

Functionele eisen DCAT-AP-DONL-2

Status: Concept

Datum: 4 oktober 2021

Inhoudsopgave

Inhoudsopgave

1. Inleiding

1.1 Vindbaarheid

1.2 Uitwisselbaarheid

1.3 Vervolgproces

2. Functionele eisen

2.1 Voldoen aan DCAT-AP-EU 2.0.1

2.2 Samenwerkende datacatalogi

Oplossingsrichting

2.3 Uitbreiding begrippen

Oplossingsrichting

2.4 Data-schema's

Oplossingsrichting

2.5 Kwaliteit

Oplossingsrichting

2.6 Gebruiksrechten

Oplossingsrichting

2.7 Registratie

Oplossingsrichting

Referenties

1. Inleiding

Deze notitie beschrijft de eisen van het DONL-team van KOOP/BZK aan enkele functionele uitbreidingen van het data.overheid.nl platform (afgekort DONL).

Deze uitbreidingen zijn er - op hoofdlijnen - op gericht om

1. de vindbaar van datasets op het portaal te verbeteren, en
2. de uitwisselingsstandaard DCAT te upgraden naar versie 2.

1.1 Vindbaarheid

Het DONL platform heeft als doel om hergebruik van gegevens (van overheden) te stimuleren. Hierbij is het noodzakelijk dat eindgebruikers de gewenste gegevens kunnen vinden en vervolgens kunnen beoordelen of deze voldoen aan hun eisen en wensen.

1.2 Uitwisselbaarheid

Het DONL platform bevat informatie over ruim 20.000 datasets. Deze informatie wordt op drie manieren verkregen:

1. Door handmatige invoer met behulp van web-formulieren.
2. Door de gegevens actief op te halen bij de bron.
3. Door push-berichten door de bronsystemen. Hierbij maken toeleveranciers gebruik van de API.

Daarnaast levert DONL deze beschrijvingen zelf ook weer door aan o.a. het Europese dataportaal. Hierbij maken afnemers gebruik van de hiervoor beschikbare API.

Uitwisselingsstandaard Data Catalog Vocabulary (DCAT) specificeert de objecten, de properties en de betekenis van de gegevens die uitgewisseld kunnen worden over catalogi, datasets, dataservices en distributies. In aanvulling op deze standaard ontstaan toepassingsprofielen (of Application Profiles). Deze bevatten **nadere afspraken over de invulling van DCAT**, bijvoorbeeld het Application Profile for data portals in Europe Version 2.0.1 [DCAT-AP-EU 2.0.1].

De huidige versie van DONL maakt nu gebruik van versie 1.1 van het toepassingsprofiel van DONL [DCAT-AP-DONL 1.1].

DONL wil dit toepassingsprofiel upgraden, zodat het

1. voldoet aan het Europese toepassingsprofiel van DCAT versie 2 [DCAT-AP-EU 2.0.1], en
2. geschikt wordt gemaakt voor uitwisseling van gegevens die de vindbaarheid van datasets verbeteren.

Hoofdstuk 2 beschrijft de eisen waaraan het nieuwe toepassingsprofiel van DONL moet voldoen.



Een DCAT toepassingsprofiel faciliteert de uitwisseling van de gegevens, dat wil zeggen de klassen en properties zoals beschreven in de DCAT specificatie. Het toepassingsprofiel stelt geen eisen aan de techniek van de uitwisseling. De enige eis is dat deze gegevens in de vorm van RDF kunnen worden geïmporteerd en geëxporteerd.

1.3 Vervolgproces

Zodra de functionele eisen in hoofdstuk 2 zijn uitgewerkt en vastgesteld, kunnen de hiervoor benodigde aanpassingen in het datamodel en de software van DONL in kaart worden gebracht, van prioriteiten worden voorzien en uiteindelijk worden gerealiseerd. Tevens zal het DONL-team het bijbehorende toepassingsprofiel verder uitwerken en publiceren, zodat de bronhouders hun data-leverantie kunnen afstemmen op de nieuwe eisen van DCAT-AP-DONL versie 2.

