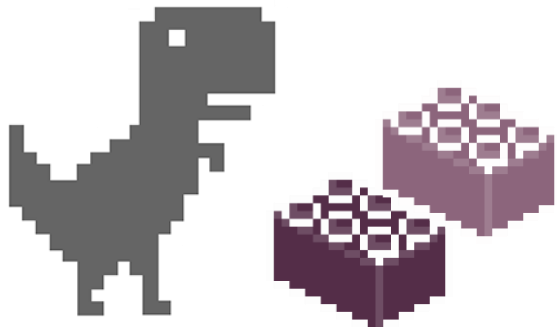
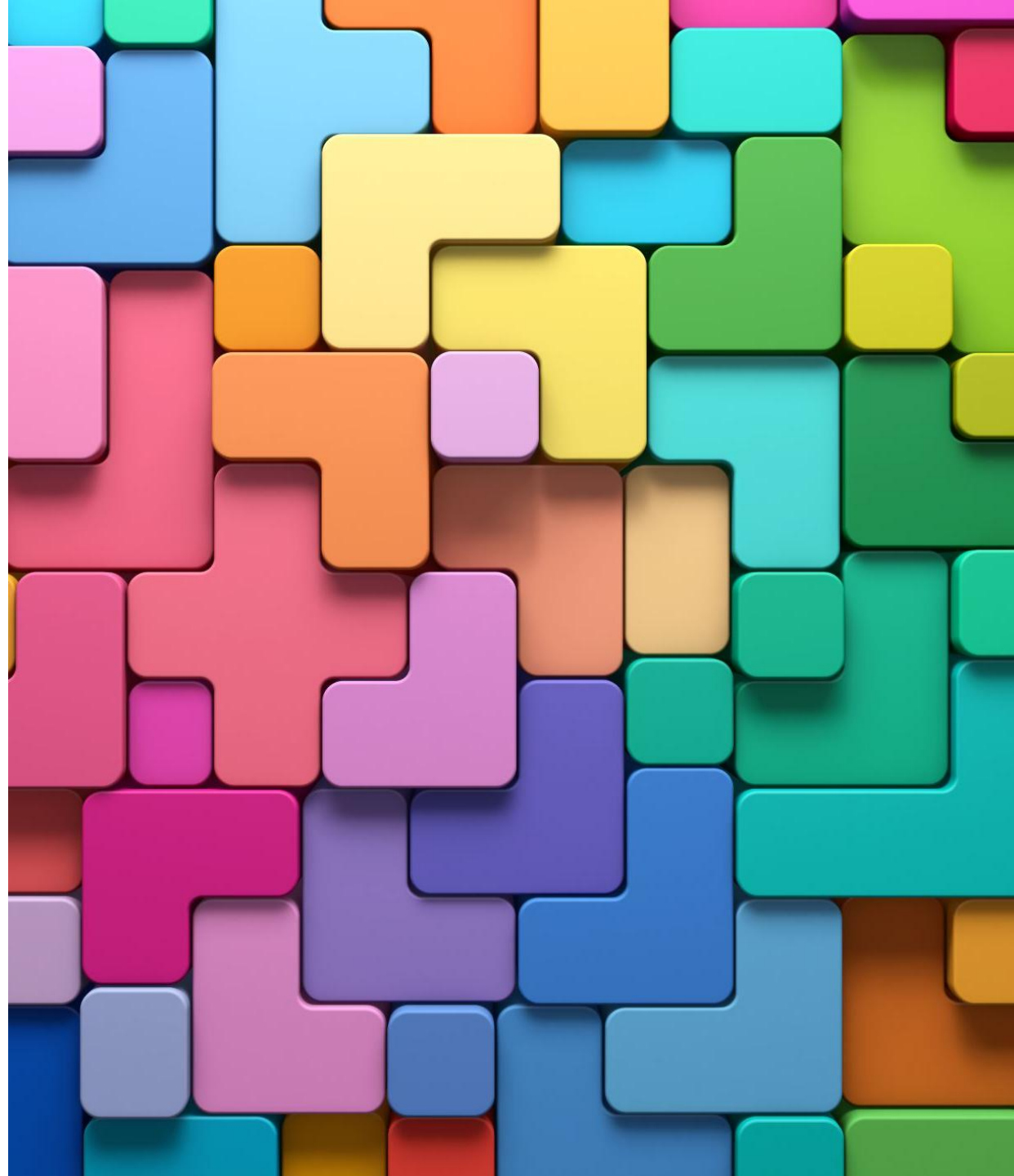

THE THESAURUS BUILDER



Shannon van Muijden
Informatiespecialist
Rijksmuseum





WORKSHOP/HANDLEIDING

1. [Waarom](#) een thesaurus?
2. [Uitdagingen](#)
3. [The Thesaurus builder](#), hoe werkt het, wat doet het en wat heb je nodig?
4. Een demo!
5. [Uitgebreide workflow](#) + instructies

WAAROM EEN THESAURUS?

