IJDC | *Research Paper*

A Maturity Model for Urban Dataset Metadata

Mark S. Fox
Urban Data Research Centre
School of Cities
University of Toronto

Bart Gajderowicz
Urban Data Research Centre
School of Cities
University of Toronto

Dishu Lyu
Urban Data Research Centre
School of Cities
University of Toronto

Abstract

The rapid increase in published datasets has intensified challenges in sourcing and integrating relevant data for analysis. Persistent obstacles include poor metadata, ineffective presentation, and difficulties in locating and integrating datasets. This paper delves into the intricacies of dataset retrieval, emphasising the pivotal role of metadata in aligning datasets with user queries. Through an exploration of existing literature, it highlights prevailing issues, such as identifying valuable metadata and developing tools to maintain and annotate them effectively. The paper proposes a dataset metadata maturity model, inspired by software engineering frameworks, to guide dataset creators from basic to advanced documentation. The model encompasses seven pivotal dimensions, spanning content to quality information, each stratified across five maturity levels to guide the optimal documentation of datasets, ensuring ease of discovery, accurate relevance assessment, and comprehensive understanding of datasets. This paper also incorporates the maturity model into a data cataloguing tool called CKAN through a custom plugin, CKANext-udc. The plugin introduces custom fields based on different maturity levels, allows for user interface customisation, and integrates with a graph database, converting catalogue data into a knowledge graph based on the Maturity Model ontology.

Submitted 9 September 2023 ~ *Accepted* 8 July 2025

Correspondence should be addressed to Bart Gajderowicz, Email: bartg@mie.utoronto.ca

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Copyright rests with the authors. This work is released under a Creative Commons Attribution License, version 4.0. For details please see https://creativecommons.org/licenses/by/4.0/

Introduction

The world is awash in data, but we cannot seem to find the data we need. Although open data platforms aim to simplify the search for relevant data, Ojo et al. (2016) note that common obstacles to finding relevant data include “poor metadata, failure to present data appropriately to different audiences and difficulty in locating data of interest.” Metadata, defined as data describing other data, includes key descriptors such as title, author, keywords, and provenance. The process of dataset exploration, which involves distinguishing between direct content examination and metadata inspection[1] (Koesten et al., 2017; Kunze & Auer, 2013; Chiu, Chen, & Cline, 2023), is a core information-seeking stage. Chapman et al. (2020) identify three persistent metadata challenges: determining the most valuable metadata, automating metadata creation, and automatically linking datasets to ontologies. Approaches such as Datasheets for Datasets (Gebru et al., 2021) offer structured metadata guidance but pose a burden on cataloguers, especially when information is limited or unstructured. Existing keyword-based searches on unstructured text may help narrow down the target datasets, but with possibly poor precision and recall (Berkley et al., 2009).

To improve discoverability and metadata quality, ontology methods have introduced structured vocabularies, such as DCAT (Albertoni et al., 2023), Schema.org, PROV (Lebo, Sahoo, & McGuinness, 2013), and DQV (Albertoni & Isaac, 2016). Tools like Open Data Portal Watch (Neumaier, Umbrich, & Polleres, 2017) and Google Dataset Search (Noy, Burgess, & Brickley, 2019) automate the harvesting and organisation of metadata into searchable knowledge graphs. However, cataloguers still face overwhelming complexity due to vocabularies such as DCAT, which has over 70 properties, and datasheets that require 50+ metadata fields.

To address this, the Dataset Metadata Capability Maturity Model (DMCMM) is proposed, following the structure of the Capability Maturity Model for Software (Paulk et al., 1993), and offering a tiered system that balances cataloguing effort with metadata utility. The DMCMM is developed through a literature review of search behaviours and metadata usage frequencies and includes metadata fields that support frameworks such as FAIR (Findable, Accessible, Interoperable, Reusable) and OCAP[2] (Ownership, Control, Access, Possession) for indigenous data (Mecredy, Sutherland, & Jones, 2018). The model guides cataloguers through progressive metadata specification, enabling improved dataset discovery, relevance evaluation, and alignment with data governance principles.

Dataset Metadata Requirements

Requirements for dataset metadata stem from multiple sources. One arises from analyses of dataset search behaviours, which reveal what metadata users frequently seek. Another is the emerging consensus among data platforms regarding essential metadata attributes. A third involves metadata necessary to evaluate alignment with FAIR principles. A fourth concerns metadata requirements specific to Indigenous data. This section reviews these requirements to ensure that those stakeholders needing access to urban datasets, whether as publishers, authors, data managers, or users, can find them easily according to their respective needs and behaviours.

Requirements Based on Search Behaviour

How can the DMCMM support the search for relevant datasets? In this section, we review the literature on dataset searching to understand what metadata is used to search for datasets, and their frequency of use, to determine the level in the DMCMM where they should appear.

[1] Which falls into the area of *sensemaking* (Russell et al., 1993).

[2] The First Nations Principles of OCAP: https://fnigc.ca/ocap-training/

Koesten et al. (2017) studied information-seeking behaviour. The following are examples of search:

> "someone trying to find the number of schools in a given post code area would need to extract the answer from a larger dataset containing all entries for all schools in a country in 2016. Someone studying how the number of schools across different regions has changed over time would need to process and aggregate several versions of the same data, published year after year. Finally, school data could be mixed with house prices statistics to understand how one aspect influences the other."

Koesten et al. (2017) identified three categories of metadata to describe a dataset: Relevance, Usability and Quality. This work resulted in the proposal of a five-pillar model for how people seek information, namely Task, Search, Evaluate, Explore, and Use. Task includes goal or process-oriented tasks, linking, time series analysis, summarising, presenting, and exporting. Search involves using tools such as web search engines, data portals, or FOI requests. This is followed by evaluation of relevance, usability, and quality; exploration via visual scans and metadata; and usage for tasks including linking, summarising, analysis, and export. Dataset metadata plays a key role, encompassing relevance, which includes context, coverage, purpose, granularity, summary, and timeframe. Usability involves documentation, licensing, access, format, and shareability. Quality relates to collection methods, provenance, consistency, completeness, and any omissions.

Kacprzak et al. (2019) analysed web portal search logs and written requests, highlighting the importance of topic, geospatial, temporal, and format metadata, with granularity varying across geospatial and temporal dimensions. Their reproduced requests underscore the significance of domain expertise in forming queries with the use of a domain-specific vocabulary. Similarly, Chen et al. (2019) examined nearly 2,000 dataset queries from multiple online communities. Their analysis (Table 1) distinguishes between metadata-related and content-related queries, revealing that 94% reference the dataset's domain or topic, 50% concern dataset concepts and properties, 20% mention geospatial information, 16% dataset format, and 10% temporal attributes.

Table 1. Analysis of dataset queries.

Category	Title	% of queries	Example query
Metadata	Name	3.54%	HUST-ASL Dataset
	Domain/topic	94.45%	weather dataset with solar radiance and solar energy production
	Data format	16.23%	jpg images for all Unicode characters
	Language	3.90%	annotated movie review dataset in German
	Accessibility	7.40%	open source handwritten English alphabets dataset
	Provenance	0.21%	FDA datasets about medicine name and the result has adverse events
	Statistics	2.98%	dataset contains at least 1000 examples of opinion articles
	Overall	96.05%	
Content	Concept	50.59%	dataset about people, include gender, ethnicity, name
	Geospatial	19.21%	judicial decisions in France
	Other entities	0.41%	datasets with nutrition data for many commercial food products (i.e., Lucky Charms, Monster Energy, Nutella, etc.)
	Temporal	9.35%	2011-2013 MoT failure rates on passenger cars

Category	Title	% of queries	Example query
	Other Numbers	1.59%	businesses that employ over 1000 people in Yorkshire region
	Overall	63.79%	

Kacprzak et al. (2019) identified several metadata attributes as relevant in their analysis of query logs from four national open data platforms, including geospatial, temporal, topic taxonomy, price, licence, format, and size. Chua et al. (2020) analysed the information-seeking behaviours of 21 people using open data portals. Spatial and temporal keywords dominated the search queries and were supplemented with format and source filters. Follow-up interviews identified dataset incompleteness and outdatedness as issues. Sharifpour, Wu, and Zhang (2023), in their analysis of search behaviours based on different levels of domain expertise, discovered that expert users used more words and succeeded with shorter sessions, confirming one of the results of White, Dumais and Teevan (2009). They also observed that dataset search is more difficult due to “the data for relevance judgement [not being] readily accessible within the metadata of datasets”.

Dataset Platform Metadata Requirements

The second source of requirements stem from the growing number of dataset platforms that are operating around the world. We determine these requirements by reviewing the literature on the metadata attributes that are found on dataset platforms. These platforms represent a growing consensus of the attributes deemed to be needed to support both search and accessibility.

Assaf, Troncy, and Senart (2015) proposed the Harmonized Data modeL (HDL), which adopts and extends key properties of schemas such as DCAT, Schema.org, or CKAN, to “ensure complete metadata coverage to enable data discovery, exploration and reuse.” Their analysis identifies eight information types to be encoded as metadata:

1. General information such as title and description.
2. Access information such as the URL and licence.
3. Ownership information such as author and maintainer.
4. Provenance information such as creation date and versioning.
5. Geospatial information such as geographic coverage.
6. Temporal information such as temporal span and granularity.
7. Statistical information such as property distribution and number of entities.
8. Quality information such as the quality of the data and metadata.

