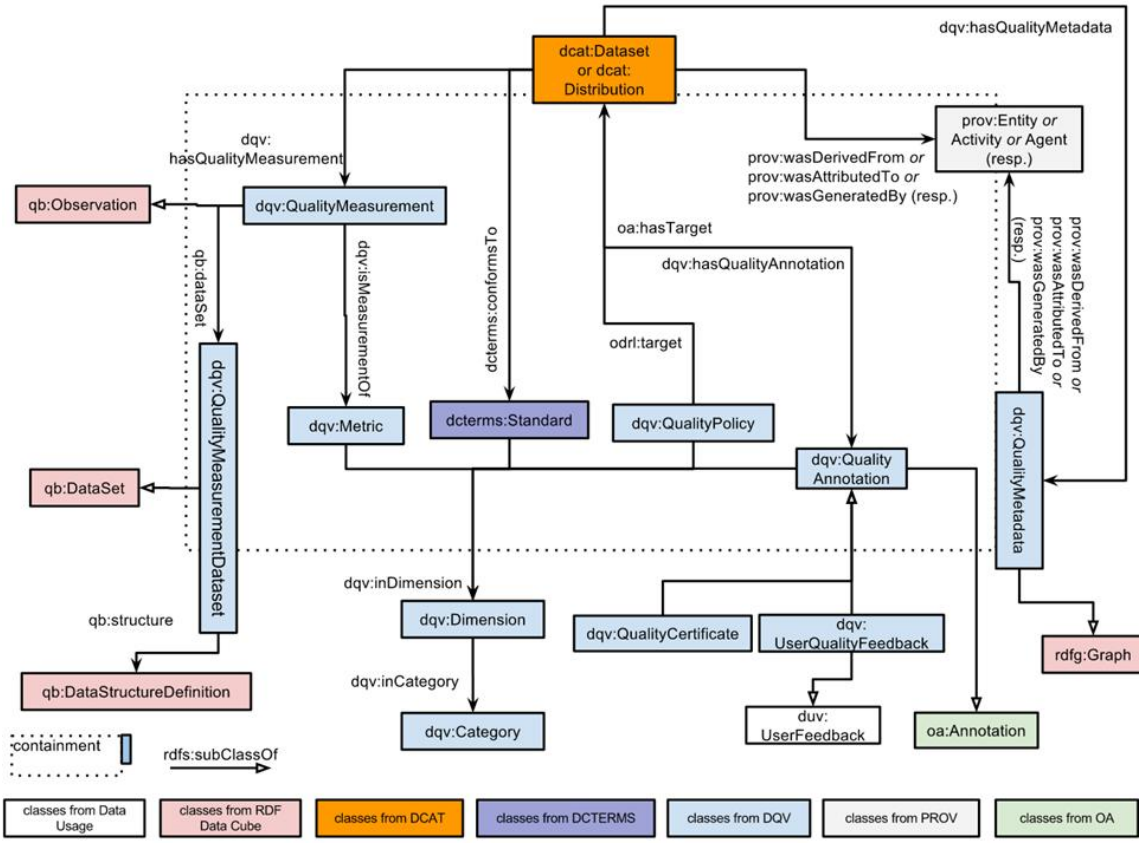


Figure 3. DQV Information Model (from Albertoni & Isaac, 2016).

[7] Data on the Web Best Practices: Data Quality Vocabulary: https://www.w3.org/TR/vocab-dqv/

DDI

Data Documentation Initiative (DDI) has developed standards for documenting social science surveys and datasets.[8] It provides a deeper dive into the properties describing the content of datasets and how they were generated (Thomas et al., 2014). Several dimensions of the content are described, including dataset provenance and analysis (DDI-Lifecycle) (Poynter & Spiegel, 2016), preservation and discovery (DDI-Codebook[9]), and a SKOS extension that includes statistical information about datasets and refinement of SKOS properties (XKOS) (Cotton, Gillman, & Joque, 2015). DDI metadata properties are viable for inclusion in the DMCMM; however, as of time of writing, DDI is not yet available in RDF or linked-data formats.

ODRL

Open Digital Rights Language[10] (Iannella & Villata, 2018) "is a policy expression language that provides a flexible and interoperable information model, vocabulary, and encoding mechanisms for representing statements about the usage of content and services." It "represents Policies that express Permissions, Prohibitions and Duties related to the usage of Asset resources. The Information Model (Figure 12) explicitly expresses what is allowed and what is not allowed by the Policy, as well as other terms, requirements, and parties involved."