Om te komen tot dit nieuwe toepassingsprofiel, wil het DONL-team haar voorstellen graag delen en afstemmen met enkele andere overheidsorganisaties die hun datasets nu aan DONL aanleveren. Dat zijn:

Gemeente Amsterdam	Gemeente Eindhoven
Gemeente Groningen	Gemeente Utrecht
Provincie Zuid-Holland	CBS
RIVM	De Nederlandsche Bank
Kadaster	Vlaamse overheid
Logius	Civity B.V.
Dexes B.V.	Esri Nederland

2. Functionele eisen

DONL wil graag de functionaliteiten van het platform en de uitwisseling van metadata op de volgende punten verbeteren:

1. Voldoen aan het Europese applicatieprofiel voor DCAT 2
2. Het initiatief "samenwerkende datacatalogi", waarbij het de bedoeling is om in DCAT/DONL ruimte te maken voor "interne datasets". Dit zijn datasets die overheidsorganisaties zelf gebruiken (en mogelijk niet openbaar zijn)

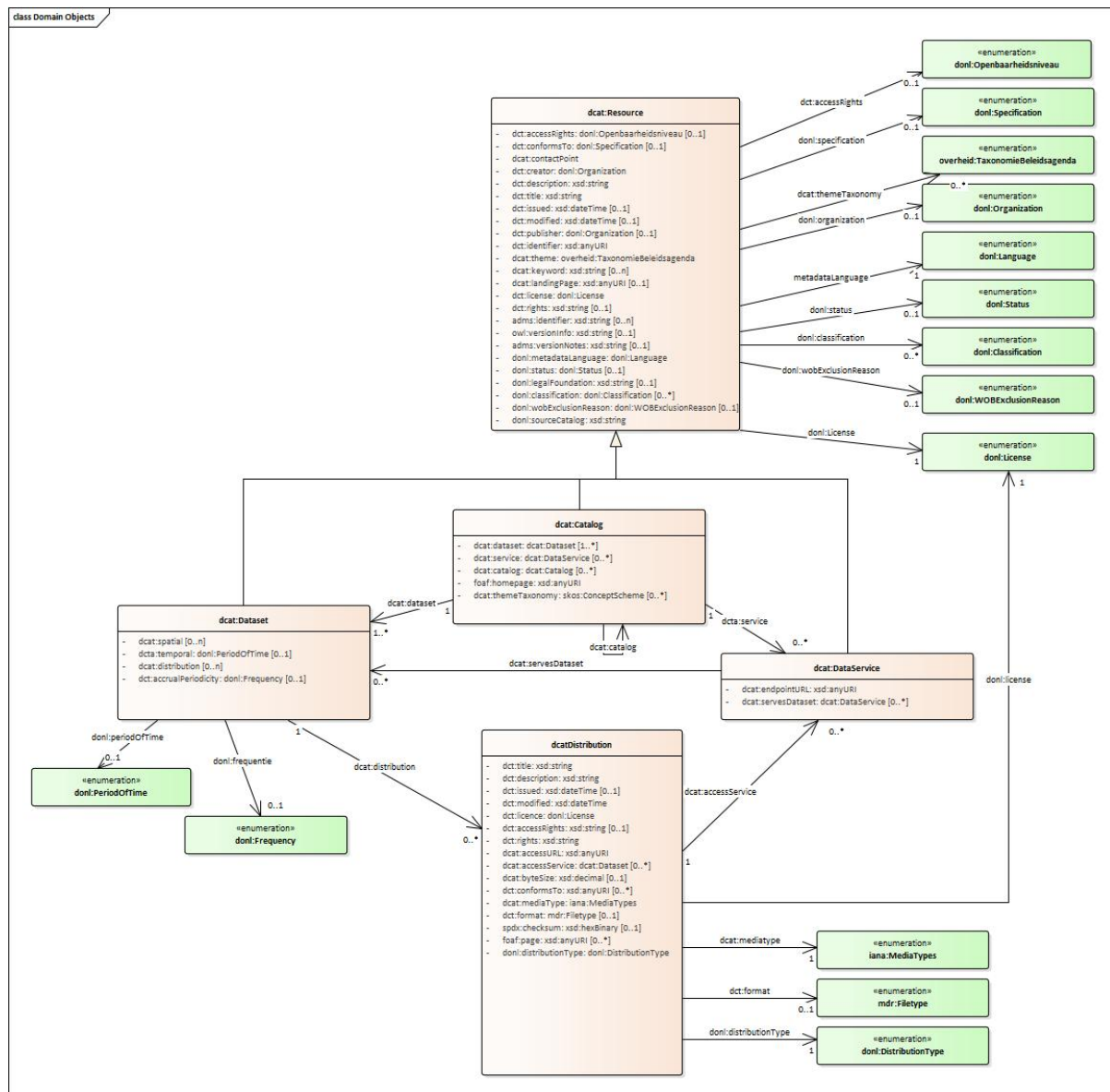
3. Uitbreiding voor koppelen van begrippen aan datasets
4. Uitbreiding voor beschrijven van data schema's van distributies
5. Uitbreiding voor beschrijven van kwaliteit van datasets
6. Uitbreiding voor het beschrijven van de gebruiksrechten van data met behulp van ODRL
7. Uitbreiding voor het beschrijven van (basis-)registraties