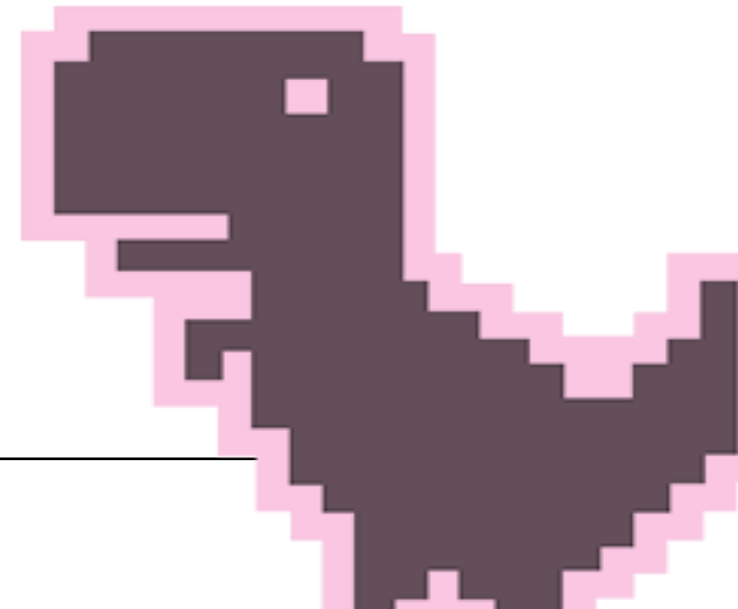
- Veel erfgoedinstellingen gebruiken *termen* om objecten te beschrijven
- Zonder *structuur* ontstaat overlap of je kunt simpelweg niets meer terugvinden
- Een thesaurus is een woordenlijst waarin begrippen *hiërarchisch* met elkaar verbonden zijn
- Dit helpt om je collectie *beter te doorzoeken en te begrijpen*.
- Voorbeeld:

Tafelgerei

bestek

lepel

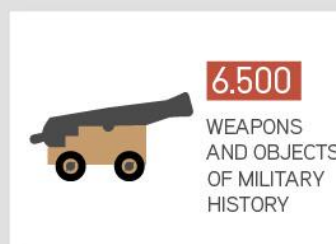
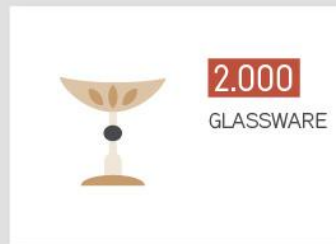
geboortelepel





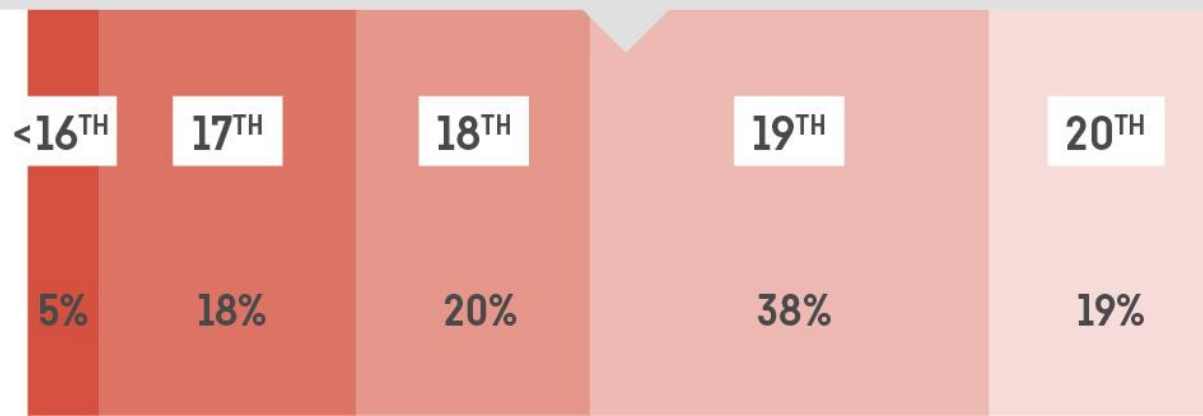


1.000.000
**OBJECTS IN
THE COLLECTION**



VISIBLE OBJECTS IN MUSEUM

DIVIDED OVER CENTURIES





UITDAGINGEN BIJ VEEL INSTELLINGEN

- Rommelige termenlijsten
- Handmatig hiërarchie aanleggen heeft laagste prio
- Fouten zijn snel gemaakt

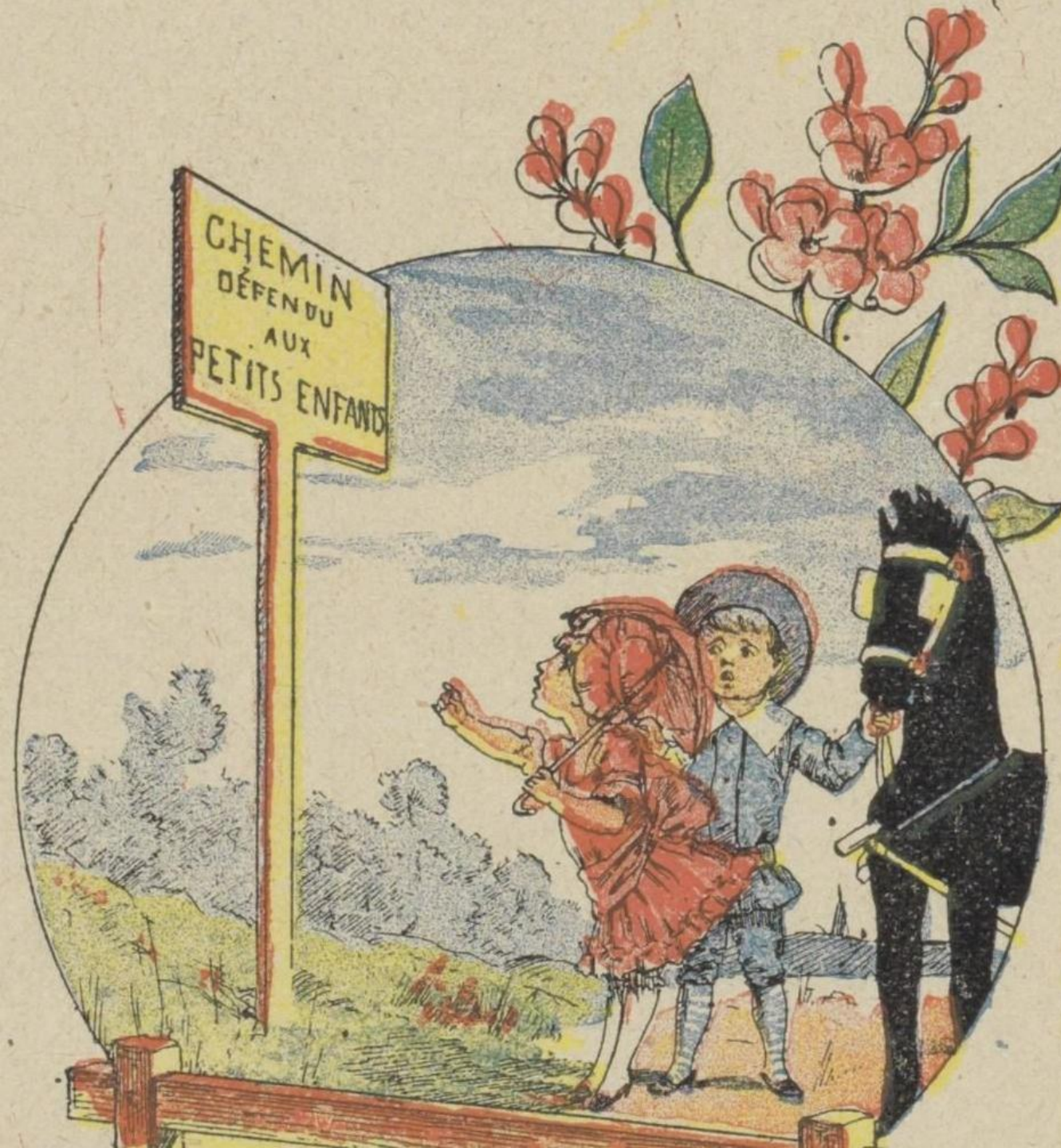
Daar komt [deze tool](#) bij kijken!



