KDI : Knowledge and Data Integration

Catalog and Metadata

Danish Cheema, Mayukh Bagchi

Metadata lecture structure

1. Metadata : definition + scope (quality, reusability)
2. Catalog : definition, services, CKAN
3. DCAT : standard definition
4. Datasets Metadata : mandatory, recommended, optional
5. Distributions Metadata : mandatory, recommended, optional
6. Examples:
   - datasets examples
   - teleologies examples
   - DCAT files examples

Contents

Metadata: Definition

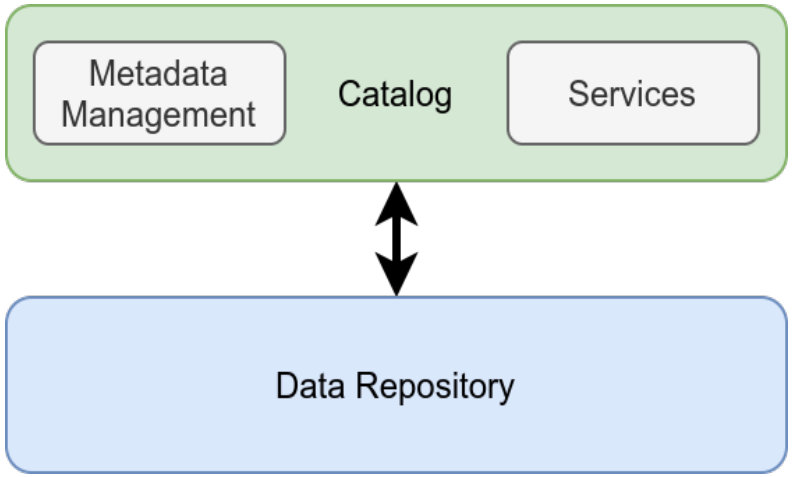
- Metadata is *"structured information that describes, explains, locates or otherwise makes it easier to retrieve, use or manage an information resource"* [NISO, 2017]
- Thus, metadata, in general, has three main purposes -
  1. information resource description
  2. information resource organization
  3. information resource discovery
- In the context of *iTelos* methodology, the information resources are *data resources* and *teleologies*
- We discuss the scope of metadata with respect to quality and reusability.

Metadata: Scope

- In the context of *iTelos*, the scope of metadata is important with respect to two dimensions - *Quality* and *Reusability*
- Firstly, metadata allows the user to determine *data quality and fitness* for their DI project by helping them assess the usefulness of a data resource or a teleology relative to their requirement specification.
- Secondly, *“iTelos assumes the existence of a repertoire of teleologies and provides a rich set of metadata for reusing them"* (Giunchiglia *et al.*, 2021)
- It is essential to always adhere to a *metadata standard* for ensuring *reusability* and *shareability* of data and knowledge resources.
- We follow a customized subset of W3C’s Data CATalog Version 2 (DCAT2) metadata vocabulary in the context of *iTelos*.

