Rolf Blijleven

IT-diensten voor musea

M 06 12 872 892

E rolf@rolfblijleven.nl

Efficiënt (??) thesaurusbeheer

Over data-vervuiling en data schonen

Een workshop met discussiepunten

Een paar ~~problemen~~ Uitdagingen!! op een rijtje

- Bibliotheek en Museum delen de Thesaurus en Personen & instellingen
  - Over auteursvermelding is (meestal) consensus. Over vervaardigersvermelding niet.
  - Contactpersoon-gegevens worden anders toegepast dan vervaardiger-gegevens, maar moeten wel uniform worden ingevoerd.
- Datum-formaten
- Vrijwilligers, personeelsverloop, voortschrijdend inzicht
- Logistieke operaties
  - “we misbruiken dit veld even voor een oormerk, OK?”
- Spelfouten op een website?
  - geen halszaak, wel onnozel

Over forceren

"Forceren" in Adlib is: een gevalideerde term toevoegen bij de invoer van stam-records.

Mag een vrijwilliger die invoert dat?

Mag een nieuwe vaste medewerker dat?

Indien niet, bij wie moet hij/zij wezen?

Indien wel, wie controleert de invoer?

Kandidaat-termen?

- Maatwerk: nieuwe thesaurus-invoer automatisch selecteerbaar

Rollen en rechten?

- Houdt het a.u.b. SIMPEL!

de uitdaging

"Wiebelige data"

- Vincent de Keijzer, Haags Gemeentemuseum

Dat zal niet veranderen

Calimero-gedrag helpt (meestal) niet

Democratisch thesaurusbeheer, werkt dat?

taalvaardige despoten gevraagd?

Data-vervuiling is mensenwerk

Data opschonen is (deels) ook mensenwerk

Voor wie er vanochtend niet bij was

RESUMÉ

FEEDBACK-LINKS

ZOEKEN EN VERVANGEN

Wat zijn feedback-links?

In “standaard” CBF Adlib is alleen een heen-verwijzing.

Geen terugverwijzing.

Verwijder je people-record 0003,

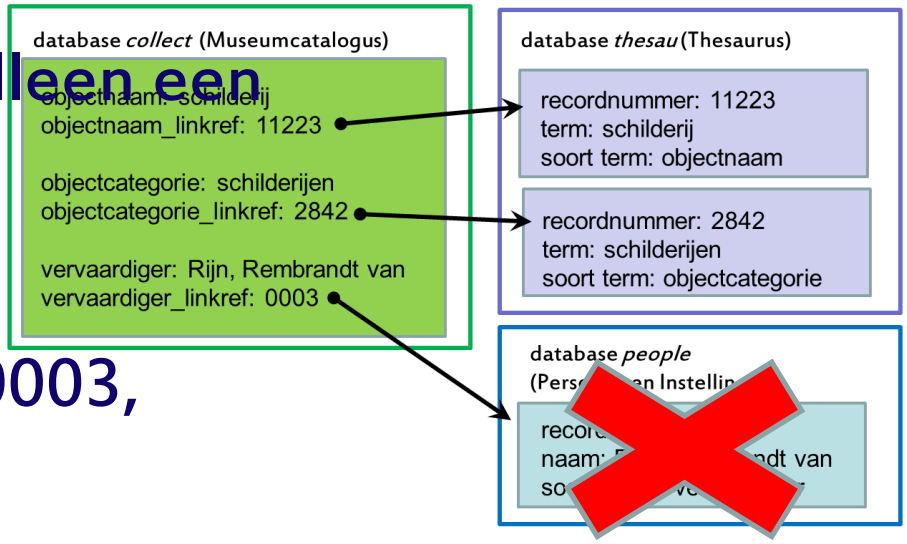
dan krijg je

- geen waarschuwing

  “record is in gebruik”

- wel een dode link

  “datacorruptie”



Wat zijn feedbacklinks? (2)

De meeste SQL-versies en sommige CBF-versies hebben *feedback-links*.

