

Hoe kun je Linked Open Data thesauri gebruiken om je collectie te beschrijven?

Een goede collectiebeschrijving is niet alleen noodzakelijk voor een goed beheer van je collectie, maar ook voor de doorzoekbaarheid ervan. Een beschrijving van het geheel en de delen geeft je reeds een overzicht van je collectie, maar hiernaast is ook een goede beschrijving per voorwerp gewenst. Het achterhalen van de gegevens die het verhaal achter elk voorwerp vertellen (bijvoorbeeld het gebruik en de herkomst), vereist echter kennis. Als deze kennis aanwezig is in kleine collectiebeherende organisaties, zit ze vaak verspreid bij verschillende medewerkers. Zij dienen samen te werken bij de beschrijving, niet alleen om ze zo volledig mogelijk te krijgen maar ook te verzekeren dat gelijkaardige voorwerpen een gelijkaardige beschrijving krijgen, en geen verschillende.

Een gelijkaardige beschrijving voor gelijkaardige voorwerpen veronderstelt het gebruik van dezelfde termen. Als je voor de beschrijving van een voorwerp bijvoorbeeld een dialectwoord in plaats van een gestandaardiseerde term gebruikt, zal je het voorwerp nog wel terugvinden in de collectiecatalogus maar mogelijk andere medewerkers niet wanneer zij bij het zoeken andere dialectwoorden gebruiken. Bovendien beperkt een dialect zich steeds tot een bepaalde regio, waardoor het gebruik van dialectwoorden in collectiebeschrijvingen de collecties ontoegankelijk maakt voor geïnteresseerden van buiten de regio. Dit is slechts één voorbeeld waarom je in je collectiebeschrijving steeds moet gebruikmaken van gestandaardiseerde termen zoals die zijn vastgelegd in een thesaurus.¹

Het normaliseren van je collectiebeschrijving met een geschikte thesaurus is echter een tijdrovende en foutgevoelige taak. Vaak moet je immers eerst voor iedere term de juiste schrijfwijze in de thesaurus opzoeken en deze vervolgens handmatig overnemen in je collectiebeheersysteem. Bovendien zijn thesauri geen statisch gegeven. Ze evolueren doordat ze verder worden uitgebreid en geactualiseerd. Hierdoor loop je al snel het risico dat je termenlijst

veroudert. Bovendien wil je je collectiedata mogelijk niet enkel doorzoekbaar maken, maar ze ook verrijken met bijkomende contextuele informatie. In dit artikel gaan we op zoek naar het antwoord op de volgende vragen: hoe kun je het gebruik van een thesaurus bij het beschrijven van je collectie vereenvoudigen, en ze minder arbeidsintensief en foutgevoelig maken? Biedt het gebruik van Linked Open Data thesauri hiervoor een oplossing? En hoe doe je dat dan in de praktijk?

1. Van loodzware thesauri naar Linked Open Data

Bij de registratie of beschrijving van objecten of concepten botst men vaak op het probleem van de dubbelzinnigheid van data. Gegevens zijn moeilijk te interpreteren als de context ontbreekt: een term kan immers meerdere betekenissen hebben en een object of een concept kan op verschillende manieren benoemd worden. Gegevens 'op zich', dus zonder context, zijn daarom vaak moeilijk uitwisselbaar en niet goed doorzoekbaar. Zo kun je de term 'bank' gebruiken om verschillende zaken te beschrijven: niet enkel een instelling die je geld beheert maar ook een zitmeubel. Terwijl in Vlaanderen een bank bijna altijd een hard en ongestoffeerd zitmeubel is voor binnen (bijvoorbeeld een kerkbank) of buiten (tuinbank), kan in Nederland een bank zowel hard als zacht zijn. Wat in Vlaanderen vaak een zetel of canapé wordt genoemd, is in Nederland een bank. Dergelijk voorbeeld illustreert dat we in onze collectiebeschrijvingen nood hebben aan gestandaardiseerde termenlijst waarin de betekenis van elke gebruikte term nauwkeurig is vastgelegd.

Om dat probleem aan te pakken, heeft men in een eerste fase afspraken in organisaties en instellingen gemaakt over de termen die worden gebruikt om objecten en concepten te benoemen. In eerste instantie werden er dus interne werkinstrumenten gecreëerd, zoals bijvoorbeeld gecontroleerde termenlijsten, en werden afspraken gemaakt over het

gebruik van de termen. Dit zorgde voor normalisering, maar loste de problemen van de doorzoekbaarheid en uitwisselbaarheid van de gegevens niet op. Deze vorm van normalisering verzekert wel dat iedereen binnen eenzelfde organisatie of instelling dezelfde term gebruikt om een object of concept te beschrijven, maar verschillende organisaties of instellingen kunnen nog steeds verschillende termenlijsten gebruiken. De interne termenlijsten kunnen zoals eerder aangegeven snel verouderen en vereisen daarom voortdurend onderhoud.

Om het gebruik van verschillende termenlijsten te voorkomen om dezelfde objecten of concepten te beschrijven heeft men in een volgende fase afspraken gemaakt over de organisaties of instellingen heen. Dit heeft geresulteerd in decimale classificatiesystemen, standaardterminologieën en thesauri voor specifieke sectoren. Zulke thesauri bestaan uit een geordende termenlijst die niet alleen een overzicht bevat van enkelvoudige begripseenheden, maar ook hun onderlinge semantische relaties vermeldt, zoals meer algemene en meer specifieke termen, synoniemen en gerelateerde termen.² Hierdoor werd het mogelijk om als collectieregistrator een term op te zoeken in een loodzware (papieren) thesaurus en vervolgens de 'afgesproken' schrijfwijze en, indien beschikbaar, een code over te nemen in de eigen beschrijving of systeem.

Met de komst van het internet en nieuwe mogelijkheden om kennis snel en wereldwijd uit te wisselen,



Het decimale classificatiesysteem ontwikkeld door de Belgen Paul Otlet en Henri La Fontaine aan het begin van de 20e eeuw. Bron: https://commons.wikimedia.org/wiki/File:Classification_D%C3%A9cimale.jpg

werden de bestaande thesauri en standaardterminologieën steeds meer online toegankelijk gemaakt. Een voorbeeld van een dergelijke online thesaurus die wordt gebruikt in de cultureel-erfgoedsector is de AAT-Ned, de Nederlandse versie van de Art & Architecture Thesaurus. Wanneer men in de AAT-



Thesaurus linguae latinae

Bron: https://commons.wikimedia.org/wiki/File:Thesaurus_linguae_latinae.JPG

Ned de term 'bank' opzoekt, krijg je als resultaat:

- term: bank (meubilair)
- termtype: niet-voorkeursterm
- concept id: 300038494
- gebruik: zitbanken (meubilair)

De AAT-Ned gebruikt dus de term 'zitbanken (meubilair)' in plaats van 'bank' om een 'zetel' of 'canapé' aan te duiden. Als men in de thesaurus de term 'zitbanken (meubilair)' opzoekt, krijg je als resultaat:

- term: zitbanken (meubilair)
- broader term: <meubilair met meerdere zitplaatsen> Preferred parent
- narrower term:
 - <banken naar locatie of context>
 - <banken naar vorm of functie>
- alternatieve term: zitbank (meubilair)
- related term: krukken
- scope note: Zitplaatsen voor twee of meer personen, meestal zonder rugleuning.
- facet: Objecten
- termtype: voorkeursterm
- record type: concept
- concept id: 300038494
- used for

banken (meubilair)

bank (meubilair)

· bron: Van Dale Engels-Nederlands (1989)

· bron volledig:

Termenlijst Meubelen

Termenlijst Meubelen

Rijksdienst voor Beeldende Kunst

· Link naar concept:

<http://browser.aat-ned.nl/300038494>

Het werd mogelijk om via het internet de nodige termen in een thesaurus sneller op te zoeken en zelfs volledige thesauri (of een deel ervan) als eigen ‘gecontroleerde termenlijsten’ over te nemen in de eigen systemen. Ook creëerden vrijwilligers nieuwe thesauri buiten wetenschappelijke instituten en bibliotheken. Die ontwikkelingen maakte het gebruik van thesauri gebruiksvriendelijker, maar het volstond echter niet om alle problemen op te lossen. Het publiceren van informatie op het internet kon immers oorspronkelijk enkel in de vorm van ‘vaste’ pagina’s die je als gebruiker wel kon bekijken, maar waaraan je niks kon aanpassen of toevoegen en die moeilijk kon worden hergebruikt. Het oude principe van ‘manueel opzoeken en overnemen’ bleef dus oorspronkelijk ook in het internettijdperk behouden. Bovendien werd het probleem van het onderhoud van verschillende verouderde versies van termenlijsten niet opgelost.