Contents

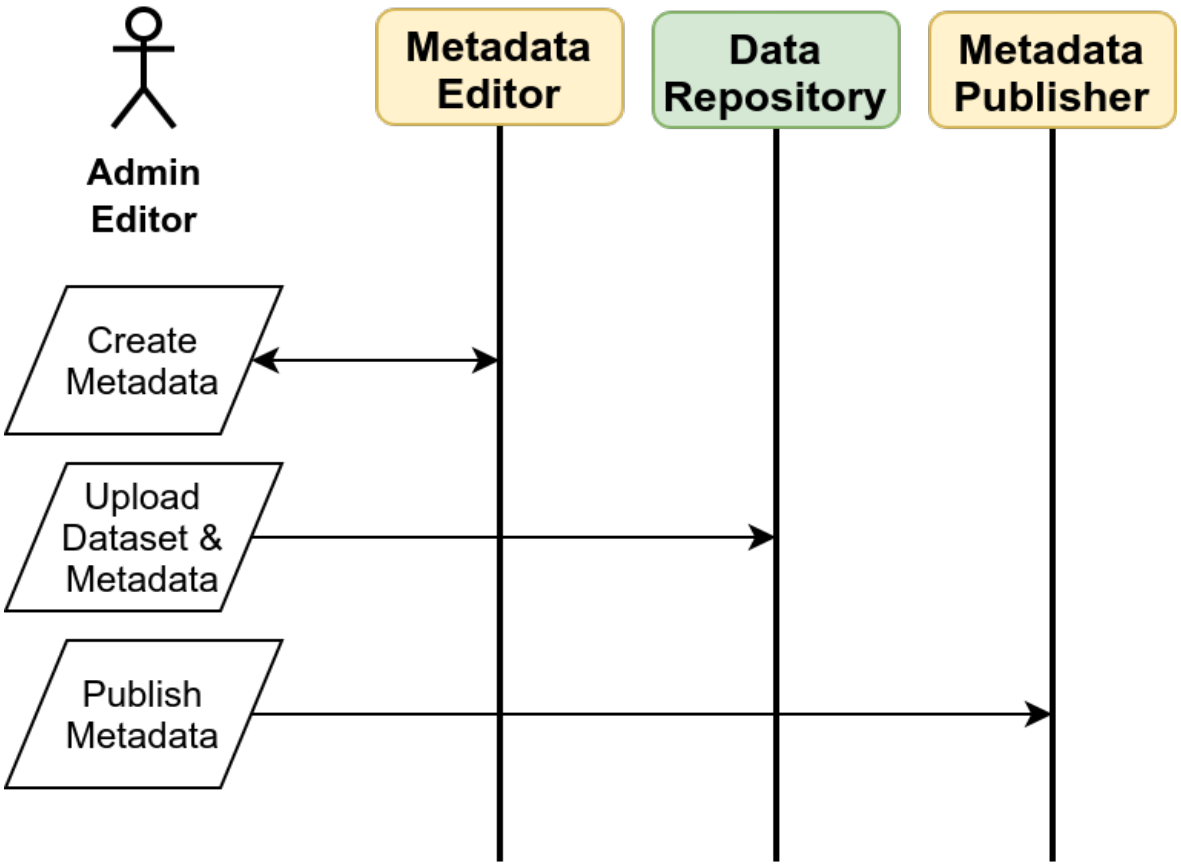
Catalog - Overview



- A web-base unique access point for data repositories
- Smart search and easier navigation for datasets
- Catalog only host dataset metadata
- Build on top of CKAN framework
- The metadata follows DCAT-AP standard

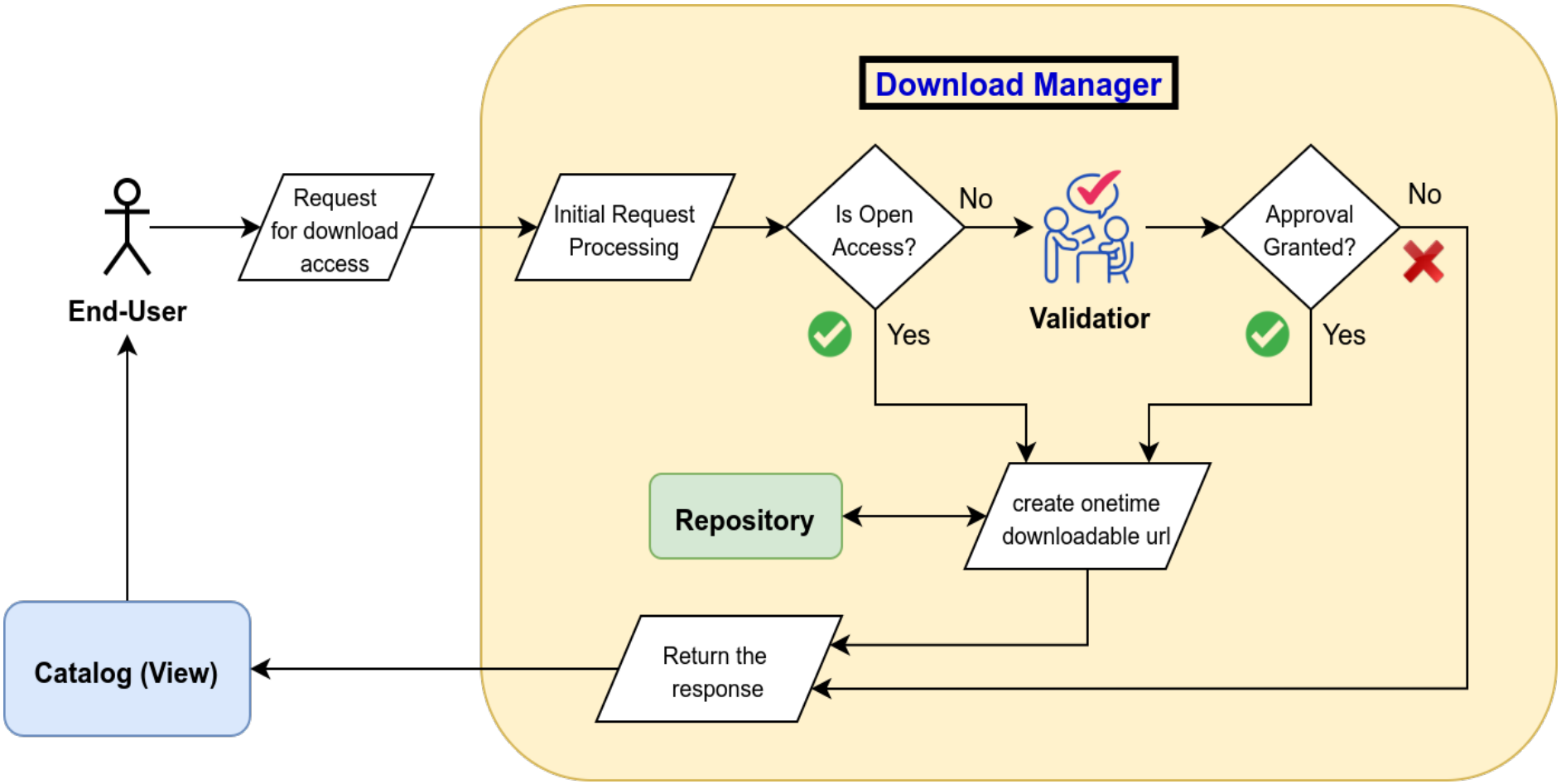
Catalog - Services (1/2)

