

Informatiemodel

E-depot

2024



Inhoud

Inhoud	2
Bijbehorend.....	2
Voorwoord	2
Versiebeheer	2
Onderliggende linked data modellen	3
MDTO, primaire ontologie	4
Belangrijkste aspecten MDTO.....	4
UML-Weergave MDTO.....	5
Afwijkingen van MDTO	6
Bijlage 1. Indeling systemen.....	7
Opbouw triple store.....	7
Opbouw object storage	7
Schematische weergave opbouw triple store	8
Schematische weergave opbouw triples	9
Schematische weergave opbouw object storage	10

Bijbehorend

- E-depot-MIM-Excel (Excel-weergave)
- E-depot-MIM-Drawio (Drawio-origineel van UML & opbouwdiagrammen)

Voorwoord

versie Februari 2024

Dit document beschrijft het informatiemodel van het e-depot van het Regionaal Archief Zuid-Utrecht. Het beschrijft primair de onderliggende linked data modellen. Daarnaast is de inrichting van de triple store (database voor metadata) en de relatie tussen objecten in de object storage (dataopslag) en de triple store opgenomen als uitwerking van het informatiemodel. Het doel van het informatiemodel is het documenteren van gemaakte inrichtingskeuzes en borgen van kennis over deze inrichtingskeuzes. Dit document is direct gekoppeld aan ARIA, de informatiearchitectuur van het RAZU dat in februari 2023 is vastgesteld.

Versiebeheer

#	Datum	Beschrijving
1.0	26-02-2024	

Onderliggende linked data modellen

Conform ARIA wordt het e-depot volledig op basis van linked data ingericht. Dit betekent dat er naast een triple store (database voor linked data) geen andere (relationele) database wordt bijgehouden voor informatieobjecten.¹ Linked data effectief inzetten vereist het opstellen en volgen van één of meerdere ontologieën, beschrijvingsstandaarden voor linked data. Bestaande ontologieën zijn nooit compleet. Om toch te kunnen voldoen aan de wensen van de verschillende (her)gebruikers van de data heeft het RAZU een keuze: op basis van het beste voorbeeld één eigen ontologie creëren waarbinnen alle data wordt vastgelegd, óf ontologieën met elkaar ‘verbinden’ zodat ze elkaar aanvullen. Het RAZU kiest voor de tweede optie, mede vanwege de flexibiliteit van een verbonden model waarbij ontologieën individueel up-to-date gehouden kunnen worden. Ook speelt mee dat een eigen RAZU-ontologie opzetten meer beheertaken en verantwoordelijkheden met zich meebrengt dan bestaande ontologieën te gebruiken.

Het RAZU kiest ervoor om MDTO als primaire ontologie te positioneren. Dit is mede op basis van het feit dat dit de de facto standaard is bij de aanleverende archiefvormers. Daarnaast is MDTO relatief klein in omvang en biedt het ‘by design’ ruimte voor uitbreidingen met andere ontologieën. Een informatieobject in ons e-depot wordt daarmee beschreven in MDTO en uitgebreid met domein-specifieke metadata. Deze specifieke metadata zijn bij voorkeur beschreven in een ontologie. In de tabel hiernaast wordt een overzicht gegeven van ontologieën die het RAZU op termijn verwacht te ontvangen. Het overzicht is niet limitatief.

De bron is de organisatie die de data conform dit model gaat aanleveren, de beheerder beheert het model.

Ontologie	Bron inhoud	Doel	Beheerder
MDTO (Metagegevens voor Duurzame Toegankelijke Overheidsinformatie)	Archiefvormer	Beschrijven, uitwisselen informatieobjecten	Nationaal Archief (NA, NL)
PREMIS (PREservation Metadata)	RAZU	Preserveren informatieobjecten	Library of Congress (LoC, VS)
RICO (Records in Context)	RAZU	Beschrijven informatieobjecten	International Council on Archives (ICA, Int.)
PICO (Persons in Context)	RAZU / Archiefvormer	Beschrijven personen	CBG Centrum voor familiegeschiedenis (CBG, NL)
ORI (Open Raadsinformatie)	Archiefvormer	Inhoudelijke data	Open State Foundation (OS, NL)
TPOD (Toepassingsprofiel Omgevingsdocumenten)	Archiefvormer	Beschrijven informatieobjecten gecreëerd uitvoering omgevingswet	GEONOVUM (NL)

¹ Sommige componenten, zoals de access & identity managers, hebben uiteraard een ingebouwde database voor hun eigen functioneren, maar deze data vallen buiten de scope van het centrale informatiemodel.