WAT DEZE TOOL DOET

Bouwt een hiërarchie voor je

- Geeft je een suggestie voor een breder term op basis van termen uit je eigen thesaurus
- Je voegt dus geen *nieuwe* termen toe aan je thesaurus, je maakt gebruik van wat je al hebt.
- Hij kijkt hiervoor naar de hiërarchie van de AAT.
- **Je termen moeten dus voorzien zijn van een URI naar de AAT.**



WAT DEZE TOOL *NIET* DOET

- Je termen verbinden met externe bronnen
- Hoe los ik dat op dan?
 - Gebruik een geautoriseerde, externe termenbron, zoals de [AAT](#), de [CHT](#), [Wikidata](#) enz. om jouw termen aan te verbinden;
 - Gebruik tools zoals [OpenRefine](#) en [handleidingen](#) om de URI's binnen te halen.
 - Vraag om hulp van datacleaners & [datawerkplaatsen van het Netwerk Digitaal Erfgoed](#).

Term

Term ambtsketen

Soort term objectnaam

Status niet gedefinieerd

Relaties

Use

Used for

Broader Term

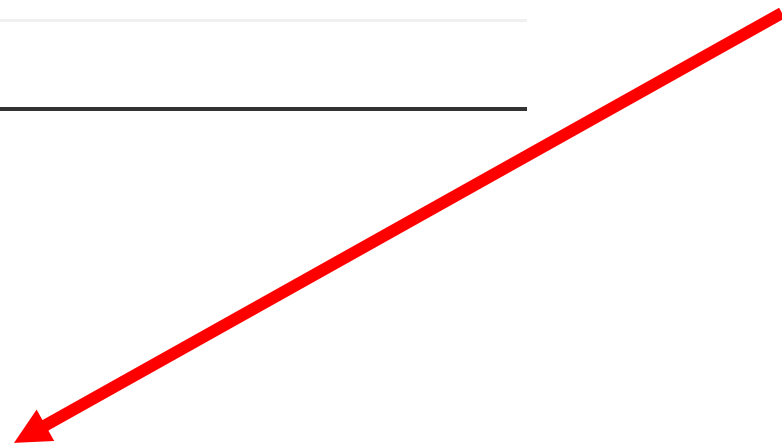
Narrower Term

Related Term

Equivalent Term

WEZEN

Termen zonder broader terms (*parents*)



THE THESAURUS BUILDER

doet drie dingen:

- **Haalt *broader terms* op** uit de AAT (via hun URI's).
- **Koppelt die** aan jouw eigen termen uit je volledige thesaurus.
- **Bouwt een hiërarchisch Excel-overzicht** met termen, matches, en eventuele mismatches.





WAT HEB JE NODIG?

- Collectiebeheersysteem of Excel document met termen die
 - ieder hun eigen unieke ID of recordnummer hebben
 - URI's naar de AAT hebben
- [Python 3.6 of hoger](#)
- Excel
- Stabiele internetverbinding



WORKFLOW

1. CSV met **subset** van termen (met AAT-URI's)
2. CSV met **volledige thesaurus**
3. Python script haalt bredere termen op via de AAT URI's
4. Output: Excel met hiërarchische structuur op basis van de AAT-hiërarchie.

Thesaurus Hierarchiebouwer

Dit Python-script (terms2broaders.py) bouwt automatisch een hiërarchische thesaurus. Het verrijkt je termen door bredere termen op te halen uit de Art & Architecture Thesaurus (AAT) met behulp van je bestaande termendatabase. Het script werkt in twee stappen: eerst geef je een subset van termen op waarvoor je bredere termen wilt; daarna geef je je volledige thesaurus om deze bredere termen tegen te matchen. Elke term in de output krijgt één bredere term toegewezen, wat zorgt voor een nette hiërarchische structuur. Het script gebruikt de URI's om de data te matchen, dus het werkt voor termen in *elke taal*, zolang ze gekoppeld zijn aan de juiste AAT-entiteit.