Gelukkig ontwikkelde men ondertussen een nieuw concept voor het toegankelijk maken van data via het internet: *Linked Open Data (LOD)*. Dit concept kan gezien worden als een poging om, net als in het *Mundaneum* van Paul Otlet en Henri La Fontaine aan het begin van de twintigste eeuw, alle kennis van de wereld samen te brengen en de problemen over de toegankelijkheid, dubbelzinnigheid en uitwisselbaarheid van data en kennis op te lossen. Het concept LOD werd in 2006 uitgewerkt door Tim Berners-Lee, de man achter het World Wide Web en het HTTP-protocol.³ LOD is data die op het web gepubliceerd wordt onder

1. een open licentie, vrij voor hergebruik,
2. in een gestructureerd standaard open formaat,
3. voorzien van persistente URL’s die het mogelijk maken om die data duurzaam te koppelen aan andere gepubliceerde data.⁴

Het doel van LOD is het ontwikkelen van een internet waar data makkelijk aan elkaar gekoppeld

kan worden en dus in een bredere context getoond kunnen worden. De principes van LOD maken het mogelijk om met verschillende onderzoekers, organisaties en instituten, én het brede publiek samen te werken in een gedistribueerd netwerk. Met LOD kunnen de verbanden tussen verschillende data veel sneller gelegd worden en kan het potentieel om nieuwe kennis en inzichten over onze wereld te creëren efficiënter worden aangesproken.

Een thesaurus is dan ook uiterst geschikt om als LOD op het web te publiceren. Gegevens uit een thesaurus die als LOD beschikbaar zijn, kunnen veel gebruiksvriendelijker geïmplementeerd worden: de data zijn goed gestructureerd, ieder concept heeft een persistente link (URI), de data wordt systematisch geactualiseerd en worden onder een open licentie vrij herbruikbaar.⁵ Steeds meer bestaande en nieuwe thesauri worden als LOD op het web toegankelijk gemaakt en aan elkaar gekoppeld. De AAT-Ned is een goed voorbeeld van een LOD-thesaurus. Je kan in je collectiebeschrijving een koppeling opnemen naar een term die onderdeel is van een LOD-thesaurus. Doordat verschillende collecties verwijzen naar deze termen, kun je ze allemaal doorzoeken met behulp van dezelfde term. In iedere collectie heeft de term dezelfde betekenis. Het stelt je, in het geval van de AAT, ook in staat om je collectie internationaal doorzoekbaar te maken want van de AAT bestaan naast een Nederlandse versie ook nog Engelse en anderstalige versies die gekoppeld zijn met elkaar. Het biedt je ook de mogelijkheid om in je collectiebeschrijving toch de term ‘zetel’ (of desgewenst ‘canapé’) te gebruiken zolang die term maar koppelt aan de term ‘zitbanken (meubilair)’ uit de AAT-Ned met behulp van de link <http://browser.aat-ned.nl/300038494>. Je zult dan het object in je collectiecatalogus vinden, zowel als je op ‘zetel’ (of desgewenst ‘canapé’) zoekt als op ‘zitbanken (meubilair)’. Veel beter is echter het opnemen van de link <http://vocab.getty.edu/aat/300038494>, waardoor het object ook gevonden wordt als men zoekt op ‘benches (furniture)’. Als men in de Engelse AAT als concept ID ‘300038494’ intikt, bekomt men immers het volgende resultaat:

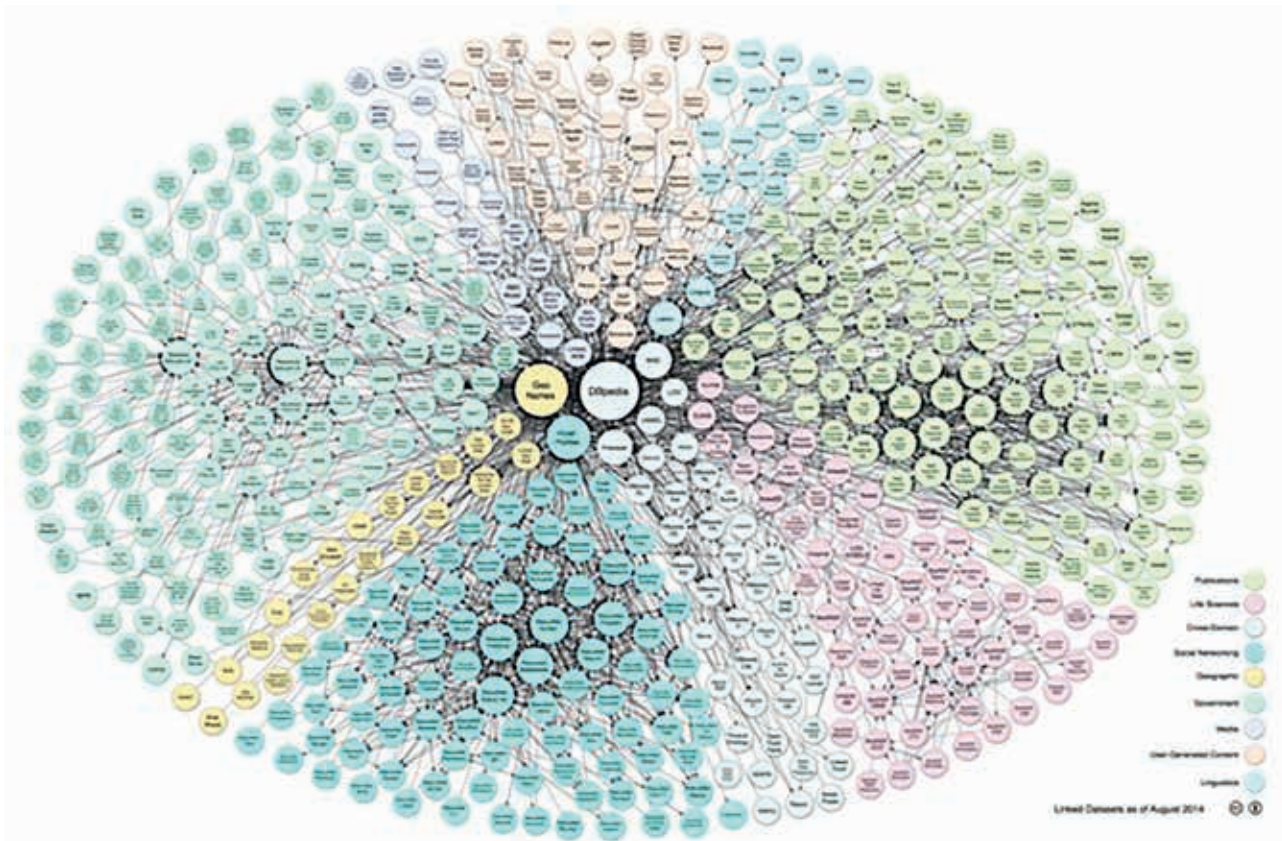
- Record Type: concept
- benches (furniture) (multiple-seating furniture, seating furniture, ... Furnishings and Equipment (hierarchy name))

