25. Oktober 2017

Ergebnisprotokoll des GND-Import-Workshops

Donnerstag, **29. Juni 2017**, von 11:00 bis 16:00
Deutsche Nationalbibliothek, Frankfurt am Main

Status: verabschiedet

Teilnehmerinnen und Teilnehmer

Verbund	Teilnehmer
BSB	Eva-Maria Gulder Angela Bernhardt
BVB	Barbara Wolf-Dahm
DNB	Petra Feilhauer Stefan Grund Sarah Hartmann Jürgen Kett Esther Scheven Sylvia Thüncher
GBV	Dr. Barbara Block
hbz	Natascha Dahmen Monika Humpertz
HeBIS	Rita Albrecht Mark Popilka
IDS	Hans Urech
KOBV	Michael Franke-Maier
OBVSG	Sebastian Aigner Josef Labner
NB	Christiane Schmidt
SWB	Annabel Feuerstein Cornelia Katz

Tagesordnung

Unterlagen

Alle Unterlagen zur Sitzung stehen im Wiki des GND-Ausschusses zur Verfügung: https://wiki.dnb.de/x/3sJ5Bw.

TOP 1. Einführung

Frau Feilhauer und Herr Kett begrüßen die Teilnehmer zum GND-Import-Workshop und führen in das Thema ein. Die Teilnehmer stellen sich in einer kurzen Vorstellungsrunde vor.

1.1 Importe in die GND

Frau Feilhauer präsentiert die Ziele des Workshops: zum Einen bestehen diese darin, den Status quo des Importprozesses, die Basisdokumente und die Werkzeuge, die dabei zum Einsatz kommen, vorzustellen; zum Anderen werden im zweiten Teil des Workshops in Kleingruppen spezifische Fragestellungen rund um das Thema GND-Importe bearbeitet, um darauf aufbauend Probleme bei den bisherigen Verfahren zu identifizieren und Verbesserungen vorzunehmen und gemeinsam die Importprozesse abzustimmen.

Für Details s. Präsentation https://wiki.dnb.de/download/attachments/125420254/GND-Import-Workshop_Einstieg.pptx

1.2 Werkzeug Match&Merge im Überblick

(TOP 2.1 Match&Merge für die Entität Person wird zur Veranschaulichung zusammen mit TOP 1.2 besprochen.)

Herr Grund stellt allgemein den Match&Merge-Prozess im CBS von PICA/OCLC sowie die Möglichkeit zur Kandidatenbearbeitung per WinIBW vor. Anhand der Entität Person (Tp-Sätze) wird beispielhaft der vollständige Ablauf eines Match&Merge-Prozesses inklusive einer Vorstellung der Vergleichselemente erläutert, um Möglichkeiten und Grenzen des Verfahrens aufzuzeigen.

Für Details s. Präsentation
https://wiki.dnb.de/download/attachments/125420254/Match%26Merge_im_PICA-CBS_2017.pdf

1.3 Fragen und Anregungen

Frage	wer	Antwort DNB
Ist es möglich zu importierende Daten im „alten" Format (vor Formatänderungen, die mit dem Export-Release 2017.03 im September wirksam werden) bereit zu stellen?	BSZ	In einer Übergangszeit ist es möglich sowohl altes als auch neues Format zu liefern und zu importieren.
Werden die als TOP500 markierten Datensätze von Änderungen bei Importen ausgeschlossen?	hbz	Ja. Änderungen an TOP500-Datensätzen werden ggf. manuell eingetragen. _Allerdings kann der Ausschluss von TOP500-Sätzen beim Matchen dazu führen, dass Dubletten eingespielt werden, die dann später händisch zusammengeführt werden müssen (Beispiele dafür gab es beim IDS-Import)
Wer entscheidet aktuell, welche Datensets in die GND eingespielt werden?		Aktuell entscheidet AfS nach Prüfung der Daten. In der Zukunft sollen aber Kriterien

