

SAS Club 2023

Der Business Analytics Club für SAS User

Wien, SAS Office Trabrennstraße
19. Oktober 2023

Gerhard Svolba, Phillip Manschek, Jens-Ole Harden,
Michael Weberberger (Premedia), Florian Stammer



Agenda

14:15 - 14:20 Uhr	Begrüßung / Intro / News Gerhard Svolba, SAS
14:20 - 14:50 Uhr	Es geht auch anders! - Erstellung analytischer Modelle mit SAS Viya Gerhard Svolba, SAS
14:50 - 15:20 Uhr	SAS und Generative AI - Überblick, Entwicklungen und Anwendungsbeispiele aus dem Marketing Michael Weberberger, Premedia // Florian Stammer & Gerhard Svolba, SAS
15:20 - 15:35 Uhr	Die SAS Explore Konferenz in Las Vegas - Ein Vor-Ort Bericht Gerhard Svolba, SAS
15:35 - 15:55 Uhr	PAUSE
15:55 - 16:25 Uhr	Fuzzy Matching von Steuernummern in externen Datenquellen mit SAS Mihai Paunescu, Bundesministerium für Finanzen
16:25 - 16:50 Uhr	SAS Studio Analyst und die Erweiterungsmöglichkeiten mit Custom Steps Phillip Manschek, SAS
16:50 - 17:15 Uhr	SAS Tipps und Tricks Session Jens Ole Harden, SAS
ab 17:15 Uhr	Gemütliches Get-Together mit Buffet



Fuzzy Matching von Steuernummern in externen Datenquellen mit SAS

Mihai Paunescu, Bundesministerium für Finanzen



Fuzzy Matching von Steuernummern in externen Datenquellen mit SAS

Wien, 19.10.2023

Fragestellung

- Matching von externen Daten (natürliche Personen und Organisationen) zur Steuernummer

Externe Daten: z.B. im Firmenbuch für einen Geschäftsführer oder Gesellschafter

Name	Vorname	Geburtsdatum	Adresse
Svolba	Gerhard	01.07.1979	ERZHERZOG-KARL-STRASSE 17/4/21, 1220 Wien

Steuerdaten des BMF

Name	Vorname	Geburtsdatum	Adresse
Svolba	Gerhard	01.07.1979	Erzherzog-Karl-Str. 17/21, 1220 Wien
Svolba	Gerhart	01.07.1979	<u>Simmeringer Hauptstr.178/Top3 1110, 1110 Wien</u>
<u>Svoboda</u>	Gerhard	01.07.1979	<u>Erzh.-Karlst. 17, 1220 Wien</u>
Svolba	Gerhard	07.01.1979	Erzherzog-Karl-Strasse 17, 8055 Graz
Svolba	Gerhard	03.09.1998	Erzherz.Karl st 17/4/21, 1220 Wien

Technische Zielsetzung

- Beispiel 1: Natürliche Personen als Geschäftsführer und Gesellschafter im Firmenbuch sollen um Steuernummer ergänzt werden.
- Ableitungen:
 - Bei wie vielen Firmen ist/war eine Person Geschäftsführer?
 - Wie viele Firmen bei der eine Person Geschäftsführer war sind in Insolvenz gegangen?
- Beispiel 2: Abgleich von Meldungen von ausländischen Banken zu Zinsen und Dividenden mit der Steuererklärung

Informations-Elemente

Matching-Ziel: Steuernummer

- mc_nachname95 (Matching Code Nachname - high sensitivity)
- mc_vorname95 (Matching Code Vorname high sensitivity)
- mc_vorname60 (Matching Code Vorname very low sensitivity)
- Geburtsdatum
- geb_tag_mon (TAG.Monat Format Geburtsdatum)
- geburt_md (Geburtsdatum mit Monat and Tag vertauscht)
- mc_adresse75 (Matchcode für bereinigte Adresse – medium sensitivity)
- mc_stadt80 (Matchcode für Stadtname medium sensitivity)
- PLZ
- Telefonnummer
- Email
- IBAN