- Metadata Editor
- Metadata Publisher



Catalog - Services (2/2)

- Download Manager

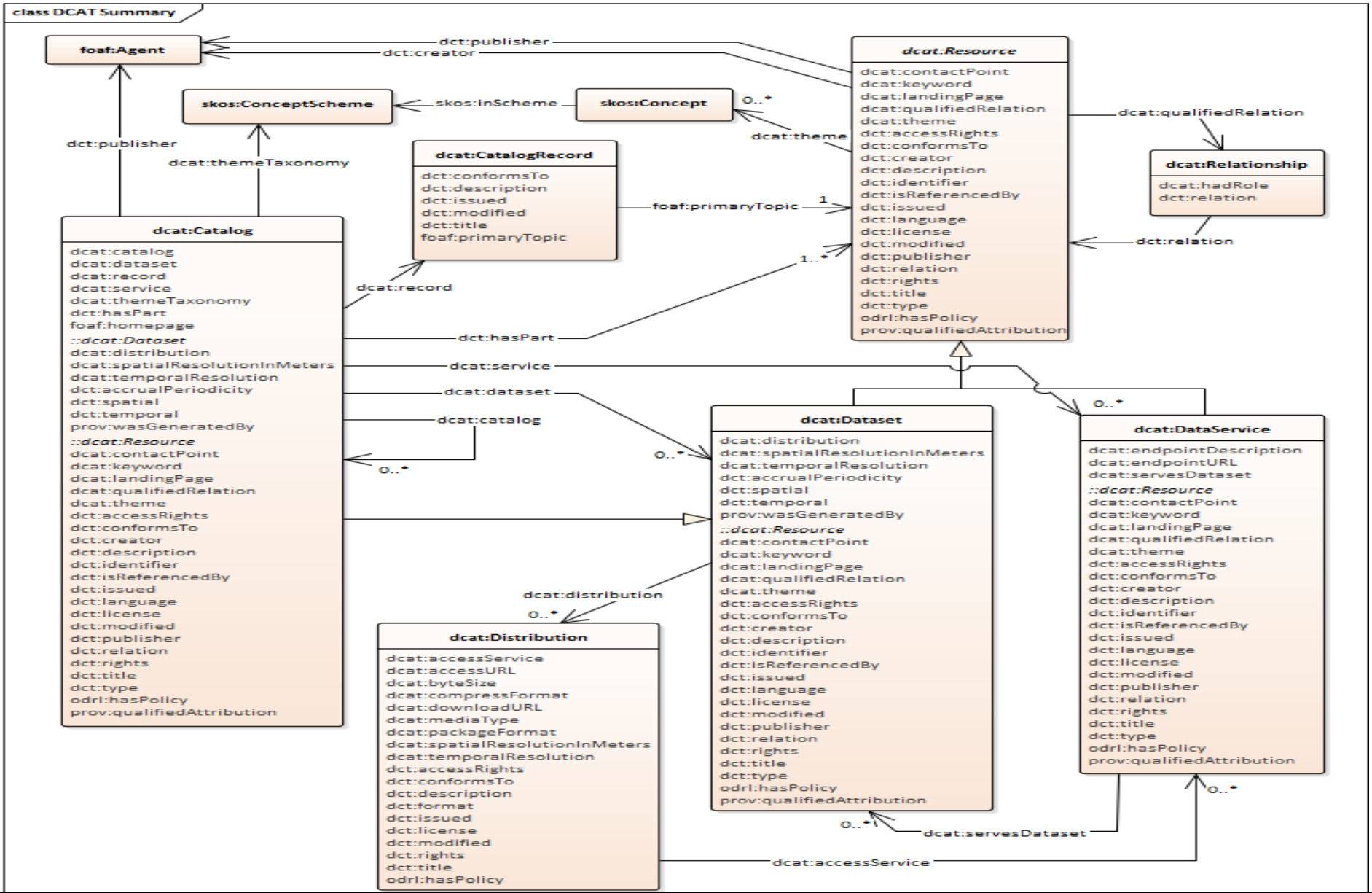


Contents

DCAT2 Introduction (1/4)

- DCAT 2 stands for Data CATalog (DCAT) vocabulary (Version 2). It is a W3C recommended metadata standard expressed as an RDF vocabulary
- DCAT provides RDF classes and properties which enables a publisher to describe datasets and data services in a catalog
- The usage of such a standard model and vocabulary increases the *discoverability* and potential *reusability* of datasets and data services
- DCAT incorporates terms from pre-existing vocabularies where stable terms with appropriate meanings could be found, such as *foaf:homepage* and *dct:title*. For detailed list, please see https://www.w3.org/TR/vocab-dcat-2/

DCAT2 Model (2/4)



DCAT Distinctions (3/4)

- The **key distinction** in DCAT metadata standard is between the *abstract dataset* and *its different manifestations or distributions*.
- DCAT Dataset: “It is a collection of data, published or curated by a single source”. In concrete terms, it is *“a conceptual entity that represents the information published”*. Ex- DBpedia Ontology
- DCAT Distribution: It is *“a physical embodiment of the Dataset in a particular format”*. Ex- DBpedia Ontology OWL file in RDF/XML serialization

Teleology-DCAT-CKAN Mapping

Teleology Distinctions	*DCAT Distinctions*	*CKAN Distinctions*
TeleologyConceptualization	Dataset	Dataset
Teleology Serialization	Distribution	Resources

References:

1. https://www.w3.org/TR/vocab-dcat-2/#dcat-scope
2. https://docs.ckan.org/en/538-package-install-docs/publishing-datasets.html

DCAT Profiles (4/4)

- An *Application Profile (AP)* is a specification that re-uses terms from one or more base standards, adding more specificity by identifying *mandatory, recommended and optional* elements
- In the context of *iTelos* methodology, we follow a selected subset of the DCAT Application Profile for Data Portals in Europe - Version 2.0.1, or, DCAT-AP in short (which is based on the DCAT2 standard)