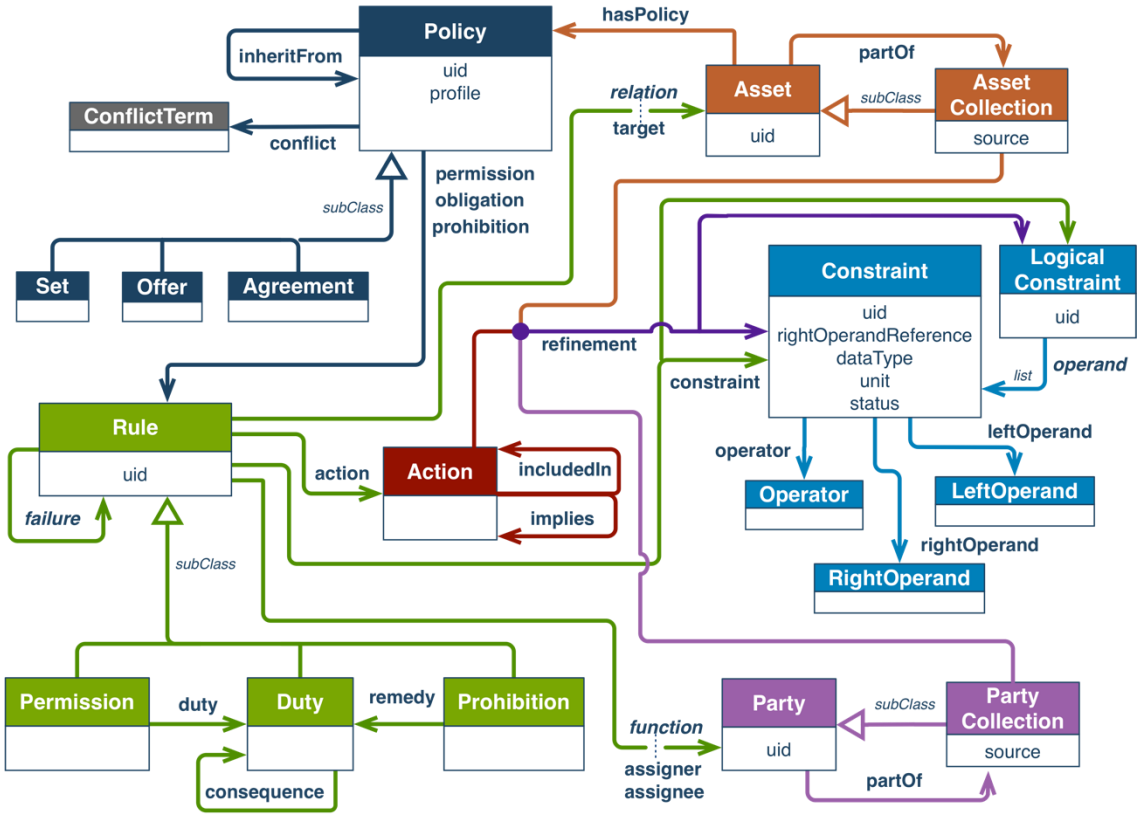


Figure 4. ODRL Information Model (from Iannella & Villata (2018)).

[8] DDI Alliance: http://www.ddialliance.org/
[9] DDI Codebook Development Work: https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/929792030/DDI+Codebook+Development+Work
[10] ODRL Information Model 2.2: https://www.w3.org/TR/odrl-model/

Appendix 4: FAIR and OCAP Requirements

FAIR

As adoption of FAIR principles continues to grow, the DMCMM FAIR attributes support the FAIR evaluation of a dataset. Bahim et al. (2020) define a FAIR (Findable, Accessible, Interoperable, Reusable) Data Maturity Model. "The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data."[11]

A set of indicators have been defined to evaluate the "FAIRness" of a dataset. The indicators are divided into Essential, Important, and Useful. Tables 26 to 28 list the indicators for each partition. The indicators allow a user to identify which datasets may be useful to them without the need to request and gain access to the dataset first. For example, a user may want to evaluate the dataset using a computer. It would be important to know whether the dataset is computer-readable.

The FAIR principle in Table 26 identifies those indicators that are essential. Indicator RDA-A1-04D indicates whether "Data is accessible through standardised protocol." An analyst might be interested to know whether data is easily accessible, or whether a custom process must be created. Similarly, Indicator RDA-A1-04M indicates whether "Metadata is accessible through a free access protocol." A data curator might need to know whether they require funding sources before pursuing to access the metadata record to add to their catalogue.

Table 10. Essential FAIR indicators.