De mogelijkheid tot uitbreiden (via MDTO:AanvullendeMetagegevens) staat toe dat in de toekomst extra ontologieën vanuit domeinen van onze archiefvormers worden toegevoegd op het moment dat de informatieobjecten uit deze domeinen worden overgebracht.

MDTO, primaire ontologie

Het MDTO is de primaire ontologie van het informatiemodel. We beschrijven dit op twee manieren:

- Een volledige weergave van MDTO, integraal onderdeel van dit informatiemodel in de Excel-weergave (E-depot-MIM-Excel). Hierbij is ook de locatie van een datapunt binnen de triple store opgenomen. Dit is tevens beschreven en schematisch uitgewerkt in bijlage 1 van dit document.
- Een UML-diagram, weergegeven op pagina 5 en als bronbestand integraal onderdeel van dit informatiemodel (E-depot-MIM-Drawio).

Belangrijkste aspecten MDTO

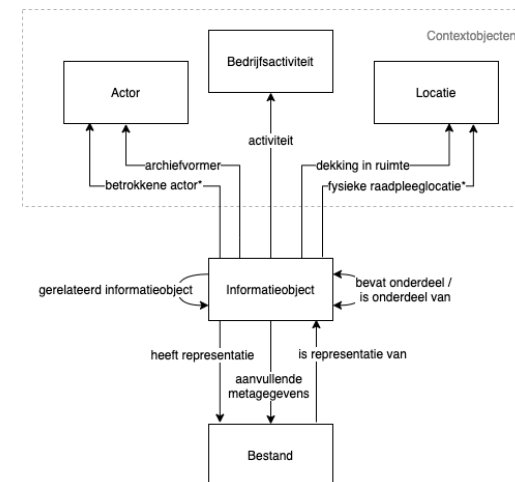
De beheerder van het MDTO, het Nationaal Archief, heeft op haar website een uitgebreide beschrijving van het model. Echter, voor het begrijpen van het RAZU informatiemodel is begrip van twee eigenschappen belangrijk. Ten eerste: MDTO beschrijft generieke informatieobjecten. Dit kunnen zaken zijn, documenten, series of hele archieven. Informatieobjecten kunnen 'nesten'. Dat betekent dat een archief een serie kan bevatten. De serie kan zaken bevatten. De zaken kunnen vervolgens documenten bevatten. Al deze onderdelen zijn in het MDTO informatieobjecten. De beschrijving van een computerbestand, bijvoorbeeld een JPEG-afbeelding, gebeurt op twee plekken: als informatieobject ('intellectueel') en als bestand ('technisch'). Dit onderscheid is met name gericht op het

fenomeen 'representaties': een document (intellectueel) kan meerdere versies hebben, bijvoorbeeld zowel een MSG-bestand als een PDF-bestand.