De volgende paragrafen beschrijven deze aanpassingen inhoudelijk en inventariseren de consequenties die deze aanpassingen hebben voor de gegevensuitwisseling volgens DCAT. Deze consequenties vormen de basis voor het nieuwe toepassingsprofiel van DCAT2 voor DONL.

2.1 Voldoen aan DCAT-AP-EU 2.0.1

We hebben - in Excel - een inventarisatie gemaakt van alle klassen en eigenschappen van DCAT2. Deze lijst hebben we aangevuld met de eigenschappen die data.overheid.nl gebruikt in hun eigen toepassingsprofiel DCAT-AP-DONL 1.1. Per eigenschap hebben we gekeken of en hoe deze wordt toegepast in DCAP-AP-EU 2.0.1.

Op basis van deze inventarisatie hebben we een selectie gemaakt van eigenschappen die we willen overnemen in DCAT-AP-DONL 2. Het onderstaande UML Class diagram toon te resultaat.



Het gebruikte spreadsheet is als bijlage bijgesloten.

2.2 Samenwerkende datacatalogi

Zowel de Data Governance Act (DGA, Eu commissie 20/11/2020) als de Interbestuurlijke datastrategie (IBDS, dec 2020) duiden op het belang van goed vindbare en herbruikbare data. Het betreft hier zowel de interne data die overheden zelf produceren, de data die ze gebruiken van anderen als ook de data die ze als open data beschikbaar stellen voor publiek hergebruik. Het initiatief "Samenwerkende datacatalogi" betreft de ontwikkeling van een federatief stelsel van datacatalogi die informatie over beschikbare data voor elke doelgroep vindbaar maakt, zowel voor niet publieke data, als voor publieke, open data.

Voor de DCAT/DONL toepassingsprofiel betekent dit, dat de metadata van een catalogus een kenmerk moet bevatten om een catalogus aan te merken als interne of beschermde datacatalogus.

Oplossingsrichting

Er zijn verschillende mogelijkheden manieren waarop "interne" catalogi kunnen worden aangeduid:

1. Door middel van een nieuwe data-property: donl:interneCatalogus, met waarde true/false, of
2. Met een of meer specifieke waarden in bestaande properties voor License and rights statements in DCAT, bijvoorbeeld dct:accessRights = 'alleen centrale overheden'.

In beide gevallen is de software van DONL in staat om te detecteren dat een catalogus een "interne catalogus" is, zodat het de vindbaarheid en of raadpleegbaarheid kan aanpassen aan de specifieke kenmerken van een ingelogde eindgebruiker.

2.3 Uitbreiding begrippen

Data.overheid.nl biedt een zoekvoorziening waarmee eindgebruikers datasets en andere informatie kunnen vinden met behulp van een of meer zoektermen. Hierbij moeten zij nu termen kiezen die mogelijk voorkomen in o.a. de titel en beschrijvingen van datasets. In aanvulling hierop wil DONL het mogelijk maken dat eindgebruikers datasets kunnen vinden op basis van termen die voorkomen in bestaande controlled vocabularies of begrippenkaders, zoals die van CBS, Stelselcatalogus, DUO, Fiscale taxonomie etc. Het is de verwachting dat de vindbaarheid van datasets en en andere informatie hiermee verbetert.

Dit betekent dat DONL de mogelijkheid moet gaan bieden om bij een dataset of data service een of meer begrippen te registreren uit een of meer bestaande begrippenkaders [BEGRIPPEN].

Oplossingsrichting

De DCAT standaard voorziet met property dcat:theme (voor thema of categorie) in de functionaliteit om begrippen te koppelen aan datasets en data services.

