# Dublettencheck Reloaded

## mithilfe der Datenanalyse-Lösung IDEA

TIPPS & TRICKS IDEA

**Autor:** Klaus Jakobi
Genossenschaftsverband e.V.

audicon
we care

Datenqualität spielt für den Unternehmenserfolg eine entscheidende Rolle. Weisen Daten Schwächen auf, kann dies den Nutzen für ein Unternehmen negativ beeinflussen, da Dubletten zu Fehlinterpretationen führen können. Die Bereinigung von Daten bindet wiederum den Einsatz von Ressourcen (Mensch, Maschine, Zeit). Diese Ressourcen stehen dadurch nicht für den eigentlichen Unternehmenszweck zur Verfügung. Auch das IT-Sicherheitsziel Integrität zielt auf richtige, vollständige und redundanzfreie Datenbestände, auf die sich der Unternehmer verlassen kann. Das Suchen nach unvollständigen oder unrichtigen Datenbeständen ist relativ einfach. Die Prüfung auf Dubletten – also dem „doppelten Lottchen" im Datenbestand – ist hingegen oftmals fehlerbehaftet und bedarf einer strukturierten Vorgehensweise. Dieser Artikel versucht anhand eines konkreten Beispiels unter Einsatz der Datenanalyse-Lösung IDEA einen multiplen Analyseansatz bei der Suche nach Dubletten darzustellen, um eine bestmögliche Trefferausbeute zu erreichen.

## Die Rohdaten

Die Rohdaten sollten möglichst alle Informationen enthalten, die für eine Dublettensuche geeignet sind. Dazu zählen naturgemäß Felder mit eindeutigen Inhalten, insbesondere einer ID (z. B. in Form einer Kunden- oder Artikelnummer) und eine Bezeichnung (z. B. Kundenname oder Artikelbezeichnung). In dem folgenden Beispiel geht es um Stammdaten von Privatkunden eines Finanzdienstleisters. Folgende „Zutaten" werden für den Dublettencheck benötigt (der Feldtyp ist in Klammern angegeben)*:

- Kundennummer (Text)
- Vorname (Text)
- Nachname (Text)
- Geburtsname (Text)
- Geburtsdatum (Datum)
- Postleitzahl (Text)
- Ort (Text)
- Straße und Hausnummer (Text)
- Optional: zuständige Filiale und/oder Sachbearbeiter (Text)

Das letztgenannte Feld ist nicht zwingend erforderlich und soll nur bei einer Ergebnisverwertung die Zuordnung zur zuständigen Betriebseinheit erleichtern. Der entsprechende Datenbestand darf nur die Daten von aktiven Privatkunden enthalten. Aktiv bedeutet, dass der Kunde in irgendeiner aktiven Geschäftsverbindung zum Finanzdienstleister steht, also Kunde, Bevollmächtigter, Sicherheitengeber oder wirtschaftlich Berechtigter ist. Weiterhin sollte es sich nur um Daten von Einzelkunden und nicht um Daten von Gemeinschaftskunden oder Gesellschaften etc. handeln. Dies kann durch entsprechende Schlüsselfelder bereits beim Datenabruf vorgefiltert werden.

*Hinweis: Bei sämtlichen in den „Tipps und Tricks IDEA" verwendeten Personendaten handelt es sich um fiktive Angaben.

**Die Datenaufbereitung**

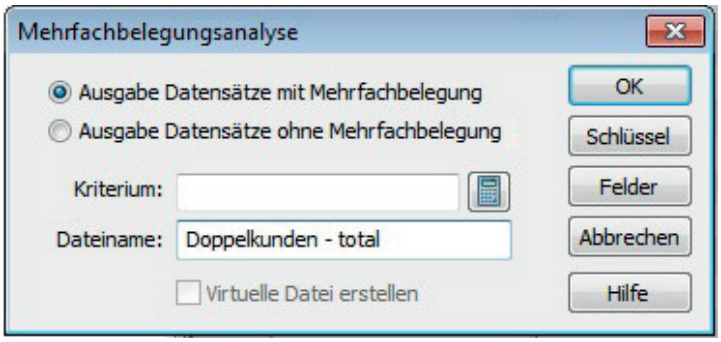
Die in IDEA importierten Rohdaten sind um einige Felder zu erweitern, die im Verlauf der Dublettenchecks noch benötigt werden. Folgende Felder müssen Sie erstellen:

- NAME_GES_NORM
  Kundenname gesamt und normalisiert (also ohne Leer- und Sonderzeichen)
  Feldgleichung: @Alltrim(@Strip(VORNAME + NACHNAME))
- VORNAME_NORM
  Vorname normalisiert
  Feldgleichung: @Alltrim(@Strip(VORNAME))
- NACHNAME_NORM
  Nachname normalisiert
  Feldgleichung: @Alltrim(@Strip(NACHNAME))
- GEB_NAME_NORM
  Geburtsname normalisiert
  Feldgleichung: @Alltrim(@Strip(GEBURTSNAME))
- NAME_ABGLEICH
  Abgleich, ob Geburtsname vorhanden ist oder nicht
  Feldgleichung: @If(.NOT.@Isblank(GEBURTSNAME); GEB_NAME_NORM; NACHNAME_NORM)
- STRAßE_NORM
  Straßenname normalisiert
  Feldgleichung: @Alltrim(@Strip(@Replace(@Replace(@Lower(STRAßE);“str.“; “straße“);“strasse“;“straße“)))
- NAME_GES_CODE
  Gesamtname mit Kölner Phonetik vercoded
  Feldgleichung: #Koelner_Phonetic(NAME_GES_NORM)

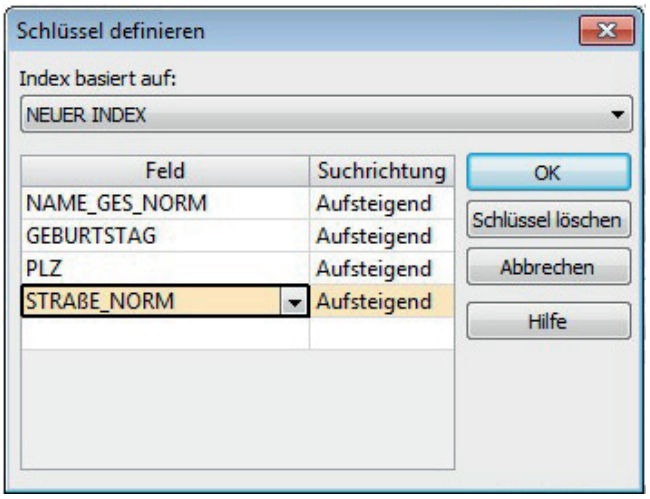
Definieren Sie alle Zusatzfelder als editierbare Zeichenfelder mit einer Mindestlänge von 50 Zeichen, um im Einzelfall manuelle Korrekturen zu ermöglichen. Nach der Definition dieser Zusatzfelder kann der eigentliche Dublettencheck beginnen.

**Dublettencheck – total**

Der erste (und einfachste) Dublettencheck ist der Abgleich aller eindeutigen Felder auf Übereinstimmung. Dies erfolgt in IDEA mit der Analyse **Mehrfachbelegung - Ermittlung**.




Verwenden Sie folgenden Schlüssel:






audicon
we care

Wählen Sie für die Ergebnisdatei mindestens die folgenden Felder aus: KUNDENNUMMER, GEBURTSTAG, VORNAME_NORM, NACHNAME_NORM und STRAßE_NORM.
Die Ergebnisdatei enthält alle Datensätze mit entsprechenden Übereinstimmungen.

