Praktijk voorbeeld
Design Museum Gent

Thesaurus opschoning en verrijking met behulp van OpenRefine

Design Museum

- 2016 – 2019: Minimale registratie
  - Aanvullen “basis” gegevens voor identificatie object
  - Aanvullen reproducties
    - >>nog geen rechten geklaard!
- **2019 – 2020: Data cleaning en aanmaken van (hiërarchische) woordenlijsten - thesaurus**
- 20?? – 2023: uitgebreide registratie
  - Waarderingstraject
  - Contextuele informatie toevoegen aan de objecten

Design Museum

- Aanmaken databank van Belgische vervaardigers
  - Dataset “vervaardigers”
  - Ontsluiten van gegevens ontwerpers / producenten van de collectiestukken
  - RKD/ULAN/Wikidata

- Publiek ontsluiten collectie via (web)interface/API
  - Veel voorwaarden waaraan voldaan moet worden
- LOD – LOUD (Linked Open Useable Data)
  - Data (her)bruikbaar maken – denken in termen van systemen
  - Doelpubliek?

Design Museum

- 1e beweging: normalisering en verrijking van de thesauri/databanken
- >>Volgens standaarden invulboek+ reconciliëren
  - Personen en instellingen
  - Thesaurus (GEO/PLAATS/LAND/…)
  - Correct wegschrijven van foutief geplaatste velden
    - Afstemmen aan periodes en homogeniseren.
- 2e beweging: richting LO(U)D - ontsluiting
  - P.URI’s
  - Creative Common Licenties
  - (open) hergebruik DING / API / Systeeminfrastructuur
  - Datadonaties (easy LOD):
    - Wikidata
    - ULAN, ..

Workflow

- Bepalen welke databanken eerst aan te pakken en analyse van de noden en verwachtingen:
  - Wat? Hoe? Wie? Waarom?
    - >> doelgericht datacleanen en normaliseren.
  - Intern-extern
  - DING
    - >> LOD mogelijk maken

Aanpak

1. GEO
   - Plaatsen en geografische termen (nodig voor nationaliteit, plaats van werkzaamheid, ea.)
2. PERS & INST
   - Opdeling subsets adhv soort term:
     - Vervaardiger, persoon, instelling, ..
     - Isoleren van deze enkel in gebruik in de bibliotheekcatalogus
       - Geavanceerd zoeken: “enkel term.soort = “auteur” AND/OR “uitgever”.
     - Opdeling ifv DING, LOUD, ..
       - Personen
       - Instellingen