Vereisten

- Een database of spreadsheet met je thesaurustermen, elk met een uniek recordnummer of ID.
- Deze termen moeten afgestemd zijn op de Art & Architecture Thesaurus (<https://www.getty.edu/research/tools/vocabularies/aat/>)
- Python 3.6 of hoger geïnstalleerd.
- Mogelijkheid om je termen uit de database te exporteren naar een CSV-bestand (puntkomma ;-gescheiden).
- Internetverbinding (om bredere termen op te halen van de AAT-service).

Wat is Python?

- Python is een programmeertaal.
- De Python-interpreter is het programma dat Python-code leest en uitvoert.
- Python kan draaien:
 - **Interactief:** commando's regel voor regel typen.
 - **Scripts:** bestanden (`.py`) met meerdere commando's die in één keer uitgevoerd worden.

De **terminal** (Command Prompt, PowerShell, of Mac/Linux Terminal) is hoe je communiceert met Python. Hiermee geef je Python opdrachten of laat je scripts uitvoeren.

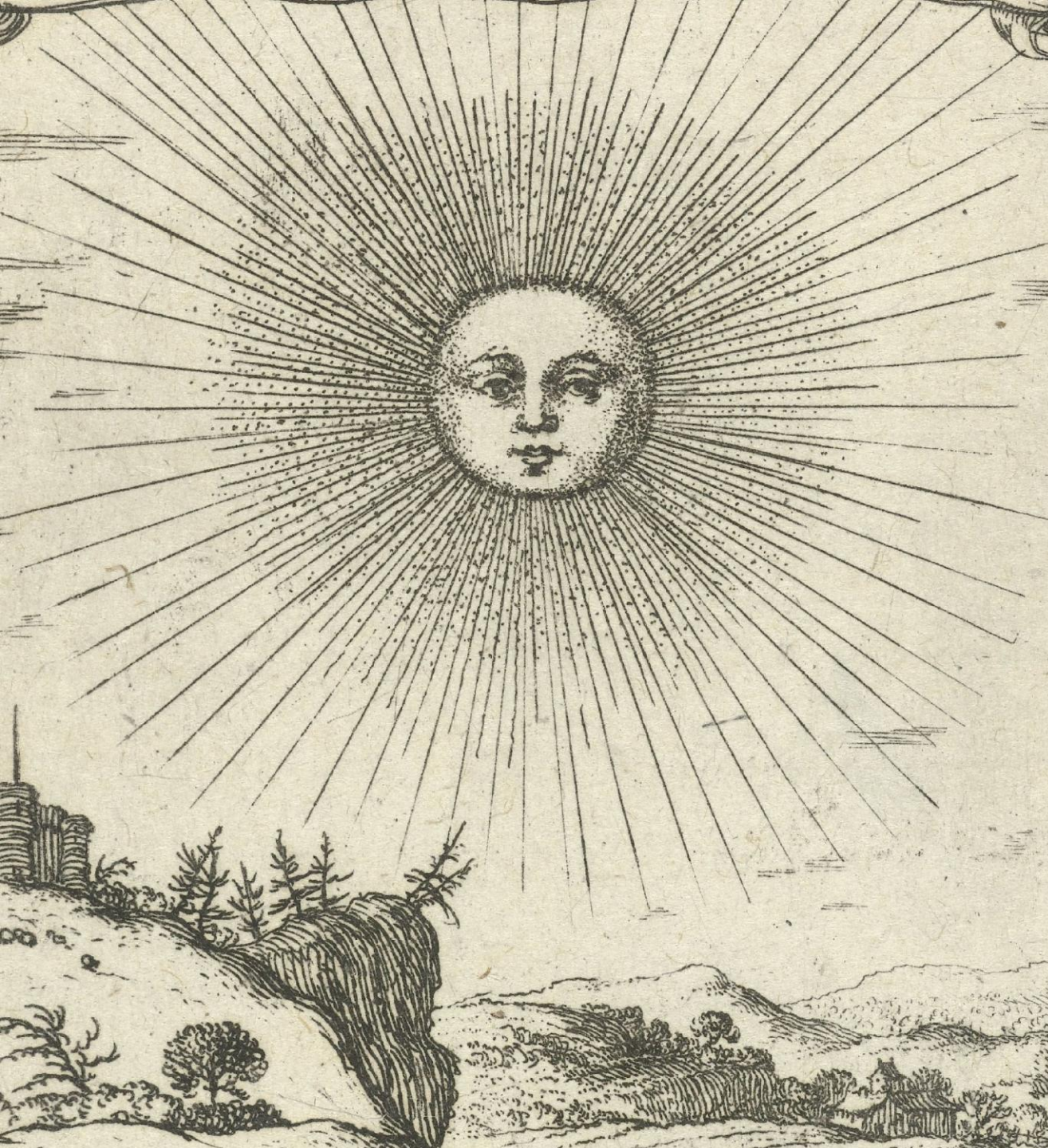
Stap 1: Installeer Python

1. Download Python van python.org.
2. Vink tijdens de installatie "Add Python to PATH" aan (Windows), zodat je Python vanuit elke terminal kunt draaien. De makkelijkste map om Python in te installeren (vooral op een werkcomputer) is C:\Users*<Je.Naam>\AppData\Local\Programs\Python\
3. Start Python in een commandowenster: Windows: Navigeer naar de map waar je Python-script is opgeslagen, Shift + Rechtsklik → "Open PowerShell window here" of "Open Command Prompt here". MacOS/Linux: Open

INSTRUCTIES

Uitgebreide instructies zijn [hier te vinden](#).