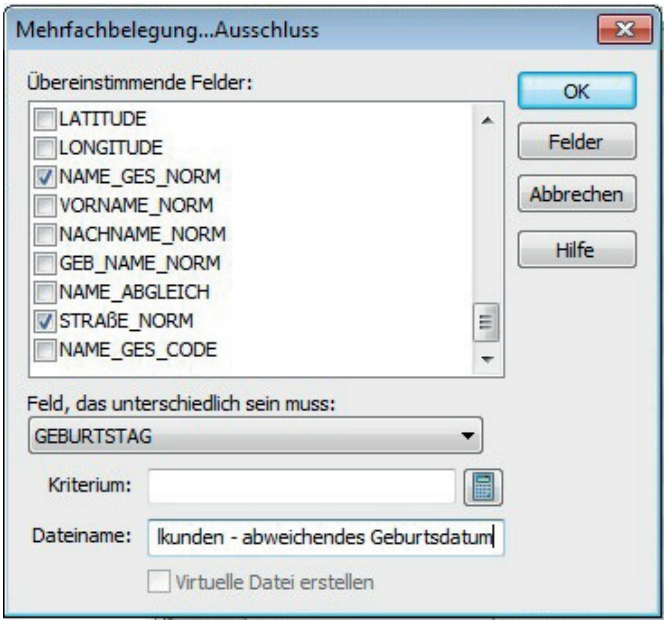
| KUNDENNUMMER | GEBURTSTAG | VORNAME_NORM | NACHNAME_NORM | STRAßE_NORM |
|---|---|---|---|---|
| 140767 | 23/08/1941 | Andreas | Lowe | mühlenstraße16 |
| 363575 | 23/08/1941 | Andreas | Lowe | mühlenstraße16 |
| 146293 | 11/11/1971 | Bernd | Müller | brandenburgischestraße5 |
| 146293 | 11/11/1971 | Bernd | Müller | brandenburgischestraße5 |
| 107271 | 08/12/1945 | Heike | Eichelberger | kirchenallee36 |
| 196024 | 08/12/1945 | Heike | Eichelberger | kirchenallee36 |
| 105139 | 23/03/1996 | Janina | Kastner | budapesterstraße97 |
| 357553 | 23/03/1996 | Janina | Kastner | budapesterstraße97 |

***Hinweis:***
Normalerweise sollte für jeden Kunden nur ein Stammdatensatz vorhanden sein. Es kann aber Sonderfälle geben, bei denen für einen Kunden mehr als ein Stammdatensatz angelegt wurde. Dies könnte der Fall sein, wenn ein Kunde sowohl als Privatperson als auch als gewerblicher Kunde geführt wird (z. B. bei einem Kaufmann oder Selbständigen). Dies müssen Sie bei der Ergebnisinterpretation individuell berücksichtigen.

**Dublettencheck – abweichendes Geburtsdatum**
Dieser Dublettencheck prüft dieselben Kundendaten auf Übereinstimmung, jedoch mit Ausnahme des Geburtsdatums. Dazu wird in IDEA die **Mehrfachbelegungsanalyse mit Ausschluss** verwendet. Folgende Eingaben sind erforderlich:

Übereinstimmende Felder:
- NAME_GES_NORM
- PLZ
- STRAßE_NORM

Feld, das unterschiedlich sein muss:
- GEBURTSTAG






audicon
we care

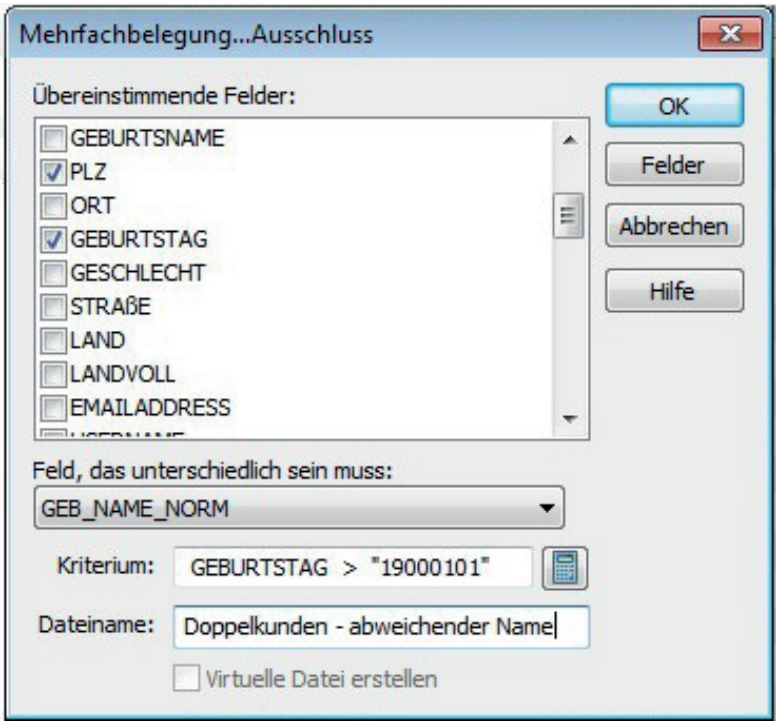
| | GEBURTSTAG | VORNAME_NORM | NACHNAME_NORM | STRAßE_NORM |
|---|---|---|---|---|
| 1 | 12/03/1948 | Martina | Fuchs | amsinckstraße4 |
| 2 | 12/04/1949 | Martina | Fuchs | amsinckstraße4 |
| 3 | 02/03/1993 | Paul | Moeller | ruschestraße4 |
| 4 | 23/12/1971 | Paul | Moeller | ruschestraße4 |
| 5 | 17/01/1940 | Lucas | Metzger | boxhagenerstraße38 |
| 6 | 28/02/1963 | Lucas | Metzger | boxhagenerstraße38 |
| 7 | 15/10/1991 | Brigitte | Hertz | fischerinsel61 |
| 8 | 15/10/1993 | Brigitte | Hertz | fischerinsel61 |