RDF Property:	<u>dcat:theme</u>
Definition:	A main category of the resource. A resource can have multiple themes.
Sub-property of:	<u>dct:subject</u>
Range:	<u>skos:Concept</u>
Usage note:	The set of <u>skos:Concepts</u> used to categorize the resources are organized in a <u>skos:ConceptScheme</u> describing all the categories and their relations in the catalog.
See also:	<u>§ 6.3.2 Property: themes</u>

DONL gebruikt deze property nu om de datasets te classificeren volgens de taxonomie beleidsagenda (overheid:TaxonomieBeleidsagenda) [BELEIDSAGENDA]. Dit moet flexibeler worden, zodat naast thema's uit de taxonomie beleidsagenda ook begrippen uit andere begrippenkaders kunnen worden toegevoegd.

Deze flexibiliteit kan op twee manieren worden gerealiseerd:

1. Aanpassen van de manier waarop DONL nu gebruikmaakt van dcat:theme.
2. Uitbreiden van het DCAT model voor DONL met nieuwe properties om begrippen te registreren.

Ad 1.

Hiervoor moet het mogelijk worden gemaakt, dat dcat:theme waarden kan bevatten die afkomstig zijn uit meerdere, verschillende begrippenkaders. De DCAT standaard biedt deze mogelijkheid met property dcat:themeTaxonomy in een catalogus. Deze property bevat een opsomming van de begrippenkaders (skos:ConceptScheme's of waardelijsten) die de begrippen bevatten voor de Datasets en DataServices die *in die Catalogus* zijn opgenomen.

RDF Property:	<u>dcat:themeTaxonomy</u>
Definition:	A knowledge organization system (KOS) used to classify catalog's datasets and services.
Domain:	<u>dcat:Catalog</u>
Range:	<u>rdfs:Resource</u>
Usage note:	It is recommended that the taxonomy is organized in a <u>skos:ConceptScheme</u> , <u>skos:Collection</u> , <u>owl:Ontology</u> or similar, which allows each member to be denoted by an IRI and published as Linked Data.

Op dit moment heeft DONL het begrippenkader voor de taxonomie beleidsagenda en andere controlled vocabularies opgenomen in waardelijsten die zij publiceert op [DCAT-AP-DONL - Waardelijsten](#). Deze lijsten worden door de software, redactie en toeleveranciers gebruikt bij de registratie van deze waarden. Dit betekent dat als DONL meerdere bestaande begrippenkaders gaat ondersteunen, deze begrippenkaders - die meestal zijn vastgelegd in een SKOS thesaurus - moeten worden omgezet naar JSON waardelijsten. Dat is technische vrij eenvoudig te realiseren.

Ad 2.

Het is ook mogelijk om een nieuwe property, bijvoorbeeld donl:begrip, toe te voegen aan datasets en data services. Deze property bevat de URI/Identificer van een gekoppeld begrip.

Deze begrippen kunnen voorkomen in verschillende taxonomieën of waardelijsten. Net als bij dcat:theme is het in deze oplossingsrichting noodzakelijk om te kunnen aangeven welke begrippenkaders worden ondersteund. Dat zou opgelost kunnen worden door de toegestane begrippenkaders te specificeren in het toepassingsprofiel.

Er kunnen eenvoudig meerdere begrippen worden gekoppeld aan een dataset of dataservice door meerdere voorkomens van donl:begrip toe te staan.

Het is de bedoeling dat bronhouders de begrippen gaan aanleveren per dataset of data service.. Dit impliceert dat begrippen onderdeel moeten worden van het nieuwe toepassingsprofiel van DCAT2 voor DONL.



Het DONL-team geeft de voorkeur aan optie 2. Op deze manier kan worden gewaarborgd dat elke dataset/dataservice verplicht tenminste een thema uit de taxonomie beleidsagenda behoudt.

2.4 Data-schema's

De distributies van datasets verwijzen naar de concrete gegevensbestanden die de afnemers van DONL uiteindelijk kunnen hergebruiken. Een voorbeeld van een distributie is een spreadsheet die bestaat uit meerdere kolommen.

Aan de hand van een data-schema kunnen bronhouders beschrijven *welke gegevens* voorkomen in de distributie, hoe ze heten en aan welke definitie ze voldoen. Dit stelt eindgebruikers van DONL in staat om de distributies beter te interpreteren. Daarnaast kunnen eindgebruikers ook zoeken op woorden die voorkomen in deze data-schema's, zodat zij datasets - die bepaalde gegevens bevatten - makkelijker kunnen vinden. Een voorbeeld hiervan is beschreven in Data on the Web Best Practices [DWBP], [Best Practice 3: Provide structural metadata](#).