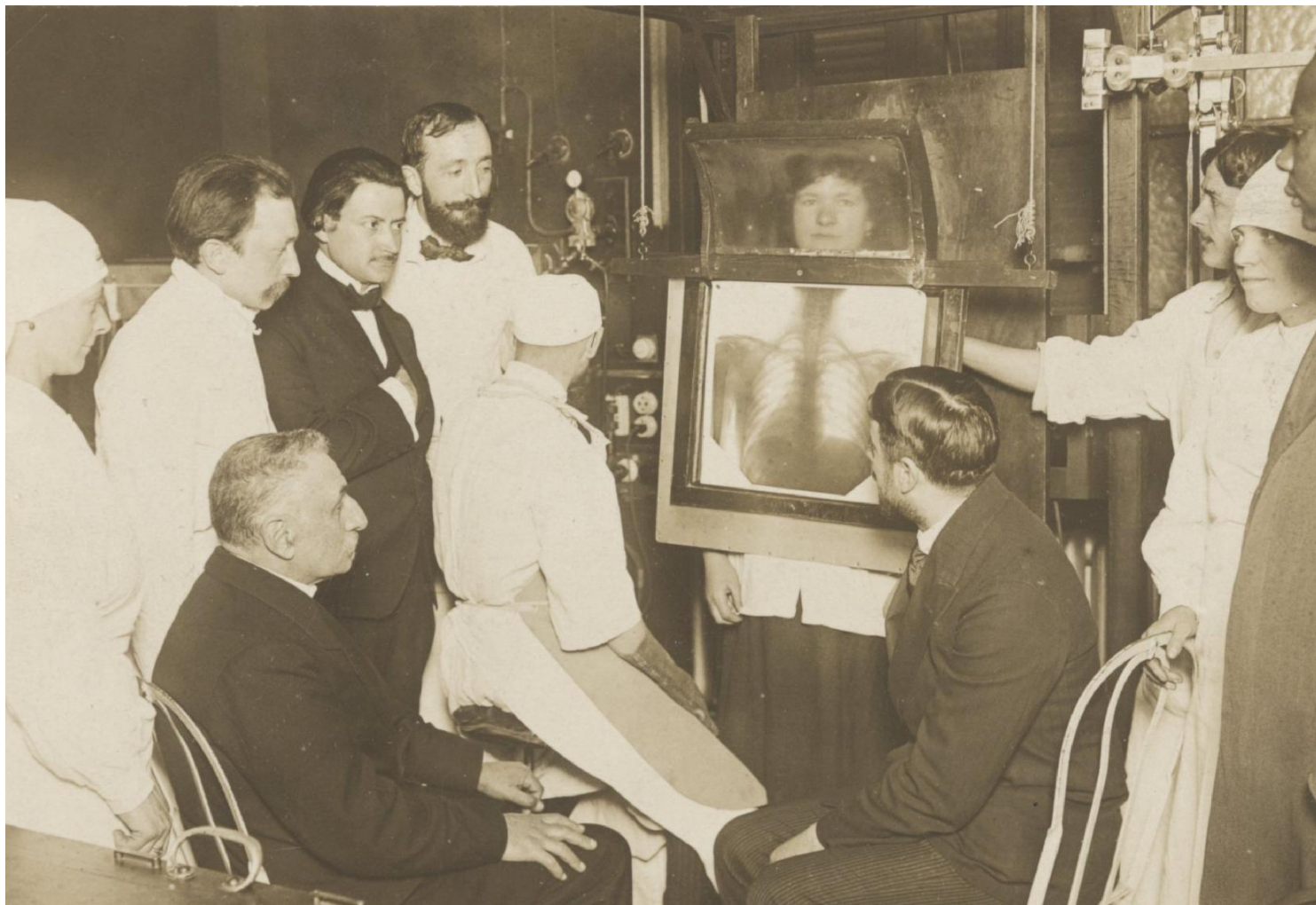
Het benodigde script (en optioneel .bat bestand) zijn [hier te vinden](#).



VOORDELEN

- Automatisch, snel en reproduceerbaar
- Geen technische voorkennis nodig dankzij het .bat-bestand
- Werkt met elke taal, zolang de AAT-URI's aanwezig zijn
- Resultaat is direct bruikbaar voor verdere verrijking of publicatie als linked data

DEMO!



Anoniem – Demonstratie van het maken van een röntgenfoto , ca. 1910. Rijksmuseum
<https://id.rijksmuseum.nl/200636592>