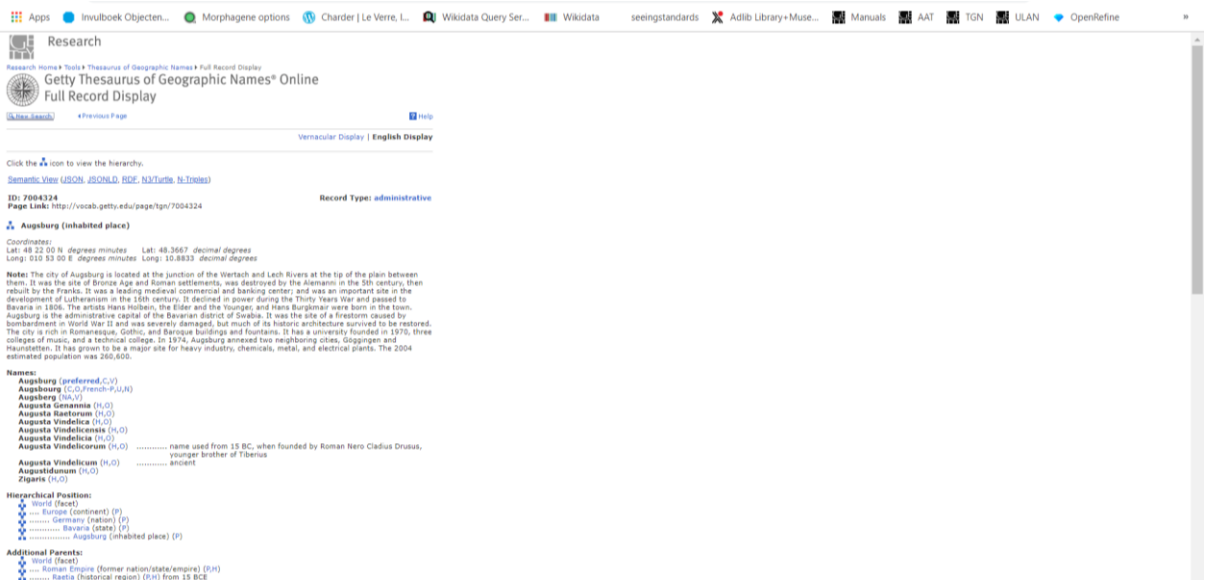
Geografische termen

Geografische trefwoord, plaats, land, ..

- Grootste vervuiling in bibliotheekmodule
  - Field overload
    - BERLIN STUTTGART NEWYORK (trefwoorden)z
      - > verworpen (status)
  - Duplicaten
    - Merge & cluster (OpenRefine)
  - Misplaatste termen
  - > migratie ALEPH (kunstenbib), intern gebruik onzeker
  - (?) > vermoedelijk
  - Nederlandse schrijfwijze (exoniem)

Geografische termen

- Hiërarchisch structureren aan de hand van Broader Term (BT) en Narrower term (NT).
  - Nieuwe zoekmogelijkheden; op plaats, streek, land, continent, werelddeel
  - Voor België ook mogelijk om te zoeken op provincie
  - Indien ambiguïteit mogelijk > plaatsen van BT tussen “()”
    - Limburg (provincie, België) x Limburg (provincie, Nederland)
    - Antwerpen (stad), Antwerpen (provincie)
    - Etc.



Geografische termen

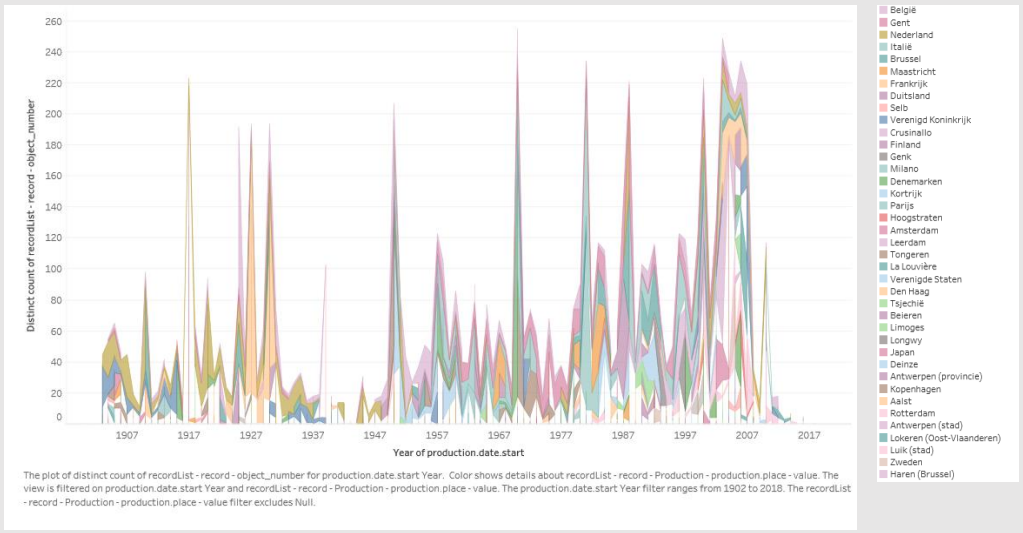
- Verrijken
  - **Reconciliëring** met **TGN** (Thesaurus of Geographic Names –Getty)
    - Mogelijkheid tot het binnentrekken van extra informatie via API Getty
      - Adhv **OpenRefine**.
        - Tool voor verrijken:
          - Data synchronisatie // koppeling API's
        - Opmeting, opschoning en normalisatie
          - Via merge & cluster
          - Splitt cells
          - GREL expressions
        - Verkennen van data dmv Facets, Tekst filters, ea