Neumaier, Umbrich and Polleres (2017) identified the following new or custom metadata properties (Table 2) in their analysis of over 749K CKAN datasets (referred to as “extra keys” in CKAN). They also add quality (DQV) (Albertoni & Isaac, 2016) and provenance (PROV) (Lebo, Sahoo, & McGuinness, 2013) information to the dataset’s metadata.

Table 2. Extra keys.

Extra key	Portals	Datasets	Mapping
spatial	29	315,652	dct:spatial

harvest_object.id	29	514,489	?
harvest_source.id	28	486,388	?
harvest_source_title	28	486,287	?
guid	21	276,144	dct:identifier
contact-email	17	272,208	dcat:contactPoint
spatial-reference-system	16	263,012	?
metadata-date	15	265,373	dct:issued

The DataCite project (Rueda, Fenner, & Cruse, 2017) aims to build an interoperable e-infrastructure for research data, emphasising the role of unique, persistent identifiers that support consistent information exchange and citation tracking. It also promotes a standardised metadata set, divided into mandatory (e.g., identifier, author), recommended (e.g., subject, date), and optional (e.g., language, format) categories. Fenner et al. (2019) outline a roadmap for data citation, distinguishing between citation metadata (e.g., identifier, title, creator, repository, publication date, version, type) and discovery metadata (e.g., description, keywords, licence, related datasets and publications), referencing standards like Dublin Core, Schema.org, DataCite, and DATS (Sansone et al., 2017). Chapman et al. (2020) emphasise the need for metadata covering provenance, annotations, data quality, schema, language, and temporal coverage. Thornton and Shiri (2021), using Dataverse North guidelines (Cooper et al., 2019) and Fenner et al.'s roadmap, assessed Canadian open health repositories, listing key metadata such as title, author, description, subject, producer, and contact information. Details of the ontologies are provided in Appendix 1.

Gebru et al. (2021), in their "Datasheets for Datasets" proposal, present 56 questions across seven categories to document machine-learning dataset provenance:

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. Preprocessing/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 includes the licences that dictate by whom and how a dataset may be used. To ascertain the metadata required for the latter, we review licences under which datasets are often published. The Creative Commons Organization has six types of licence,[3] spanning the continuum from unrestricted use of the material for both commercial and non-commercial uses, to limitations on remixing, adapting, and building upon, and for commercial

[3] Creative Commons Licenses: https://creativecommons.org/about/cclicenses/

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[4] 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.

FAIR

As adoption of FAIR principles grows, the DMCMM must incorporate attributes that support FAIR evaluation. Bahim et al. (2020) define the FAIR (Findable, Accessible, Interoperable, Reusable) Data Maturity Model, emphasising machine-actionability to manage the increasing volume, complexity, and speed of data creation.[5] Appendix 4 outlines FAIRness indicators, categorised as Essential, Important, and Useful, to help users assess the utility of a dataset before access. For example, indicators assess if data is machine-readable or accessible via standardised protocols (RDA-A1-04D), or if metadata is available through free protocols (RDA-A1-04M), which helps users determine accessibility and funding needs.

Indigenous Data Requirements

Metadata requirements for datasets containing Indigenous data are grounded in Indigenous Data Sovereignty, which protects the rights of Indigenous Peoples regarding data about themselves, their lands, and cultures (Carroll et al., 2020). Key frameworks include Canada's OCAP[6] (Mecredy, Sutherland, & Jones, 2018), the CARE Principles (Carroll et al., 2020), and Australia's national guidance (Commonwealth of Australia, 2024). CARE emphasises ethical use of data for the benefit of Indigenous communities, while OCAP and Australia's guidelines stress ownership and control by Indigenous stakeholders. This paper adopts OCAP for its relevance to the Canadian context. Developed by the First Nations Information Governance Centre,[7] OCAP governs the collection, use, and disclosure of First Nations data, addressing both individual privacy and collective rights. OCAP stands for Ownership, Control, Access, and Possession:[8] Ownership refers to collective rights to data; Control to the authority over all data management processes; Access to the right to retrieve and regulate information; and Possession to physical stewardship enabling control. Appendix 4 outlines OCAP requirements for metadata indicators that can help users assess dataset utility before accessing the data directly. Australia's guidance incorporates community collaboration and blends OCAP with FAIR principles (Commonwealth of Australia, 2024).

Dataset Metadata Vocabularies

As open data adoption increases, so does the need for standardised vocabularies to represent dataset metadata on platforms like CKAN and Dataverse. This section reviews vocabularies to

[4] Open Data Commons Licenses: https://opendatacommons.org/licenses/

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

[6] The First Nations Principles of OCAP: https://fnigc.ca/ocap-training/

[7] First Nations Information Governance Centre: https://fnigc.ca/

[8] Reproduced from Module 1 of OCAP online training participant notes, developed by Algonquin College and FNIGC.

assess included metadata attributes and identify reusable terms for the Dataset Metadata Capability Maturity Model (DMCMM). Details of the ontologies are provided in Appendix 3.

VoID[9] (Alexander et al., 2011) is an early RDF vocabulary that identifies Dublin Core terms (e.g., title, creator, source) and provides properties for licensing, access, and statistics. DCAT,[10] a W3C RDF vocabulary, defines metadata for data catalogues and datasets, with classes such as dcat:Catalog, dcat:Dataset, and dcat:DataService. DCAT-AP (Van Nuffelen, 2022) adapts DCAT for European portals, structuring metadata into mandatory, recommended, and optional fields grouped into descriptive, coverage, administrative, access, provenance, relationship, versioning, and technical categories.

Schema.org[11] includes dataset metadata grouped into descriptive, access, identification, and content-specific properties. It was published by Google, which distinguishes between different required properties.[12] DQV[13] (Albertoni & Isaac, 2016), extending DCAT, focuses on data quality through annotations, measurements, standards, and provenance. DDI[1] provides metadata for social science surveys, covering provenance, discovery, and statistical structure, but is not yet available in RDF or linked data (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[14]), 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, but as of time of writing, DDI is not yet available as RDF or linked-data formats. ODRL[15] (Iannella & Villata, 2018) provides a vocabulary to express rights, duties, and conditions associated with digital asset usage.

A Capability Maturity Model for Dataset Metadata

Many metadata vocabularies exist, offering extensive details a dataset producer could supply, yet as Gebru et al. (2021) note, not all information is readily available or easy to provide. To accommodate the varying expertise and familiarity of dataset providers, our goal is to simplify the process while assuring that the most essential metadata is captured at the outset. This is achieved through a maturity model that stratifies metadata attributes by importance, increasing the likelihood of collecting key details and limiting the perceived complexity of the task for the dataset producer.

To manage the broad range of metadata options, properties are first grouped into information categories. These categories are then integrated into successive maturity levels based on the likelihood of obtaining the most relevant attributes, following a logic similar to that used by Assaf, Troncy, and Senart (2015).

1. **Content** such as title and description
2. **Access** Information such as the URL and licence.
3. **Ownership** information such as author and maintainer.
4. **Provenance** information such as creation date and versioning.
5. **Temporal/Geospatial** information such as geographic coverage and temporal span and granularity.

[9] Describing Linked Datasets with the VoID Vocabulary: https://www.w3.org/TR/void/
[10] Data Catalog Vocabulary (DCAT) - Version 3: https://www.w3.org/TR/vocab-dcat-3/
[11] Describing a Dataset: https://github.com/ESIPFed/science-on-schema.org/blob/master/guides/Dataset.md
[12] Dataset (Dataset, DataCatalog, DataDownload) structured data: https://developers.google.com/search/docs/appearance/structured-data/dataset
[13] Data on the Web Best Practices: Data Quality Vocabulary: https://www.w3.org/TR/vocab-dqv/
[14] DDI Codebook Development Work: https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/929792030/DDI+Codebook+Development+Work
[15] ODRL Information Model 2.2: https://www.w3.org/TR/odrl-model/

6. **Statistical** information such as property distribution and number of entities.

7. **Quality** information.

Where appropriate, DCAT properties and classes are used for compatibility. Based on DCAT, a dataset is a dcat:Dataset object. A specific version of the dataset is a dct:Distribution. These are related by the property: dct:distribution (dctat:Dataset, dcat:Distribution). Depending on the domain of the property, the data resource being catalogued is either dct:Dataset or the dcat:Distribution related to the dct:Dataset.

The following prefixes are used in the proposed model.

Table 3. Maturity model prefixes.