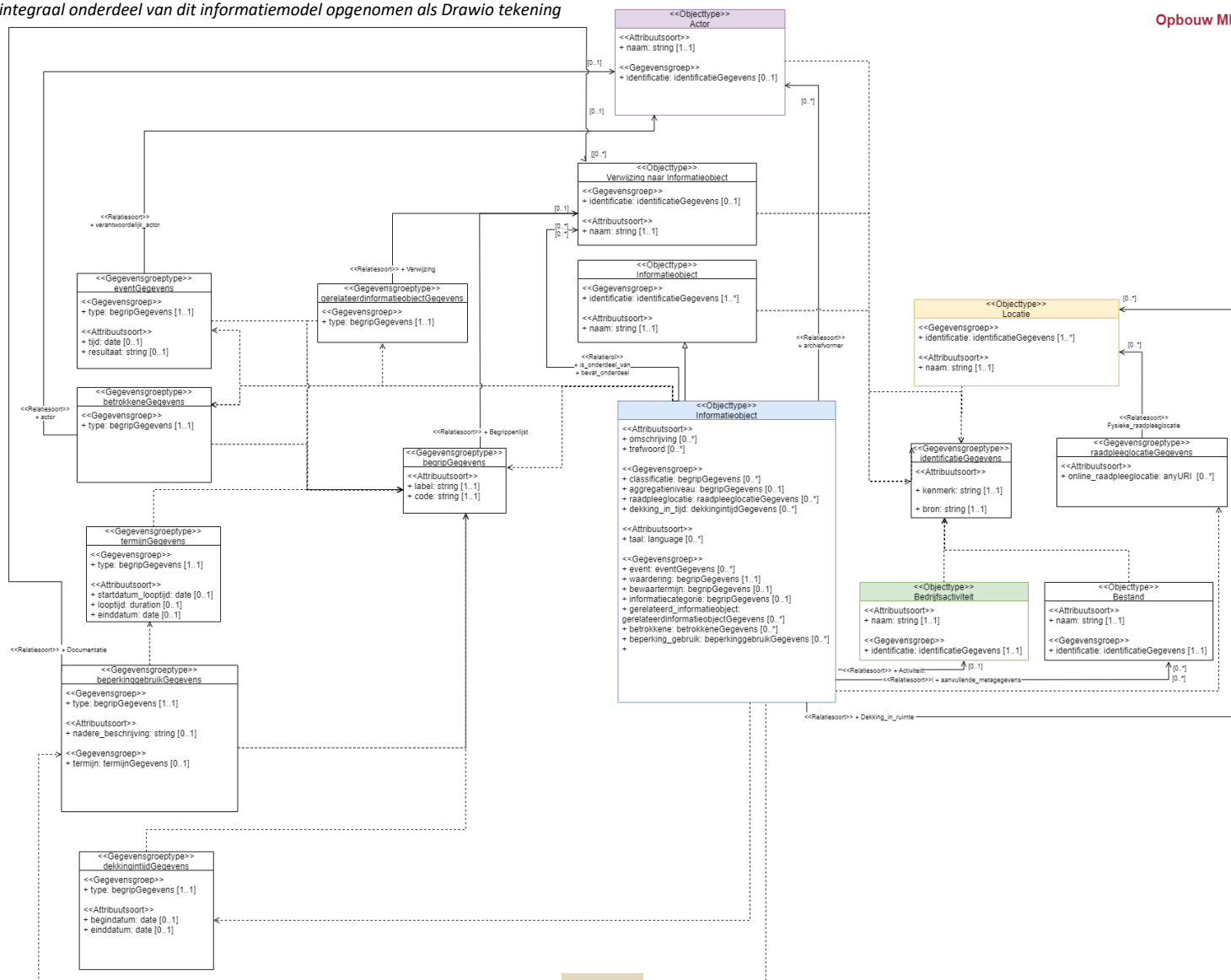
Ten tweede gaat MDTO uit van het hergebruiken van structuren. Dit is op meerdere manieren zichtbaar maar het duidelijkst in de verwijzingen naar begrippenlijsten. Begrippenlijsten (ook wel gecontroleerde vocabulaires genoemd) staan voor een vaste set aan waarden waaruit gekozen kan worden.

Deze kunnen door het RAZU, de archiefvormer of door externen beheerd worden. Denk bijvoorbeeld aan de Basisadministratie Adressen en Gebouwen, BAG, waarin alle adressen in Nederland zijn beschreven. Elke begrippenlijst in MDTO heeft een identificatie en een bron (van wie is de lijst). Vervolgens legt MDTO ook de

gekozen waarde vast, optioneel vergezeld van een nummer. De standaardisatie van dit soort 'elementgroepen' binnen MDTO geeft enkele voordelen, zoals het makkelijker begrijpen van de opbouw, het standaardiseren van verwijzingen en het gemakkelijker opvragen van gegevens door systemen. Naast begrippenlijsten gebruikt het MDTO ook verwijzingen voor zogenaamde contextobjecten, entiteiten zoals locaties, personen en activiteiten die elders beschreven staan. Het MDTO neemt bijvoorbeeld geen volledig adres op, maar verwacht een verwijzing, bijvoorbeeld naar de BAG.



