OpenRefine – een tool voor data-cleaning

Alina Saenko

Een van de kerntaken van een collectiebeherende instelling is informatie over objecten bijhouden en toegankelijk maken. Die data zijn een waardevolle bron van kennis. Niet gestandaardiseerde en rommelige data zijn echter zeer lastig voor zowel de eindgebruiker als voor intern gebruik.

Om de kwaliteit van de ontsluiting en vindbaarheid van een collectie te verhogen moet men aan data-cleaning doen. Dit vergemakkelijkt interne werking met data en biedt mogelijkheden voor externe publicatie en uitwisseling van gegevens. Data-cleaning is vaak arbeidsintensief, zeker als het handmatig moet uitgevoerd worden. Om de opdracht van het opschonen op een semi-automatische en dus snellere manier te laten verlopen heeft men een gespecialiseerde softwaretools nodig. Zo'n tool is OpenRefine. OpenRefine (vroeger ondersteund door Google) is een volledig opensourcesoftware waarmee men gemakkelijk grote hoeveelheden van data kan visualiseren, analyseren, manipuleren en corrigeren.

Met de reconciliation service koppelt OpenRefine kunstenaarsnamen aan VIAF-records.

DRIE TROEVEN

Het principe van werking van OpenRefine is vergelijkbaar met een Exceltabel, maar is veel geavanceerder en specifiek ontwikkeld om te werken met tekstuele waarden. Het is een alleenstaand programma dat op de computer wordt geïnstalleerd en via een browser lokaal en zonder internetverbinding kan worden gebruikt. Er wordt dus met data buiten het eigenlijke databeheersysteem gewerkt. De data worden eerst geëxporteerd uit het gebruikte systeem (in csv, xml, xls of een ander formaat) en dan opgeladen in OpenRefine. De opgeschoonde data kunnen achteraf vanuit OpenRefine terug geïmporteerd worden naar het oorspronkelijke systeem.

Wat kan je nu precies met data doen in OpenRefine? De belangrijkste mogelijkheden van dit programma kan je in drie grote groepen verdelen:

- **Analyseren van data**.
  Data worden op een overzichtelijk manier in verschillende kolommen als een tabel voorgesteld. Op elke kolom kan je specifieke filters toepassen om op verschillende manieren na te gaan welke waarden er voorkomen en of er fouten aanwezig zijn. Zo kan je een text filter gebruiken om te controleren of er in de waarden onnodige spaties of komma's zitten.
- **Transformeren van data**.
  De gevonden fouten hoeven niet één voor één gecorrigeerd te worden. OpenRefine biedt verschillende functies om in één keer grote hoeveelheden data aan te passen. Eén functie laat bijvoorbeeld toe data in een bepaalde kolom te clusteren, waarbij een lijst van spellingsvarianten worden opgesteld. Daarna kan men voor een bepaalde variant kiezen die al de andere waarden vervangt. Daarnaast kan je gebruik maken van verschillende functies (uitgedrukt in GREL expression language), die rijke mogelijkheden tot datamanipulatie biedt. Zo kan je alle overbodige leestekens op het einde van termen zoeken en meteen verwijderen. Alle stappen worden bewaard en indien nodig kan je de oorspronkelijke waarde herstellen.
- **Verrijking van de data**.
  De zogenaamde reconciliation service kan gebruikt worden om je eigen data te linken naar en te verrijken met waarden vanuit externe bronnen en standaard terminologieën. Zo kan je namen van auteurs of kunstenaars koppelen met het overeenkomstige record in VIAF (Virtual International Authority File – een internationale standaard terminologie voor personen en instellingen), 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 in het opschonen van rommelige data. Meer informatie kan je vinden op de website openrefine.org. Er bestaat ook een praktische handleiding: *Using OpenRefine*, door Ruben Verborgh en Max De Wilde (Packt Publishing, 2013).

"OPENREFINE IS BOVENDIEN EEN GEBRUIKSVRIENDELIJK PROGRAMMA DAT JE HELPT IN HET OPSCHONEN VAN ROMMELIGE DATA."

> Dit artikel werd bezorgd door PACKED vzw dat als expertisecentrum digitaal erfgoed kennis, ervaring en deskundigheid omtrent digitaal erfgoed centraliseert en verspreidt. Vragen voor PACKED vzw zijn welkom via info@packed.be.