Prefix	URI
adms	http://www.w3.org/ns/adms#
cc	http://creativecommons.org/ns#
cudr	http://data.urbandatacentre.ca/
dc	http://purl.org/dc/elements/1.1/
dcat	http://www.w3.org/ns/dcat#
dct	http://purl.org/dc/terms/
dqv	http://www.w3.org/ns/dqv#
fair	http://ontology.eil.utoronto.ca/fair#
foaf	http://xmlns.com/foaf/0.1/
oa	http://www.w3.org/ns/oa#
odrl	http://www.w3.org/ns/odrl/2/
owl	http://www.w3.org/2002/07/owl#
prov	http://www.w3.org/ns/prov#
rdfs	http://www.w3.org/2000/01/rdf-schema#
sc	https://schema.org/
skos	http://www.w3.org/2004/02/skos/core#
vann	http://purl.org/vocab/vann/
vcard	http://www.w3.org/2006/vcard/ns#
void	http://rdfs.org/ns/void
xsd	http://www.w3.org/2001/XMLSchema#

Maturity Level 1

Maturity Level 1 emphasises dataset findability, based on typical search behaviours identified in the dataset search literature. This level restricts cataloguer input to the most frequently used search information. According to Chen et al. (2019), the most sought-after attributes include domain information, specifically title, description, and keywords, followed by geospatial data, format, and temporal details, with publication date also noted. DCAT-AP's mandatory attributes are encompassed within Level 1.

Table 4. Dataset Maturity Level 1: Findability.

Category	Description	Property	Value restriction
Content	Domain/topic	dcat:theme	skos:Concept
	Title	dct:title	rdfs:Literal
	Description	dct:description	rdfs:Literal
	Keywords	dcat:keyword	rdfs:Literal
	Format (file type if relevant)	dct:format	dct:MediaType
	Dataset size in megabytes	datasetSize	xsd:integer
	Metadata identifier – to be used as a unique identifier for the catalogue entry	catalogueEntryIdentifier	rdfs:Literal
Provenance	Published date	dct:issued	xsd:datetime
Temporal/ geospatial	Time period data spans	dct:temporal	dct:PeriodOfTime
	Geospatial area data spans	dct:spatial	dct:Location

Maturity Level 2

Maturity Level 2 focuses on characteristics of access, including ownership. Once a dataset is "found", understanding who, how and where to access the dataset is the next most important metadata.

Table 5. Dataset Maturity Level 2: Access.

Category	Description	Property	Value restriction
Access	Access category: open, closed, service	accessCategory	{open, closed, service}
	Licence	dct:license	dct:LicenseDocument
	Limits on use (e.g., academic purposes, going beyond licence)	odrl:hasPolicy	odrl:Policy
	Location of dataset: where it can be accessed	dcat:accessURL	rdfs:Resource
	Access service specification	dcat:accessService	dcat:DataService
	URL for a downloadable file	dcat:downloadURL	rdfs:Resource
Ownership	Owner	dct:rightsHolder	foaf:Agent
	Contact point	dcat:contactPoint	vcard:Kind
	Publisher	dct:publisher	foaf:Agent

Category	Description	Property	Value restriction
	Creator	dct:creator	foaf:Agent

Notes:

- **odrl:Policy**[16]: A non-empty group of permissions (via the permission property) and/or prohibitions (via the prohibition property) and/or duties (via the obligation property). The Policy class is the parent class to the Set, Offer, and Agreement subclasses. For odrl: properties, the data resource being catalogued is assumed to be an instance of the odrl:Asset class.

Maturity Level 3

Maturity level 3 delves deeper into the dataset's content. It covers identification, language, and documentation, as well as whether the dataset contains synthetic data. It also supports the specification of whether the dataset contains information about individuals and indigenous data. Secondly, it expands on the temporal/geospatial aspects of the dataset.

Table 6. Dataset Maturity Level 3: Content.

Category	Description	Property	Value restriction
Content	Unique identifier for the dataset. Often assigned by creator or publisher. Not necessarily persistent or globally unique.	dct:identifier	rdfs:Literal
	Language	dct:language	dct:LinguisticSystem
	Documentation	dcat:landingPage	foaf:Document
	Contains data about individuals	containsIndividualData	xsd:boolean
	Contains data about identifiable individuals	containsIdentifiable IndividualData	xsd:boolean
	Contains Indigenous data	containsIndigenousData	xsd:boolean
	Contains synthetic data	containsSyntheticData	xsd:boolean
	Location of synthetic data generation documentation	syntheticData Documentation	rdfs:Resource
Temporal/ geospatial	Temporal resolution	dcat:temporalResolution	xsd:duration
	Spatial resolution in	dcat:spatialResolution	xsd:decimal

[16] https://www.w3.org/TR/odrl-model/#policy

Category	Description	Property	Value restriction
	metres	InMeters	
	Spatial resolution by administrative area (e.g., a city or neighbourhood)	:spatialResolution ByAdminArea	sc:AdminstrativeArea

Maturity Level 4

Maturity Level 4 focuses primarily on provenance of a dataset, which includes versioning information, and linkages to other versions. Secondly, it expands on the temporal/geospatial aspects of the dataset.

Table 7. Dataset Maturity Level 4: Provenance.

Category	Description	Property	Value restriction
Provenance	Version of the dataset	owl:versionInfo	rdfs:Literal
	Version notes	adms:versionNotes	rdfs:Literal
	Link to dataset that it is a version of	dct:isVersionOf	dcat:Dataset
	Link to datasets that are versions of it	dct:hasVersion	dcat:Dataset
	Provenance of the data	dct:provenance	dct:ProvenanceStatement
	Provenance document location	prov:wasQuoted From	prov:Entity
Temporal/ geospatial	Temporal resolution	dcat:temporal Resolution	xsd:duration
	Spatial resolution in metres	dcat:spatialResolution InMeters	xsd:decimal
	Spatial resolution by administrative area (e.g., a city or neighbourhood)	:spatialResolution ByAdminArea	sc:AdminstrativeArea

Maturity Level 5

Maturity level 5 focuses on attributes relevant to Indigenous data management policies. It includes properties relevant to determining whether the dataset contains data about communities that require additional policies with stricter privacy rules. For example, when a community as a whole owns a dataset, a steward acts as the rights holder on behalf of that community.

Table 8. Dataset Maturity Level 5: Indigenous.

Category	Description	Property	Value restriction
Access	Stewardship by an organisation that is accountable to the community.	hasSteward	foaf:Organization
Ownership	Community permission (who gave permission?)	communityRightsHolder	foaf:CommunityGroup
Temporal/ geospatial	Communities from which data is derived	spatialCommunity (sub-property of dct:spatial)	dct:Location

Notes:

- **odrl:Policy**[17]: A non-empty group of permissions (via the permission property) and/or prohibitions (via the prohibition property) and/or duties (via the obligation property). The Policy class is the parent class to the Set, Offer, and Agreement subclasses. For odrl: properties, the data resource being catalogued is assumed to be an instance of the odrl:Asset class.
- **containsIdentifiableIndividualData**: Does the data hold identifiable individual data that can be used to uniquely identify the individual data was collected about? If yes, the dataset is not anonymised.

Table 9. CommunityGroup class definition.

ClassProperty	Property	Value restriction
CommunityGroup	rdfs:subClassOf	foaf:Group

- **CommunityGroup**: A subclass of foaf:Group that is a group of agents in a community.

Table 10. communityRightsHolder property definition.

ClassProperty	Property	Value restriction
communityRightsHolder	rdfs:subPropertyOf	dct:rightsHolder

- **communityRightsHolder**: An agent that has the rights to manage access rights to Indigenous data. That person can be Indigenous themselves or a non-Indigenous agent that acts as the steward for access rights to the data.

Table 11. spatialCommunity property definition.

ClassProperty	Property	Value restriction
spatialCommunity	rdfs:subPropertyOf	dct:spatial

- **spatialCommunity**: A geospatial area occupied by or representative of a community.

[17] https://www.w3.org/TR/odrl-model/#policy

Maturity Level 6

Maturity Level 6 includes data quality and basic statistics about the dataset within the scope of the metadata provided. These are found less often in the literature but are relevant to the searcher in ascertaining relevance. The attributes are derived from VOID and DQV.

Table 12. Dataset Maturity Level 6: Statistics and Quality.

Category	Description	Property	Value restriction
Statistical	If tabular dataset, number of rows	void:rows	xsd:positiveInteger
	If tabular dataset, number of columns	void:columns	xsd:positiveInteger
	If tabular dataset, the number of filled-in data cells	void:cells	xsd:positiveInteger
	If RDF dataset, total number of triples	void:triples	xsd:postiveInteger
	If RDF dataset, total number of entities in the dataset	void:classes	xsd:postiveInteger
	If RDF dataset, total number of properties in the dataset	void:properties	xsd:postiveInteger
Quality	Description of data quality.	dqv:hasQualityAnnotation	dqv:QualityAnnotation
	Metrics for data quality property, like completeness, accuracy, etc.[18]	dqv:inDimension	dqv:Dimension

Notes:

- **dvq:inDimension** : Represents the dimensions a quality metric, certificate and annotation allow a measurement of.
- **dqv:QualityAnnotation**[19]: Represents quality annotations, including ratings, quality certificates or feedback that can be associated to datasets or distributions. Quality annotations must have one oa:motivatedBy statement with an instance of oa:Motivation (and skos:Concept) that reflects a quality assessment purpose. This instance is defined as dqv:qualityAssessment.
- **dvq:Dimension**[20]: Represents criteria relevant for assessing quality. Each quality dimension must have one or more metric to measure it. A dimension is linked with a category using the dqv:inCategory[21] property.