*Hinweis:*
Die Ergebnisdatei enthält namensgleiche Kunden mit gleicher Anschrift und einem abweichenden Geburtsdatum. Dies kann in vielen Fällen auch korrekt sein, wenn beispielsweise Vater und Sohn denselben Vor- und Nachnamen haben und unter der gleichen Anschrift wohnen. Bei einigen Treffern wird dies jedoch nicht der Fall sein. Dies müssen Sie im Einzelfall invididuell untersuchen.

**Dublettencheck – abweichender Name**
Hier wird ähnlich wie bei der vorhergehenden Analyse verfahren. Die **IDEA Mehrfachbelegungs-analyse mit Ausschluss** wird mit folgenden Eingaben verwendet:

Übereinstimmende Felder:
- GEBURTSTAG
- PLZ
- STRAßE_NORM

Feld, das unterschiedlich sein muss:
- GEB_NAME_NORM






audicon
we care

Die Ergebnisdatei enthält alle Kunden, die bei gleichem Geburtsdatum und gleicher Anschrift verschiedene Namen haben.

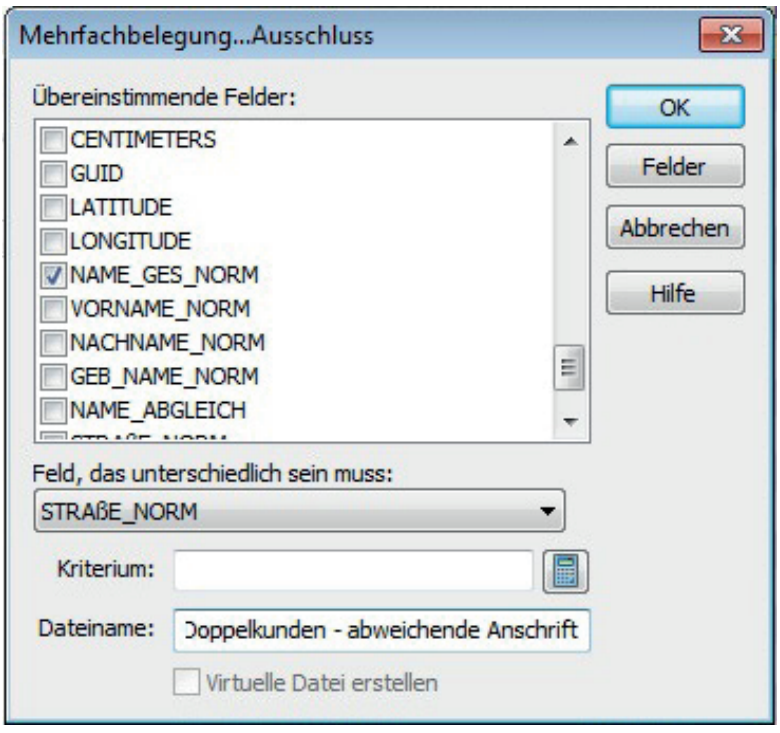
| | GEBURTSTAG | VORNAME_NORM | NACHNAME_NORM | STRAßE_NORM |
|---|---|---|---|---|
| 1 | 06/07/1964 | Sabrina | Kaestner | ingebeisheimplatz2 |
| 2 | 06/07/1964 | Sabrina | Kästner | ingebeisheimplatz2 |
| 3 | 11/04/1992 | Luca | Fried | jenaerstraße30 |
| 4 | 11/04/1992 | Lukas | Fried | jenaerstraße30 |
| 5 | 18/02/1970 | JensUwe | Hoch | meinekestraße5 |
| 6 | 18/02/1970 | Uwe | Hoch | meinekestraße5 |
| 7 | 29/11/1936 | Ines | Fischer | lützowplatz42 |
| 8 | 29/11/1936 | Ines | Kuester | lützowplatz42 |

*Hinweis:*
In einigen Fällen können die Treffer auch durch Zwillinge bedingt sein, die im selben Ort leben. Hier muss eine Prüfung im Einzelfall erfolgen.