TERMEN LIJSTEN VERKRIJGEN

1. Open **Collections** of een ander collectieregistratiesysteem. De onderstaande voorbeelden gaan uit van Axiell Collections.
2. Ga naar de **thesaurus / termenlijst / trefwoordenlijst**.
3. Open **geavanceerd zoeken**. Let op: de onderstaande veldnamen kunnen per collectiebeheersysteem verschillen.
4. Zoek op alle termen *behalve* de niet-voorkeurstermen en de termen waarbij al een bredere term is ingevuld:
[all AND NOT use = * AND NOT broader_term = *].
5. Zoek op het **URI-veld**: geef alleen termen waarbij *aat* voorkomt in de URI:
[PID_overige.URI = *aat*].
6. Combineer die zoekacties met:
[set [nr] AND set [nr]] of
[(all AND NOT use = * AND NOT broader_term = *) AND (PID_overige.URI = *aat*)].
7. Maak eventueel een **subset** om de tool sneller te laten werken. Als je tijd hebt, kun je ook de volledige set gebruiken.
8. Je kunt subsets ook per domein maken, bijvoorbeeld voor objectnamen:
[(all AND NOT use = * AND NOT broader_term = * AND term.soort = objectnaam) AND (PID_overige.URI = *aat*)].
9. Open deze set en selecteer **alle objecten**.
10. Kies **Exporteren**.
11. Selecteer het **.CSV**-formaat en controleer bij de instellingen dat het **scheidingsteken een puntkomma (;)** is.
Neem in elk geval de velden **recordnummer**, **term** en **URI** mee.
Optioneel (maar aanbevolen): **term.soort** voor het domein (zoals materiaal, objectnaam, techniek, enz.), zodat je later kunt controleren of de termen correct aan de juiste domeinen zijn gekoppeld.
12. Ga vervolgens terug naar geavanceerd zoeken en haal de **volledige thesaurus** op met alle voorkeurstermen met AAT-URI's:
[(all AND NOT use = *) AND (PID_overige.URI = *aat*)].
13. Exporteer ook deze volledige lijst naar CSV met dezelfde kolommen (recordnummer, term, URI en eventueel term.soort).
14. Plaats de twee CSV-bestanden samen in een map waar je ze later makkelijk kunt terugvinden.
15. Open beide bestanden. Zorg dat de kolommen exact als volgt heten:
 - **recordnr** → voor recordnummer
 - **term** → voor de term
 - **URI** → voor de AAT-URI
De overige kolomnamen kun je laten zoals ze zijn.
 - **Domain** → voor term.soort in je volledige thesaurusbestand, als je daar de domeinen in hebt meegenomen.
16. Geef de bestanden duidelijke namen zodat je kunt zien welk bestand de **volledige thesaurus** is en welk de **subset**.

PYTHON INSTALLEREN EN GEBRUIKEN

1. Download **Python** via python.org/downloads.
2. Als je op een werklaptop werkt, installeer Python dan in de standaardlocatie die Python zelf voorstelt (meestal: C:\Users\<jouw.naam>\AppData\Local\Programs\Python\).
3. Zet een vinkje bij **“Add Python to PATH”**.
4. Ga naar de GitHub-pagina van **The Thesaurus Builder**:
https://github.com/Nondenon/The_Thesaurus_Builder
5. Download de bestanden **run_terms2broaders.bat** en **terms2broaders.py**.
6. Plaats beide bestanden samen in een map (de locatie maakt niet uit).
7. Voer het script uit: Dubbelklik op **run_terms2broaders.bat**.
Dit bestand installeert automatisch de benodigde libraries en start daarna direct het script.
8. Zodra het script start, opent er een **Verkenner-venster**. Selecteer hierin het bestand met de **subset van termen**.
9. Je ziet nu een **voortgangsbalk**. Afhankelijk van de grootte van je set kan dit even duren; het script haalt voor elke AAT-URI gegevens op via de AAT-webservice (ongeveer 5 minuten voor 800 termen).
10. Na afloop opent opnieuw een Verkenner-venster. Selecteer nu het bestand met de **volledige thesaurus**.
11. Deze stap verloopt snel. Daarna verschijnt het bericht dat er een Excel-bestand is aangemaakt: **thesaurus_with_hierarchy.xlsx**

DE OUTPUT (1)

1. Open de Excel. Onderaan zie je vier tabbladen:
 - matched_broaders
 - no_broader_match
 - no_AAT_URI
 - complete_hierarchy
2. In **matched_broaders** staan de termen waarvoor een geschikte bredere term is gevonden binnen je eigen thesaurus, op basis van de AAT-hiërarchie.
 - Het script kijkt naar de URI van de term en zoekt naar bredere termen (parents) binnen de AAT die ook in jouw thesaurus voorkomen.
 - De gevonden bredere term hoeft niet altijd direct één niveau hoger in de AAT te staan; het kan ook een hogere term zijn binnen dezelfde boomstructuur.
 - Jij bepaalt zelf of de voorgestelde term geschikt is om als broader toe te voegen.
 - In de sheet complete_hierarchy kun je de volledige parent-structuur bekijken om eventueel een geschiktere term te kiezen (bijvoorbeeld een meer specifieke gids-term).

Als je het veld **term.soort** hebt meegenomen, controleer dan of de domeinen van de broader terms overeenkomen met die van de oorspronkelijke termen.

In het voorbeeld hieronder wijkt het domein van de term af van het domein van de broader match.



recordnr	term.soort	term	URI	broader_term	AAT_ID	T-parentstri	Broader_term	domain
45	objectnaam	lood (gewicht)	http://vocab.getty.edu/aat/300011022	66618\$500	300011022	alloys\$http://vocab.getty.edu/aat/300241452	lood en loodlegering\$http://vocab.getty.edu/aat/300241452	materiaal

Hier zie je dat de oorspronkelijke term een *objectnaam* is, maar de bredere term verwijst naar een *materiaal*.

Dat betekent dat de gekoppelde AAT-URI voor *lood (gewicht)* eigenlijk naar *lood (materiaal)* verwijst en dus onjuist is.

DE OUTPUT (2)