Capability Maturity Model Evaluation

In this section, the DMCMM metadata is evaluated as to its adequacy to support three uses: 1) dataset search; 2) FAIR evaluation; and 3) OCAP compliance.

[18] https://www.w3.org/TR/vocab-dqv/#examples
[19] https://www.w3.org/TR/vocab-dqv/#dqv:QualityAnnotation
[20] https://www.w3.org/TR/vocab-dqv/#dqv:Dimension
[21] https://www.w3.org/TR/vocab-dqv/#dqv:inCategory

Search

Chen et al.'s (2019) analysis of search logs clearly identifies a stratification of metadata used to search for datasets, based on frequency. The topic of the dataset is the most used metadata for searching, far exceeding any other. Kacprzak et al. (2019) highlight the importance of a rich taxonomy of topics to support the searcher. We note that this is also important in the cataloguing process. Topics are captured at Level 1 by the dcat:theme property, that has a value restriction of skos:Concept. It is recommended that the platform used to catalogue and search for datasets provide for taxonomies of topics across multiple domains.

Geospatial and temporal information are the next most commonly used terms for searching a dataset. Consequently, they are included at Level 1 as dct:spatial with its values restricted to dct:PeriodOfTIme, and dct:temporal with its values restricted to dct:Location. Data format is the next highest metadata used in search. This is included at Level 1 as dct:format with value restriction of dct:MediaType. Along with format, we included a datasetSize property with a value of xsd:integer, as identified by Kacprzak et al. (2019). In summary, from a search perspective, the most often used metadata attributes are included at Level 1. Assuming that Level 1 will be the level most often completed by a cataloguer, basic search will be supported.

FAIR

To support FAIR evaluation, each FAIR indicator is mapped to a corresponding DMCMM attribute. Metadata references in FAIR are assumed to align with DMCMM attributes. A maturity level of 0 indicates the indicator is inherently satisfied—for instance, the presence of rich metadata is guaranteed if DMCMM is used. When the level is P, the indicator's satisfaction depends on the underlying platform; in this study, the platform is CKAN with a DMCMM plugin. Some FAIR indicators lack direct DMCMM attribute matches but can still be evaluated through metadata or dataset examination, designated as Level E (Tables 13–15). Certain attributes may only partially fulfil indicators. For example, DMCMM's *dct:identifier* ensures uniqueness but not necessarily persistence, requiring additional validation by FAIR evaluation software.

Table 13. Essential FAIR properties and corresponding DMCMM properties.

FAIR ID		Essential indicators	Level	DMCMM attribute
F1	RDA-F1-01M	Metadata is identified by a persistent identifier	1	catalogueEntryIdentifier
	RDA-F1-01D	Data is identified by a persistent identifier	2	dct:identifier
	RDA-F1-02M	Metadata is identified by a globally unique identifier	1	catalogueEntryIdentifier
	RDA-F1-02D	Data is identified by a globally unique identifier	2	dct:identifier
F2	RDA-F2-01M	Rich metadata is provided to allow discovery	0	DMCMM
F3	RDA-F3-01M	Metadata includes the identifier for the data	2	dct:identifier
F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	P	CKAN+DMCMM
A1	RDA-A1-02M	Metadata can be accessed manually (i.e., with human intervention)	P	CKAN+DMCMM

FAIR ID		Essential indicators	Level	DMCMM attribute
	RDA-A1-02D	Data can be accessed manually (i.e., with human intervention)	P	CKAN+DMCMM
	RDA-A1-03M	Metadata identifier resolves to a metadata record	P	CKAN+DMCMM
	RDA-A1-03D	Data identifier resolves to a digital object	E	dct:identifier
	RDA-A1-04M	Metadata is accessed through standardised protocol	P	CKAN+DMCMM
	RDA-A1-04D	Data is accessible through standardised protocol	2	dcat:accessService
A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	P	CKAN+DMCMM
A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	P	CKAN+DMCMM
R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	E	dcat:accessURL dcat:accessService
R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	2	dct:LicenseDocument
R1.3	RDA-R1.3-01M	Metadata complies with a community standard	0	DMCMM is defined in terms of dc, dcat, etc.
R1.3	RDA-R1.3-01D	Data complies with a community standard	E	dcat:accessURL dcat:accessService
R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	0	DMCMM is defined in terms of dc, dcat, etc.

Table 14. Important FAIR properties and corresponding DMCMM properties.

FAIRID		Important indicators	Level	DMCMM attribute
A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	2	accessCategory
			2	dct:license
			2	dcat:accessURL
			2	dct:rightsHolder
			2	dcat:contactPoint
	RDA-A1-05D	Data can be accessed automatically (i.e., by a computer programme)	2	dcat:accessService

FAIRID		Important indicators	Level	DMCMM attribute
A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	2	dcat:accessService
I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	0	DMCMM
	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	E	
	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	0	DMCMM
	RDA-I1-02D	Data uses machine-understandable knowledge representation	E	
I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	0	DMCMM
I3	RDA-I3-01M	Metadata includes references to other metadata	E	
	RDA-I3-03M	Metadata includes qualified references to other metadata	E	
R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	2	dct:license
			2	accessCategory
			5	odrl:hasPolicy
	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	2	dct:license
R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	4	owl:versionInfo
			4	adms:versionNotes
			4	dct:isVersionOf
			4	dct:hasVersion
			4	dct:provenance
			4	prov:wasQuotedFrom
R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	2	dcat:accessURL dcat:accessService

Table 15. Useful FAIR properties and corresponding DMCMM properties.

FAIR ID		Useful indicators	Level	DMCMM attribute
A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	2	dcat:accessService
I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	2	dcat:accessURL

FAIR ID		Useful indicators	Level	DMCMM attribute
I3	RDA-I3-01D	Data includes references to other data	2	dcat:accessURL
	RDA-I3-02M	Metadata includes references to other data	0	Determined by evaluator
	RDA-I3-02D	Data includes qualified references to other data	2	dcat:accessURL
	RDA-I3-04M	Metadata include qualified references to other data	0	Determined by evaluator
R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	4	owl:versionInfo
			4	adms:versionNotes
			4	dct:isVersionOf
			4	dct:hasVersion
			4	dct:provenance
			4	prov:wasQuotedFrom

OCAP

In Table 16, we identify possible indicators for each OCAP theme. For each indicator, we identify the DMCMM attributes that can be used to evaluate the indicator.

Table 16. OCAP properties and corresponding DMCMM properties.

	Indicator	Level	DMCMM attribute
Ownership	Identify the community from whom the data is drawn.	5	spatialCommunity
	Identify the organisation that “owns” the data.	5	communityRightsHolder
	Identify if the dataset contains iIndigenous data.	3	contansIndigenousData
Control	Licence to be agreed to by user.	2	dct:license
	Data-sharing agreement that defines who, what, and how data is to be shared, including beyond the terms of the licence.	2	odrl:hasPolicy
Access	Access methods and limitations.	2	accessCategory
		2	dcat:accessURL
		1	dct:format
		2	dcat:downloadURL
Possession	Identify the steward who manages the data.	5	hasSteward

Implementation

The Maturity Model has been implemented as a CKAN[22] plugin (CKANext-udc[23]) to integrate the model into the CKAN dataset cataloguing process. The plugin is intended for publishers and managers of open data portals, specifically those utilising the CKAN architecture. The backend integration with the CKAN architecture has been designed in accordance with existing CKAN guidelines to ensure seamless integration and maintenance. The interface has been reviewed by several urban data curators, with changes implemented to ensure an easy dataset curation and search processes. The plugin facilitates the inclusion of custom fields, allows for their reordering, and categorises them into distinct maturity levels. It also allows for integration with a graph database to store each catalogue entry as a knowledge graph built on top of the ontology. The maturity model itself is defined as an ontology[24] and implemented in OWL.