		vom GND-Ausschuss für die Importe aufgestellt werden und in Zweifelsfällen alle GND-Partner befragt werden.
Kann der Datenlieferant die ins Approval-System geladenen Testdatenlieferungen sichten?	GBV	Ja. Im Approvalsystem kann jeder mit einem WinIBW-Zugang die Daten sichten und prüfen. Auch per ONS können die Inhalte des Approvalsystem geharvestet werden.
Gibt es Felder, die zwar geliefert werden, aber nicht ins Produktivsystem importiert werden?	BSZ	Ja. Lokale Felder werden weder in das Approvalsystem, noch in das Produktivsystem importiert. Absprachen hierzu erfolgen jeweils mit den Lieferanten
Werden die als Kandidaten in der GND markierten Datensätze auch in weiteren Importprozessen für das Matching berücksichtigt?	hbz	Ja, alle GND-Datensätze werden in den Matchingprozess mit einbezogen.
Wie kann es sein, dass buchstabenidentische Körperschaften nicht zusammengefunden haben?	BSB	Wahrscheinliche Ursache: Vorgänger / Nachfolge-Relationen müssen ebenfalls identisch sein Offene Frage: ist es sinnvoll, Vorgänger/Nachfolge-Relationen in den Abgleich mit einzubeziehen? Problem bei Nichtberücksichtigung: Auflösung von Ketten von Vorgängern/Nachfolgern bei den gelieferten Daten, Übernahme von 510er Feldern in die GND, die u.U. für den gefundenen Datensatz nicht korrekt sind
Wie funktioniert die Entdopplungsfunktion in der WinIBW?	hbz	Der Entdoppelungsprozess verwendet die ursprünglich für den Import-Prozess eingerichtete Match&Merge Konfiguration. Der Prozess ist in Feld 169 codiert angegeben. Näheres zur Kandidatenbearbeitung s. https://wiki.dnb.de/x/tAbeBg
Der Match&Merge-Prozess gibt als Ergebnis immer nur einen Kandidaten aus?	GBV	Ja. Es wird lediglich der nach dem Vergleichsalgorithmus als bester erkannte Kandidat im Feld 169 angegeben.
Ist der genutzte Match&Merge-Prozess der übliche Standardprozess von Pica?	Hebis	Es wird das Standard-Match&Merge-Programm von PICA/OCLC genutzt. Der vollständige Workflow wird mit Shellskripten jobgesteuert. Eine kleine Programmentwicklung verschmilzt beim Mergen

		entstandene doppelte Felder, die laut Datenformat nur einmal vorhanden sein dürfen, da dies zur Zeit mit dem M&M-Programm nicht geht (sollte allerdings in der nächsten CBS-Version möglich sein)

Aufgaben:

Nr.	Aufgabe	Deadline	zuständig
	Aktualisierung der Import-Checkliste (z. B. Anpassung neue Felder/Unterfelder) ggf. als Wikiseite / FAQ-Liste zum Thema Importe in die GND anlegen		DNB
	Rechtzeitig über neues Level, das eingeführt wird, um Datensätze von einer Bearbeitung zu sperren (z. B. TOP-500-Datensätze oder Tp-Datensätze, in denen keine individualisierenden Angaben ergänzt werden dürfen)informieren, da diese Änderung von den GND-Partnern (z. B. Alma) eingeplant werden muss.		DNB
	Workflow für die Kandidatenbearbeitung (Abbau von Kandidatenpool) z. B. über ONS konzipieren		DNB
	Für künftige Importe: Statistik der insgesamt importierten Datensätze im Verhältnis zu Kandidaten und Neueinspielungen bereitstellen		DNB
	Vorgaben bzw. Erläuterung der spezifischen Besonderheiten der Daten müssen im Vorfeld des Imports benannt werden		jeweilige Datenlieferanten

TOP 2. Orientierung

2.1 Match&Merge für die Entität Person

s. TOP 1.2

2.2 Fragen und Anregungen zu Match&Merge für die Entität Person

Frage	wer	Antwort DNB
Werden die Match-Kriterien nur in den dafür vorgesehenen		Ja, es werden auch Angaben im Feld 678, wenn sie mit