- Note: Long seats for two or more people, with or without a back, usually of wood, but also of stone, metal, or another material. A bench is distinguished from a stool by its greater length.
- Terms:
 - benches (furniture) (preferred, C,U,LC, English-P,D,U,PN)
 - bench (furniture) (C,U,English,AD,U,SN)
 - bancos (benches) (C,U,English,UF,U,N)
 - bancos (benches) (Spanish,UF,U,PN)
 - banks (benches) (C,U,English,UF,U,N)
 - zitbanken (meubilair) (C,U,Dutch-P,D,U,U)
 - zitbank (meubilair) (C,U,Dutch,AD,U,U)
 - banken (meubilair) (C,U,Dutch,UF,U,U)
 - bank (meubilair) (C,U,Dutch,UF,U,U)
 - Bänke (Sitzmöbel) (C,U,German,D,PN)
 - Bank (Sitzmöbel) (C,U,German-P,AD,SN)
 - banco (C,U,Spanish,UF,U,SN)
 - banco (arredamento) (Italian,AD,U,SN)
 - banquetas (C,U,Spanish-P,D,U,PN)
 - banqueta (C,U,Spanish,AD,U,SN)
- Facet/Hierarchy Code: V.TC
- Hierarchical Position:
 - Objects Facet
 - Furnishings and Equipment
 - (hierarchy name) (G)
 - Furnishings (hierarchy name) (G)
 - furnishings (artifacts) (G)
 - <furnishings by form or function> (G)
 - furniture (G)
 - <furniture by form or function> (G)
 - seating furniture (G)
 - multiple-seating furniture (G)
 - benches
 - (furniture) (G)
- Additional Notes:
 - Dutch Zitplaatsen voor twee of meer personen, meestal zonder rugleuning.
 - German Lange Sitzgelegenheit für zwei oder mehr Personen, mit oder ohne Rückenlehne, in der Regel aus Holz, aber auch aus Stein, Metall oder anderen Materialien. Eine Bank wird von einem Hocker durch ihre größere Länge unterschieden.
 - Spanish Asientos para dos o más personas, generalmente sin respaldo.

- Related concepts:
 - distinguished from
 - › stools (seating furniture)
 - › (single seating furniture, seating furniture, ... Furnishings and Equipment (hierarchy name)) [300038418]
- Sources and Contributors:
 - banco (arredamento)
 - › [CDBP-DIBAM, ICCD Preferred, VP]
 - › ICCD database (1988-)
 - › Miller, Enciclopedia de Antigüedades (1999) 44
 - bancos (benches)
 - › [VP]
 - › Fairbanks and Bates, American Furniture (1981)
 - › Webster's Third New International Dictionary (1961)
 - bank (meubilair)
 - › [RKD, AAT-Ned]
 - › AAT-Ned (1994-)
 - › Termenlijst Meubelen
 - › Van Dale Engels-Nederlands (1989)
 - Bank (Sitzmöbel)
 - › [IfM-SMB-PK Preferred]
 - › Böth, Möbel: eine Typologie für Museen und Sammlungen (2005) 71
 - › Duden [online] (2011) http://www.duden.de/rechtschreibung/Bank_Sitzgelegenheit
 - › GND - Gemeinsame Normdatei <http://d-nb.info/gnd/4144005-5>
 - Bänke (Sitzmöbel)
 - › [IfM-SMB-PK]
 - › Duden [online] (2011) http://www.duden.de/rechtschreibung/Bank_Sitzgelegenheit
 - banken (meubilair)
 - › [RKD, AAT-Ned Preferred]
 - › AAT-Ned (1994-)
 - › Termenlijst Meubelen
 - › Van Dale Engels-Nederlands (1989)
 - banks (benches)
 - › [VP]
 - › Aronson, Encyclopedia of Furniture (1959) 35; banks
 - › Getty Vocabulary Program rules
 - banqueta
 - › [CDBP-DIBAM Preferred]
 - › Meyer, Manual de Ornamentación

- (2004) 564
- › Rodríguez Bernis, Diccionario de Mobiliario (2006) 58
- banquetas
 - › [CDBP-DIBAM]
 - › Comité, Plural del término en singular
- bench (furniture)
 - › [VP]
 - › Chenhall, Revised Nomenclature (1988)
- benches (furniture)
 - › [GCI Preferred, VP Preferred]
 - › AATA database (2002-) 120966 checked 26 January 2012
 - › Aronson, Encyclopedia of Furniture (1959) illustration, 48
 - › CDMARC Subjects: LCSH (1988-)
 - › Comstock, American Furniture (1962) 15
 - › Eames, Medieval Furniture (1977) 181
 - › Gilbert, Back-stairs Furniture (1977) illustration, ob13
 - › Gloag, Short Dictionary of Furniture (1965) 138
 - › Kane, 300 Years of American Seating Furniture (1976)
 - › RIBA, Architectural Keywords (1982)
 - › RILA, Subject Headings (1975-1990)
 - › Schwartz, Chairs, Tables, Sofas, Beds (1982) 160
 - › Sheraton, Cabinet Dictionary (1970) 45
 - › Winterthur Museum, Object name list, unpub. (1987)
- zitbank (meubilair)
 - › [RKD, AAT-Ned]
 - › AAT-Ned (1994-)
 - › Termenlijst Meubelen
 - › Van Dale Engels-Nederlands (1989)
- zitbanken (meubilair)
 - › [RKD, AAT-Ned]
 - › AAT-Ned (1994-)
 - › Termenlijst Meubelen
 - › Van Dale Engels-Nederlands (1989)
- Subject:
 - [CDBP-DIBAM, GCI, ICCD, IfM-SMB-PK, RKD, AAT-Ned, VP]
 - AATA database (2002-) 120966 checked 26 January 2012
 - RILA, Subject Headings (1975-1990) precoor.; Medieval benches

- Termenlijst Meubelen
- Van Dale Engels-Nederlands (1989)
- Note:
 - English
 - › [VP]
 - › Oxford English Dictionary Online (2002)
 - Dutch
 - › [RKD, AAT-Ned]
 - › AAT-Ned (1994-)
 - › Termenlijst Meubelen
 - › Van Dale Engels-Nederlands (1989)
 - German
 - › [IfM-SMB-PK]
 - › IfM Berlin
 - Spanish
 - › [CDBP-DIBAM]
 - › TAA database (2000-)



Linking Open Data cloud diagram 2014 (Max Schmachtenberg, Christian Bizer, Anja Jentzsch en Richard Cyganiak).
Bron: <http://lod-cloud.net/>

De AAT is maar een van de vele standaardterminologieën en thesauri die op dit moment als LOD zijn gepubliceerd. Het volledige overzicht van de standaardterminologieën en thesauri die deel uitmaken van de 'LOD cloud' kan men online bekijken:

In eerste instantie werden vooral de *crowdsourcing*-standaardterminologieën zoals DBpedia⁶ en Geonames⁷ volgens de principes van LOD online gecreëerd. Zulke thesauri zijn vaak erg groot en bieden een algemeen overzicht van objecten en concepten. *Crowdsourcing*-standaardterminologieën hebben echter soms minder wetenschappelijke autoriteit dan de klassieke thesauri, terwijl de cultureel-erfgoedsector daar net vragende partij voor is.