(Stap over naar SQL)

Er is dan een heenverwijzing én een terugverwijzing.

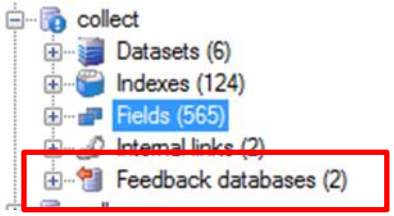
niet elke Adlib heeft dit, vanwege 'performance'

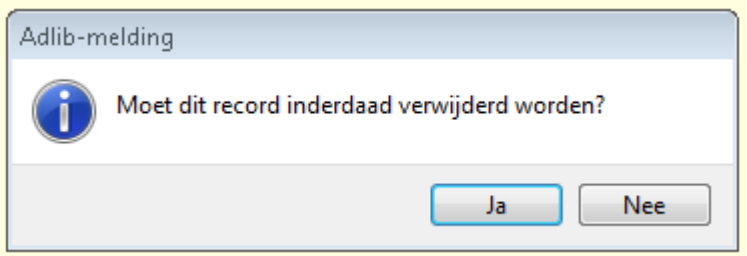
record opslaan vergt meer administratie

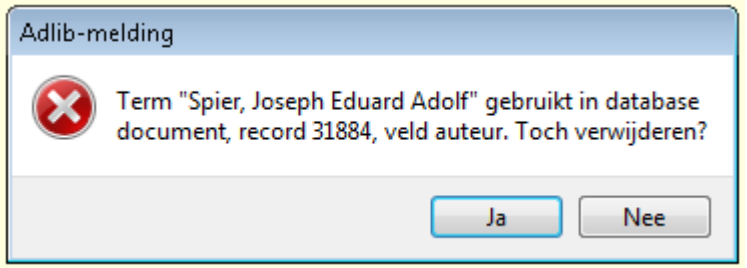
Je krijgt wél een waarschuwing als het record in gebruik is. Lees die!

Hoe weet ik van tevoren of ik feedback-links heb?

Kijk in Designer







Wat als ik geen feedbacklinks heb?

Hoe kun je nagaan of het record in gebruik is?

Zoek in de selectietaal met record contains ..

.. in de volledige museumcatalogus

.. en in de volledige boekencatalogus

Vind je de term daar niet,

dan is-ie misschien nog wel in gebruik in loans, exhibit, enzovoort enzovoort...

Overweeg overstappen naar SQL..

..want feedbacklinks zijn onmisbaar bij schonen.

Schonen van losse records

Benut de mogelijkheden:

- Gebruik
- Gebruikt voor
- Equivalente term

“zelfreinigend”

kies je bij invoeren een niet-voorkeursterm, dan neemt Adlib automatisch de juiste voorkeursterm

Schonen ‘in bulk’ met Zoeken en Vervangen

Algemene werkwijze

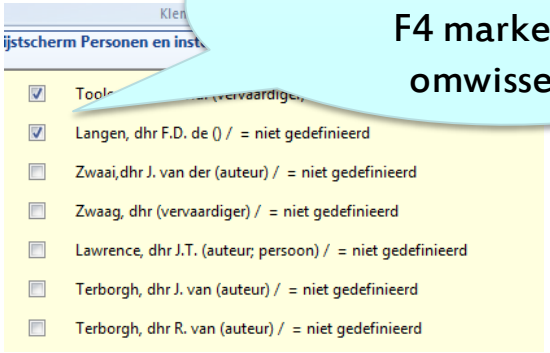
1. Zoek op: de (te schonen) veldnaam ‘onder water’
2. In welke database woont dat veld?
3. Selecteer in die database de te wijzigen records
4. Test je zoek-en-vervang-actie in 1 of 2 records