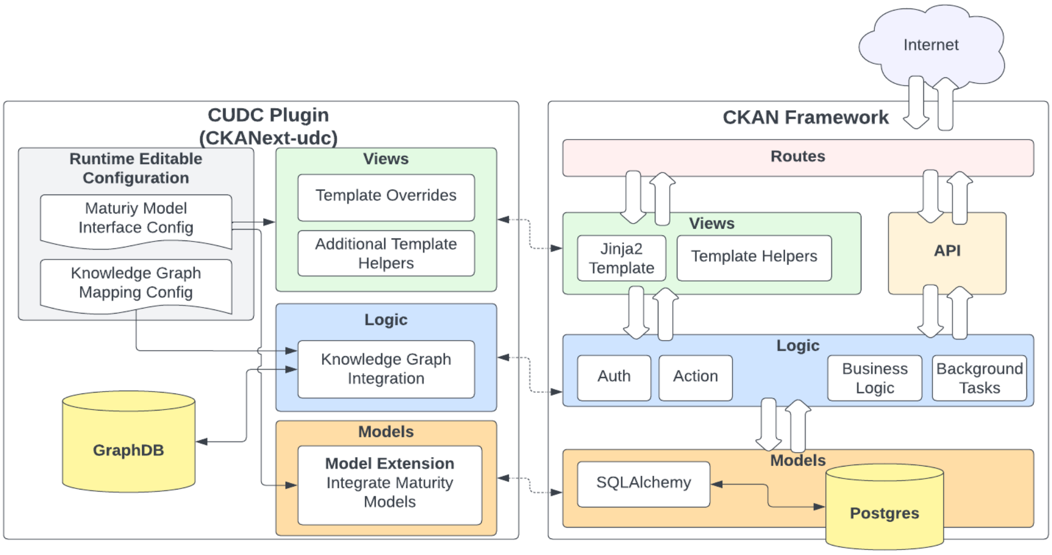

Internet
CUDC Plugin (CKANext-udc)
Runtime Editable Configuration
Maturiy Model Interface Config
Knowledge Graph Mapping Config
Views
Template Overrides
Additional Template Helpers
Logic
Knowledge Graph Integration
Models
Model Extension
Integrate Maturity Models
GraphDB
CKAN Framework
Routes
Views
Jinja2 Template
Template Helpers
API
Logic
Auth
Action
Business Logic
Background Tasks
Models
SQLAlchemy
Postgres

Figure 1. CUDC plugin (CKANext-udc) architecture.

The CKAN extension architecture, shown in Figure 1, is developed as a CKAN plugin. The plugin interacts with the CKAN architecture by modifying how catalogue metadata is collected from the user, how it is displayed to the user for data entry and viewing, and how data is stored in the database. Data entry views enable the seamless grouping and entry of maturity model properties in CKAN.

The plugin first refines terminology by renaming “Dataset” to “Catalogue Entry” and “Resource” to “Dataset”, aligning with the maturity model. It then integrates the maturity model into the edit and view pages of each catalogue entry and adds an advanced filter for improved catalogue search. Additionally, it overrides CKAN’s default logic to ensure the knowledge graph is updated during create, update, or delete operations, based on a predefined field-to-ontology mapping. The plugin also extends CKAN’s default metadata model by incorporating maturity model fields and storing data in both the knowledge graph and CKAN’s Postgres database. This dual storage enables backward compatibility with CKAN up to Version 2.11, while maintaining all core functionalities, including API access for data import/export. Advanced users can continue managing catalogue entries via CKAN’s Python interface and scripts.

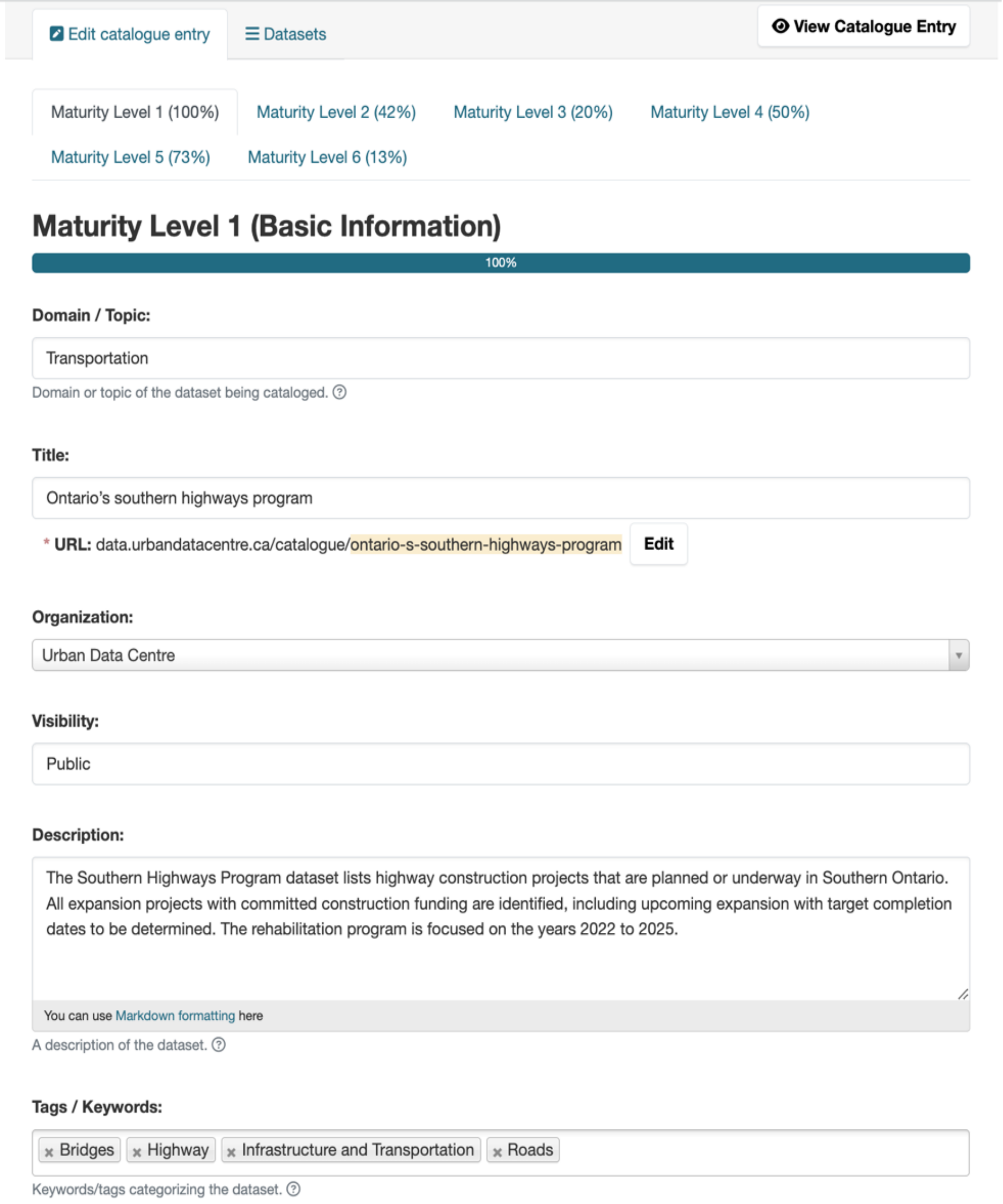
[22] CKAN: https://docs.ckan.org/

[23] CKAN plugin: https://github.com/csse-uoft/ckanext-udc

[24] Dataset Maturity Model Ontology: https://github.com/csse-uoft/maturity-model-ontology

Maturity Model Views

A user can enter a new catalogue entry and its maturity levels properties, as shown in Figure 2. Each maturity level can be completed either partially or entirely. The plugin calculates and displays the completion percentage of the maturity levels for a catalogue entry. Maturity Model properties defined in the ontology are represented as text fields, date-time fields, or dropdown-select fields. Values in dropdown-select fields can either be entered manually or fetched from an ontology that defines the property, ensuring that the metadata aligns with imported ontologies. For example, the dataset file type on Level 3 is represented as an instance of the class *dct:MediaType*. As such, it may include any of the types in IANA,[25] including simple formats such as "json" and "csv", application formats such as "pdf" and "vnd.ms-excel", or complex formats, such as "ace+json", "csvm+json", "csv-schema", and so on. The formats are displayed to a user during the data entry step as one of the possible options.



Maturity Level 1 (Basic Information)

Domain / Topic: Transportation

Domain or topic of the dataset being cataloged.

Title: Ontario's southern highways program

* URL: data.urbandatacentre.ca/catalogue/ontario-s-southern-highways-program

Organization: Urban Data Centre

Visibility: Public

Description: The Southern Highways Program dataset lists highway construction projects that are planned or underway in Southern Ontario. All expansion projects with committed construction funding are identified, including upcoming expansion with target completion dates to be determined. The rehabilitation program is focused on the years 2022 to 2025.

You can use Markdown formatting here

A description of the dataset.

Tags / Keywords: × Bridges × Highway × Infrastructure and Transportation × Roads

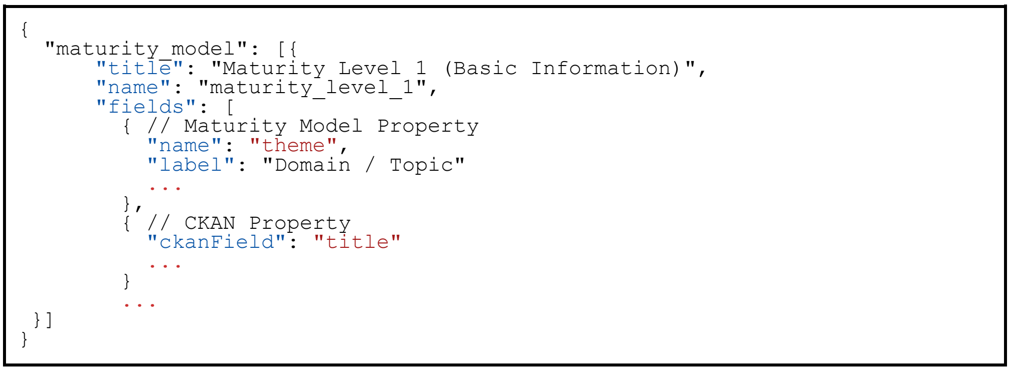
Keywords/tags categorizing the dataset.