FAIR	ID	Indicator
F1	RDA-F1-01M	Metadata is identified by a persistent identifier
F1	RDA-F1-01D	Data is identified by a persistent identifier
F1	RDA-F1-02M	Metadata is identified by a globally unique identifier
F1	RDA-F1-02D	Data is identified by a globally unique identifier
F2	RDA-F2-01M	Rich metadata is provided to allow discovery
F3	RDA-F3-01M	Metadata includes the identifier for the data
F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed
A1	RDA-A1-02M	Metadata can be accessed manually (i.e., with human intervention)
A1	RDA-A1-02D	Data can be accessed manually (i.e., with human intervention)
A1	RDA-A1-03M	Metadata identifier resolves to a metadata record
A1	RDA-A1-03D	Data identifier resolves to a digital object

[11] FAIR Principles: https://www.go-fair.org/fair-principles/

FAIR	ID	Indicator
A1	RDA-A1-04M	Metadata is accessed through standardised protocol
A1	RDA-A1-04D	Data is accessible through standardised protocol
A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol
A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available
R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse
R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused
R1.3	RDA-R1.3-01M	Metadata complies with a community standard
R1.3	RDA-R1.3-01D	Data complies with a community standard
R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard

Table 11. Important FAIR indicators.

FAIR	ID	Indicator
A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data
A1	RDA-A1-05D	Data can be accessed automatically (i.e., by a computer programme)
A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol
I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format
I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format
I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation
I1	RDA-I1-02D	Data uses machine-understandable knowledge representation
I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies
I3	RDA-I3-01M	Metadata includes references to other metadata
I3	RDA-I3-03M	Metadata includes qualified references to other metadata
R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence
R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence

FAIR	ID	Indicator
R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards
R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard

Table 12. Useful FAIR indicators.

FAIR	ID	Indicator
A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation
I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies
I3	RDA-I3-01D	Data includes references to other data
I3	RDA-I3-02M	Metadata includes references to other data
I3	RDA-I3-02D	Data includes qualified references to other data
I3	RDA-I3-04M	Metadata includes qualified references to other data
R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language

Indigenous Data Requirements

Metadata requirements for datasets containing Indigenous data stem from Indigenous Data Sovereignty, which addresses "the rights and interests of Indigenous Peoples in relation to data about them, their territories, and their ways of life" (Carroll et al., 2020). Several frameworks have been proposed, including Canada's OCAP[12] (Mecredy, Sutherland, & Jones, 2018), CARE principles (Carrol et al., 2020), and Australia's guidance for Indigenous data (Commonwealth of Australia, 2024). CARE focuses on the ethical usage of data to ensure it is used to the benefit of the Indigenous communities whose data it is about. OCAP and Australia's principles focus on ensuring ownership of the data and control over it by the intended stakeholders (Indigenous community). While complementary, OCAP principles are selected due to their Canadian origins and context. OCAP, developed by the First Nations Information Governance Centre,[13] is a set of principles "regarding the collection, use and disclosure of data or information regarding first nations." It focuses on protecting Indigenous individual privacy rights as well as the collective rights of communities. OCAP is an acronym for Ownership, Control, Access, and Possession.[14] Each are defined in the following excerpts from the OCAP training module:

- **Ownership:** "The notion of ownership refers to the relationship of a First Nations community to its cultural knowledge/data/information. The principle states that a community or group owns information collectively in the same way that an individual

[12] The First Nations Principles of OCAP: https://fnigc.ca/ocap-training/
[13] First Nations Information Governance Centre: https://fnigc.ca/
[14] Reproduced from Module 1 of OCAP online training participant notes, developed by Algonquin College and FNIGC.

owns their personal information. Ownership is distinct from stewardship. The stewardship or custodianship of data or information by an institution that is accountable to the group is a mechanism through which ownership may be maintained. This can be done with data-sharing agreements and other legal instruments."

- **Control:** "The aspirations and inherent rights of First Nations to maintain and regain control of all aspects of their lives and institutions extend to information and data. The principle of 'control' asserts that First Nations people, their communities and their representative bodies must control how information about them is collected, used and disclosed. The element of control extends to all aspects of information management, from collection of data to the use, disclosure, and ultimate destruction of data."

- **Access: "**First Nations must have access to information and data about themselves and their communities, regardless of where it is held. The principle also refers to the right of First Nations communities and organizations to manage and make decisions regarding who can access their collective information."

- **Possession:** "While 'ownership' identifies the relationship between a people and their data, possession reflects the state of stewardship of data. First Nations possession puts data within First Nations jurisdiction and therefore, within First Nations control. Possession is the mechanism to assert and protect ownership and control. First Nations generally exercise little or no control over data that is in possession of others, particularly other governments."

The guidance provided for the management of First Nations data in Australia contains guidance on how to work with Indigenous communities along with elements of OCAP and FAIR (Commonwealth of Australia, 2024).