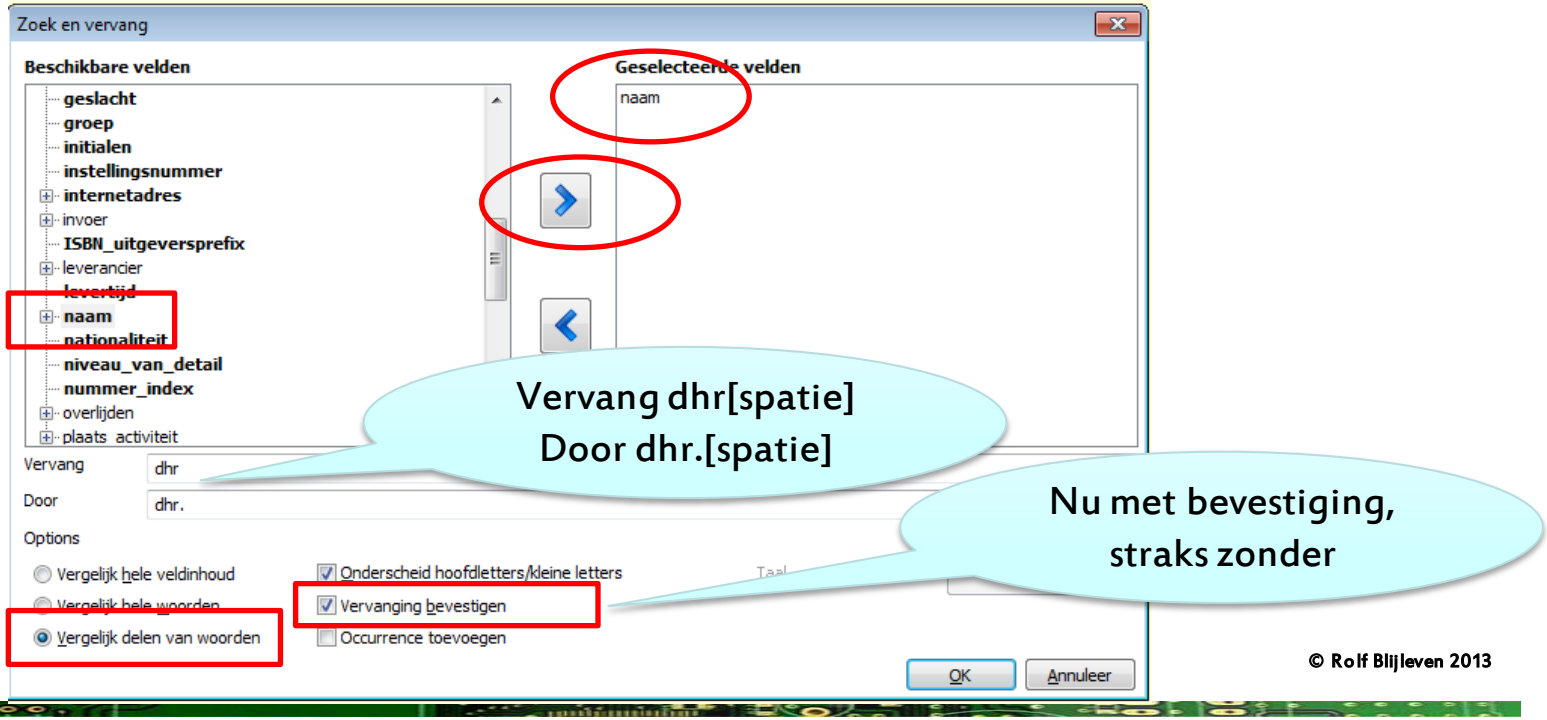
   met ‘bevestigen’ AAN
5. Draai selectie om (F4)en zoek-en-vervang de rest

   met ‘bevestigen’ UIT

Oefening: is “dhr” overal goed ingevoerd?

1. In Personen en instellingen, zoek alle records die wel “dhr” in de naam hebben maar niet “dhr.”.
   Gebruik ‘contains’ of ‘_’ (niet =).
2. Vervang alle “dhr “ en “dhr,“ door “dhr.”