Opbouw MDTO



De hoofdelementen van MDTO zijn:

Informatieobject	Bestand
Identificatie	Identificatie
Naam	Naam
Omschrijving	URL bestand
Trefwoord	Omvang
Classificatie	Bestandsformaat
Aggregatieniveau	Checksum
Raadpleeglocatie	Is representatie van (verwijzing naar Informatieobject)
Dekking in tijd	
Dekking in ruimte (verwijzing naar Locatie)	
Taal	
Event	
Waardering	
Bewaartermijn	
Informatiecategorie	
Is onderdeel van / bevat onderdeel (verwijzing naar Informatieobject)	
Heeft representatie (verwijzing naar Bestand)	
Gerelateerd informatieobject (verwijzing naar Informatieobject)	
Aanvullende metagegevens	
Archiefvormer (verwijzing naar Actor)	
Betrokkene (verwijzing naar Actor)	
Activiteit (verwijzing naar Bedrijfsactiviteit)	
Beperking gebruik	

Niet elk element is verplicht. Welke elementen geregistreerd zijn zal dus variëren, afhankelijk van de archiefvormer en de aard van de dataset.

Afwijkingen van MDTO

Het RAZU volgt de standaard van MDTO zo secuur mogelijk. Toch wijken we op twee aspecten af:

- De type-indicatie van een gerelateerd informatieobject is volgens MDTO verplicht. Het RAZU beschouwt dit element voorlopig als 'verplicht indien bekend'. Dit doen we op basis van huidige implementaties van relaties in systemen van onze archiefvormers, waar dit soort informatie niet altijd wordt vastgelegd.
- Repeterende elementen binnen een dataset nemen we op in een losse dataset genaamd 'gedeelde metadata', mits dit tenminste 6 waarden (triples) bevat. Onze huidige triple store baseert een deel van de kosten op het aantal triples dat geregistreerd staat. Het loont dus om metadata die voor elk informatieobject hetzelfde is slechts eenmalig vast te leggen. Denk bijvoorbeeld aan de archiefvormer, de informatiecategorie, de waardering en de bewaartermijn. De voorwaarde dat er tenminste 6 waarden gedeeld moeten zijn tussen alle informatieobjecten komt voort uit het feit dat het aanleggen van een relatie tussen een informatieobject en de gedeelde waarden via 'additionele metadata' 5 triples kost. Pas vanaf 6 waarden wordt er een besparing gerealiseerd, die wel al heel snel kan oplopen. Zo kan een dataset al gauw 40.000 informatieobjecten bevatten, wat per bespaarde waarde dus 40.000 triples scheelt. In het model zijn deze waarden al onder gedeeld geschaard, maar per aanlevering zal worden bepaald of de onderverdeling toegepast kan worden.

Via SPARQL-queries kunnen deze afwijkingen voor de afnemer 'onzichtbaar' worden gemaakt, indien dit noodzakelijk blijkt.

Bijlage 1. Indeling systemen

Opbouw triple store

Elke archiefvormer staat in de triple store als organisatie geregistreerd. Datasets van de archiefvormer worden aan de organisatie gekoppeld. Binnen deze datasets is onderscheid in 3 'virtuele bakjes' (graphs): de informatieobjecten, de gedeelde metadata en het MDTO-model. Deze laatste is enkel als referentie voor systemen die de data opvragen en heeft geen inhoudelijke meerwaarde. Het RAZU, als archiefvormer en beheerder van gedeelde collecties staat ook als organisatie geregistreerd.

Gedeelde (meta)data (Begrippenlijsten, actoren en locaties), voor zover deze niet door een externe verwijzing beschreven worden, staan centraal opgeslagen onder de virtuele organisatie 'gedeeld'.

Beheer (meta)data, zoals logging en preserveringsdata, staan centraal opgeslagen onder de virtuele organisatie 'beheer'.