**Dublettencheck – abweichende Anschrift**
Auch hier kommt die **IDEA Mehrfachbelegungsanalyse mit Ausschluss** zum Einsatz. Folgende Eingaben sind erforderlich:

Übereinstimmende Felder:
- NAME_GES_NORM
- GEBURTSTAG

Feld, das unterschiedlich sein muss:
- STRAßE_NORM






audicon
we care

| | GEBURTSTAG | VORNAME_NORM | NACHNAME_NORM | STRAßE_NORM |
|---|---|---|---|---|
| 1 | 02/05/1944 | Sandra | Hertzog | billwerderneuerdeich80 |
| 2 | 02/05/1944 | Sandra | Hertzog | mohrenstraße36 |
| 3 | 02/06/1961 | Stefan | Jager | brandenburgischestraße70 |
| 4 | 02/06/1961 | Stefan | Jager | kurfürstendamm53 |
| 5 | 03/07/1961 | Martina | Dreher | kielerstraße26 |
| 6 | 03/07/1961 | Martina | Dreher | schmarjestraße49 |
| 7 | 03/07/1993 | Jörg | Fiedler | amsinckstraße61 |
| 8 | 03/07/1993 | Jörg | Fiedler | hollanderstraße60 |

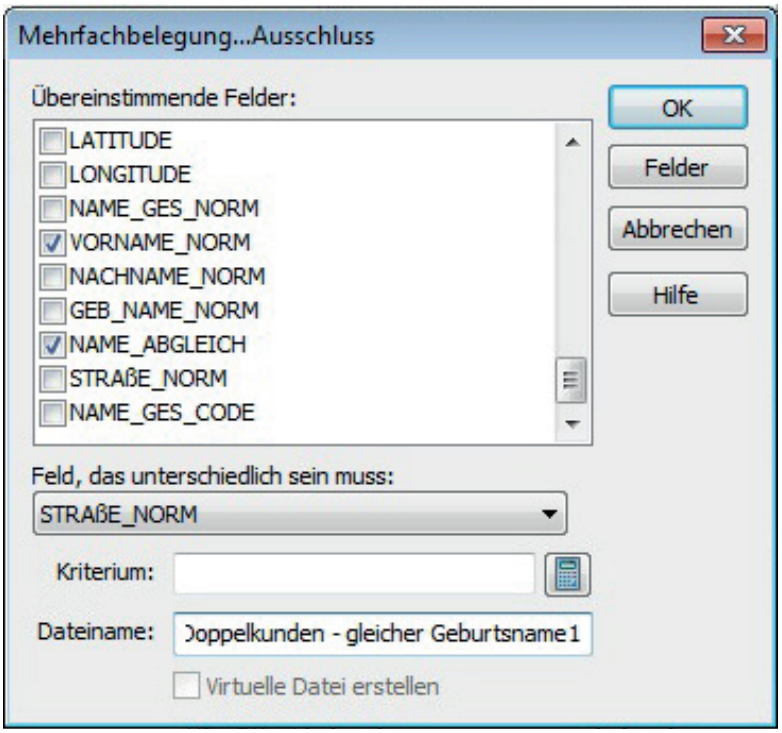
*Hinweis:*
Unechte Treffer können bei Personen vorkommen, für die jeweils ein Kundenstamm für private und gewerbliche Zwecke (z. B. bei Kaufleuten oder Selbständigen) angelegt wurde.

### Dublettencheck – gleicher Geburtsname

Bei diesem Dublettencheck wird auf die Übereinstimmung des Geburtsnamens (sofern vorhanden) bei gleichem Vornamen und Geburtsdatum, jedoch abweichendem Nachnamen und Anschrift, geachtet. Dies erfordert eine zweistufige Vorgehensweise. Führen Sie zuerst eine **Mehrfachbelegungsanalyse mit Ausschluss** mit folgenden Angaben durch:

Übereinstimmende Felder:
- VORNAME_NORM
- NAME_ABGLEICH
- GEBURTSTAG

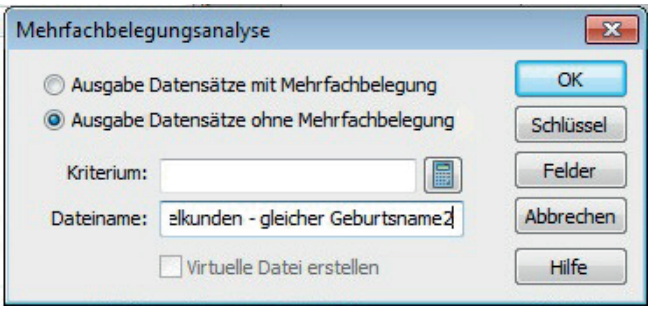
Feld, das unterschiedlich sein muss:
- STRAßE_NORM






audicon
we care

Die Ergebnisdatei enthält alle Treffer mit gleichem Vornamen, gleichem Geburtsdatum sowie gleichem Geburtsnamen bei abweichender Anschrift. Verwenden Sie als zweite Stufe die Analyse **Mehrfachbelegung - Ermittlung** mit der Option **Ausgabe Datensätze ohne Mehrfachbelegung**.



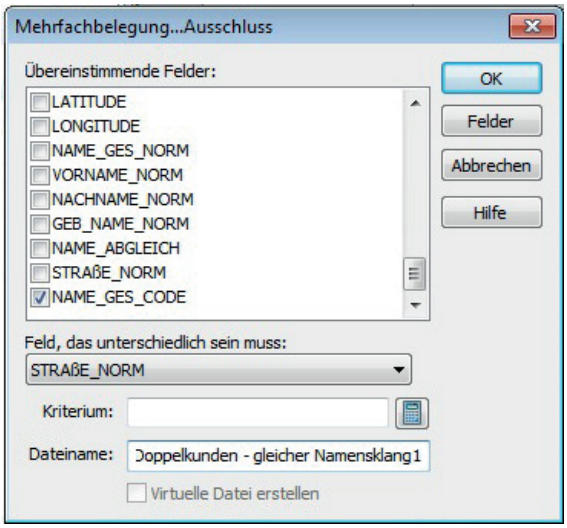

Wählen Sie als Schlüssel die Felder GEBURTSTAG und NACHNAHME_NORM. Damit werden alle Kunden gefiltert, die den gleichen Vor- und Geburtsnamen sowie das gleiche Geburtsdatum bei abweichendem Nachnamen und abweichender Anschrift haben. Es werden also zwei abweichende Elemente im Ergebnis berücksichtigt.

| | GEBURTSTAG | VORNAME_NORM | NACHNAME_NORM | GEB_NAME_NORM | STRAßE_NORM |
|---|---|---|---|---|---|
| 1 | 01/08/1968 | Jessika | Wurfel | | reeperbahn39 |
| 2 | 01/08/1968 | Jessika | Zimmer | Wurfel | augsburgerstraße31 |
| 3 | 04/06/1996 | Silke | Ackermann | Lang | prenzlauerallee10 |
| 4 | 04/06/1996 | Silke | Lang | | güntzelstraße8 |
| 5 | 08/08/1951 | Heike | Freitag | | kurfürstendamm98 |
| 6 | 08/08/1951 | Heike | Schulze | Freitag | halleschesufer17 |
| 7 | 14/08/1992 | Anna | Meister | Schwarz | storkowerstraße55 |
| 8 | 14/08/1992 | Anna | Schwarz | | stresemannstraße54 |
| 9 | 15/12/1969 | Sabine | Abt | | kielerstraße51 |
| 10 | 15/12/1969 | Sabine | Drescher | Abt | neueroßstraße34 |
| 11 | 19/01/1989 | Michelle | Herrmann | Oster | hoheluftchaussee84 |
| 12 | 19/01/1989 | Michelle | Oster | | fasanenstraße37 |
| 13 | 24/10/1925 | Leah | Schroeder | Wannemaker | rosenstraße10 |
| 14 | 24/10/1925 | Leah | Wannemaker | | fontenay7 |
| 15 | 25/03/1943 | Sara | Sommer | Waechter | rhinstraße83 |
| 16 | 25/03/1943 | Sara | Waechter | | wallstraße56 |
| 17 | 26/12/1928 | Jennifer | Scholz | | paderbornerstraße47 |
| 18 | 26/12/1928 | Jennifer | Schweizer | Scholz | schönhauserallee91 |
| 19 | 28/09/1978 | Stefanie | Konig | Wannemaker | parkstraße62 |
| 20 | 28/09/1978 | Stefanie | Wannemaker | | jenaerstraße32 |
| 21 | 31/10/1984 | Katja | Bieber | Schmitz | joachimstalerstraße40 |
| 22 | 31/10/1984 | Katja | Schmitz | | leipzigerstraße19 |