Een goed voorbeeld van een thesaurus met wetenschappelijke autoriteit die recent als LOD beschikbaar werd gemaakt door het beherende instituut zelf is de *Art & Architecture Thesaurus* (AAT). De AAT is een thesaurus die wereldwijd wordt gebruikt voor het toegankelijk maken van architectuur-, kunst- en cultuurhistorische collecties in onder andere musea, bibliotheken, archieven en documentatiecentra. De thesaurus bevat termen over kunst en architectuur, maar ook over kunstnijverheid, archeologie, archiefmaterialen en materiële

cultuur. De AAT bevat zowel termen die objecten aanduiden als termen voor materialen, technieken, personen en organisaties, stijlen, gebeurtenissen en abstracte begrippen.

In het najaar van 2014 heeft The Getty Institute de AAT als LOD online beschikbaar gemaakt via het *Getty-portaal*: onder een Open Data Commons Attribution Licentie ODC BY 1.0, in een goed gestructureerde vorm en voorzien van persistente URI's. Vroeger kon de data van de AAT enkel op een niet-commerciële basis geraadpleegd, geciteerd en gedeeld worden met vermelding van de bron, maar mocht de gehele structuur niet zonder toestemming worden overgenomen. Sinds 2014 zijn de mogelijkheden en de vrijheid voor het hergebruik van de data uitgebreid.

Op het *Getty-portaal* is ook een API (*SPARQL Endpoint*) beschikbaar. Dit is een webservice waarmee je grootschalige en ingewikkelde zoekopdrachten kunt uitvoeren en die je de resultaten in een gestructureerde vorm bezorgt. Daarnaast blijft het mogelijk om de volledige dataset te exporteren en in je eigen systeem te verwerken.

The screenshot shows the homepage of the Getty Vocabularies website. The header includes the logo and navigation links for LOD, SPARQL, and Queries. The main heading is 'The Getty Vocabularies' with a sub-heading 'Welcome to the SPARQL endpoint'. Below this, there is a paragraph explaining the vocabularies and their purpose. A section titled 'NOW Available:' lists three vocabularies: The Art & Architecture Thesaurus (AAT), The Getty Thesaurus of Geographic Names (TGN), and The Union List of Artist Names (ULAN). Each entry includes a brief description and a link to the vocabulary. There is also a 'Documentation and Downloads:' section with links to various resources like GVP Semantic Representations, GVP ontology, and datasets. On the right side, there is a 'Tweets' section showing recent tweets from Michelle Chesner and ASeim #478. At the bottom right, there is a 'Community Support Forum' link.

Bron: <http://vocab.getty.edu>

Deze vernieuwingen van The Getty Institute maken het mogelijk om:

- altijd de meest recente versie van de AAT voor alle taalvarianten te raadplegen via de webservice;
- een koppeling te leggen tussen je collectiebeheersysteem en de webservice van de AAT om vanuit je collectiebeheersysteem *live* zoekopdrachten uit te voeren in de AAT en tijdens de het invoeren van data suggesties te krijgen voor de 'juiste' schrijfwijze van termen. Op die manier moet je minder tijd besteden aan het opzoeken op welke manier een gelijkaardig voorwerp in de eigen of een andere collectie beschreven werd, om zo tot een uniforme beschrijving te komen. Dit vraagt echter bijkomende aanpassingen in je collectiebeheersysteem;
- updates van al gebruikte concepten en gegevens *live* en automatisch in je collectiebeheersysteem binnen te halen door een koppeling te leggen tussen je eigen data en de AAT. Een koppeling leggen kan door een persistente URI van een concept in je eigen data op te nemen. Via zo'n koppeling kan je systeem de *SPARQL Endpoint* aanspreken voor updates;
- gerichte ingewikkelde zoekopdrachten uit te voeren op de data uit de thesauri en zo meer context te krijgen over gebruikte concepten. Met behulp van de persistente URI van een term kun je vanuit een computerprogramma als

OpenRefine, je collectiebeheer- of CMS -systeem de *SPARQL Endpoint* aanspreken en vragen stellen (in *SPARQL*-zoektaal) die voor jou relevant zijn. Voorbeeld: 'Zoek een bepaalde term a.d.h.v. de persistente URI en geef me alle vertalingen terug'. Je ontvangt het resultaat in een formaat dat je eigen systeem zal begrijpen en kan verwerken, o.a. RDF/XML en JSON. De publicatie van de AAT als LOD creëert dus mogelijkheden voor een automatische maar gecontroleerde overname van gestandaardiseerde concepten en contextuele gegevens in je beschrijvingen die resulteert in een snellere, betere en meer correcte manier van werken. Je maakt je data niet louter beter doorzoekbaar, maar voorziet ook meertaligheid en een rijke inhoudelijke context. In het volgende hoofdstuk tonen we hoe je gebruik kan maken van een LOD-thesaurus. Hierbij leggen we de nadruk op hoe je een bestaande inventaris kunt aanpassen en dus verbeteren.

2. Hoe kan ik zelf gebruik maken van LOD-thesauri, zoals bijvoorbeeld de AAT?

De standaardmogelijkheid om een recente en correcte term online zelf op te zoeken in de AAT is natuurlijk nog altijd beschikbaar. Voor AAT dien je dan naar <http://vocab.getty.edu/> te surfen en de

gewenste term in het Engels, Nederlands of een andere taal op zoeken:

Om het volledige potentieel van LOD-thesauri aan te spreken heb je echter een ‘machine’ nodig die voor jou automatisch grootschalige zoekopdrachten kan uitvoeren en de koppelingen tussen jouw data en de data in een LOD-thesaurus kan ontdekken. Die ‘machine’ kan verschillende vormen aannemen, zoals bijvoorbeeld je collectiebeheersysteem, de CMS van een website, een applicatie of een handige tool zoals *OpenRefine*.

The screenshot shows the 'Getty Vocabularies: LOD SPARQL Queries' interface. The search query is 'olieverf' and it shows 4 results. The table below represents the data shown in the screenshot:

Subject	Term	Parents	Description
est:300030690	oil sketches	oil paintings (visual works), <paintings by material or technique>, ... Objects Facet	Paintings executed
est:300015050	oil paint (paint)	<paint by composition or origin>, paint (coating), ... Materials Facet	A paint made by oil paint to provide fa
est:300178604	oil painting (technique)	<painting techniques by medium>, painting techniques, ... Activities Facet	The art or practice from at least the 19
est:300037738	oil paintings (visual works)	<paintings by material or technique>, paintings (visual works), ... Objects Facet	Creative works in

Screenshot van een zoekopdracht voor ‘olieverf’ in de AAT op het Getty-portaal.

OpenRefine is een gratis en volledig *opensource software* die in het verleden werd ondersteund door Google en waarmee men makkelijk grote hoeveelheden data kan visualiseren, analyseren, manipuleren en corrigeren. Het is een tool die ontwikkeld werd voor *datacleaning*, het opschonen van data. Als collectiebeherende organisatie of instelling hou je gegevens bij over objecten en maak je die toegankelijk. *Datacleaning* doe je om de kwaliteit van de ontsluiting en vindbaarheid van een collectie te verhogen. Het vergemakkelijkt de interne omgang met de data en biedt mogelijkheden voor externe publicatie en uitwisseling van de gegevens. OpenRefine kan dus ook voor andere doeleinden interessant zijn, maar we focussen ons hier vooral op hoe je het kan aanwenden om in je bestaande collectiegegevens relatief snel te verbeteren.