Zoeken en vervangen

Gebruik ‘delen van woorden’ voor een *vaste* tekenreeks *korter* dan de hele veldinhoud

- Vervang [iets] door [iets anders]
  - o.a. om gewijzigde padnamen aan te passen
  - Yabba [dibi] doe ⇨ Yabba [dabba] doe
- Vervang [iets] door [niets]
  - Yabba [dibi] dabba doe ⇨ Yabba dabba doe

Vervang * om elke willekeurige *volledige veldinhoud* compleet te vervangen door iets anders

- Niet mogelijk: iets nieuws voor of achter een willekeurige veldinhoud plakken
- * ⇨ *[afstoten] werkt niet

Vervang “” om een leeg veld in een serie records te vullen

Versie 7: occurrence toevoegen

- Voorbeeld: alle Verwervingsbronnen in één klap tot Persoon bombarderen

Adapl als opschoon-hulp

Praktijkgeval: “het is niet presentabel genoeg voor de website”

- (deze twee voorbeelden hebben overigens niets met Thesaurusbeheer te maken)

records met 1e plaatje-occurrence blanco

- oorzaak: neveneffect van foto-import
- vaststellen welke records:
  - selectietaal, pointerfiles
  - reconstrueren wat er gebeurd is
  - evt een adapl om records met blanco occ. te vinden

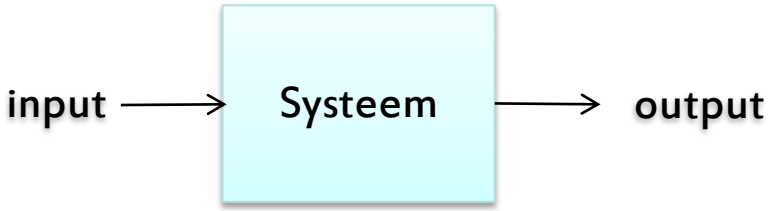
```
FN[i] = ''
```

- 350 lege occurrences verwijderen ~ 1 uur handwerk
  - Automatisch schonen is lang niet altijd de aangewezen oplossing!

lege vervaardiger-occurrences

- met opschoon-adapl die op een pointerfile werkt.
  - als occurrence X bij alle leden van de veldgroep Vervaardiger leeg is, wis die occurrence

Export - Schonen - Import



Wat exporteren?
In welk formaat?
XML, CSV of
Adlib Tagged?

Wat voor Opschoner?
Waar moet ik op letten?

Hoe importeren?
Met Import.exe
of Designer
of Adlwin?

Strategie

1. Adlib-data is meerdimensionaal, maar normale mensen werken liever met platte data.
   - Werk veld voor veld in selecties records
     - niet heel record schonen, next, etc
   - Houdt altijd de priref bij je record-data
2. “Een import moet HERHAALBAAR zijn. Liefst met één druk op de knop.”
   - Houdt het controleerbaar en overzichtelijk.
   - Het zal wel eens misgaan. Zorg dat je altijd een vluchtweg hebt: maak reservekopiën en herstel-importjobs (Designer).

export tagged - editor - import tagged

Stel, tig objecten zijn van standplaats Y naar standplaats Y-depot verplaatst.

Aanpak, grofweg:

1. Selecteer je records.
2. Exporteer Priref en Huidige standplaats naar Adlib Tagged.
3. Met je editor: wijzig tag en voeg depot toe.
4. Importeer.

export tagged - editor - import tagged (2)

Dezelfde aanpak, in detail:

1. Selecteer je records.
    1. Huidige Standplaats = * AND not Huidige standplaats = -
    2. Tab Standplaats | Toekomstige verplaatsingen:
        1. hoe heet Huidige Standplaats onder water?

           wijzig record, rechtsklik in Huidige Standplaats, Tab Data Dictionary, 2e is Nederlands
        2. hoe heet het tag voor Toekomstige Standplaats?

           2L
2. Exporteer Priref en Huidige standplaats naar Adlib Tagged.
3. Met je editor: wijzig tag en voeg depot toe.
    1. In (bijvoorbeeld) Notepad++
        1. wijzig 2A in 2L
        2. wijzig CR-LF** in –depotCR-LF**
4. Importeer er eerst twee, ter controle
5. Importeer dan de rest.