OpenRefine

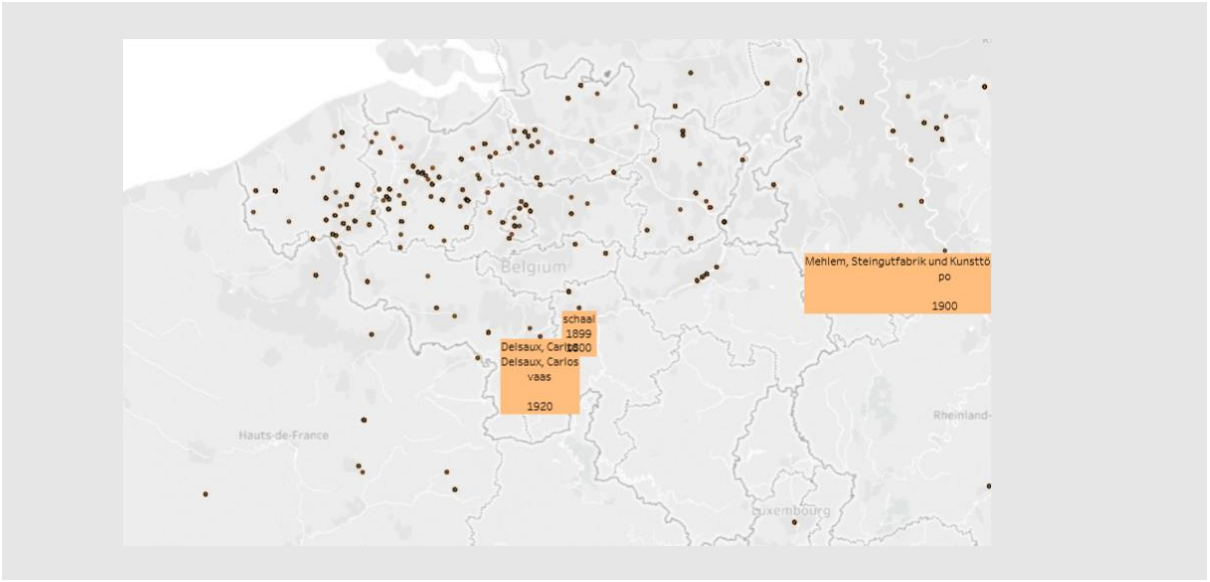
- Voorbeeld van binnentrekken metadata via SPARQL:

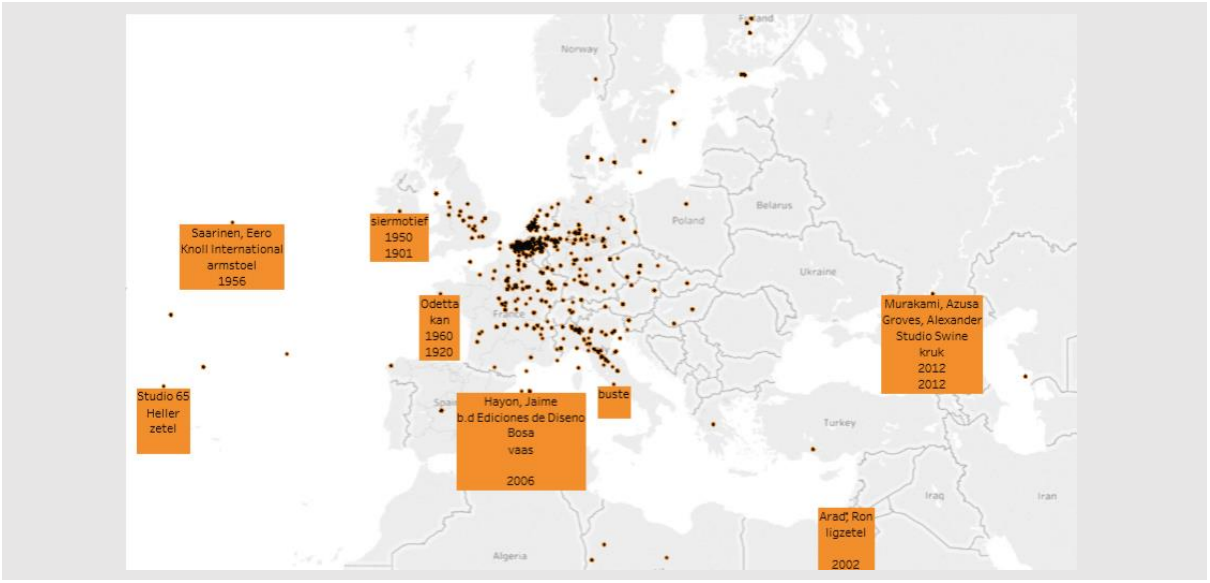
1. >> Create a new column based on the ID column
   - 'http://vocab.getty.edu/sparql.json?query=select+?lat+?long+{tgn:'+value+'+foaf:focus/wgs:lat+?lat.+tgn:'+value+'+foaf:focus/wgs:long+?long}'