[25] Internet Assigned Numbers Authority (IANA) – Media Types: https://www.iana.org/assignments/media-types/media-types.xhtml

Figure 2. Maturity Model entry form, with six maturity levels and showing completion percentages for each level.

Maturity Model View Configuration

Administrators have flexibility to adjust the plugin settings directly from the web interface, including on the data entry screen and by mapping between entry fields and the maturity model properties. CKAN provides several predefined dataset properties, such as “title”, “tags”, “note”, “author”, and more. The DMCMM used by the plugin is defined in a configuration file (JSON format), as illustrated in Figure 3, to organise the layout and properties when they are entered and displayed. Each maturity level tab, as shown in Figure 2, has a configuration entry under the “maturity_model” dictionary key. Each maturity level has a “title”, “name”, and an array of “fields" that specify the level’s properties. Existing CKAN properties that are also specified in the maturity model are mapped to the model’s ontology properties. If the property is a CKAN property, “ckanField” is used, and the property name is the value in “ckanField”. If the maturity model introduced a new property, the configuration requires a “name” and “label” entries in the “fields” key.



```
{
  "maturity_model": [{
      "title": "Maturity Level 1 (Basic Information)",
      "name": "maturity_level_1",
      "fields": [
        { // Maturity Model Property
          "name": "theme",
          "label": "Domain / Topic"
          ...
        },
        { // CKAN Property
          "ckanField": "title"
          ...
        }
        ...
 }]
}
```

Figure 3. CKAN Maturity Model configuration for data entry.

Users can search for catalogue entries using a text search field or a filter applicable to several key maturity model properties, as shown in Figure 4a. Search functionality utilises CKAN’s built-in support for Apache Solr library[26] for indexing and searching text data. Results are shown in the view list screen. For catalogue entries that match a filter or search query, a side panel, shown in Figure 4b, displays an aggregate sum for indexed properties. The aggregate results are limited to the subset of results that match the original query.

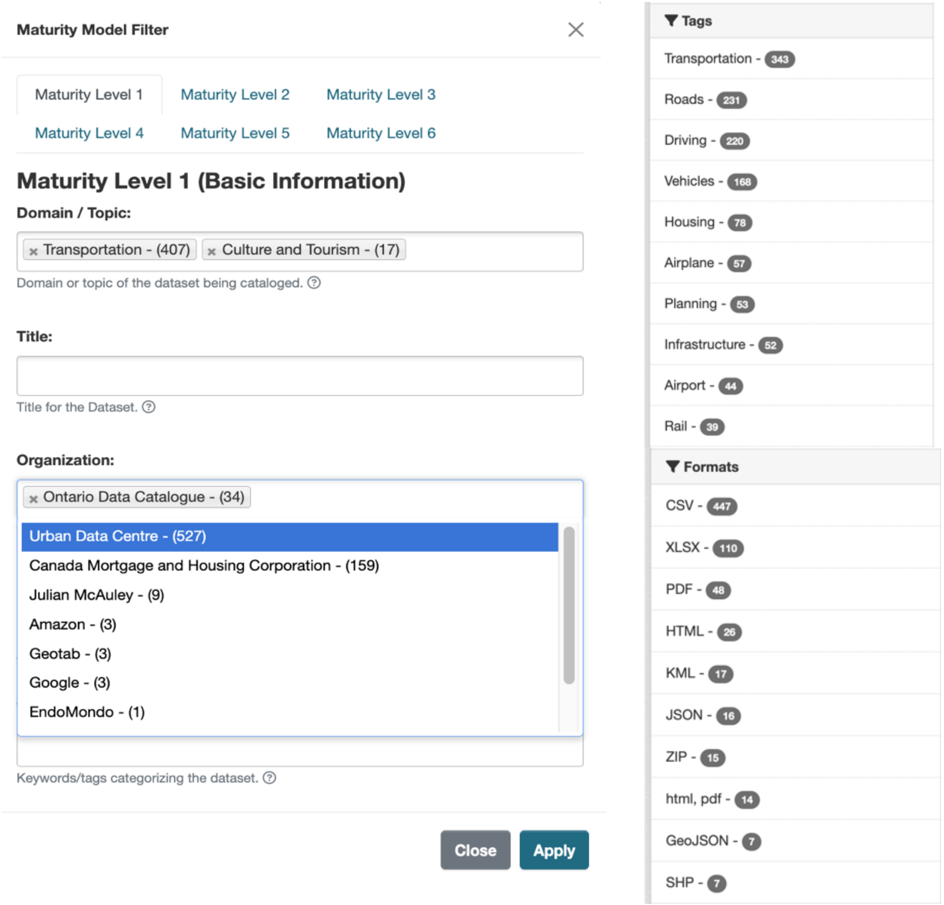
By utilising the text search and filter search, the plugin allows users of the catalogue to search for datasets. For example, free text search provides insights into how users refer to or spell various maturity model properties. Filter search provides us with similar monitoring capabilities but on a catalogue property level.

Storing Dataset Metadata Capability Maturity Models in a Knowledge Graph

All catalogue entries, and their maturity model data, are stored in Postgres by CKAN. There is also an option to configure the CKAN plugin to connect with a graph database. The graph database uses the DMCMM ontology as its schema to store catalogue entries. To do this, one must supply the necessary mappings to the data present in the knowledge graph. As catalogue entries are updated or deleted, the plugin produces corresponding SPARQL queries based on these mappings, ensuring that the knowledge graph remains in sync with CKAN’s database. An illustrative example of this mapping configuration utilises a structure reminiscent of JSON-LD in

[26] Apache Solr library: https://solr.apache.org/

Figure 5. For dynamic elements, such as generating an UUID value, the syntax allows one to provide a JavaScript helper function[27] call, such as “generate_uuid()”, as demonstrated below.



Maturity Level 1 (Basic Information)

Domain / Topic: × Transportation - (407) × Culture and Tourism - (17)

Domain or topic of the dataset being cataloged.

Title:

Title for the Dataset.

Organization: × Ontario Data Catalogue - (34)

Urban Data Centre - (527)
Canada Mortgage and Housing Corporation - (159)
Julian McAuley - (9)
Amazon - (3)
Geotab - (3)
Google - (3)
EndoMondo - (1)

Keywords/tags categorizing the dataset.

Tags:
Transportation - 343
Roads - 231
Driving - 220
Vehicles - 168
Housing - 78
Airplane - 57
Planning - 53
Infrastructure - 52
Airport - 44
Rail - 39

Formats:
CSV - 447
XLSX - 110
PDF - 48
HTML - 26
KML - 17
JSON - 16
ZIP - 15
html, pdf - 14
GeoJSON - 7
SHP - 7

a) Filter search screen (entry screen)

b) Filter aggregation (side panel)

Figure 4. Maturity Model search capability.

```
{
  "mappings": {
    "@context": {
        // Define various namespaces that are used below.
        "xsd": "http://www.w3.org/2001/XMLSchema#",
        "dcat": "http://www.w3.org/ns/dcat#",
        "foaf": "http://xmlns.com/foaf/0.1/",
        "dct": "http://purl.org/dc/terms/"
    },
    // The URI of the catalogue entry, values in the curly bracket { }
    // will be evaluated at runtime. `ckanField` is a dictionary
    // and `ckanField.id` is a unique id of this catalogue entry.
    "@id": "http://data.urbandatacentre.ca/catalogue/{ckanField.id}",
    // The RDF Type of the catalogue
    "@type": "http://data.urbandatacentre.ca/Catalogue",
    // Author name and email are mapped into a `foaf:Agent` instance. Contents in the
    // curly bracket { } will be evaluated in the runtime.
    "dct:creator": {
      "@id": "http://data.urbandatacentre.ca/creator/{generate_uuid()}",
      "@type": "foaf:Agent",
      "foaf:mbox": "{ckanField.author_email}",
      "foaf:name": "{ckanField.author}"
    },
    // title is mapped to `dct:title`
    "dct:title": {
      "@type": "xsd:string",
      "@value": "{ckanField.title}"
    },
    // Published Date is mapped to `dct:issued`
    "dct:issued": {
      "@type": "xsd:date",
      "@value": "{to_date(published_date)}" // to_date(…) is a helper function
    }
  }
}
```

[27] The plugin allows for custom helper functions to be defined in “ckanext/udc/graph/mapping_helpers.py”.

Figure 5. CKAN Maturity Model configuration for data entry.

Metadata Availability Evaluation

Over a period of 12 months, a team of eight cataloguers scanned the web for Canadian urban-related datasets, whether open, closed, and through a web service, searching primarily for the themes of “transportation”, “housing”, “bylaws”, “homelessness”, and “culture and tourism”.[28] In total, 1,162 datasets were catalogued, of which 83% related to these five themes. Figure 6 shows a word cloud of keywords associated with each catalogue entry. The keywords have a strong correlation with the selected themes. For each dataset, the cataloguers extracted as much information as available to complete the metadata properties in the maturity model.