Er worden verschillende termen gebruikt voor "data-schema", waaronder structuurbeschrijving, structural metadata, informatiemodel, gegevensmodel of datamodel.

Een data-schema beschrijft de kolommen in een dataset met behulp van enkele kenmerken, waaronder

- Naam van de kolom
- Beschrijving
- Range (wat de mogelijke waarden zijn die in de betreffende kolom mogen voorkomen), bijvoorbeeld datatype
- Cardinaliteit (of deze altijd mis gevuld, en hoe vaak deze kan voorkomen), bijvoorbeeld [1], of [0..*].

Het is de bedoeling dat bronhouders de data-schema's gaan aanleveren per distributie. Dit impliceert dat data-schema's onderdeel moeten worden van het nieuwe toepassingsprofiel van DCAT2 voor DONL.

Het CBS maakt gebruik van de standaard Statistical Data and Metadata eXchange [SDMX]. Dit is een ISO 17369:2013 standaard voor o.a. het uitwisselen en beschrijven van datasets en de data zelf. Het gaat echter veel verder dan DCAT en

is gericht op statistische data. Deze standaard is naar verwachting te complex voor het onderhavige vraagstuk.

Oplossingsrichting

Er zijn diverse manieren om data-schema's op te nemen bij een distributie:

1. Door een nieuwe klasse of tabel toe te voegen aan het DCAT model waarin het data-schema kan worden geregistreerd (donl:DataSchema). Vervolgens kan met behulp van een nieuwe property donl:dataSchema in dcat:Distribution een relatie worden gelegd naar deze structuurbeschrijving.
2. Door een data-property (donl:dataSchema) toe te voegen aan dcat:Distribution waarin een XML of JSON structuur kan worden opgenomen die de datastructuur beschrijft.
3. Door een data-property (donl:accessURLDataSchema) toe te voegen aan dcat:Distribution die middels een URL verwijst naar een file die de structuur beschrijft.

Ad 2.

Deze oplossing kan alleen werken als ook de XML of JSON structuur, die het data-schema beschrijft, wordt gestandaardiseerd.

Ad 3.

Ook dit alternatief vereist dat de structurering van de file voldoet aan een afgesproken standaard, bijvoorbeeld een JSON of XML schema.



Het DONL-team geeft de voorkeur aan optie 3, waarbij leveranciers de distributies van hun datasets beschrijven volgens een afgesproken JSON Schema en vastleggen in een bestand, zodat deze via een URL kan worden opgevraagd.

Dit betekent in praktijk dat een dataleverancier bij elk databestand ook een data-schema-bestand beschikbaar stelt (met een vaste URL).

2.5 Kwaliteit

DONL wil haar eindgebruikers inzicht bieden in de kwaliteit van datasets, zodat zij beter in staat zijn om de voor hen gewenste dataset te selecteren. Hierbij kan

worden gedacht aan een filter-optie die alleen de datasets toont die bepaalde kwaliteitskenmerken heeft.

Het Europese toepassingsprofiel van DCAT-2 heeft geen uitwerking voor dit punt. De DCAT-2 standaard geeft wel de volgende opmerking:

The Data Quality Vocabulary (DQV), offers common modelling patterns for different aspects of Data Quality. It can relate DCAT datasets and distributions with different types of quality information including:

- dqv:QualityAnnotation, which represents feedback and quality certificates given about the dataset or its distribution.
- dqv:QualityPolicy, which represents a policy or agreement that is chiefly governed by data quality concerns.
- dqv:QualityMeasurement, which represents a metric value providing quantitative or qualitative information about the dataset or distribution.

Each type of quality information can pertain to one or more quality dimensions, namely, quality characteristics relevant to the consumer. The practice to see the quality as a multi-dimensional space is consolidated in the field of quality management to split the quality management into addressable chunks. DQV does not define a normative list of quality dimensions. It offers the quality dimensions proposed in ISO/IEC 25012 and [ZaveriEtAl] as two possible starting points. It also provides an RDF representation for the quality dimensions and categories defined in the latter. **Ultimately, implementers will need to choose themselves the collection of quality dimensions that best fits their needs.**

W3C heeft in 2016 een notitie gepubliceerd waarin ze een framework beschrijft om de kwaliteit van datasets te beschrijven [VOCAB-DQV]. Zie ook <https://www.w3.org/TR/dwbp/#quality>.