2. >> Parse JSON
   - value.parseJson().results.bindings[0].lat.value+', '+value.parseJson().results.bindings[0].long.value



The plot of distinct count of recordList - record - object_number for production.date.start Year. Color shows details about recordList - record - Production - production.place - value. The view is filtered on production.date.start Year and recordList - record - Production - production.place - value. The production.date.start Year filter ranges from 1902 to 2018. The recordList - record - Production - production.place - value filter excludes Null.


Belgium
Hauts-de-France
Rheinland-
Luxembourg
Mehlem, Steingutfabrik und Kunsttö
po
1900
schaal
1899
1900
Delsaux, Carlos
Delsaux, Carlos
vaas
1920

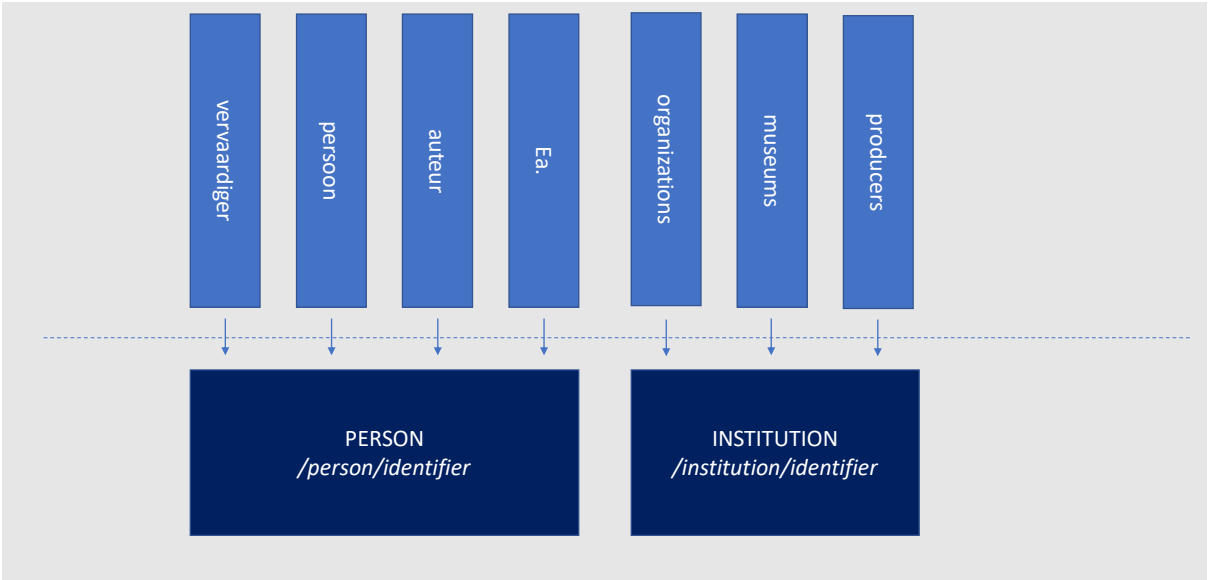

Norway
Belarus
Poland
Ukraine
Turkey
Iraq
Iran
Algeria
Saarinen, Eero
Knoll International
armstoel
1956
siermotief
1950
1901
Odetta
kan
1960
1920
Murakami, Azusa
Groves, Alexander
Studio Swine
kruk
2012
2012
Studio 65
Heller
zetel
Hayon, Jaime
b.d Ediciones de Diseno
Bosa
vaas
2006
buste
Arad, Ron
ligzetel
2002

OpenRefine

- Verrijken van data m.b.v. ext. Authorities via Reconciliation service:
  - (1) Getty (AAT (objectnamen/materialen), ULAN (personen en instellingen), TGN (geografische entiteiten)
  - (2) RKD (via Wikidata)
  - (3) VIAF (persons/instit.)
  - (4) Wikidata
    - Belang van het raadplegen van meerdere authorities in het kader van desambiguering.

Personen en Instellingen

- Normalisatie
  - Victor Horta (?) > Horta, Victor *+ precisie > vermoedelijk*
  - Horta, Victor > PT
  - Victor Horta
- Gebruik maken van de velden voornaam, achternaam en naamtoevoeging (ipv. Veld NAAM)
  - “automatische” normalisering door Adlib/Axiell Collections
    - Naam, voornaam
  - Eenvoudig voor hergebruik; reconstructie voor ontlsuiting, etc.. >> structured data
  - Maakt het mogelijk voor het zoeken op voornaam als op achternaam