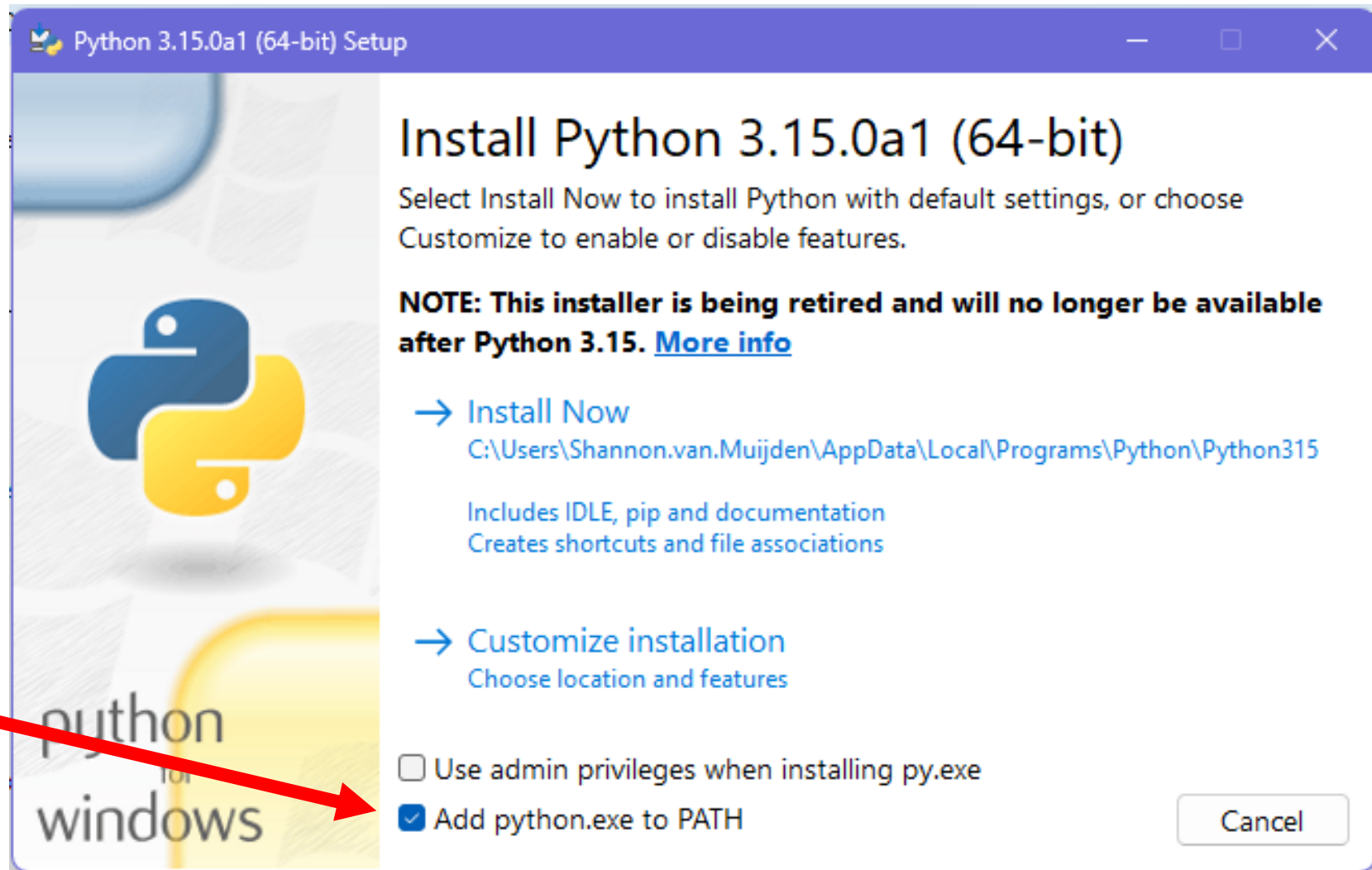
3. Sheet **no_broader_match** bevat termen waarvoor geen geschikte bredere term in je eigen thesaurus is gevonden.
4. Sheet **no_AAT_URI** toont termen zonder AAT-URI (het veld is leeg of bevat een andere URI).
5. Sheet **complete_hierarchy** toont van alle termen de volledige AAT-hiërarchie.

IMPORTEREN (IN AXIELL COLLECTIONS)

1. In de kolom **Broader_term** in `matched_broaders` vind je regels in de vorm:
`recordnummer$term$URI`.
2. Kopieer uit **matched_broaders** de kolommen **recordnr** en **Broader_term**. En plak deze in een nieuw Excel document.
3. Splits `Broader_term` op het `$`-teken, zodat de drie waarden elk in hun eigen kolom komen.
4. Maak een importbestand waarin:
 - Kolom A = `recordnummer` (of `priref / %0`)
 - Kolom B = `broader_term.lref`

Met het `recordnummer` dat verwijst naar de `broader term` (`broader_term.lref`) importeer je de gevonden `broader terms` bij de juiste records in Adlib of Axiell Collections.

PYTHON INSTALLEREN



 **The_Thesaurus_Builder** Public

Pin Unwatch 2

main 1 Branch 0 Tags Add file <> Code

 **Nondenon**

Update README.md c16971f · 15 minutes ago 🕒 31 Commits

 README.md	Update README.md	15 minutes ago
 run_terms2broaders.bat	Add files via upload	19 hours ago
 terms2broaders.py	Update terms2broaders.py	18 minutes ago

README 📄 🔗

Voor Nederlands [klik hier](#)