DQ MATCH - Funktionen

SAS® Help Center

Customer Support | SAS Documentation

SAS® Viya® Platform Programming Documentation | 2023.09

PDF | EPUB | Feedback

DQEXTTOKENPUT Function

DQGENDER Function

DQGENDERINFOGET Function

DQGENDERPARSED Function

DQIDENTIFY Function

DQIDENTIFYIDGET Function

DQIDENTIFYINFOGET Function

DQIDENTIFYMULTI Function

DQLOCALEGUESS Function

DQLOCALEINFOGET Function

DQLOCALEINFOLIST Function

DQLOCALESCORE Function

DQMATCH Function

DQMATCHINFOGET Function

DQMATCHPARSED Function

DQOPTSURFACE Function

DQPARSE Function

CALL DQPARSE Routine

DQPARSEINFOGET Function

DQPARSEINPUTLEN Function

DQPARSERESLIMIT Function

DQPARSESCORDEPTH Function

DQPARSETOKENGET Function

DQPARSETOKENPUT Function

DQPATTERN Function

DQSCHEMEAPPLY Function

CALL DQSCHEMEAPPLY Routine

DQSTANDARDIZE Function

SAS Data Quality Language Reference

DQMATCHPARSED Function

Returns a match code from a delimited string of parse token values.

Valid in: SAS Viya CAS, SAS Viya Compute Server, SAS Viya DS2 CAS, SAS Viya DS2 Compute Server, and SAS 9. Also valid in PROC SQL and SAS Component Language.

Table of Contents

Syntax

Required Arguments

Optional Arguments

Example

See Also

Syntax

DQMATCHPARSED(*delimited-string*, '*match-definition*' < ,*sensitivity* > < ,'*locale*' >)

Required Arguments

delimited-string

specifies the input variable, which must have a value that is a delimited string of parse token values. The delimited string must have been generated with a parse definition that is associated with the *match-definition* in the *locale*.

match-definition

specifies the name of the match definition that is referenced to generate the return value. The definition must be supported by the *locale*.

Optional Arguments


```
%dqload(DQLocale=(&locale.), DQSETUPLOC="E:\SAS\sashome\SASQualityKnowledgeBase\CI\28"); /*Lade deutsche Knowledge Base*/







%dqputloc(DEDEU); /*Zeige parsing definitionen verfügbar für deutschen locale*/

data test;
  length name $60;
  name='Svolba, Gerhard'; output;
  name='Svolba Gerhard'; output;
  name='Gerhard Svolba'; output;
  name='Gerhard Michael Svolba'; output;
  name='Paunescu Mihai'; output;
  name='Dr. Mihai Paunescu'; output;
run;

data test2;
  set test;

  parsedname=dqParse(name, 'NAME', 'DEDEU'); /*Zerlege string gemäß Parsing Defintion "Name" in seine Tokens: Prefix, Vorname, Nachname, Suffix */
  Vorname=dqParseTokenGet(parsedname, 'Given Name', 'NAME', 'DEDEU'); /*Extrahiere Vorname */
  Nachname=dqParseTokenGet(parsedname, 'Family Name', 'NAME', 'DEDEU'); /*Extrahiere Nachname */
  MC_NAME95 = dqmatchparsed(parsedname,"Name",95,"DEDEU"); /*Generiere Matchcode für den gesamten Namen mit einer Sensitivität von 95*/
  MC_NAME70 = dqmatchparsed(parsedname,"Name",70,"DEDEU");

run;
```

 name	 parsedname	 Vorname	 Nachname	 MC_NAME95	 MC_NAME70
Svolba, Gerhard	/=//=/Gerhard/=/Svolba/=//=/	Gerhard	Svolba	4L@WM&\$\$\$\$\$F_Y&Y~\$\$\$	4L@WM&\$\$\$\$\$F_Y&\$\$\$\$\$
Svolba Gerhard	/=//=/Gerhard/=/Svolba/=//=/	Gerhard	Svolba	4L@WM&\$\$\$\$\$F_Y&Y~\$\$\$	4L@WM&\$\$\$\$\$F_Y&\$\$\$\$\$
Gerhard Svolba	/=//=/Gerhard/=/Svolba/=//=/	Gerhard	Svolba	4L@WM&\$\$\$\$\$F_Y&Y~\$\$\$	4L@WM&\$\$\$\$\$F_Y&\$\$\$\$\$
Gerhard Michael Svo...	/=//=/Gerhard Michael/=/Svolb...	Gerhard Mich...	Svolba	4L@WM&\$\$\$\$\$F_Y&Y~\$B\$	4L@WM&\$\$\$\$\$F_Y&\$\$\$\$\$
Paunescu Mihai	/=//=/Mihai/=/Paunescu/=//=/	Mihai	Paunescu	N&#P_43#\$\$\$\$B7_7\$\$\$\$\$	N&#P_43\$\$\$\$B7_7\$\$\$\$\$
Dr. Mihai Paunescu	/=/Dr./=/Mihai/=/Paunescu/=//...	Mihai	Paunescu	N&#P_43#\$\$\$\$B7_7\$\$\$\$\$	N&#P_43\$\$\$\$B7_7\$\$\$\$\$