De werking van OpenRefine is vergelijkbaar met die van een Excel-tabel, maar ze is veel geavanceerder en specifiek ontwikkeld voor tekstuele waarden. Het is een zelfstandig programma dat je op een computer kunt installeren en vervolgens via een browser lokaal en zonder internetverbinding kan gebruiken. Het werkt met je data buiten het eigenlijke databeheersysteem. De data moeten dus eerst worden geëxporteerd uit het gebruikte systeem (in

CSV, XML, XLS of een ander bestandsformaat) en vervolgens worden opgeladen in OpenRefine. De opgeschoonde data kunnen achteraf vanuit OpenRefine terug geïmporteerd worden in het oorspronkelijke databeheersysteem.

Met behulp van OpenRefine kun je de volgende dingen doen:

1. Data analyseren

Data worden op een overzichtelijke manier in verschillende kolommen als een tabel voorgesteld. Op elke kolom kun je specifieke filters toepassen om op verschillende manieren na te gaan welke waarden er voorkomen en of er fouten aanwezig zijn. Je kunt bijvoorbeeld een *text filter* gebruiken om te controleren of er in de waarden onnodige spaties of komma's voorkomen.

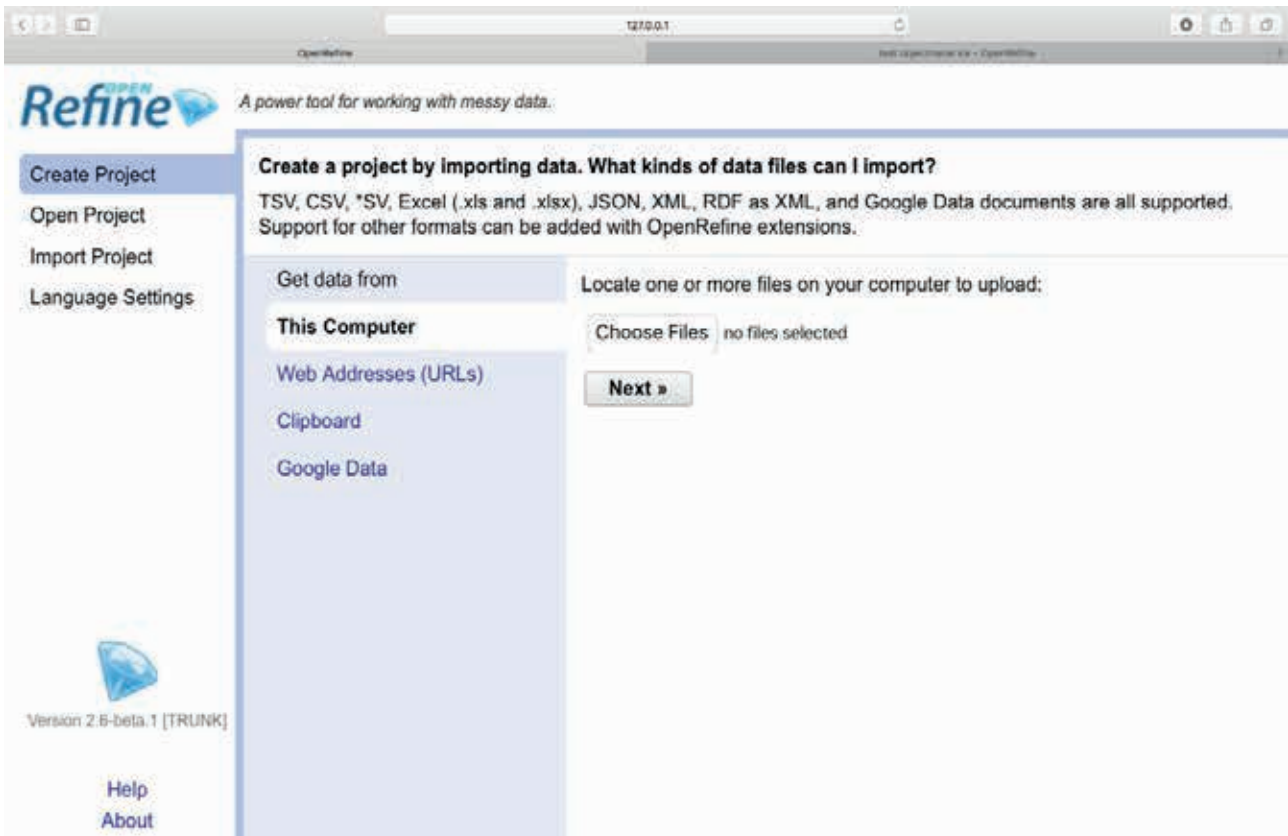
2. Data transformeren

De gevonden fouten hoeven niet apart gecorrigeerd te worden. OpenRefine biedt verschillende functies om grote hoeveelheden data tegelijkertijd aan te passen. Het is bijvoorbeeld mogelijk om data in een bepaalde kolom te *clusteren*, waarbij een lijst van spellingsvarianten wordt opgesteld. Daarna kun je voor een bepaalde variant kiezen die alle andere waarden vervangt. Daarnaast kun je gebruik maken van verschillende functies (uitgedrukt in *GREL expression language*), die rijke mogelijkheden tot datamanipulatie bieden. Zo kun je alle overbodige leestekens op het einde van termen zoeken en meteen verwijderen. Alle stappen worden bewaard, en indien nodig kun je de oorspronkelijke waarde herstellen.

3. Data verrijken

De zogenaamde *reconciliation service* kun je gebruiken om je eigen data te koppelen aan en te verrijken met waarden uit externe bronnen en standaardterminologieën. Zo kun je namen van bijvoorbeeld auteurs of kunstenaars koppelen met het overeenkomstige record in VIAF waarbij het identificatienummer van het VIAF-record wordt overgenomen, samen met de verschillende naamspellingsvarianten en biografische informatie.

OpenRefine is bovendien een gebruiksvriendelijk programma dat je helpt bij het opschonen van rommelige data. Meer informatie kun je vinden op de websites <http://openrefine.org/>, <https://github.com>.



Screenshot startscherm van Open Refine.

com/OpenRefine/OpenRefine/wiki en <http://freeyourmetadata.org/>. Er bestaat ook een praktische handleiding: *Using OpenRefine*, door Ruben Verborgh en Max De Wilde (Packt Publishing, 2013).

Een van de belangrijkste troeven van OpenRefine is dat de software je een mogelijkheid biedt om live te communiceren met LOD-thesauri, zoals de AAT. Je dient dan de volgende stappen te overlopen:

1. installeer OpenRefine op je computer;
2. exporteer de bestaande data die je wil verbeteren uit je collectiebeheersysteem en importeer ze in OpenRefine;
3. bewerk de data in OpenRefine (opschonen indien nodig);
4. voer een zoekopdracht voor alle data uit in AAT vanuit je OpenRefine, waarna je de resultaten ontvangt;
5. haal bijkomende data binnen uit AAT in je OpenRefine;
6. exporteer de opgeschoonde en genormaliseerde data uit OpenRefine en importeer ze in je collectiebeheersysteem.

Het resultaat is dat je data zijn opgeschoond en klaar voor hergebruik, uitwisseling en juiste interpretatie door je collega's en externe gebruikers.