- We provide in the following slides the DCAT-AP metadata properties which we recommend for -
  1. *Datasets* - Teleology Conceptualization, AND, Dataset as a conceptual entity (collection of data)
  2. *Distributions* - Teleology File (ex- OWL RDF/XML), AND, Data Resources (dataset file, for ex, in CSV)

Contents

Mandatory Metadata

DCAT prescribes two mandatory metadata properties for datasets (by which we mean the teleology/data resource conceptualization)

- **description**: This property contains a free-text account of the Dataset. Ex - Schema.org is a shared vocabulary for structured data on the Internet.
- **title**: This property contains a name given to the Dataset . Ex - Schema.org vocabulary

NOTE (1): Both can be repeated for parallel language versions
NOTE (2): For detailed understanding of each metadata property for all categories, please consult : https://www.w3.org/TR/vocab-dcat-2/

Recommended Metadata

We prescribe three DCAT recommended metadata properties for datasets

- **dataset distribution**: This property links the Dataset to an available Distribution. Ex - lov_schema.ttl
- **keyword** / **tag**: This property contains a keyword or tag describing the Dataset. Ex - semantic annotation etc.
- **theme** / **category** - This property refers to a category of the Dataset. Ex - General and Upper Ontologies.

Optional Metadata (1/3)

We prescribe fourteen DCAT optional metadata properties for datasets

- **other identifier**: This property refers to a secondary identifier of the Dataset. Ex - https://w3id.org/
- **version notes**: This property contains a description of the differences between this version and a previous version of the Dataset.
- **landing page**: This property refers to a web page that provides access to the Dataset, and/or additional information.
- **creator**: This property refers to the entity primarily responsible for producing the dataset. Ex - KnowDive Research Group
- **has version**: This property refers to a related Dataset that is a version, edition, or adaptation of the described Dataset.

