# Plan van aanpak analyse Axiell thesauri RCE

Binnen RCE collecties Bibliotheek en Archieven en Kunstcollecties zijn de meeste thesaurustermen binnen Axiell Collections nog niet gestandaardiseerd en verrijkt met URI’s van gedeelde thesauri (CHT, AAT, etc.). Hierdoor zijn objecten niet goed vindbaar, slecht verwijsbaar en daardoor niet te koppelen met objecten uit andere collecties. Zoals aanbevolen in de DERA moeten bronhouders ervoor zorgen dat objecten in digitale collecties verwijzen naar gedeelde termen en daarbij horende URI’s (webadressen), zodat erfgoeddata gebruiksvriendelijker wordt.

Voor nieuwe termen kan in de toekomst binnen Axiell Collections het Termenentwerk worden aangeroepen, wat dit proces vereenvoudigt. Voor al bestaande termen is dit geen oplossing, want dat vereist het handmatig selecteren van de juiste URI voor meer dan honderdduizend termen. Bovendien moeten hiervoor eerst de thesauri in Axiell Collections worden opgeschoond: termen moeten zoveel mogelijk gestandaardiseerd worden voordat ze kunnen worden voorzien van URI’s van gedeelde thesauri.

Om te kunnen inschatten hoeveel werk het is om onze Axiell thesauri op te schonen en uiteindelijk te voorzien van URI’s – en hoeveel hiervan geautomatiseerd kan worden – dient deze analyse.

*Beperkingen*

Voor het geautomatiseerd standaardiseren en toevoegen van URI’s aan termen binnen Axiell Collections gelden twee belangrijke beperkingen:

1. Zolang er binnen de thesaurus termen zijn die of identiek zijn ofwel hetzelfde betekenen maar anders zijn geschreven (denk aan ‘textiel’ en ‘Textiel’), dan moeten deze eerst worden samengevoegd binnen Axiell. Dit kan vooralsnog niet automatisch omdat we dan de koppeling met objecten kwijtraken.
   * Wat we wel kunnen doen is analyseren om hoeveel termen het hier gaat en hoeveel objecten hieraan gekoppeld zijn.
   * Eventueel zou je aan beide spellingsvarianten van textiel een URI kunnen toevoegen en hopen dat Axiell in de toekomst thesaurustermen en daaraan gekoppelde objecten kan samenvoegen a.d.h.v. identieke URI’s. Maar dit zorgt op dit moment niet voor een geharmoniseerde thesaurus.
2. Gedeelde termen (‘textiel, knoopwerk’) kunnen programmatisch wel worden gescheiden, maar (nog) niet als separate termen worden teruggeladen in Axiell. Bovendien spelen hier inhoudelijke keuzes: welke term dient in elk geval te worden verkozen boven de andere? Dat is een keuze die niet geautomatiseerd te maken is.

* Wat we wel kunnen doen is analyseren om hoeveel termen het hier gaat en hoeveel objecten hieraan gekoppeld zijn.

Wat betreft punt 1 zou ingezet kunnen worden op een technische oplossing (vanuit de leverancier), want technisch zou dit geen grote uitdaging moeten zijn – het is alleen iets dat wij als gebruiker niet zelf kunnen. Wat betreft punt 2 moet er idealiter vanuit collectiebeheer een (handmatig) werkproces opgezet worden om dergelijke termen systematisch te evalueren en te wijzigen binnen Axiell Collections, om daarna URI’s toe te kunnen voegen.

*Analyse*

Uitgangspunt van deze analyse is om voor de interne thesauri van B&AC en KC te onderzoeken:

1. hoeveel overeenkomstige termen er voorkomen;
2. hoeveel gedeelde termen er voorkomen;
3. hoeveel termen er overblijven die wel geautomatiseerd gestandaardiseerd kunnen worden (unieke, niet-gedeelde termen);
4. om hoeveel koppelingen met objecten het gaat voor ieder punt.