  - Sleutel voor desambiguering persoon/instelling (find and replace)



Personen en Instellingen

- Traceren van duplicaten mbv OpenRefine
  - >>Merge & Cluster
    - Key Collision
    - Nearest Neighbour

Verrijking:

- Toekennen van externe identificatiecodes
  - ULAN
  - WIKIDATA (meertalige API's)
  - VIAF (persons/institutions)

  Bron: (domein)
  Nummer: (identifier)
  - Vb; http://vocab.getty.edu/page/ulan/500131469

Personen en intstellingen

- Indien niet reconcilieerbaar
  - >> minimale registratie (zo volledig mogelijk)
  - Status: partieel/minimaal

  - In **tweede beweging** deze termen toevoegen aan Wikidata en evt. donatie ULAN. >> externe identifiers toevoegen >> status: volledig.

Andere toepassingen OpenRefine

- Toekennen van auteursrechtenstatus aan de hand van GREL expressie. (+70 jaar)
- Genereren van lijsten voor gebruik publiekswerking
  - vb. verjaardagen van vervaardigers (delen op Sociale media, ea.)

OpenRefine > Adlib/Axiell?

- Mogelijk maar niet hiërarchisch
- Via adlib tagged formaat
  - Conversie: Export OpenRefine> excell > .dat (ASCII, UTF)
  - Priref als sleutel, inventarisnummer als tweede sleutel
  - Kolomnamen >> adlib tags
  - Overschrijft alles!

```
%0 134
%1 Brown's Print Shop
%2 Brown, J.
%3 DE236JX
%4 Derby
A1 131 Bonsall Avenue
BE The house was built in 1910,
   while it was destroyed by fire in 1923.
BE The house was rebuilt in 1925.
A3 England
A4 01510 - 752582
A5 PRINTERS
A9 Ron
**
```

OpenRefine

Pro's

- Reconciliation service
  - VIAF/WIKIDATA/GETTY
  - >> binnenhalen vertalingen (niet geverifieerd)
- Easy mangling, facet en normalisatie data
- Verbinden van verschillende thesauri, databanken doormiddel van gedeelde sleutel

Contra's

- No easy way back (Adlib>OpenRefine>Adlib)
- Statische lijst (veranderingen niet zichtbaar)
- Beperkt werkgeheugen
  - Niet mogelijk om met volledige exports te werken
  - >> axiell collections excells.