Optional Metadata (2/3)

We prescribe fourteen DCAT optional metadata properties for datasets

- **is version of**: This property refers to a related Dataset of which the described Dataset is a version, edition, or adaptation.
- **identifier**: This property contains the main identifier for the Dataset (in the context of the catalog)
- **release date**: This property contains the date of formal issuance (e.g., publication) of the Dataset.
- **update** / **modification date**: This property contains the most recent date on which the Dataset was changed or modified.

Optional Metadata (3/3)

We prescribe fifteen DCAT optional metadata properties for datasets

- **language**: This property refers to a language of the Dataset. Ex - en, it
- **provenance**: This property contains a statement about the lineage of a Dataset.
- **documentation**: This property refers to a page or document about this Dataset
- **was generated by**: his property refers to an activity that generated, or provides the business context for, the creation of the dataset.
- **version**: This property contains a version number or other version designation of the Dataset

Contents

Mandatory Metadata

DCAT prescribes one mandatory metadata properties for distribution (by which we mean the actual manifestation of the teleology/ data resource via, for example, an OWL RDF/XML file)

- **access URL**: This property contains a URL that gives access to a Distribution of the Dataset. Ex - http://liveschema.eu/dataset/lov_schema/resource/57247809-b0af-4448-ac8c-62db403d9aaa

NOTE: For detailed understanding of each metadata property for all categories, please consult : https://www.w3.org/TR/vocab-dcat-2/

Recommended Metadata

We prescribe three DCAT recommended metadata properties for distributions

- **description**: This property contains a free-text account of the Distribution. Ex- Serialized rdf format of the schema.org vocabulary
- **format**: This property refers to the file format of the Distribution. Ex - RDF
- **license**: This property refers to the licence under which the Distribution is made available. Ex - Creative Commons Attribution 4.0

Optional Metadata (1/2)

We prescribe nine DCAT optional metadata properties for distributions

- **status**: This property refers to the maturity of the Distribution. It MUST take one of the values Completed, Deprecated, Under Development, Withdrawn
- **access service**: This property refers to a data service that gives access to the distribution of the dataset
- **byte size**: This property contains the size of a Distribution in bytes.
- **download URL**: This property contains a URL that is a direct link to a downloadable file in a given format.
- **release date**: This property contains the date of formal issuance (e.g., publication) of the Distribution

Optional Metadata (2/2)

We prescribe nine DCAT optional metadata properties for distributions

- **language**: This property refers to a language used in the Distribution
- **update** / **modification date**: This property contains the most recent date on which the Distribution was changed or modified
- **title**: This property contains a name given to the Distribution
- **documentation**: This property refers to a page or document about this Distribution.

Contents

Liveschema Catalog Metadata

- LiveSchema is a high-quality catalog of reference teleologies.
- It aggregates schemas from several state-of-the-art repositories such as Linked Open Vocabulary, FINTO etc.
- Being powered by CKAN, Liveschema, by design, is fully compliant with the DCAT distinctions between Dataset and Distribution
- We now see some examples in Liveschema which makes the DCAT distinctions more clear (from the perspective of knowledge resources).

Open Data Trentino

- Open Data Trentino is a single point of access , a catalog of reusable data, which allows the search , access , preview and download of open data and some services of the Trentino system.
- Being powered by CKAN, Open Data Trentino, by design, is fully compliant with the DCAT distinctions between Dataset and Distribution
- We now see some examples in Open Data Trentino which makes the DCAT distinctions more clear (from the perspective of data resources).

Contents

SHAPEness Metadata Editor

- The SHAPEness Metadata Editor is a desktop application conceived to help users creating and updating metadata descriptions
- It provides a rich user interface which allows users to easy populate and validate metadata, structured as graphs, against a set of DCAT-AP properties (for Datasets and Distributions)
- Downloads for Windows/Mac/Linux- SHAPEness Metadata Editor
- We now show a demo of how the application can be used in the context of the *iTelos* methodology.

Contents

Summary

- We learnt about the importance of catalog and metadata in semantic data management in the context of *iTelos* methodology
- We learnt about the metadata properties relevant for the KDI DI project with respect to - (i) Data Resources and (ii) Knowledge Resources
- We saw examples of how DCAT is used in practice
- THANK YOU !!!

KDI : Knowledge and Data Integration

Danish Cheema, Mayukh Bagchi

Catalog and Metadata