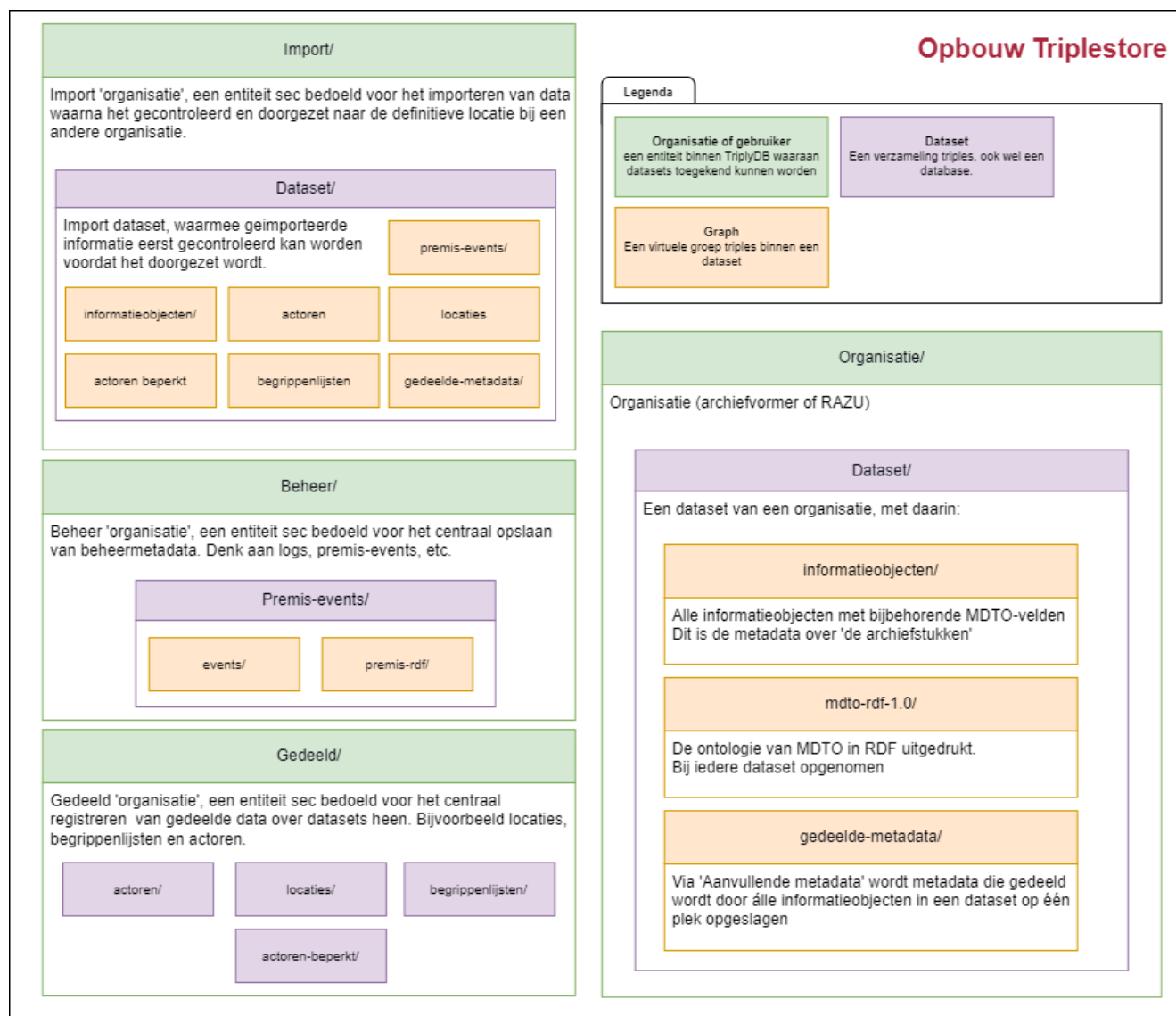
Daarnaast is er een 'import' organisatie, die enkel gebruikt wordt voor het importeren en vervolgens controleren van nieuwe datasets. Vanuit de import worden de onderdelen van de dataset verdeeld over de verschillende andere organisaties..

Op de volgende pagina's is deze opbouw schematisch weergegeven, alsook de vertaling naar de individuele triple. Een triple is een 'statement', waarmee metadata wordt gekoppeld aan, in ons geval, een informatieobject. Het schema geeft hier concrete voorbeelden van.

Opbouw object storage

Op het moment van schrijven (jan. 2024) is de object storage nog niet geheel opgeleverd. De opbouw is daarmee nog niet definitief. We zijn voornemens de object storage te spiegelen aan de indeling van organisaties in de Triple Store. Omdat de bestanden die in de object storage staan beperkt in openbaarheid kunnen zijn is ervoor gekozen om deze beperking ook door te voeren in de 'buckets'. Hieronder staat een schematische weergave van de objectstorage. Hierbij is ook het content delivery network (CDN) ingetekend, dat voorziet in de persistente link naar objecten, gekoppeld aan de domeinnaam van het RAZU (opslag.razu.nl).

Schematische weergave opbouw triple store



Schematische weergave opbouw triples

Opbouw Triples

Triples bestaan uit drie delen: **Subject**, **Predicaat**, **Object**.

Voorbeelden:

Bunnik ligt in Provincie Utrecht

Dossier overgebracht op 12 december 1963

In het geval van het eerste voorbeeld is het object (Provincie Utrecht) ook een subject. Als dit niet het geval is dan spreken we van een property, zoals de datum in het tweede voorbeeld.

