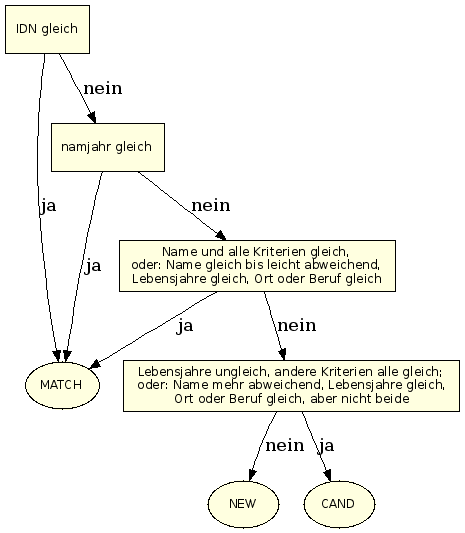
Match-Vorgaben Personen

#### Festlegungen M&M für die Importe von Tp-Sätzen in die GND-Matchkriterien

Tp-Sätze werden beim Import immer auf Dubletten überprüft. Dazu werden folgende Felder verglichen:

|  |  |  |
| --- | --- | --- |
| Feld/Aspekt | Erläuterung IT | Bemerkungen/Fragen |
| **IDN**  PICA 003@ 007K 007N 071A  MARC 001 035 $a 035 $z | Vergleich von mitgelieferten   * DNB-IDNs * GND-IDNs (auch bei umgelenkten Datensätzen) * alten SWD/PND/GKD-IDNs (auch bei umgelenkten Datensätzen)   Wenn eine der Identnummern matcht, ist der ganze Vergleich ein Match, unabhängig von den anderen Kriterien. Wenn die mitgelieferte(n) Identnummer(n) nicht match(t)(en), hat dies keine Auswirkungen auf den endgültigen Matchwert, es werden dann nur die Werte der übrigen Vergleichefür den Gesamtmatchwert herangezogen. |  |
| **namjahr**  Schlüssel aus:  Pica  028A/P/@ $a $c $d(1) oder $P $l und  050R $a $b oder $d (nur wenn "Jahr - Jahr")  MARC 100/400/700 $a[Nachname - 1. Buchst. Vorname] und 548 $a $b | gr, 05.10.15: Neu für IDS-Import: Es wird ein Schlüssel gebildet aus dem 1. Buchstaben Vornamen, dem Nachnamen (Vorzugsbenennung) und dem Geburts- und Sterbejahr (letztere müssen zwingend vorhanden sein, sonst kann der Schlüssel keinen Match ergeben).  Wenn der Schlüssel matcht, gilt der ganze Vergleich des Paares als Match.  gr, Mitte Dezember 2015: Anpassung: Schlüssel wird jetzt aus Nachname - 1.  Buchstabe Vorname - Geburtsjahr - Sterbejahr gebildet. Es werden auch Schlüssel mit Verweisungformen gebildet.  gr, September 2016: Neu: Wenn es sich um einen Doppelnamen handelt (2 oder mehrere Vornamen oder Nachnamen, durch Bindestrich oder Blank getrennt) ist auch eine Jahresangabe ausreichend |  |
| Name  PICA 028A 028@  MARC 100 400 | Es werden nicht nur AF mit AF verglichen, sondern auch AF mit VW. Dies sollte möglich sein, da noch weitere Kriterien herangezogen werden, also eine zusätzliche Absicherung stattfindet |  |
| Beruf  PICA 041R 050G $b"Beruf:"  MARC 550 678 $b"Beruf:" | Hier geht es vor allem um Berufe; $4 "ber[uc]" muss vorhanden sein bzw. 050G/678 $b mit dem String "Beruf:" beginnen (neu mit NB Import, Sept. 2016)  gr, September 2016: Neu: Es werden auch Angaben aus 678/050G mit herangezogen, wenn sie mit "Beruf:" eingeleitet werden. Wenn mehrere mit Komma getrennte Angaben vorhanden sind, werden die ersten drei für den Abgleich genutzt | gr, 20.1.: Ist es wichtig, dass auch $4 identisch ist? Muss ein Beruf also zwingend "beru" oder "berc" sein?  10.02.2014: $4 muss nicht identisch sein, beru und berc gleich behandeln  26.9.2014: 041R wird jetzt nur mit herangezogen, wenn $4 mit "ber" beginnt, damit nicht akademische Titel etc. auch mit ausgewertet werden (sind zu unspezifisch) |
| **Jahr**  PICA 050R  MARC 548 | Lebensjahre $a und $b werden geprüft. Zur Zeit ist es so, dass es ausreicht, wenn eines von beiden Subfeldern identisch ist. D.h. wenn das Geburtsjahr identisch ist, aber dasTodesjahr nicht, gelten die Lebensjahre dennoch als identisch (s. Diskussion in der Nebenspalte) | gr, 20.1.: reicht das so aus? Oder müssen zwingend beide Jahresangaben identisch sein? Ich habe gerade bei älteren Personen öfters den Fall gefunden, dass eine Jahresangabe um 2-3 Jahre abwich, aber das andere Jahr identisch war.  10.02.2014: eine der Angaben sollte identisch sein; andere "Jahresangabe" kann abweichen |
| **Ort**  PICA 065R $9 $6 $A+$0 $a $4  MARC 551 $0 $a $4 | Orte; $4 darf nicht "ortw" sein | gr, 20.1.: Ist es wichtig, dass auch $4 identisch ist? Oder reicht u.U. der Hinweis auf einen gemeinsamen Ort, ohne die Information, ob es sich um einen Geburts-, Sterbe- oder Wirkungsort handelt?  ~~10.02.2014: $4 muss nicht identisch sein~~  AfS, 2.4.: Neuer Matchwert 0.84000: im Prinzip in Ordnung, allerdings matchen mit einem Wert von 0.84615 zwei verschiedene Axel Müller (Nr. 2858), der eine hat den Wirkungsort Koblenz und der andere ist in Koblenz geboren -> Ist es möglich, bei den Orten doch $4 zu vergleichen?  **Besprechung am 9.4.:** Ja, die Konfiguration wird entsprechend angepasst. Update 19.12.2015: "ortw" wird ausgeschlossen + $4 muss gleich sein. Verweisungsformen des verknüpften Ortes werden mit einbezogen |
| Berechnung des Matchwertes | Für einen Match müssen gleich sein:   * namjahr * Name (Ähnlichkeitsvergleich, identisch bis leicht abweichend), Lebensjahr und Beruf oder Ort |  |
| Matchwert | alt: 0.96000   * wird 0.96000 erreicht, aber 050R ist ungleich bzw. nicht vorhanden, wird nicht gematcht   gr, neu **21.03.**: 0.84000   * ergibt sich aus der Bedingung, dass zwei Kriterien gleich sein müssen. 050R ist dabei auf jeden Fall gleich | AfS, 26.2.: Der Matchwert sollte in diesem Fall mit 0.96 in Ordnung gehen - ändert sich dieser bei weiteren Einspielungen entsprechend der Anzahl der einzuspielenden Sätze, so müsste die Festlegung entsprechend angepasst werden.  AfS, 2.4.: s. Matchwert |
| Kandidaten |  | (Frage)sollen Kandidaten gebildet werden? Wenn ja, bis zu welchem Matchwert?  (Frage)gr, 21.03.: Vorschlag Kandidaten 0.83999 - 0.71000  AfS, 26.2.: Hier müssten Lebensdaten gleich wie die anderen Kriterien gewichtet werden; die Festlegung des Matchwertes auf 0.00 bei fehlenden Lebensdaten ist dabei nicht mehr anwendbar, da es z.B. gleichnamige Personen mit dem gleichen Ort oder Beruf geben könnte, die sich in der intellektuellen Kandidatenbearbeitung als eindeutig dublett erweisen. Ist dafür ein weiterer Durchlauf notwendig? (Info)gr, 21.03.: Ist jetzt erledigt, daher neue Matchwerte, s.Ergebnisse Testdatei 21. März |