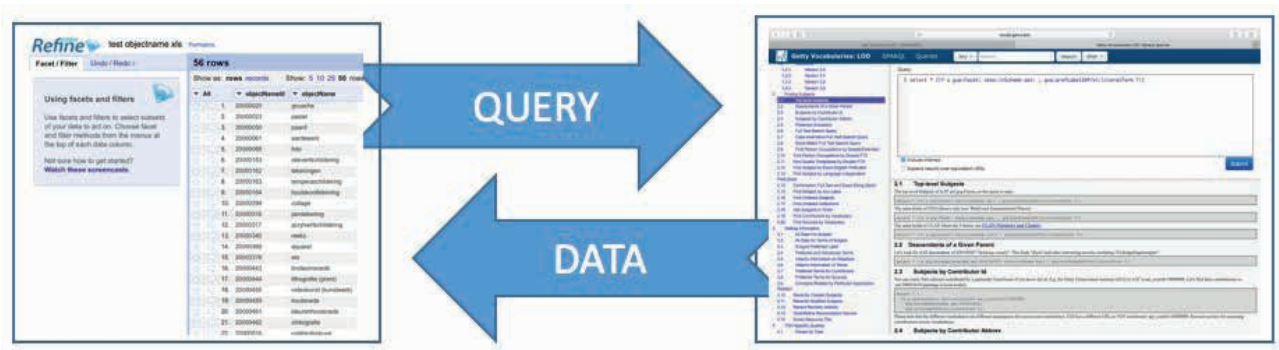
Deze stappen worden hieronder in detail overlopen en geïllustreerd met screenshots:

1. Installeer OpenRefine

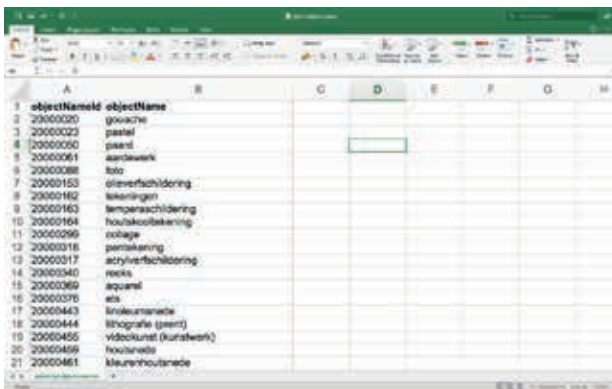
- a) Download het nodige installatiebestand OpenRefine 2.6: <http://openrefine.org/download.html>;
- b) Volg de instructies op <https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions>:
 1. installeer OpenRefine;
 2. dubbelklik op het icoontje en laat OpenRefine aan staan (visueel gebeurt er dan niet veel);
 3. open je webbrowser (best Firefox of Chrome) en ga naar <http://127.0.0.1:3333/>;
 4. je krijgt het startscherm te zien.

2. Importeer data in OpenRefine

- a) In OpenRefine kun je datasets in verschillende formaten importeren. Je kunt data uit je Excel-tabel nemen of data exporteren uit je collectiebeheersysteem. Voor dit voorbeeld bekijken we een eenvoudige tabel bestaande uit twee kolommen: 'objectNameId' (databank nummer van de term), 'objectName' (de term zelf):



Via OpenRefine spreek je de API van de Getty Vocabularies aan en ontvang je de gevraagde data.



b) Ga naar OpenRefine, dat je al open hebt staan in je webbrowser, en klik op 'Create project' op het startscherm. Klik vervolgens op 'Choose Files' en kies het Excel bestand op je computer dat je wil importeren en klik op 'NEXT'. In de volgende stap kun je een titel aan je project geven en functies aanpassen die invloed hebben op hoe je data wordt 'verwerkt' door OpenRefine (dit is handig voor bijvoorbeeld XML of CSV).

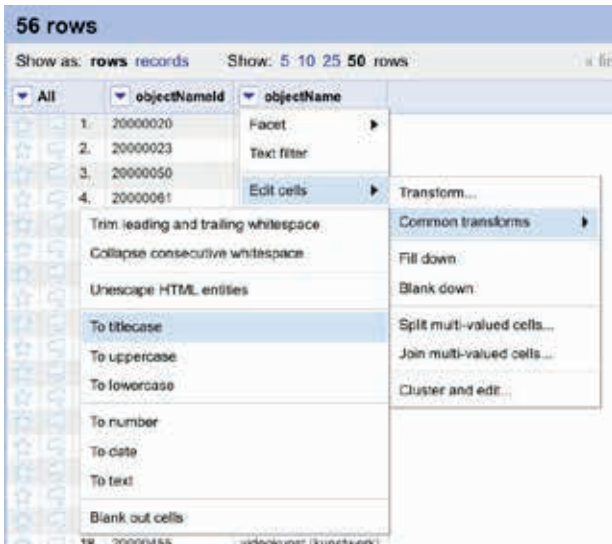


c) Druk op 'Create Project' rechtsboven en je tabel zal in OpenRefine verschijnen als een project.



3. Creëer een overzicht van je data en schoon de data op

Automatische opzoeking van de juiste schrijfwijze van een term in een externe thesaurus zoals de AAT zal vanuit OpenRefine makkelijker verlopen als je oorspronkelijke termen goed opgeschoond zijn en dus ontdaan zijn van tikfouten en onnodige leestekens. Zoals eerder vermeld, is OpenRefine een geschikte tool voor *datacleaning*. Met behulp van verschillende functies en filters kun je zicht krijgen op je data en die indien nodig opschonen (bijvoorbeeld door gebruik te maken van functies 'Tekst Facet', 'Cluster' en 'Edit') . Je kunt aanpassingen uitvoeren voor alle waarden in een bepaald kolom of per cel:



In dit voorbeeld zijn de waardes al opgeschoond.

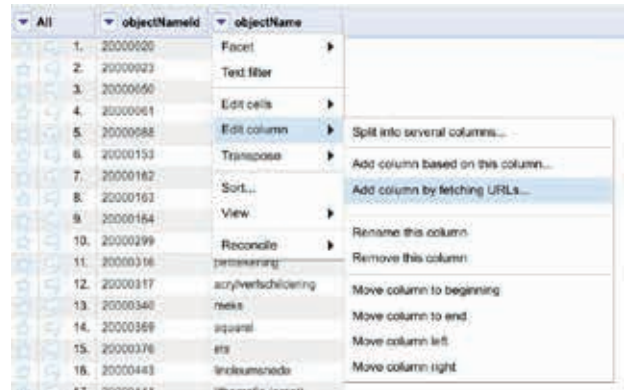


4. Zoek je termen op in de AAT met behulp van de API

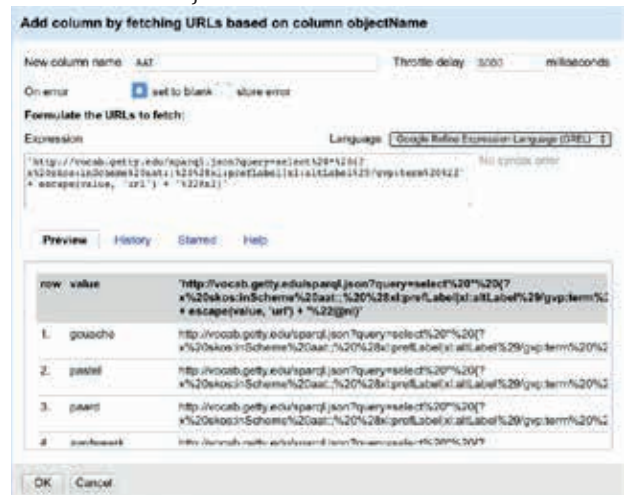
De API (webservice) van de *Getty-Portaal* laat toe om al je termen tegelijkertijd op te zoeken in de AAT. Daarvoor dien je een zoekopdracht (*query*) correct op te stellen en vanuit OpenRefine 'versturen' naar de API van Getty-Portaal. De query die in dit voorbeeld wordt verstuurd, vraagt aan de API om al onze termen op te zoeken in de AAT en, indien ze worden gevonden, een ID-nummer van de *preferred term* (voorkeursterm) terug te geven. Die query wordt in een *SPARQL-taal* geschreven, zodat de machines de query kunnen verstaan en voor ons de zoekopdracht kunnen uitvoeren.

a) Om een query te lanceren vanuit OpenRefine moet je de volgende stappen doorlopen:

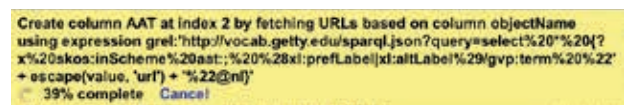
1. Klik op het pijltje boven de kolom 'objectName' -> kies voor de optie Edit column -> Add column by fetching URLs...