Elk subject heeft een uniek kenmerk (unieke identifier). Deze verwijzen óf binnen de datasets van het RAZU, óf naar een externe bron zoals het kadaster.

Voor interne verwijzingen laat de unieke identifier ook zien in welke **organisatie** en **dataset** het **subject** staat:

data.razu.nl/**Gem-UH**/**bouwdossiers2006**/a221kkd2

Graphs zijn virtuele scheidingen tussen triples en daarom niet zichtbaar in de identifier. Daarmee kan op basis van een triple de gehele structuur uitgelegd worden:

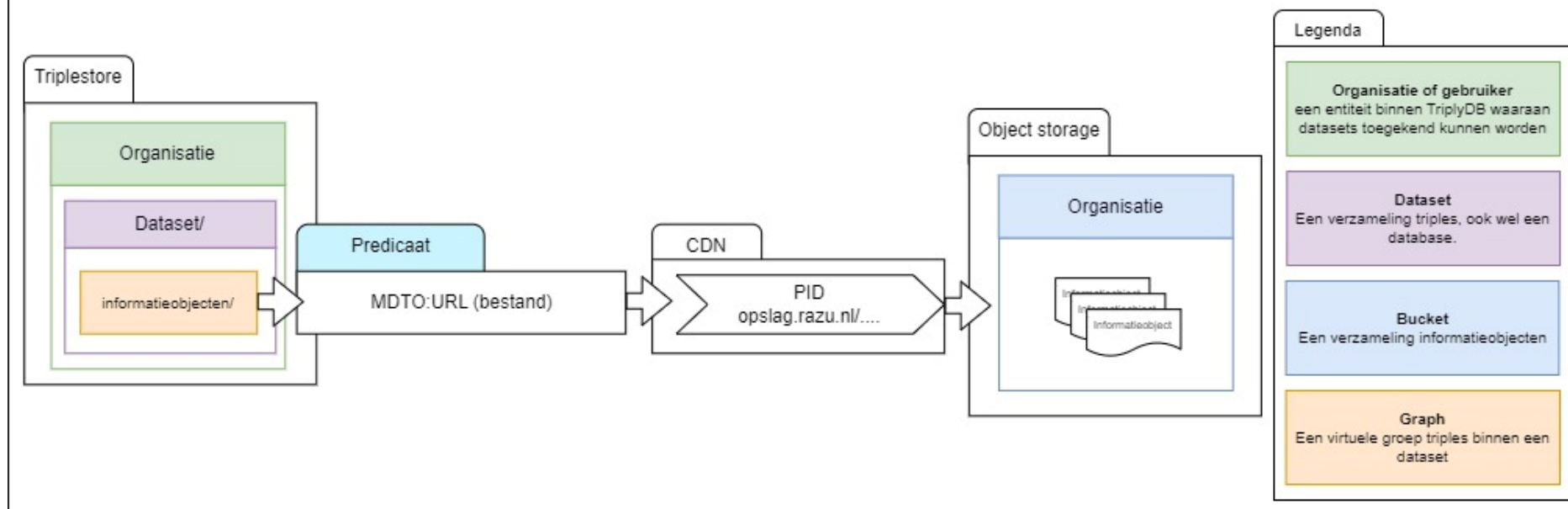


Schematische weergave opbouw object storage

Concept

De objectstorage volgt de organisaties van de Triple store. Elke organisatie krijgt een eigen bucket toegewezen. Binnen de buckets worden datasets gescheiden middels folder-structuren. Toegang tot de buckets is niet open, maar gebeurt via API-sleutels die het RAZU afgeeft. De toegang verloopt daarnaast via de CDN, waarmee de werkelijke locatie van de bucket wordt verborgen. Het CDN-adres van een object is de PID die als triple staat geregistreerd.

Opbouw Object storage



Colofon

Informatiemodel e-depot

26 februari 2024

Foto voorpagina: Burgemeester Bransen (links) bij een model van het beeld 'Mijn baas en ik' dat op Het Rond is geplaatst, THA Houten, Schalkwijk en Tull en 't Waal (353), inv. 223-105871. Fotograaf: Rob Glastra.

Regionaal Archief Zuid-Utrecht

Karel de Grotestraat 30

3962CL Wijk bij Duurstede

<https://www.razu.nl>