Editors, Editors-mét en verder..

Belangrijk van een editor:

- Ondersteuning van diacriten en ligaturen (UTF-8)
- Ondersteuning van controle-karakters (CR, LF, etc)
- Ondersteuning van Regular Expression
- Notepad++ (notepad-plus-plus.org) is open source (gratis) donation ware, maar er zijn vele andere mogelijkheden

Regular Expression

- Een mini-taal voor patroonherkenning
- Oogt héél cryptisch en dat is wennen
- is héél krachtig, maar niet zaligmakend

Zelfgemaakte Opschoners

- Herhaalbaar: maak een script van je opschoon-actie
- Python: Open Source, draait onder interpreter, volwassen, leuker dan C# en C++, veel krachtiger dan Adapl

Praktijk: contactperSchonen

- Selectie: People- database, Naam = * and do = PERSON
- Python-script herkent 25 patronen in de vorm

  Achternaam, [titels] [Voornaam][Initialen][tussenvoegsels]
- Splitst uit naar desbetreffende tags voor import in Adlib

  zodat P&I-gegevens ook te gebruiken zijn voor brieven, mailings e.d.

  voorbeeld:

  Mondriaan, Piet ⇨ AN Mondriaan; VN Piet

  Diepenhorst, Mr. Isaac A. ⇨ AN Diepenhorst; VN Isaac; sn I.A.; sg Mr.
- Levert output klaar voor import in tagged-formaat
- Levert dezelfde output ter controle in CSV-formaat->Excel
- idem voor non-matches, ter correctie (in Excel)

  namen hebben altijd weer andere patronen
- Getallen:
  - 25 patronen gebaseerd op 766 naam-records: 80% herkenning
  - losgelaten op ~10.000 namen: 5500 automatisch gesplitst

Screenshots contactperSchonen

De input: namen in de vorm achternaam, [titels] [voornaam] [initialen] [tussenvoegsel]

```
Meer-Mohr, P. van der
Klein Essing, Coen (Josef)
Klein Essing, Coen van
Lely, dr. ir C.
Minnaert, prof. dr. ir. Marcel
Kleijne, prof. mr. Isaac de
Groeningen, de heer drs. J.J.W.
Klein Essing, dr. ir. C.P.
Koekoek, mevrouw L. de
Dijk, Anthony Q.L.M. van
Brantsen-van Hasselt, Maria Lec
Maas-van der Moer, A.
Schampers, de heer dr. Karel
```

output 1: gesplitste namen in Adlib tagged-formaat, klaar om te importeren

```
**
%0 1415
BA Klaassen, Nel
AN Klaassen
VN Nel
**
%0 1416
BA Goedeljee, C.R.
AN Goedeljee
sf C.R.
**
%0 1417
BA Kantelberg, Peter
AN Kantelberg
VN Peter
**
%0 1429
BA Lambers, Mr. B.J.
AN Lambers
sf B.J.
sg Mr.
**
%0 1430
BA Van Lier, Amsterdam
AN Van Lier
VN Amsterdam
**
```

output 2: non-match in Excel-formaat voor handmatig schonen

A	
0%	BA
1	Gemeentemusea Arnhem
2	Anoniem
6	Veiling Gouda Quint, Arnhem
8	erven van mr. W.E.J. baron van Balveren
9	Mollerus-van Eck, baronesse
13	Genootschap voor Oudheden, Arnhem
18	J.C.J. baron Brantsen
21	Veiling Sotheby Mak van Waay

Screenshots contactperSchonen (2)

Output 3: Matches in excel. Inhoudelijk hetzelfde als Adlib Tagged output 1, maar makkelijker door mensen te controleren.