Thesaurus Hierarchy Builder

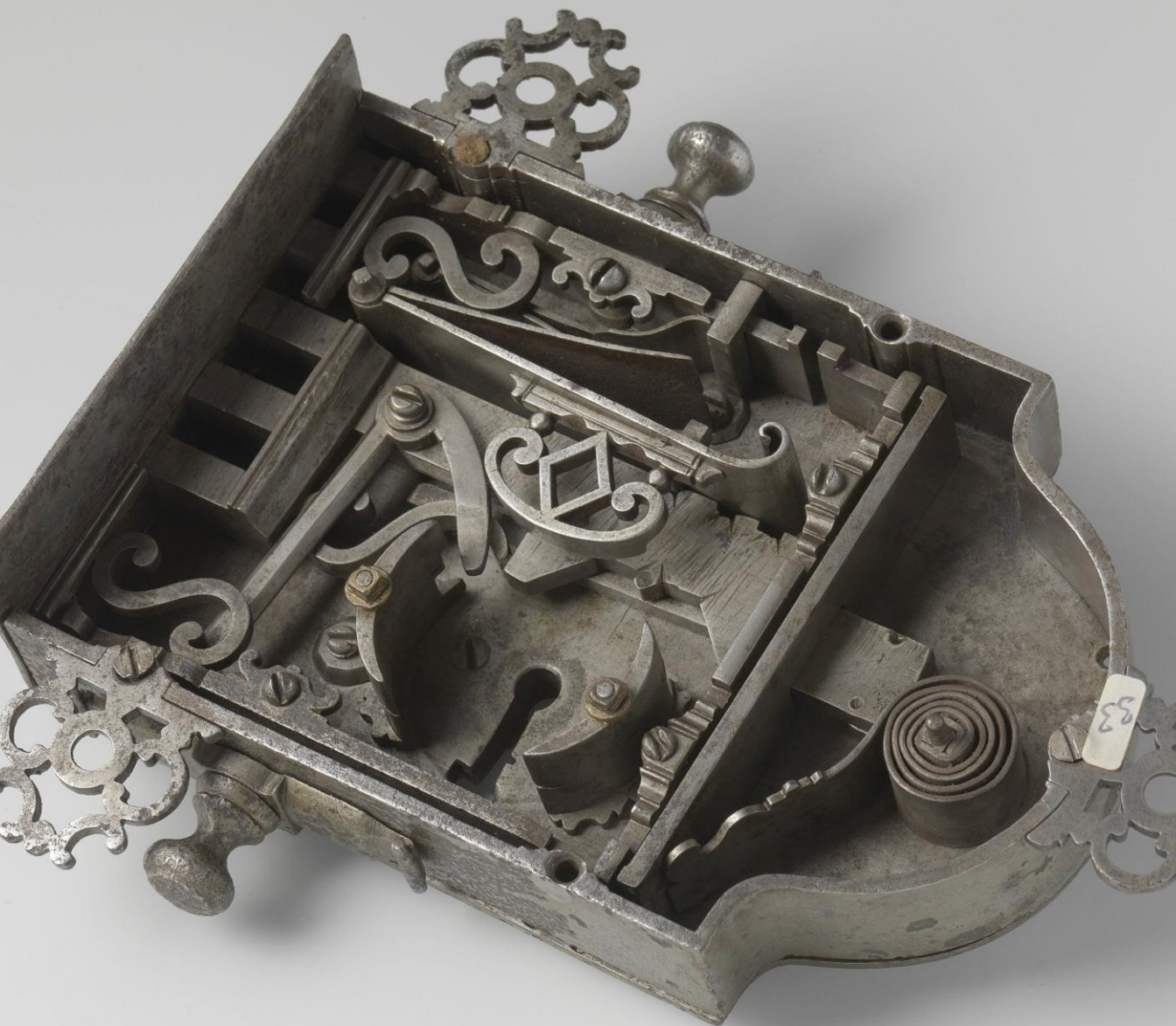
This Python script (terms2broaders.py) automatically builds a hierarchical thesaurus. It enriches your terms by fetching broader terms from the Art & Architecture Thesaurus (AAT) using your existing term database. The script works in two steps: first, you provide a subset of terms for which you want broader terms; then, you provide your full thesaurus to match these broader terms against. Each term in the output is assigned one broader term, ensuring a clean, hierarchical structure. The script uses the URI's to match the data so it will work on terms in *any language*, as long as they are matched with their respective AAT entity.

Requirements

- A database or spreadsheet with your thesaurus terms, each with a unique record number or ID.
- These terms must be aligned with the Art & Architecture Thesaurus (<https://www.getty.edu/research/tools/vocabularies/aat/>)
- Python 3.6 or higher installed.
- Python libraries:

GITHUB

[The Thesaurus Builder github repository](#)



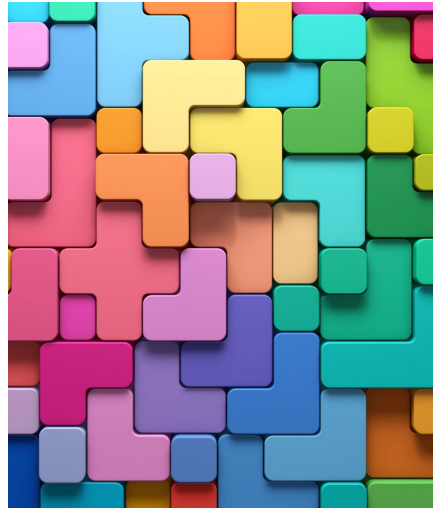
AFLSUITEND...

Met deze tool breng je **structuur** aan in je termen en maak je de collectiedata **beter bruikbaar**, zowel intern als online op je collectiewebsite.



EN DAN NOG IETS...

- **Bespaart tijd:** je hoeft niet zelf een volledige thesaurus uit te denken of te hiërarchiseren.
- **Flexibiliteit:** kies zelf hoeveel hiërarchische stappen je wilt overnemen; dit hangt af van wat past bij jouw collectie.
- **Omgaan met ontbrekende broader terms:** termen zonder match geven inzicht in waar de thesaurus kan worden uitgebreid.
- Geeft intern betere zoekfunctionaliteit in collectiebeheersystemen.
- **Extern / website:** complexere vragen kunnen beantwoord worden, bv.:
 - “Geef me alle wapens en alles gerelateerd aan wapentuig in de collectie van het Rijksmuseum.”
 - “Geef me alle objecten met ‘kant’ als techniek.” → van 709 naar 4236 objecten als bredere term wordt toegepast.



BEDANKT!
VRAGEN?



Shannon van Muijden

Informatiespecialist

Rijksmuseum

Shannon.van.Muijden@rijksmuseum.nl