Oplossingsrichting

Wat willen we precies met kwaliteitskenmerken?



Het DONL-team geeft er de voorkeur aan om de kwaliteitskenmerken nog niet op te nemen in het applicatieprofiel.




DONL wil haar eindgebruikers wel graag inzicht bieden in de kwaliteit van datasets, bijvoorbeeld door de mogelijkheid te bieden dat eindgebruikers feedback kunnen geven over de kwaliteit van datasets.

2.6 Gebruiksrechten

DCAT heeft verschillende kenmerken om gebruiksrechten te beschrijven. Hieronder volgt informatie die is overgenomen uit de DCAT 2 standaard.

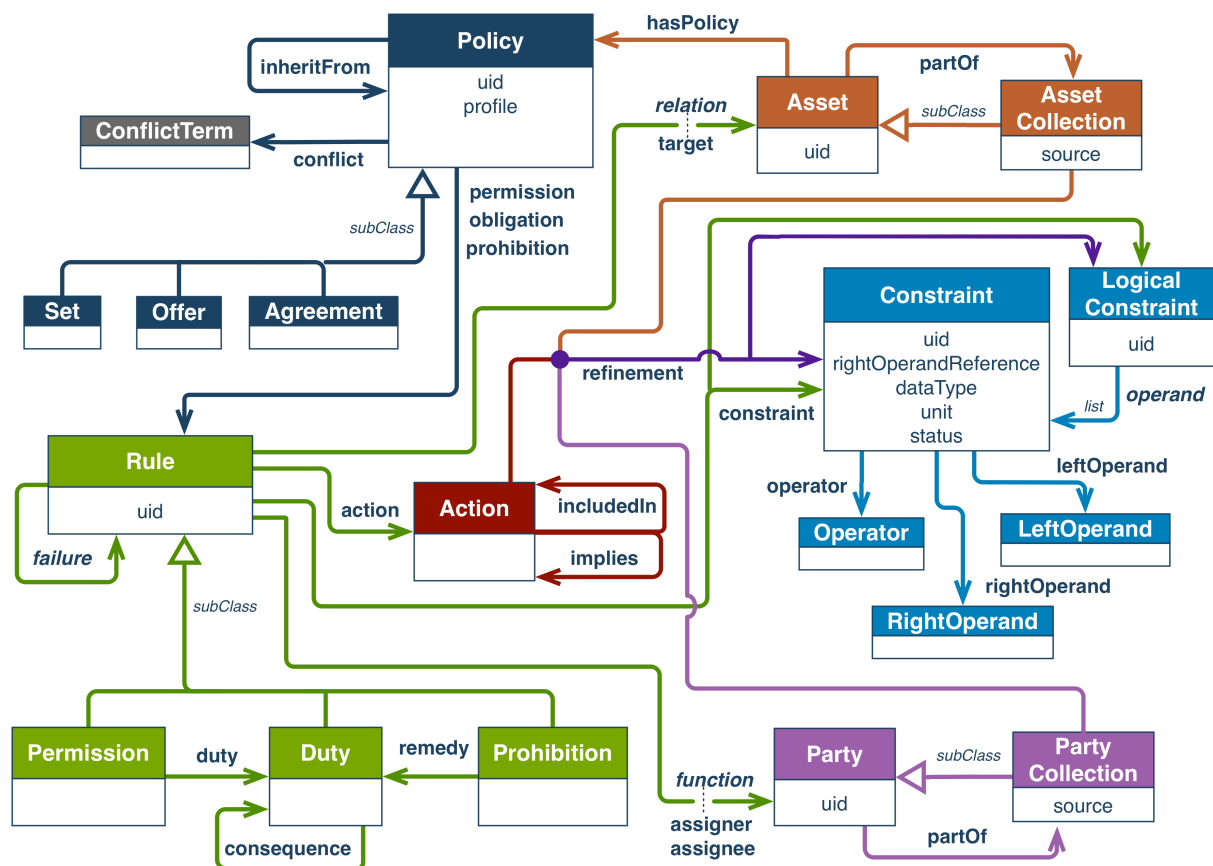
DCAT 2014 handling of license and rights do not appear to satisfy all requirements. The recently completed W3C ODRL model and vocabulary provide a rich language for describing many kinds of rights and obligations.

