

Een reconciler voor de erfgoedsector?

TEKST Jeroen Cortvriendt, BibliothecairErfgoed.be

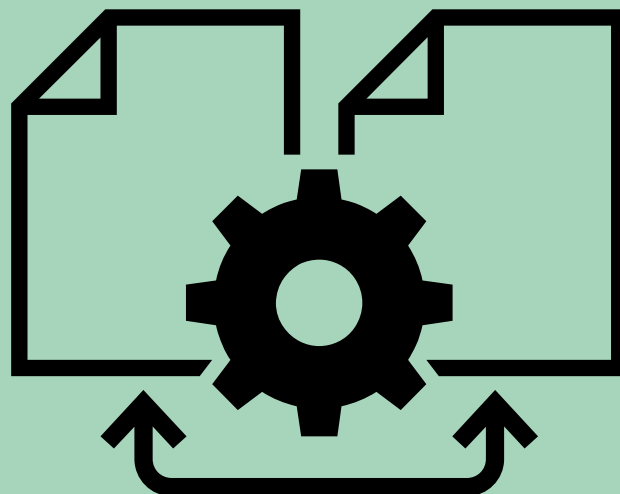
In de wereld van open data speelt interoperabiliteit een cruciale rol. Gegevens afkomstig van verschillende bronnen moeten met elkaar vergeleken kunnen worden, gecombineerd en hergebruikt. In de praktijk blijkt dat echter vaak lastig: namen van personen, trefwoorden of organisaties verschillen van dataset tot dataset. Een reconciler biedt hier een oplossing.

Een reconciler is een hulpmiddel dat helpt om gegevens uit verschillende bronnen op elkaar af te stemmen. Het vergelijkt termen, namen of identifiers uit een dataset met die in een referentiedatabank (zoals Wikidata, Virtual International Authority File (VIAF) en dergelijke) en stelt voor welke entiteiten overeenkomen. Zo kunnen 'Aristotle', 'Aristoteles' en 'Aristotelis' allemaal herkend worden als dezelfde entiteit. Dat proces – reconciliatie – vormt een essentiële stap richting goed verbonden en herbruikbare open data.

Het gebruik van een reconciler vergroot de kwaliteit en bruikbaarheid van open data aanzienlijk. Ten eerste vermindert het dubbele of inconsistente informatie, waardoor analyses betrouwbaarder worden. Ten tweede maakt het datasets semantisch rijker: door entiteiten te koppelen aan gestandaardiseerde identifiers, worden gegevens machineleesbaar en beter integreerbaar in het web van gelinkte data.

Bovendien draagt reconciliatie bij aan duurzame data-ecosystemen. Organisaties hoeven niet telkens opnieuw handmatig verbanden te leggen tussen datasets: eenmaal data gelinkt zijn, kan er eenvoudiger geüpdatet en gedeeld worden. Dat bevordert samenwerking tussen overheden, onderzoeksinstellingen en burgers, en stimuleert innovatie op basis van open data.

Een reconciler moet dus een probleem oplossen: daar verschillende instellingen verschillende termen gebruiken, is het niet zonder meer duidelijk wie naar wat refereert. Een echt geïntegreerd termennetwerk voor de cultureel-erfgoedsector zou een structurelere oplossing bieden. Dat zou de vindbaarheid over de deelsectoren (musea, archieven en bibliotheken) heen ten goede komen. De haalbaarheid is zowel conceptueel (is één bestand denkbaar?) als technisch (data-conversie?) niet evident. Uiteraard komt zo'n systeem met een beheerskost.



Daarom ontwikkelde BibliothecairErfgoed.be zelf een reconciler (services.vebhosting.net/reconciler). De software berekent de Levenshteinratio van de zoekvraag ten opzichte van al de termen in de referentiedatabank. Dat betekent dat er nagegaan wordt hoeveel bewerkingen er nodig zijn om de databaseterm in de zoekvraag te veranderen, rekening houdend met de lengte van de zoekterm. Hoe dichter bij 1, hoe beter. Zo is de afstand tussen water en pater 1 (enkel de w moet ingeruild worden voor een p), en er zijn vijf letters, wat betekent dat de ratio 0.8 is.

Hoewel er een interface beschikbaar is, is het allereerst de bedoeling dat de JSON-output (JavaScript Object Notation) rechtstreeks bevestigd wordt via de Application Programming Interface (API). Daarnaast is er ook de mogelijkheid om te zoeken op externe ID's: je zoekt bijvoorbeeld op VIAF 7524651 en je krijgt de ID's die BibliothecairErfgoed.be in zijn projecten hanteert.

Deze software is opgevat als een *Minimal Proof of Concept*. Zeker aan de performantie kan worden gewerkt. Wie wil meewerken aan een betere versie, mag contact opnemen met jeroen@bibliothecairerfgoed.be. ■