Feldern berücksichtigt, also z. B. Berufe in Feld 550 oder z. B. auch als Textstrings in 678 „Beruf:"		"Beruf:" eingeleitet werden, herangezogen. Weiteres wird u.U. herangezogen, wenn es eindeutig zu identifizieren ist (hängt jeweils von der Datenlieferung ab)
Werden die alten GKD-Namensformen (Feld 913) beim Matching berücksichtigt?	Pitz	Bisher wurden die alten GKD-Formen nicht berücksichtigt. Sollte zu den Kriterien hinzugenommen werden.
Wurde bereits ein Import von Sachschlagwörtern (Ts) durchgeführt?		Nein, seit GND (2012) wurden keine Ts-Datensätze importiert. Dies war lediglich vor GND der Fall.

Aufgaben:

Nr.	Aufgabe	Deadline	zuständig
	Häufigkeit von Personennamen im Match&Merge-Verfahren berücksichtigen bzw. gesondert behandeln		
	Externe Informationen (z. B. VIAF) mit in den M&M-Prozess einbeziehen		
	Verknüpfte Titel mit im M&M-Prozess berücksichtigen		
	Es sollen Kriterien entwickelt werden, um Datenimporte zu akzeptieren bzw. abzulehnen.		
	Importe, Statistiken, Approval-System sollen öffentlich gemacht werden, um Transparenz zu schaffen		
	Kann Datenbereinigung ein Geschäftsmodell sein? Gibt es Beratungsangebote von Agenturen für Datenimporte?		

TOP 3. Arbeit in Kleingruppen

Die Teilnehmer diskutieren in Kleingruppen folgende drei Fragestellungen, halten ihre Ergebnisse auf Pinnwänden fest und stellen die Ergebnisse und offenen Fragen im Plenum vor (siehe auch Fotoprotokoll im Wiki).

a) Kriterien für Entität Körperschaft ermitteln

In der Kleingruppe wird über die Matchingkriterien für Körperschaften beraten. Es wird z. B. kritisch gesehen, dass die Vorgänger-/Nachfolgerbeziehungen einen so großen Einfluss auf das Matching von Datensätzen haben; es wird vorgeschlagen, dass Datensätze, die aufgrund von unterschiedlichen Vorgänger-/Nachfolgerbeziehungen nicht gemergt werden, als Kandidaten gekennzeichnet werden. Ebenso sollten Datensätze, deren Vorzugsbenennungen identisch sind, aber sich durch fehlende Orte unterscheiden, als Kandidaten gekennzeichnet werden.

Insgesamt muss das Problem behandelt werden, wie zu verhindern ist, dass maschinelle Prozesse ein erreichtes Qualitätsniveau nicht wieder senken. Dies kann zum Beispiel eintreten, wenn bereits qualitätsgesicherte Informationen durch einen Merge um unpräzise Merkmale ergänzt werden. Das kann dazu führen, dass Datensätze mehrfach manuell korrigiert werden müssen. Als Lösung wurde vorgeschlagen, Datensätze für automatische Ergänzungen sperren zu können. Dies wurde kontrovers diskutiert. Pauschale Sperrungen ganzer Datensätze für die Anreicherung aus Mergeprozessen anhand von Qualitätskennzeichen werden von einigen als zu restriktiv betrachtet. Eine möglicherweise bessere Variante wäre es, wenn automatische Ergänzungen durch Markierung von den bereits qualitätsgesicherten Informationen unterschieden werden könnten. Dies ist auf dem aktuellen technischen Stand aber nicht möglich. Daher sollte der Punkt als Anforderung an die künftige Umgebung aufgenommen werden. Gleichzeitig sollten praktikable Lösung für die aktuellen Importe geprüft werden.
Es wird die Empfehlung ausgesprochen, Feld 667 beim Merge nicht zu berücksichtigen, da die Inhalte meist von lokalem Interesse sind. Die genannten Punkte sollen bei der Prüfung der bisherigen Match&Merge-Kriterien überdacht werden.
Die Kleingruppe empfiehlt im Rahmen des Qualitätsmanagements der GND Textstrings in den 5XX-Feldern in Links zu GND-Datensätzen umzusetzen und den Umgang mit uneindeutigen Verweisungsformen zu klären.