Matchcodes für Namen mit unterschiedlichen Sensitivitäten

MC_VORNAME95	MC_VORNAME75	MC_VORNAME60
F_Y&Y~	F_Y&Y	F_
GERHARD	GERHARD	GHEORGHE
GERARD	GERARD	GERTRUDE
GERHART	GERHART	GEORG
	GERARDO	GEORGE
	GERARDUS	GERDA
		GERLINDE
		GERALD
		GERNOT

Strasse	Matchcode medium-low Sensitivity (70)
KUBIN-PLATZ	3#M7PNW&~
KUBIN-PL	3#M7PNW&~
KUBIN-PL.	3#M7PNW&~
KUBINPL	3#M7PNW&~
KUBINPL.	3#M7PNW&~
A-KUBIN-PLATZ	&3#M7PNW
A-KUBIN-PL	&3#M7PNW
A-KUBIN-PL.	&3#M7PNW
A-KUBINPL	&3#M7PNW
A-KUBINPL.	&3#M7PNW
A.-KUBIN-PLATZ	&3#M7PNW
A.-KUBIN-PL	&3#M7PNW
A.-KUBIN-PL.	&3#M7PNW
A.-KUBINPL	&3#M7PNW
A.-KUBINPL.	&3#M7PNW
A. KUBIN-PLATZ	&3#M7PNW
A. KUBIN-PL	&3#M7PNW
A. KUBIN-PL.	&3#M7PNW
A. KUBINPL	&3#M7PNW
A. KUBINPL.	&3#M7PNW
A KUBIN-PLATZ	&3#M7PNW
A KUBIN-PL	&3#M7PNW
A KUBIN-PL.	&3#M7PNW
A KUBINPL	&3#M7PNW
A KUBINPL.	&3#M7PNW
ALFRED-KUBIN-PLATZ	&WGY_~3#M
ALFRED-KUBIN-PL	&WGY_~3#M
ALFRED-KUBIN-PL.	&WGY_~3#M
ALFRED-KUBINPL	&WGY_~3#M
ALFRED-KUBINPL.	&WGY_~3#M

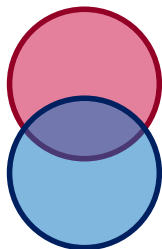
mc_adresse70 = dqmatch(Strasse,'Address',70,'DEDEU');

1. Nutzen von Matchcodes mit unterschiedlichen Sensitivitäten

Network links – Komponenten

Komponente Name-Geburtsdatum:








- mc_nachname95 (Matchcode high sensitivity)
- mc_vorname95 (Matchcode high sensitivity)
- Geburtsdatum



Komponente Name - Adresse:

- mc_nachname95 (Matchcode high sensitivity)
- mc_vorname60 (Matchcode low sensitivity)
- mc_adresse75 (Matchcode medium Sensitivity)
- plz_cl (Cleansed PLZ)

2. Nutzen von PROC HPENG um
Personen mittels Komponenten
von Informationen zu matchen.

 VORNAME	 NACHNAME	 GEBURTSD ATUM_DAT	 STRASSE	 PLZ	 STADT	 FREIE_ADRESSE
Dinu	Bleban	14JUL1967:00:...			NA	Kaplanweg 9 A-8071...
DINU	BLEBAN	19JUL1969:00:...	TUCHSTRASSE 154/2	A - 8055	GRAZ-PUNTIG...	
DINU	BLEBAN	19JUL1969:00:...	TUCHSTRASSE 154/2	A - 8055	GRAZ-PUNTIG...	
Dinu	Bleban	14JUL1967:00:...			NA	Kaplanweg 9 A-8071...
Dinu	Bleban	14JUL1967:00:...			NA	Kaplanweg 9 A-8071...
Dinu	Bleban	14JUL1967:00:...	Tuchstra. 154/2	8055	Graz	