2. Je krijgt een scherm te zien waar je de functie kan definiëren die OpenRefine moet uitvoeren. In het veld *New column name* geef je een naam aan de nieuwe kolom, bijvoorbeeld *AAT*, waaruit de resultaten van de query zullen binnengehaald worden. In veld *Expression* dien je de *SPARQL*-query zelf te plaatsen. De query die we voor dit voorbeeld gebruiken is: `'http://vocab.getty.edu/sparql.json?query=select%20*%20{x%20skos:inScheme%20aat:;%20%28xl:prefLabel|xl:altLabel%29/gvp:term%20%22' + escape(value, 'url') + '%22@nl}'`

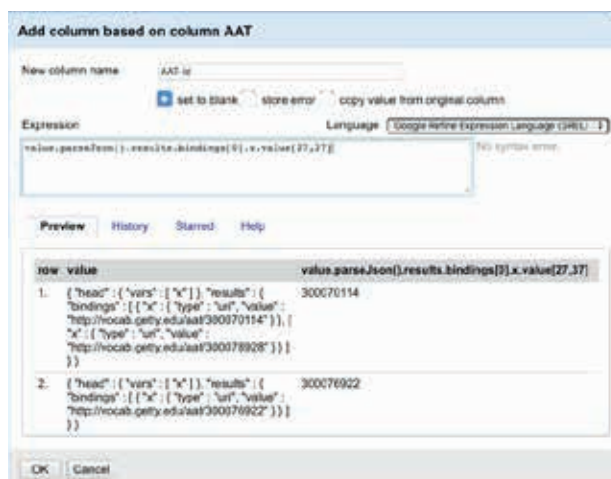


3. Druk op OK en wacht een paar minuten terwijl OpenRefine voor jou de zoekopdracht uitvoert.



4. De resultaten die je ontvangt worden in het JSON-formaat in OpenRefine weergegeven:

objectNameId	objectName	AAT
1. 2000000	gusache	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300070114" }, { "type": "uri", "value": "http://vocab.getty.edu/aat/300070114" }]] }] } }
2. 2000023	jester	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300076922" }]] }] } }
3. 2000090	paard	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300023378" }, { "type": "uri", "value": "http://vocab.getty.edu/aat/300023378" }, { "type": "uri", "value": "http://vocab.getty.edu/aat/300023378" }, { "type": "uri", "value": "http://vocab.getty.edu/aat/300023378" }]] }] } }
4. 2000091	antworb	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300140803" }]] }] } }
5. 2000088	foto	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300046300" }]] }] } }
6. 20000151	slaverfchabering	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300032799" }]] }] } }
7. 20000152	skeningen	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300038872" }]] }] } }
8. 20000161	lampenschabering	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300038872" }]] }] } }
9. 20000164	houtsoortbenaming	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300038872" }]] }] } }
10. 2000029	stijl	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300038872" }]] }] } }
11. 20000316	perskleding	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300038872" }]] }] } }
12. 20000317	anyvelschabering	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300181918" }]] }] } }
13. 2000040	reeks	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300046300" }]] }] } }
14. 2000038	spoor	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300046300" }]] }] } }



Je krijgt twee nieuwe kolommen met een AAT-ID en een AAT-URI voor elke gevonden term:

b) Nu kun je uit de resultaten in het JSON-formaat de nodige data halen door de volgende twee stappen uit te voeren:

1. Het ID-nummer van een AAT-term uithalen: klik op het pijltje boven de kolom met de JSON-resultaten 'AAT' -> kies voor de optie Edit column -> Add column based on this column -> Titel nieuwe kolom: AAT-id, Expression (zonder spaties): value.parseJson().results.bindings[0].x.value[27,37]

objectNameId	objectName	AAT	AAT-URI	AAT-id
1. 2000000	gusache	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300070114" }, { "type": "uri", "value": "http://vocab.getty.edu/aat/300070114" }]] }] } }	http://vocab.getty.edu/aat/300070114	300070114
2. 2000023	jester	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300076922" }]] }] } }	http://vocab.getty.edu/aat/300076922	300076922
3. 2000090	paard	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300023378" }, { "type": "uri", "value": "http://vocab.getty.edu/aat/300023378" }, { "type": "uri", "value": "http://vocab.getty.edu/aat/300023378" }, { "type": "uri", "value": "http://vocab.getty.edu/aat/300023378" }]] }] } }	http://vocab.getty.edu/aat/300023378	300023378
4. 2000091	antworb	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300140803" }]] }] } }	http://vocab.getty.edu/aat/300140803	300140803
5. 2000088	foto	{ "head": { "vars": ["x"], "results": [{ "bindings": [[{ "type": "uri", "value": "http://vocab.getty.edu/aat/300046300" }]] }] } }	http://vocab.getty.edu/aat/300046300	300046300

Als je op zo'n AAT-URI klikt, zal je naar de overzichtspagina van de term op het Getty-portaal worden doorverwezen. Deze links zijn persistent en zullen dus altijd actief blijven en informatie weergeven over een bepaalde term. Je kunt deze links dus overnemen in je tabel of je collectiebeheersysteem voor toekomstige referenties. Op die manier zullen andere gebruikers die je data bekijken of die met andere data willen vergelijken, weten wat je bedoelt met je termen: ze kunnen namelijk via de AAT-URI de term altijd opzoeken in de AAT-thesaurus en de context ervan bekijken.

5. Verrijk je data met contextuele data uit de AAT m.b.v. de API

Met behulp van de AAT-IDs van de termen, verkregen in de vorige stappen, kun je de API van de AAT weer aanspreken en bijkomende data binnenhalen zoals bv. voorkeurterm, alternatieve term, bovenliggende term en vertalingen van de term in andere

Term	chine-collé	Goedgekeurd	<input type="checkbox"/>
Soort term	techniek		
	objectnaam		
Term code			
Status	kandidaat		
Rubriek			
Relaties			
Use			
Used for			
Broader Term		termnummer	
Narrower Term			
Related Term			
Equivalent Term			
Bron en definitie			
Bron	Wikidata	Nummer	Q3674992 AAT <input type="checkbox"/>
URL	https://www.wikidata.org/wiki/Q3674992		
Bron	AAT	Nummer	300256021
URL	http://vocab.getty.edu/aat/300256021		
Scope Note			
Inhoudelijke datering			

Naam	Kreitz, Willy	Goedgekeurd	<input checked="" type="checkbox"/>
Soort naam	vervaardiger		<input type="checkbox"/>
Referentienummer			
Opmerking naam			
Relaties met andere namen			
Gebruik			
Gebruikt voor			
Equivalent		Categorie	
Naamdetails			
Titel		Geslacht	
Achternaam		Voornamen	
Tussenvoegsels		Naamtoevoegingen	
Voorletters			
Aanhef			
Bronnen en andere informatie over het record			
Bron	RKDartists	Nummer	46378
URL	https://rkd.nl/explore/artists/46378		
Bron	VIAF	Nummer	27933602
URL	http://viaf.org/viaf/27933602		
Bron	wikidata	Nummer	Q2228306
URL	http://www.wikidata.org/wiki/Q2228306		
Bron	ODIS	Nummer	PS_13256
URL	http://www.odis.be/lnk/PS_13256		