Wahrscheinlich könnte das Matching verbessert werden, wenn mit dem Datensatz verknüpfte Titel berücksichtigt werden.

Aufgaben:

Nr.	Aufgabe	Deadline	zuständig
	Alte GKD-Namensformen (Feld 913) beim Matching berücksichtigen	kurzfristig	Arbeitsgruppe, die sich um die Prüfung und Weiterentwicklung des M&M-Prozesses für Körperschaften kümmert
	Häufigkeit von Namen als Kriterium berücksichtigen	kurzfristig	Arbeitsgruppe, die sich um die Prüfung und Weiterentwicklung des M&M-Prozesses für Körperschaften kümmert
	Folgende Anforderung aufnehmen: Das Markieren von Datenelementen hinsichtlich ihres Ursprungs und Status („maschinelles erzeugt durch Verfahren XY", „aus Fremddatenquelle XY gewonnen" / intellektuell geprüft durch XY", …) muss möglich sein.		
	Kurzfristige Lösung für das Schützen bestimmter Datensätze und Felder prüfen		
	Feld 667 beim Merge nicht zu berücksichtigen, da die Inhalte meist von lokalem Interesse sind		
	Rahmen des Qualitätsmanagements der GND Textstrings in den 5XX-Feldern in Links zu GND-Datensätzen umzusetzen		

b) Was darf bei Match&Merge NICHT geschehen?

Die Kleingruppe spricht sich dafür aus, externe Eigenschaften von Entitäten (z. B. aus externen Quellen) und verknüpfte Datensätze wie z. B. Titeldaten mit in das Matching einzubeziehen und die jeweiligen Matchwerte und Prozesse transparent zu machen. Ein Trigger zur intellektuellen Bearbeitung von Datensätzen soll das Kriterium der Häufigkeit sein: wie häufig wird ein GND-Datensatz verwendet. Dem Testen der zu importierenden Daten und des M&M-Prozesses wird eine besondere Wichtigkeit zugeschrieben, so dass erst nach stabilen Testergebnissen die Daten importiert werden sollen. Weitere Anforderungen an die abzuliefernden Daten bzw. an den Datenlieferanten sind, dass der Erschließungsstandard und die Besonderheiten der zu importierenden Daten durch den Datenlieferanten transparent gemacht werden und dass die Struktur der Daten dem GND-Format entspricht. Nicht mehr gültige Informationen, wie z. B. Angaben, die im lokalen Kontext der Daten sinnvoll sind, aber nicht im Kontext der Integration in die GND, sollen vom Merge ausgeschlossen werden. Dies erfordert u. a. eine intensive Kommunikation zwischen dem Datenlieferanten und dem technischen Betreiber der GND, der DNB.

Im Laufe des Importprozesses müssen die TOP500-Datensätze und die indexrelevanten Felder und deren Änderungen berücksichtigt werden. Dies wirft auch die Frage auf, wie der Workflow zum manuellen Nacharbeiten der Merges (TOP500) verbessert werden kann.

Die Vorgehensweise bei der Definition der Matching-Kriterien werden unterschiedlich bewertet: bei einem Teil der Teilnehmer wird Wert darauf gelegt, bei Importen die Anzahl der möglichen Kandidaten möglichst gering zu halten und somit eher „weichen" Matchingkriterien zu folgen - auch auf die Gefahr hin, falsche Datensätze zu mergen. Andere Teilnehmer sprechen sich für eine sichere Variante aus, d.h. im Zweifelsfall eher Datensätze als Kandidaten zu markieren, als Datensätze fälschlicherweise zu mergen.