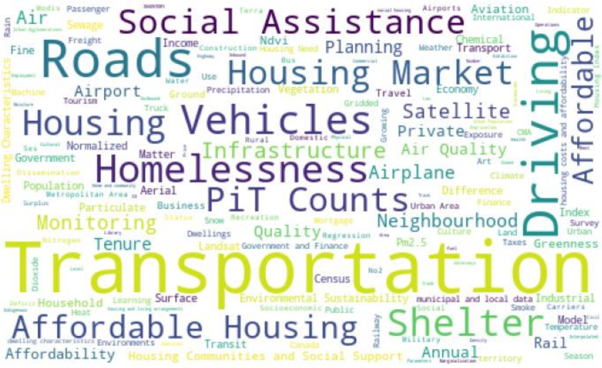


Figure 6. Word cloud of catalogue keywords.

The evaluation identifies which dataset properties are most readily available and ranks maturity levels accordingly. Completion rates across seven information categories were assessed to understand what cataloguers can typically access (Figure 7). “Content” and “Access” properties are the most common, each exceeding 60%, although “Content” is lowered by missing values for “file size” and “Metadata Identifier.” “Temp-geo” (temporal and geospatial coverage) appears in 60% of cases, while “Ownership” properties show 37% completion. Though “Ownership” and “Access” relate to licensing, access details are 33% more complete, suggesting usability is prioritised over ownership. “Quality” metrics, which rely on publisher-provided information, appear only 24% of the time. “Provenance” is similarly rare at 17%, indicating it may be seen as nonessential. The “Statistics” category, reflecting dataset size, has the lowest completion at 12% and the highest variability, making it the hardest information to obtain.

[28] Prior to cataloguing Canadian urban datasets, two studies were undertaken to ascertain data requirements for research on the themes of transportation (Pandya, 2023a) and housing (Pandya, 2023b).

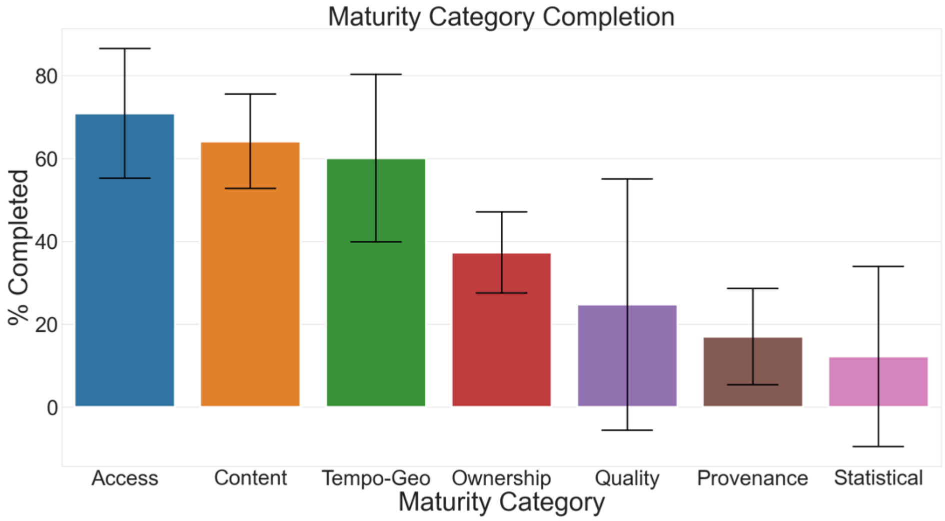


Figure 7. Percentage of category completion.

Finally, our evaluation focuses on ascertaining, for each level of the maturity model, the percentage of metadata properties that are available. The percentage of completed properties on each level are given in Figure 8. Maturity Level 1 emphasises the metadata predominantly employed for dataset searches. The completion rate for this level averages 75%, with a standard deviation of 11. Maturity Level 2 focuses on access and ownership of metadata. This level records a completion rate of 64% and a standard deviation of 12. Level 3 expands on content, provenance and temporal/geospatial information. It registered a 56% completion rate and a standard deviation of 17. Level 4 assesses the existence of data on individuals versus aggregates, any limits on use, and whether data pertinent to certain communities is captured or not. Level 4 has a completion rate of 32% and associated standard deviation of 14%. Maturity Level 5 focuses on attributes relevant to indigenous data management policies, having the lowest completion level at 0.5% and a high standard deviation of 4%. Finally, Maturity Level 6 is centred on the statistical and quality properties of the data, including the number of triples and concepts in triple stores or the row and column count in tabular datasets. The completion metric for this level is low at 15%, with a high standard deviation relative to the mean of 18.

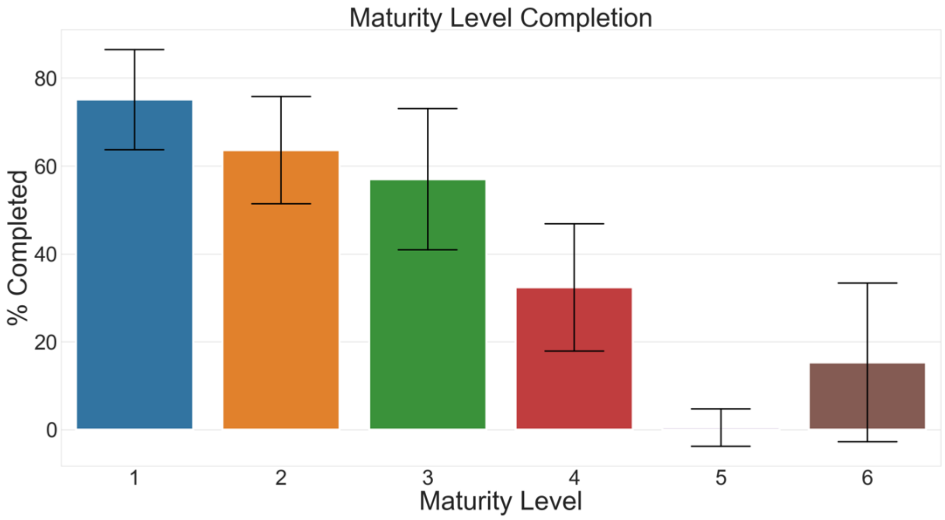


Figure 8. Percentage of maturity level completion.

As expected, property availability generally followed the maturity levels, except for Maturity Level 5. While Maturity Level 1 was anticipated to exceed 80%, its low completion rate stems from missing file size and metadata in the "Content" category. High standard deviations in Levels 3–6 indicate significant variability that requires further analysis. Level 4's low rate is attributed to missing data on related datasets and versioning issues. Level 6 showed the lowest completion rate, due to the division of statistical properties across triple stores and tabular datasets, where full completion is assumed for multimodal datasets. Identifying such data also requires counting mode-specific data points, which was not always feasible. Adding "reviews" to quality metrics may improve completion, and this effort is ongoing.

Basic descriptive properties at Level 1 remain the easiest to identify and serve as effective search criteria. Level 2 and 3 properties are similarly obtainable, showing that locating access information is as difficult as finding aggregation-level details. Level 4 provenance and versioning details are harder to identify. Level 5's low completion highlights the need for better identification of Indigenous data and policies on access and ownership. Finally, Level 6's low completion rate reflects the complexity of its dual-modality structure and the lack of readily available, high-quality descriptions.

Conclusion

This paper addresses the difficulty of finding relevant data despite its abundance, citing issues such as poor metadata, inadequate presentation, and the dataset creators' lack of knowledge about specifying appropriate metadata. It emphasises the complexity of metadata documentation, given the wide range of properties in datasheets and vocabularies like DCAT, Schema.org, PROV, and DQV. A review of dataset search literature identifies key types of information that searchers use to locate relevant data.

To address these issues, the paper introduces the Dataset Metadata Capability Maturity Model (DMCMM), structured into levels to balance documentation effort with the need for sufficient metadata for discovery, relevance, and content understanding. Attribute selection in the model is based on requirements such as dataset discovery, access conditions, content understanding, derivation, FAIRness evaluation, and support for Indigenous data sovereignty. Metadata levels are organised progressively, starting with discovery, followed by access, content, provenance, Indigenous data, and data quality and statistics.