License and rights statements in DCAT-2

 Property	 Definition	 Use
<u>licence</u>	A legal document under which the resource is made available.	Use dct:license to refer to licenses. For interoperability, it is recommended to use canonical URIs of well-known licenses such as those defined by Creative Commons.
<u>access rights</u>	Information about who can access the resource or an indication of its security status.	Use dct:accessRights to express statements concerning only access rights (e.g., whether data can be accessed by anyone or just by authorized parties); Access rights can also be expressed as code lists / taxonomies.
<u>rights</u>	A statement that concerns all rights not addressed with dct:licence or dct:accessRights.	Use dct:rights for all the other types of rights statements - those which are not covered by dct:license and dct:accessRights, such as copyright statements. A more sophisticated approach to express rights, based on and extending [DCTERMS], is provided by the Open Data Rights Statement Vocabulary (ODRS), which defines properties for specifying, among others, copyright statements and copyright notices.

Aa Property	Definition	Use
<u>has policy</u>	An ODRL conformant policy expressing the rights associated with the resource.	In the particular case when rights are expressed via ODRL policies, it is recommended to use the odrl:hasPolicy property as the link from the description of the cataloged resource or distribution to the ODRL policy, in addition to the corresponding [DCTERMS] property that matches the same ODRL policy type. The Open Digital Rights Language (ODRL) is a policy expression language that provides a flexible and interoperable information model, vocabulary, and encoding mechanisms for representing statements about usage (i.e. permissions, prohibitions, and obligations) of content and services.

The ODRL Information Model defines the underlying **semantic model for permission, prohibition, and obligation statements describing content usage**. The information model covers the core concepts, entities and relationships that provide the foundational model for content usage statements. These **machine-readable policies** may be linked directly with the content they are associated to with the aim to allow consumers to easily retrieve this information.



ODRL Information Model

Het bereik van dcat:hasPolicy is odrl:Policy.

Oplossingsrichting

1. ODRL is behoorlijk complex en vereist tooling om Rules samen te stellen. Hoe willen we deze Rules gaan maken binnen DONL?

2.7 Registratie

Deze uitbreiding biedt eindgebruikers inzicht in de namen van registraties die de datasets hebben voortgebracht. Het biedt vervolgens ook de mogelijkheid om te zoeken op registratie, zodat eindgebruikers datasets (of -services) kunnen vinden die afkomstig zijn uit bijvoorbeeld de Basisregistratie Gebouwen (BAG).

Als eindgebruikers de bron van datasets of -services kennen, geeft het hen mogelijk een indicatie van de kwaliteit en/of betrouwbaarheid van de gegevens in de dataset.

Naast de naam van de registratie, kan ook een relatie worden gelegd naar websites die nader informatie verschaffen over de desbetreffende registratie, zoals de Logius website [Platform Stelsel van Basisregistraties](#).

Oplossingsrichting

Nader uit te werken.

Referenties

[BEGRIPPEN]

We willen ons hier beperken tot bestaande begrippenkaders, omdat de ontwikkeling van begrippenkaders buiten de scope van DONL valt. Bovendien zullen eindgebruikers vaak bekend zijn met de bestaande begrippenkaders uit hun kennisdomein. Zie bijvoorbeeld de websites <https://www.begrippenxl.nl/nl/> of <https://onderwijsbegrippen.nl/nl/> voor overzichten van deze begrippenkaders.

[BELEIDSAGENDA]

De begrippen in de taxonomie beleidsagenda worden vervangen door de nieuwe Toplijst, zoals gedefinieerd in de TOOI ontologie van KOOP. Het is vooralsnog niet bekend wanneer TOOI operationeel wordt. Tot die tijd maakt DONL gebruik van de taxonomie beleidsagenda.

[Conformance to DCAT-AP]

What does conformance to DCAT-AP mean for a data portal? zie <https://github.com/SEMICeu/DCAT-AP/issues/198>

[DCAT-AP-DONL 1.1]

[DCAT application profile for data.overheid.nl](#), [DCAT-AP-DONL 1.1](#)

[DCAT-AP-EU 2.0.1]

[Application Profile for data portals in Europe Version 2.0.1](#)

[DWBP]

Data on the Web Best Practices, zie <https://www.w3.org/TR/dwbp/>.

[SDMX]

Statistical Data and Metadata eXchange, zie bijvoorbeeld https://sdmx.org/?page_id=2555/.

[VOCAB-DQV]

Data Quality Vocabulary, zie <https://www.w3.org/TR/vocab-dqv/>