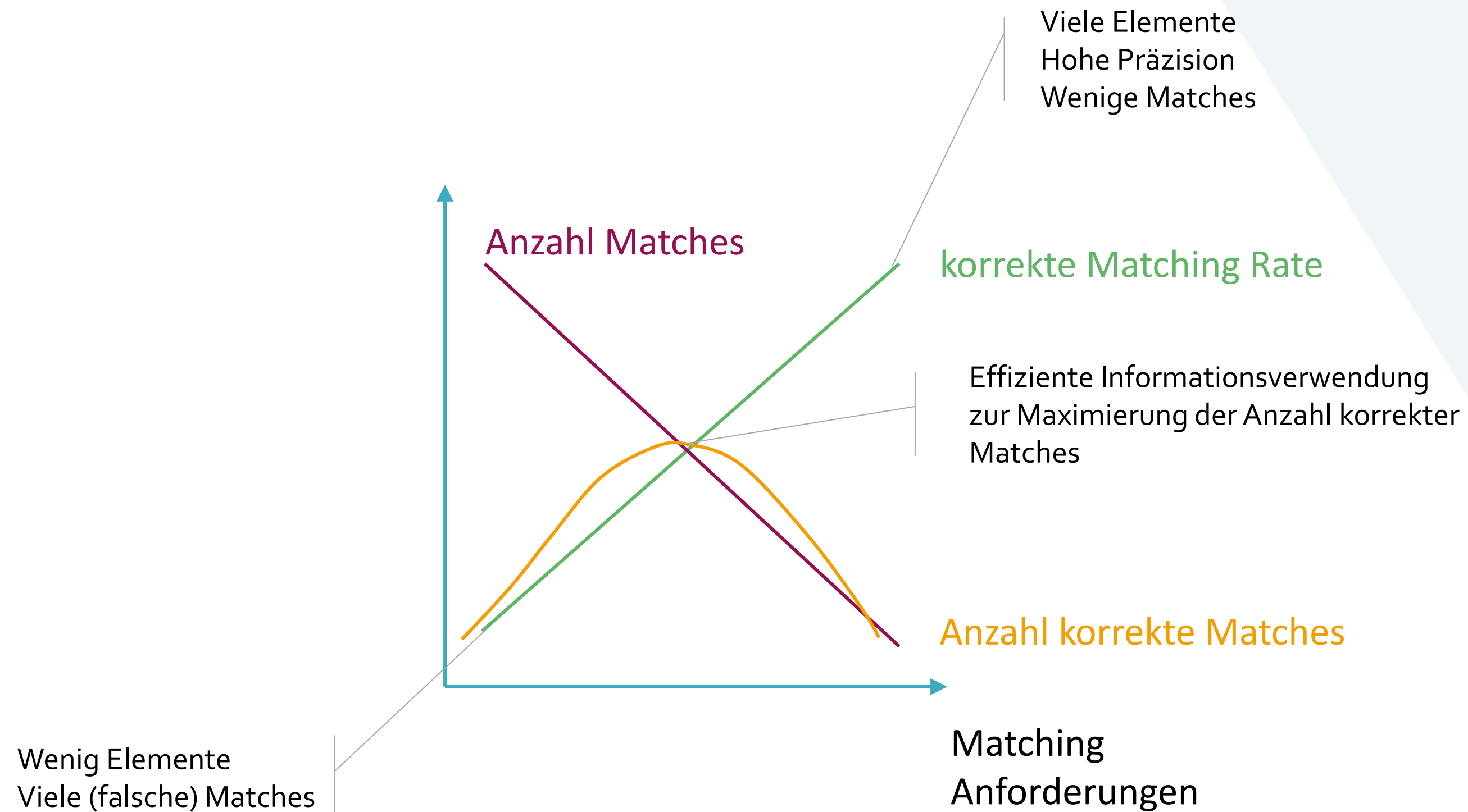
Vielfalt an Komponenten

- nm_geb mc_nachname95
- nm_geb mc_vorname95
- nm_geb geburt_txt
- nm_geb_ort mc_nachname95
- nm_geb_ort mc_vorname95
- nm_geb_ort geburt_txt
- nm_geb_ort plz_cl
- nm_geb_ort2 mc_nachname95
- nm_geb_ort2 mc_vorname95
- nm_geb_ort2 geburt_md
- nm_geb_ort2 plz_cl
- nm_geb_ort3 mc_nachname95
- nm_geb_ort3 mc_vorname95
- nm_geb_ort3 geburt_txt
- nm_geb_ort3 mc_stadt80
- nnm_geb_ort mc_nachname95
- nnm_geb_ort geburt_txt
- nnm_geb_ort plz_cl
- vnm_geb_ort mc_vorname95
- vnm_geb_ort geburt_txt
- vnm_geb_ort plz_cl