#### Schematischer Ablauf



#### Genaue Verteilung der Matchwerte

(aktualisiert am 16.9.2016)

* Matchwert 1.00: alle oben beschriebenen Kriterien sind exakt gleich oder IDN ist gleich oder namjahr ist gleich (MATCH)
* Matchwerte 0.99 - 0.85: wie oben, aber Name ist nicht 100%ig identisch (MATCH)
* Matchwert 0.84615: Name ist gleich, von Beruf oder Ort ist eines gleich, Lebensjahre sind gleich (MATCH)
* Matchwerte 0.83 - 0.70: wie oben, aber Name ist nicht 100%ig identisch (MATCH)
* Matchwert 0.69231: Name ist gleich, Lebensjahre sind gleich (aber namjahr kann nicht gebildet werden), Beruf und Ort sind ungleich oder nicht vorhanden (CAND)
* Matchwerte 0.68 - 0.61: wie oben, aber Name ist nicht 100%ig identisch (CAND)
* Matchwert 0.61538: Name ist gleich, Lebensjahre sind abweichend oder nicht vorhanden, von Beruf oder Ort ist eines gleich (CAND)
* Matchwerte 0.61 - 0.47: wie oben, aber Name ist nicht 100%ig identisch (NEW)
* Matchwert 0.46154: Name ist gleich, alle anderen Kriterien weichen ab oder sind nicht vorhanden (NEW)
* Matchwerte unter 0.46: wie oben, aber Name ist nicht 100%ig identisch (NEW)