c) Rechte & Pflichten eines Datenlieferanten

Die Kleingruppe schlägt vor, den Datenlieferanten zu verpflichten, dublettenfreie Daten und Daten von hoher Qualität zu liefern. Des Weiteren soll eine Voraussetzung für einen Import in die GND sein, dass der Datenlieferant auch nach dem Datenimport die GND weiterhin nutzt und pflegt (z. B. Bereitschaft Dubletten zu bereinigen) und die Rahmenbedingungen der GND (z. B. Format, Datenstruktur) akzeptiert. Die Erwartungen (Pflichten) an die Datenlieferanten sollten entsprechend formuliert werden. Die Pflichten der DNB werden darin gesehen, die Infrastruktur für eine kooperative Mitarbeit bei der Abarbeitung von Dubletten bereitzustellen und Selfmatches[1] durchzuführen. Die Teilnehmer sprechen sich dafür aus, dass Datenimporte auch abgelehnt werden können. Für diese Entscheidungsgrundlage müssen durch den GND-Ausschuss zunächst die Kriterien entwickelt werden. Der Datenlieferant sollte sich mit den Match&Merge-Kriterien auseinandersetzen und in den Prozess der Abnahme (vor der finalen Dateneinspielung) einbezogen werden. Als offene Frage wird formuliert, wie zukünftig Kandidatenpools aufgelöst oder abgearbeitet werden können.

Es wird folgendes Vorgehen bzw. folgende Arbeitspakete beschlossen:
Aufgaben:

Nr.	Aufgabe	Deadline	zuständig
	Prüfung und Weiterentwicklung des M&M-Prozesses für	kurzfristig / mittelfristig	Humpertz Urech

[1] Dublettenbereinigung innerhalb der GND

	Körperschaften		Pitz
	Sammlung von Projektideen insbesondere zum Workflow und Werkzeugen für Importe in die GND (z. B. wie kann der Workflow zur Mitarbeit des Datenlieferanten an der Matchingkonfiguration aussehen, welche Vorprozessierung beim Datenlieferanten ist notwendig oder kann beratend angeboten werden?		Franke-Meier Block Albrecht (Feuerstein/Katz) Bernhardt/Gulder
	Minimallevel für den Import definieren		
	Verbesserung des Workflows zum manuellen Nacharbeiten der Merges (TOP500) erarbeiten		
	Organisation: Rechte und Pflichten von Datenlieferanten formulieren (z. B. Kandidatenbearbeitung)	Eher später	

TOP 4. Feedback

Dieser Workshop wird von den Teilnehmern als ein erster Schritt gesehen, auf den aber weitere Aktivitäten und Diskussionen folgen müssen. Frau Katz schlägt vor, im Wiki noch mehr Transparenz bezüglich der Importe zu schaffen, indem die Mengengerüste je Import und - konsequenter als bisher - die Anzahl der Merges, Kandidaten und der neueingespielten Datensätze veröffentlicht werden.

TOP 5. Sonstiges

Frau Albrecht regt an, das Thema Importe von Geografika in einem separaten Termin oder in einer Arbeitsgruppe zu besprechen. Eine Diskussion darüber sollte geführt werden, ob aus externen Quellen oder anderen Systemen Geografika in die GND importiert werden können und wie dies automatisiert realisiert werden kann, z. B. fehlende hessische Orte. Möglicherweise haben auch andere GND-Partner vor dem Hintergrund von Regionalbibliografien diesen Bedarf. Hintergrund dieser Frage ist die Tatsache, dass die Hessische Bibliografie ihre inhaltliche Erschließung umstellt und dafür die GND nutzt.

Nr.	**Aufgabe**	**Deadline**	**zuständig**
	Diskussion Import von Geografika aus anderen Quellen (Trigger „Regionalbibliografie“)		Albrecht Katz Kett Scheven
	Regelmäßige Selfmatches innerhalb der GND durchführen		DNB
	Einspielung von Sachbegriffen unter Nachnutzung von Thesauri und Wörterbüchern		