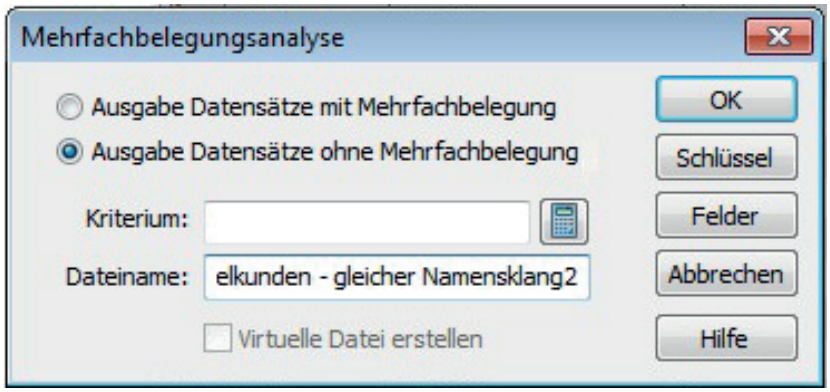
**Dublettencheck – gleicher Namensklang**
Bei diesem Dublettencheck ist ebenfalls eine zweistufige Vorgehensweise erforderlich. Führen Sie zuerst eine **Mehrfachbelegung mit Ausschluss** durch:

Übereinstimmende Felder:
- NAME_GES_CODE
- GEBURTSTAG

Feld, das unterschiedlich sein muss:
- STRAßE_NORM




Untersuchen Sie in der zweiten Stufe diese Ergebnisdatei mit der Analyse **Mehrfachbelegung – Ermittlung** und der Option **Ausgabe Datensätze ohne Mehrfachbelegung**.




Verwenden Sie als Schlüsselfelder VORNAME und NACHNAME. Somit werden alle Kunden gefiltert, die den gleichen Namensklang sowie das gleiche Geburtsdatum aufweisen, bei denen aber abweichende Schreibweisen des (Gesamt-)Namens sowie eine abweichende Anschrift vorhanden sind. Auch hier werden also zwei abweichende Elemente im Ergebnis berücksichtigt.

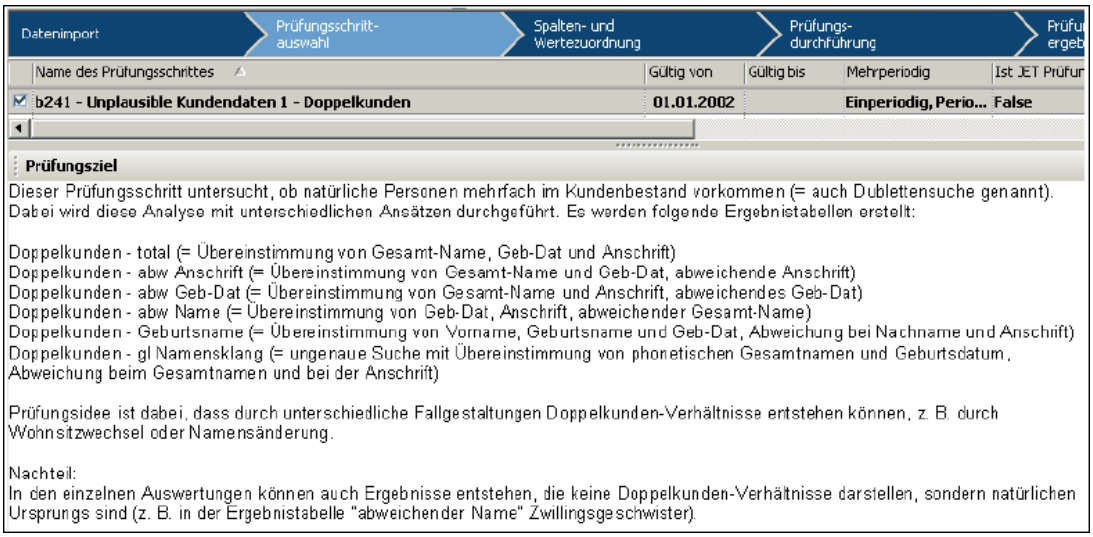
| | GEBURTSTAG | VORNAME_NORM | NACHNAME_NORM | STRAßE_NORM | NAME_GES_CODE |
|---|---|---|---|---|---|
| 1 | 07/07/1940 | Alexander | Oster | bissingzeile12 | 0548627827 |
| 2 | 07/07/1940 | Alexandra | Oster | fischerinsel59 | 0548627827 |
| 3 | 03/01/1988 | Anna | Braun | schillerstraße51 | 06176 |
| 4 | 03/01/1988 | Anne | Braun | leipzigerstraße30 | 06176 |
| 5 | 29/10/1951 | Kerstin | Saenger | paderbornerstraße51 | 478268647 |
| 6 | 29/10/1951 | Kerstin | Sänger | luebeckerstraße17 | 478268647 |
| 7 | 22/02/2005 | Lea | Baader | augsburgerstraße52 | 5127 |
| 8 | 22/02/2005 | Lea | Bader | schaarsteinweg12 | 5127 |