1	BA	TV	(	AN	sg	sf	VN
528	Lambers, Mr. B.J.		1429	Lambers	Mr.	B.J.	
073	Verrijp, dr. C.D.		2696	Verrijp	dr.	C.D.	
278	Scheidius, Mr. E.P.A.M.		3439	Scheidius	Mr.	E.P.A.M.	
529	Staring, mr. A.		4169	Staring	mr.	A.	
229	Herfst, mr. H.P.		10918	Herfst	mr.	H.P.	
375	Akkerboom, mr. ir. B.P.J.		12902	Akkerboom	mr. ir.	B.P.J.	
536	Verhuell, mr. A.W.M.C.		13206	Verhuell	mr.	A.W.M.C.	
537	Bloemers, Mr J.H.F.		13214	Bloemers	Mr	J.H.F.	
908	Brouwer, drs M.		13894	Brouwer	drs	M.	
909	Schoon, drs P.J.		13895	Schoon	drs	P.J.	
042	Groeningen, de heer drs. J.J.W.P. van	van	14076	Groeningen	drs.	J.J.W.P.	
063	Camps, de heer drs. R.		14109	Camps	drs.	R.	
064	Lansberg, Dr. M.P.		14110	Lansberg	Dr.	M.P.	
104	Boezer, dr. R.E.K.		14180	Boezer	dr.	R.E.K.	
250	Asser, Mr. R.W.		14415	Asser	Mr.	R.W.	
262	Schellingerhout, mr. M		14430	Schellingerhout	mr.	M	
352	Maresch, dr. J.		14577	Maresch	dr.	J.	
430	Zweite, Prof. Dr. A.		14753	Zweite	Prof. Dr.	A.	

..nog iets uit de praktijk..

30.000 trefwoorden te controleren van een nieuw ingevaren deelcollectie?

- AAT staat online.
- binnen domein staat je Thesaurus online via Adlib API

Maak een apart bestand van die trefwoorden

Zoek, geautomatiseerd met een script alle trefwoorden in dat bestand, op in je thesaurus via API en in AAT-online

Het script stuurt steeds een URL naar de online-AAT en naar je API:

`http://service.aat-ned.nl/api/wwwopac.ashx?&database=aat-xml&search=te=`**`schilderijen`**
`http://service.aat-ned.nl/api/wwwopac.ashx?&database=aat-xml&search=te=`**`schilderij`**
`http://service.aat-ned.nl/api/wwwopac.ashx?&database=aat-xml&search=te=`**`schoeisel`**
`http://service.aat-ned.nl/api/wwwopac.ashx?&database=aat-xml&search=te=`**`schoen`**

De API en AAT-online geven XML terug.

Het script ziet of het trefwoord bestaat of niet.

- Als 't bestaat
  - OK, koosjer trefwoord -> sla op met priref in Adlib tagged & importeer
- Als 't niet bestaat moet je 't nader bekijken
  - sla op in CSV, zet in Excel en schoon de rest handmatig

toekomst.. *dromen?*

Raad voor de Cultuur:

“Grote musea moeten kleine musea helpen”

De Raad doet het voorstel dat de minst door bezuiniging getroffen musea een grote verantwoordelijkheid hebben voor de minder sterke musea (het delen van kennis, vaardigheden en netwerken). De instellingen dienen zich te richten op duurzame samenwerkingsvormen.

bron: Samenvatting Advies Bezuiniging Cultuur 2013 – 2016, p 28.

Meer hulpbronnen gewenst (zoals AAT-online)

- Veel Erfgoed is weinig kunst en veel Cultuur
- Bibliotheken en musea hadden weinig aan de AAT

NOM werkt al aan uitbreiding van de AAT-online. Streekmusea kunnen daarvan profiteren

- Rijkmuseum-thesaurus online?
- En zo voort?

‘t zou eenvoudig kunnen m.b.v. (onder andere) de Adlib API...

Wordt hier misschien al aan gewerkt?

Hoe dan ook..

Je kunt veel opschoonwerk automatiseren..

..maar controle en correctie door mensen blijft altijd nodig.

Vragen?

dank u voor uw aandacht