The model is operationalised via the CKANext-udc plugin for CKAN, which adds maturity-level fields, reorganises the user interface, and integrates a knowledge graph based on an OWL-defined ontology. It updates the knowledge graph when catalogue entries change and supports advanced filtering, while retaining CKAN's core functions, including its API and Python interface. The model is in use in the Canadian Urban Data Catalogue (CUDC), which hosts over 1,200 datasets on CKAN with the CKANext-udc plugin.[29]

References

Albertoni, R., & Isaac, A. (2016). *Data on the Web Best Practices: Data Quality Vocabulary* (W3C Working Group Note 14 December 2016). Retrieved from https://www.w3.org/TR/vocab-dqv/

Albertoni, R., Browning, D., Cox, S., Beltran, A. G., Perego, A., & Winstanley, P. (2023). *Data Catalog Vocabulary (DCAT) – Version 3* (W3C Working Draft 07 March 2023). Retrieved from https://www.w3.org/TR/vocab-dcat-3/

[29] Canadian Urban Data Catalogue (CUDC): data.urbandatacentre.ca

Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2011). *Describing Linked Datasets with the VoID Vocabulary* (W3C Interest Group Note 03 March 2011). Retrieved from https://www.w3.org/TR/void/

Assaf, A., Troncy, R., & Senart, A. (2015). HDL - Towards a harmonized dataset model for open data portals. In *Proceedings of the 2nd International Workshop on Dataset Profiling & Federated Search for Linked Data* (pp. 62–74). Retrieved from https://ceur-ws.org/Vol-1362/PROFILES2015_paper3.pdf

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., & Stall, S. (2020), The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. *Data Science Journal* 19, 41. https://doi.org/10.5334/dsj-2020-041

Berkley, C., Bowers, S., Jones, M. B., Madin, J. S., & Schildhauer, M. (2009). Improving data discovery for metadata repositories through semantic search. In *2009 International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 1152–1159). https://doi.org/10.1109/CISIS.2009.122

Carroll, S., Garba, I., Figueroa-Rodríguez, O., Holbrook, J., Lovett, R., Materechera, S., ... Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal* 19, 43. https://doi.org/10.5334/dsj-2020-043

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L. D., Kacprzak, E., & Groth, P. (2020). Dataset search: a survey. *The VLDB Journal* 29(1), 251–272. https://doi.org/10.1007/s00778-019-00564-x

Chen, J., Wang, X., Cheng, G., Kharlamov, E., & Qu, Y. (2019). Towards more usable dataset search: from query characterization to snippet generation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2445–2448). https://doi.org/10.1145/3357384.3358096

Chiu, T. H., Chen, H. L., & Cline, E. (2023). Metadata implementation and data discoverability: A survey on university libraries' Dataverse portals. *The Journal of Academic Librarianship* 49(4), 102722. https://doi.org/10.1016/j.acalib.2023.102722

Chua, U. C., Santiago, K. L., Ona, I. B. M., Peña, R. M. N., Marasigan, G. Z. S., Delos Reyes, P. G. A., & Samson, B. P. V. (2020). From Access to Effective Use: Open Data Portals for Everyday Citizens. In *AsianCHI '20: Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures* (pp. 61–64). https://doi.org/10.1145/3391203.3391219

Commonwealth of Australia. (2024). *Framework for Governance of Indigenous Data.* Retrieved from https://www.niaa.gov.au/resource-centre/framework-governance-indigenous-data

Cooper, A., Gagnon, M., Leahey, A., Paquette-Bigras, E., Perrier, L., Steeleworthy, M., & Taylor, S. (2019). *Dataverse North Metadata Best Practices Guide.* Retrieved from http://hdl.handle.net/2429/72537

Cotton, F., Gillman, D. W., & Joque, Y. (2015). XKOS - An RDF Vocabulary for Describing Statistical Classifications. *IASSIST Quarterly* 38(4), 47–57. https://doi.org/10.29173/iq900

Elmqvist, N. (2011). Embodied human-data interaction. In *ACM CHI 2011 Workshop "Embodied Interaction: Theory and Practice in HCI"* (Vol. 1, pp. 104–107). Retrieved from http://www.elisevandenhoven.com/publications/antle-chi11wp.pdf

Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., … Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data* 6(1), 1–9. https://doi.org/10.1038/s41597-019-0031-8

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM* 64(12), 86–92. https://doi.org/10.1145/3458723

Hodgson, R. (2022). *Quantities, Units, Dimensions and Types (QUDT) Schea – Version 2.1.21.* Retrieved from https://www.qudt.org/doc/DOC_SCHEMA-QUDT.html

Iannella, R., & Villata, S. (2018). *ODRL Information Model 2.2* (W3C Recommendation, 15 February 2018). Retrieved from https://www.w3.org/TR/odrl-model/

Kacprzak, E., Koesten, L., Ibáñez, L. D., Blount, T., Tennison, J., & Simperl, E. (2019). Characterising dataset search—An analysis of search logs and data requests. *Journal of Web Semantics* 55, 37–55. https://doi.org/10.1016/j.websem.2018.11.003

Koesten, L. M., Kacprzak, E., Tennison, J. F., & Simperl, E. (2017). The Trials and Tribulations of Working with Structured Data: a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 1277–1289). https://doi.org/10.1145/3025453.3025838

Kunze, S. R., & Auer, S. (2013). Dataset retrieval. In *ICSC '13: Proceedings of the 2013 IEEE Seventh International Conference on Semantic Computing* (pp. 1–8). https://doi.org/10.1109/ICSC.2013.12

Lebo, T., Sahoo, S., & McGuinness, D. (2013). *PROV-O: The PROV Ontology* (W3C Recommendation, 30 April 2013). Retrieved from https://www.w3.org/TR/prov-o/

Mecredy, G., Sutherland, R., & Jones, C. (2018). First Nations data governance, privacy, and the importance of the OCAP® principles. *International Journal of Population Data Science* 3(4). https://doi.org/10.23889/ijpds.v3i4.911

Neumaier, S., Umbrich, J., & Polleres, A. (2017). Lifting data portals to the web of data. In *Proceedings of the Workshop on Linked Data on the Web* (pp. 1–10). Retrieved from https://ceur-ws.org/Vol-1809/article-03.pdf

Noy, N., Burgess, M., & Brickley, D. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In L. Liu & R. White (Eds.), *The World Wide Web Conference* (pp. 1365–1375). https://doi.org/10.1145/3308558.3313685

Ojo, A., Porwol, L., Waqar, M., Stasiewicz, A., Osagie, E., Hogan, M., … Zeleti, F. A. (2016). Realizing the innovation potentials from open data: Stakeholders' perspectives on the desired affordances of open data environment. In *Working Conference on Virtual Enterprises* (pp. 48–59). https://doi.org/10.1007/978-3-319-45390-3_5

Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., & Ojo, A. (2017). Usability evaluation of an open data platform. In C. C. Hinnant & O. Adegboyega (Eds.), *Proceedings of the 18th annual international conference on digital government research* (pp. 495–504). https://doi.org/10.1145/3085228.3085315

Pandya, M. (2023a). *Transportation problems and data requirements: Report of the Transportation Panel.* Retrieve from https://storage.googleapis.com/wzukusers/user-12947767/documents/c4609af45a0546deb1a4468616c808ad/UDC%20Transportation%20Panel%20Report%20v3.pdf

Pandya, M. (2023b). *Affordable housing problems and data requirements: Report of the Affordable Housing Panel.* Retrieved from https://storage.googleapis.com/wzukusers/user-12947767/documents/32479acd560c4ed1be6d1f8a393d8846/UDC%20-%20Affordable%20Housing%20Panel%20Report%20-%20v3.pdf

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). Capability maturity model, version 1.1. *IEEE Software* 10(4), 18–27. https://doi.org/10.1109/52.219617

Poynter, W., & Spiegel, J. (2016). Protocol Development for Large-Scale Metadata Archiving using DDI-Lifecycle. *IASSIST Quarterly* 39(3), 23–29. https://doi.org/10.29173/iq128

Project Open Data. (2014). *DCAT-US Schema v1.1 (Project Open Data Metadata Schema).* Retrieved from https://project-open-data.cio.gov/v1.1/schema

Rueda, L., Fenner, M., & Cruse, P. (2017). Datacite: Lessons learned on persistent identifiers for research data. *The International Journal of Digital Curation* 11(2), pp. 39–47. https://doi.org/10.2218/ijdc.v11i2.421

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 269–276). https://doi.org/10.1145/169059.169209

Sansone, S. A., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J. S., Xu, H., ... Ohno-Machado, L. (2017). DATS, the data tag suite to enable discoverability of datasets. *Scientific Data* 4(1), 1–8. https://doi.org/10.1038/sdata.2017.59

Sharifpour, R., Wu, M., & Zhang, X. (2023). Large-scale analysis of query logs to profile users for dataset search. *Journal of Documentation* 79(1), 66–85. https://doi.org/10.1108/JD-12-2021-0245

Thomas, W., Gregory, A., Gager, J., Johnson, J., and Wackerow, J. (2014). *Technical Document for DDI 3.2* (Data Documentation Initiative). Retrieved from https://ddialliance.org/hubfs/Specification/DDI-Lifecycle/3.2/XMLSchema/HighLevelDocumentation/DDI_Part_I_TechnicalDocument.pdf

Thornton, G. M., & Shiri, A. (2021). Challenges with organization, discoverability and access in Canadian open health data repositories. *Journal of the Canadian Health Libraries Association* 42(1). https://doi.org/10.29173/jchla29457

Van Nuffelen, B. (2022). *DCAT Application Profile for data portals in Europe Version 2.1.1.* Retrieved from https://github.com/SEMICeu/DCAT-AP/releases/tag/v2.1.1

White, R. W., Dumais, S. T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In R. Baeza-Yates & P. Boldi (Eds.), *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 132–141). https://doi.org/10.1145/1498759.1498819