## Fazit

Durch einen multiplen Ansatz bei der Dublettensuche kann das Trefferspektrum spürbar erhöht und somit ein wichtiger Beitrag zur Herstellung einer besseren Datenqualität geleistet werden. Im vorstehend geschilderten Fall wurden insgesamt sechs unterschiedliche Dublettensuchansätze auf demselben Datenbestand durchgeführt. Hier noch einmal die Ansätze im Überblick:

- Doppelte Kunden – total
- Doppelte Kunden – abweichendes Geburtsdatum
- Doppelte Kunden – abweichender Name
- Doppelte Kunden – abweichende Anschrift
- Doppelte Kunden – gleicher Geburtsname
- Doppelte Kunden – gleicher Namensklang

Um diese Ansätze zu standardisieren, könnte über AIS TaxAudit bzw. SmartAnalyzer ein entsprechender Prüfungsschritt erstellt werden.




Den Prüfungsschritt können Sie über eine Entwicklungsumgebung (für AIS TaxAudit) bzw. die App SDK ab der Version 9.2 (für SmartAnalyzer) erstellen.

## Über Audicon

Die Audicon GmbH ist der führende Anbieter von Software-Lösungen, methodischem und fachlichem Know-how sowie Dienstleistungen rund um Audit, Risk und Compliance. Die Lösungen richten sich an Wirtschaftsprüfer und Steuerberater, Compliance- und Risiko-Manager sowie Revisoren und Rechnungsprüfer/Kämmerer.

Die Audicon Software-Lösungen werden eingesetzt von

- 23 der 25 in der Lünendonk®-Liste 2013 genannten führenden Wirtschaftsprüfungs- und Steuerberatungsgesellschaften in Deutschland
- 90 der 120 umsatzstärksten deutschen Firmen
- rund 14.000 Steuerprüfern der Finanzverwaltung
- den Big Four, den vier weltweit größten Wirtschaftsprüfungsgesellschaften

Weitere Informationen: www.audicon.net

## Über den Verfasser

**Klaus Jakobi** arbeitet für den Genossenschaftsverband e.V. (Prüfungs- und Beratungsverband, Bildungsträger und Interessenvertretung für rund 2.400 Mitgliedsgenossenschaften in 13 Bundesländern) sowie für die Wirtschaftsprüfungsgesellschaft AWADO Deutsche Audit GmbH. Der Schwerpunkt seiner Tätigkeit liegt seit einigen Jahren in den Bereichen IT-Prüfung und dabei insbesondere auf dem Einsatz von Datenanalyse mithilfe von IDEA und AIS TaxAudit Professional, einschließlich der Entwicklung von Standardprüfungsanalysen bei genossenschaftlichen Banken im Rahmen von Jahresabschlussprüfungen oder Sonderuntersuchungen (z. B. Fraud Detection).

Kontakt: klaus.jakobi@genossenschaftsverband.de

## Sie haben Fragen?

**Sprechen Sie uns an – wir helfen Ihnen gerne weiter!**

| Telefonisch: | Per E-Mail: | Im Internet: |
|---|---|---|
| +49 211 5 20 59 - 430 | sales@audicon.net | www.audicon.net |