Bibliotheek en Archief Collecties

Het totaal aantal termen in Axiell is 106.984, waarvan 102.466 uniek. Deze zijn onderverdeeld in verschillende soorten termen (top-10). Sommige termen komen voor onder meerdere soorten.

*Soorten termen (top-10)*

|  |  |
| --- | --- |
| **Soort term** | **Aantal termen** |
| geografisch trefwoord | 47136 |
| onderwerp | 28193 |
| plaats | 11483 |
| classificatie | 4085 |
| objectnaam | 3968 |
| geografisch trefwoord#plaats | 2871 |
| geografisch trefwoord#plaats#onderwerp | 1234 |
| geografisch trefwoord#onderwerp | 905 |
| gemeente | 618 |
| plaats#geografisch trefwoord | 598 |

*Soorten termen en aantal objecten per soort (top-10)*

|  |  |
| --- | --- |
| **Soort term** | **Aantal objecten** |
| onderwerp | 655040 |
| rechten#onderwerp | 620318 |
| geografisch trefwoord#gemeente#plaats#filiaal#onderwerp | 599995 |
| rol#onderwerp#beroep | 564470 |
| collectie | 438585 |
| geografisch trefwoord#land#plaats | 420021 |
| techniek | 394354 |
| uitleencategorie#materiaalsoort#onderwerp#objectnaam#doelgroep#verwervingsmethode | 394144 |
| techniek#rol#objectnaam#onderwerp | 249029 |
| onderwerp#objectnaam | 217270 |

De rechtenthesaurus kent maar één term: ‘CC-BY’. De plaats-thesaurus staat grotendeels los van de geografische thesaurus: van deze termen komen er slechts 117 voor in de geografische thesaurus en moet dus apart worden behandeld.

*Ongebruikte termen*

Om te beginnen kunnen we thesaurustermen verwijderen die geheel niet gebruikt worden binnen Axiell Collections B&AC: dit zijn 12,166 termen, opgeslagen in bestand thesaurus\_zero\_use.csv.

*Dubbele termen*

EERST DUPLICATED TERMS ERUIT > 494 see thesaurus\_onderwerp\_dubbel.csv

*Gedeelde termen – onderwerp*

Voor nu richten we ons op de onderwerp en geografische thesaurus, omdat deze zowel qua termen als koppelingen met objecten het meest numeriek zijn.

Gedeelde termen worden binnen onderwerp aangegeven met een “;”. Het gaat om 595 termen, opgeslagen in bestand thesaurus\_onderwerp\_gedeeld.csv.

*Overeenkomstige termen – onderwerp*

Csv thesaurus\_onderwerp\_matchingset.csv

Beginnen met 22,939 termen. Uitvinden of ze uniek zijn doen we d.m.v. fuzzy matching: als een x-percentage van een term overeenkomt, classificeren we de term als niet uniek. Deze parameter zetten we vooralsnog op 85%.

NB: er zijn nu twee vergelijkingen: een met het fuzzywuzzy algoritme en 1 met levenshtein. De eerste is ruimer en vindt dus meer overeenkomsten (false positives: “Bierbrouwerij op NOM-terrein VS Boerderij op NOM-terrein”), de andere zal deze termen niet vinden maar daardoor ook potentieel dezelfde termen missen (false negatives).

Eerste match methode: thesaurus\_onderwerp\_matches\_v2.csv

Tweede match methode: thesaurus\_onderwerp\_matches\_v2.csv

Volgende stap: wellicht de levenshtein gebruiken maar op 80% zetten. Dan wordt vermoedelijk wel het fuzzy voorbeeld van hierboven gevonden, maar heb je minder false positives dan fuzzy matching op 85%.

Uitkomst = zie twee csv bestanden, maar komt neer (methode 1) op bijna 17k van de bijna 23k terms die vermoedelijk uniek zijn. Wat uitkomt op 51% van de objecten die we op die manier via Openrefine van een URI zouden kunnen voorzien (als die term bestaat natuurlijk).