- geb_hadr geburt_txt
- geb_hadr mc_adresse75
- geb_hadr mc_nachname95
- geb_hadr mc_vorname60
- geb_hadr plz_cl
- geb_hadr2 geburt_txt
- geb_hadr2 mc_adresse75
- geb_hadr2 mc_nachname95
- geb_hadr3 geburt_txt
- geb_hadr3 mc_adresse75
- geb_hadr3 mc_vorname95
- hnm_hadr mc_nachname95
- hnm_hadr mc_vorname60
- hnm_hadr mc_adresse75
- hnm_hadr geburtsjahr
- hnm_hadr2 mc_nachname95
- hnm_hadr2 mc_vorname60
- hnm_hadr2 mc_adresse75
- hnm_hadr2 geb_tag_mon
- hnm_hadr3 mc_nachname95
- hnm_hadr3 mc_vorname60
- hnm_hadr3 mc_adresse75
- hnm_hadr3 plz_cl
- hnm_hadr4 mc_nachname95
- hnm_hadr4 mc_vorname60
- hnm_hadr4 mc_adresse75
- hnm_hadr4 mc_stadt80
- hnm_hadr5 mc_nachname95
- hnm_hadr5 mc_vorname60
- hnm_hadr5 mc_adresse75

- hnm_plz mc_nachname95
- hnm_plz mc_vorname95
- hnm_plz plz_cl
- sid steuer_nr
- tin_nm_plz mc_nachname95
- tin_nm_plz mc_vorname60
- tin_nm_plz steuer_nr
- tin_nm_plz plz_cl
- tin_nm_fa mc_nachname95
- tin_nm_fa mc_vorname60
- tin_nm_fa steuer_nr
- tin_nm_fa finanzamt_nr
- tin_geb_fa geburt_txt
- tin_geb_fa steuer_nr
- tin_geb_fa finanzamt_nr
- tin_geb_plz geburt_txt
- tin_geb_plz steuer_nr
- tin_geb_plz plz_cl
- tin_geb_adr geburt_txt
- tin_geb_adr steuer_nr
- tin_geb_adr mc_adresse75
- tin_geb_adr plz_cl

Effiziente Informationsverwendung



Exclusion Lists: Ausschluss von mehrdeutigen Komponenten

- Keine Komponenten verwenden, wenn für diese Komponente mehr als zwei Steuernummer existieren.

Name	Vorname	Geburtsdatum	Steuernummer
Svolba	Gerhard	01.07.1979	09 336/4206
Svolba	Gerhard	01.07.1979	13 517/1901
Mihai	Paunescu	19.05.1983	72 966/2107



- PROC HPENG unterstützt Datasets mit exclusions für jede einzelne Komponente

Exclusion Lists: Beispiele

Compounds	Elements	Exclusion
MAN_EXCL_NMADR3	mc_nachname95, mc_vorname60, mc_adresse75, plz_cl	25976
MAN_EXCL_NMGEb	mc_nachname95, mc_vorname95, geburt_txt	824
MAN_EXCL_NMPLZ	mc_nachname95, mc_vorname95, plz_cl	42257

3. Effiziente Informationsnutzung
mittels Exclusion Lists

MC_NACHNAME95	NACHNAME_CL	MC_VORNAME60	VORNAME_CL	MC_ADRESSE75	ADRESSE_CL	PLZ_CL	GEBURTDATUM
#1#P_Y	UZUNER	@B	ÖMER	7@&PB_7YW_MZ	FASANGASSE 19/9	4050	21.Apr.88
#1#P_Y	UZUNER	@B	ÖMÜR	7@&PB_7YW_MZ	FASANGASSE 19/9	4050	31.Jul.84

MC_NACHNAME95	NACHNAME_CL	MC_VORNAME95	VORNAME_CL	PLZ_CL	ADRESSE_CL	Geburtsdatum
#P3_Y	UNGER	7@4_G	JOSEF	1307	FASANGASSE 33	01.Mai.34
#P3_Y	UNGER	7@4_G	JOSEF	1307	HOFÄCKER 7	23.Feb.82
#P3_Y	UNGER	7@4_G	JOSEF	1307	HOFÄCKER 7	28.Apr.54
#P3_Y	UNGER	7@4_G	JOSEF	1307	SCHÖNAUSTRASSE 131	22.Sep.70
#P3_Y	UNGER	7@4_G	JOSEF	1307	WELGERSDORF 19	19.Mär.22
#P3_Y	UNGER	7@4_G	JOSEF	1307	WILDENTENGRABEN 7	24.Mär.21
#P3_Y	UNGER	7@4_G	JOSEF	1307	ZIEGLERGASSE 21	28.Dez.83
#P3_Y	UNGER	7@4_G	JOSEF	1307	ZIEGLERGASSE 21	28.Jul.58

Lessons Learned

- Nutzen von Matchcodes (DQMTACH – Funktionen) für Fuzzy Matching.
- Aufbau von Netzwerken mit Links auf Basis von definierten Komponenten
PROC HPENG /RTENG: Entities – Compounds – Elements
- Effiziente Informationsnutzung mittels Exclusion Lists