Eindnoten

- 1 Een thesaurus is een geordende lijst van termen die in principe enkelvoudige begripseenheden bestrijkt, met vermelding van hun onderlinge semantische relaties, zoals meer algemene en meer specifieke termen, synoniemen en gerelateerde termen. Een voorbeeld van een thesaurus die wordt gebruikt in de cultureel-erfgoedsector is de AAT-Ned, de Nederlandse versie van de Art & Architecture Thesaurus. Voor meer informatie, zie: <http://website.aat-ned.nl/home>.
- 2 Zie: <https://www.projectceest.be/wiki/Glossarium:Thesaurus>.
- 3 HTTP (HyperText Transfer Protocol) is het protocol voor de communicatie tussen een webclient (meestal een webbrowser) en een webserver dat de uitwisseling van documenten over netwerken mogelijk maakt. Het protocol wordt niet alleen veel op het World Wide Web gebruikt, maar ook op lokale netwerken (intranet).
- 4 Een URL is de 'locatie' op een webserver waar een browser een bepaald document kan terugvinden. Een persistente URL is een webadres dat ongewijzigd blijft doorheen de tijd. Door het gebruik van persistente URL's wordt het fenomeen van *linkrot* voorkomen dat optreedt wanneer de pagina waarnaar een URL verwijst niet meer bestaat.
- 5 Een webadres (URI) heeft twee functies: Het geeft een document een 'naam' (URN) waarmee het uniek geïdentificeerd wordt op het web en het bepaalt de 'locatie' (URL) op een webserver waar je browser het document kan terugvinden. Het HTTP-protocol biedt de mogelijkheid een onderscheid te maken tussen 'naam' en 'locatie' van een bestand. Je kunt (met hulp van je webbeheerder) je webserver zo instellen dat hij, op basis van een onveranderlijk webadres, het document op een ander webadres terugvindt. Op die manier kunnen je documenten steeds hetzelfde webadres behouden, wanneer je ze verplaatst of nieuwe versies publiceert.
- 6 DBpedia is een kennisdatabank die informatie uit Wikipedia op een gestructureerd manier als LOD toegankelijk maakt, waardoor het hergebruik ervan veel gemakkelijker wordt. Zie: <http://wiki.dbpedia.org/>
- 7 Geonames is een meertalige databank die meer dan 8 miljoen geografische namen bevat die ruim 7 miljoen entiteiten beschrijven, waarvan 2,6 miljoen bewoonde locaties en 2,8 miljoen alternatieve benamingen. Geonames is vrij beschikbaar als download of als webservice. Gebruikers kunnen data aanpassen in een wiki-interface. Zie: <http://www.geonames.org/> Met behulp van Geonames kun je op een eenvoudige manier een onderscheid maken tussen gemeenten die eenzelfde naam hebben, bv. het Kempisch grensdorp Putte en Putte in het zuiden van de provincie Antwerpen.
- 8 The Getty Institute maakte naast de AAT ook andere Getty Vocabularies beschikbaar als LOD: de Getty Thesaurus of Geographic Names (TGN) en de Union List of Artist Names (ULAN)).
- 9 Zie: <http://vocab.getty.edu/doc/>
- 10 Een CMS of Content Management Systeem is een applicatie die de pagina's, afbeeldingen en andere bestanden beheert die samen een website vormen. Een CMS biedt de mogelijkheid om webpagina's aan te maken en te bewerken, afbeeldingen toe te voegen en de presentatie, in de vorm van sjablonen, direct aan te passen. In toenemende mate is een CMS gekoppeld aan een databank (in bv. SQL) waarin de teksten worden opgeslagen. In andere gevallen gebruikt een CMS een XML-bestand om gegevens in te bewaren.
- 11 RDF is een voor het web ontwikkeld standaardmodel voor het coderen van metadata. Het model, dat uitgedrukt wordt in XML, gaat uit van drie onderdelen: object-eigenschap-waarde (een 'RDF-triple'). Het object is de bron die beschreven wordt, de eigenschappen zijn generieke kenmerken van die bron, de waarde geeft invulling aan een eigenschap. Voorbeeld: tafel-kleur-rood. Door objecten op deze wijze te beschrijven, kunnen grote en complexe dataverzamelingen op een uniforme wijze doorzocht worden.
- 12 XML of Extensible Markup Language is, zoals HTML een subset van Standard Generalized Markup Language (SGML). XML maakt de functionaliteit van SGML voor internet mogelijk. Met behulp van XML kan de structuur van tekstdocumenten gedefinieerd worden met behulp van tags en attributen. XML kan ook gebruikt worden om data te beschrijven die normaal gesproken in relationele databasesystemen worden opgenomen.
- 13 JSON staat voor JavaScript Object Notation. Het is een deelverzameling van de programmeertaal JavaScript. Deze wordt gebruikt voor het uitwisselen van datastructuren, met name in webapplicaties die asynchroon gegevens ophalen van de webserver zoals AJAX. De eenvoud van JSON heeft geleid tot een grote populariteit ervan, met name als een alternatief voor XML.
- 14 Zie: <http://openrefine.org/>
- 15 Een koma-gescheiden bestand, of CSV-bestand, in het Engels 'Comma Separated Values', is een specificatie voor tabelbestanden. Het CSV-formaat is het eenvoudigste databaseformaat dat er bestaat. Het bestaat enkel uit tekstgegevens, waardoor het gemakkelijk geïmplementeerd (lezen en/of schrijven) kan worden en het een brede verspreiding kent. Waarden worden in principe gescheiden door komma's, en regels door het nieuwe-regelteken ('\n'). Andere scheidingstekens komen echter ook voor waardoor er problemen kunnen ontstaan als CSV wordt gebruikt om gegevens uit te wisselen. CSV-gegevens kunnen in een rekenblad- of een databaseprogramma worden ingelezen en vervolgens als tabel worden gepresenteerd. Er bestaat geen algemene standaardspecificatie voor CSV.
- 16 Microsoft Excel (volledige naam Microsoft Office Excel) is een spreadsheet- of rekenbladtoepassing. Het .XLS-bestandsformaat is een proprietary bestandsformaat, maar wordt op zeer grote schaal toegepast in Windows- en Mac-platforms. Tot versie 2003 gebruikte het programma een binair formaat, dat in de meeste andere toepassingen kan worden ingelezen. Vanaf versie 2007 wordt een XML-gebaseerd formaat gebruikt dat echter niet door oudere Office-toepassingen kan herkend worden.
- 17 VIAF of Virtual International Authority File is een internationale standaardterminologie voor personen en instellingen. Zie: <https://viaf.org/>
- 18 Meer informatie over alle mogelijke functies in OpenRefine kun je vinden op de website <http://openrefine.org/>, <https://github.com/OpenRefine/OpenRefine/wiki> of <http://freeyourmetadata.org/>. Er bestaat ook een praktische handleiding: Using OpenRefine, door Ruben Verborgh en Max De Wilde (Packt Publishing, 2013).
- 19 SPARQL is een zoektaal (query language) die RDF graphs gebruikt. SPARQL is ontworpen om op webniveau te gebruiken en maakt daarom zoekopdrachten over verspreide bronnen mogelijk ongeacht het formaat. Het creëren van een enkele zoekopdracht over diverse dataopslagplaatsen is makkelijker dan het creëren van meervoudige zoekopdrachten. Het kost ook minder

en levert rijkere resultaten op.

- 20 Veel van de gebruikte query's zijn ook vindbaar online op CEST en kunnen vandaar overgenomen worden in je OpenRefine: https://www.projectcest.be/wiki/Publicatie:Collectiedata_normaliseren_met_Open_Refine en in de handleiding: https://www.projectcest.be/wiki/Bestand:20150430_Handout_Open_Refine_workshop.pdf