

Verfahren für die Dublettenzusammenführung bei der Einspielung von Daten in den B3Kat (Match&Merge-Verfahren)

Stand: 11.09.2017

Im Folgenden werden die wesentlichen Schritte des Verfahrens zur Dublettenzusammenführung von Titeldaten im B3Kat beschrieben. Das Verfahren ist sehr vereinfacht dargestellt. Auf eine detaillierte Beschreibung wird aus Gründen der Übersichtlichkeit bewusst verzichtet.

1. Datenumfang

Der Umfang der Titeldatenlieferung an die BVB-Verbundzentrale wird in Absprache zwischen der BVB-Verbundzentrale, der Bibliothek und der von der Bibliothek gegebenenfalls mit der Datenaufbereitung beauftragten Firma festgelegt (z.B. sollen keine Zeitschriftenaufnahmen ohne ZDB-ID bzw. BVB-ID geliefert werden; rein lokale Titelaufnahmen werden von der Lieferung ausgenommen).

Die folgenden Lieferformate sind zulässig:

- MAB2-Bandformat im MAB Zeichensatz
- MAB2-Bandformat in UTF8-Codierung des UNICODE-Zeichensatzes
- MARC-Format (MARC/UTF8)
- MARC-XML (MARC/UTF8)
- ALEPH-DOCSEQuential (ASEQ)

Im B3Kat wird der MAB2-Satztyp "y" nicht verwendet. Sollte eine Bibliothek eigene Abteilungsätze verwendet haben, sind diese vorab zu löschen. Auch Pauschalverweisungen (Titel, Personennamen, Körperschaften und Schlagwörter) werden nicht in den B3Kat geladen und müssen daher aus dem Lieferbestand entfernt werden.

2. Datenaufbereitung

Die Datenaufbereitung kann von der Bibliothek selbst oder durch einen externen Dienstleister nach Vorgaben der BVB-Verbundzentrale und der Bibliothek durchgeführt werden. Dabei müssen die geltenden Katalogisierungskonventionen, auf denen die Plausibilitätsprüfungen für den B3Kat beruhen, beachtet werden (s. [KKB online](#)).

2.1. Normdatenverknüpfung

Es ist erstrebenswert, aus den Angaben in den Titeldaten heraus in der Verbund-Normdatei (GND) passende Normdatensätze zu finden. Finden sich keine passenden Normdatensätze in der GND, werden die Titeldaten ohne vollständige Normdatenverknüpfung übernommen. In den Titeldaten steht im Verknüpfungsfeld dann nur der Textstring des Normdatennamens - ohne ID.

2.2. Vorab-Dublettenzusammenführung

Die zu migrierenden Titeldaten sollten vor der Übergabe an die BVB-Verbundzentrale einer Prüfung auf intrinsische Dubletten unterzogen und diese gegebenenfalls zusammengeführt worden sein.

In den Lieferdateien an die BVB-Verbundzentrale dürfen neben den „echten“ zu migrierenden Daten auch die ggf. bereits im B3Kat erfassten und in das Lokalsystem übernommenen Daten enthalten sein. Anhand der in den letzteren bereits enthaltenen Verbund-ID („BV-Nummer“) lassen sie sich zielsicher erkennen und ausfiltern.

2.3. Konsistenz der gelieferten Daten

Im Rahmen der Datenaufbereitung sollte die Konsistenz der Lieferdaten geprüft werden. Fehlende hierarchische Verknüpfungen sind so weit wie möglich herzustellen bzw. nicht mehr aktuelle Verknüpfungen zu löschen:

- Zu hierarchisch untergeordneten Titeln sollen die übergeordneten Datensätze im Lieferumfang enthalten sein.

- Die in den gelieferten Titeldatensätzen eingetragenen Bezüge zu Identifikationsnummern der B3Kat-Titeldatenbank (BVB01) müssen aktuell sein.

Stücktitelaufnahmen, zu denen weder eine passende Überordnung geliefert wird, noch im B3Kat ermittelt wurde, können in einer separaten Datei an die BVB-Verbundzentrale geliefert werden. Das ermöglicht, dass diese Titeldaten einen Sonderlauf innerhalb des Match&Merge-Verfahrens durchlaufen, um trotz fehlender Überordnung noch Dubletten innerhalb des B3Kat zu finden. Ohne diesen Sonderlauf werden unverknüpfte Stücktitel zu Neuaufnahmen, da im regulären Match&Merge-Lauf die Zusammenführung der Überordnungen Voraussetzung für die Dublettenprüfung der Stücktitel ist.

2.4. Vergabe der Aleph-Satzkennung und -Satztypen (Aleph-Feld LDR, Pos. 23; Aleph-Feld 051/052, Pos. 0)

Zu Beginn des Match&Merge-Verfahrens werden die gelieferten Titeldaten in das Aleph-Internformat ASEQ konvertiert. Die ASEQ-Feldstruktur basiert im B3Kat auf MAB2. Gemäß den Vorgaben des Internformats werden folgende Satzkennungen und Satztypen für die gelieferten Titeldaten vergeben:

- **Monografie** -> Aleph-Feld **LDR**, Pos. 23: **h**; Aleph-Feld **051**, Pos. 0: **m**

- **Stücktitel** (RDA: Teil mit unabhängigem Titel) -> Aleph-Feld **LDR**, Pos. 23: **h**; Aleph-Feld **051**, Pos. 0: **s**

- **Mehrbändiges begrenztes Werk** (RDA: mehrteilige Monografie) -> Aleph-Feld **LDR**, Pos. 23: **h**; Aleph-Feld **051**, Pos. 0: **n**

- **Mehrbändiges begrenztes Werk** (RDA: Mehrteiliges Werk) **mit Zählung innerhalb einer Reihe** -> Aleph-Feld **LDR**, Pos. 23: **h**; Aleph-Feld **051**, Pos. 0: **t**

- **Serie** (RDA: monografische Reihe) -> Aleph-Feld **LDR**, Pos. 23: **h**; Aleph-Feld **052**, Pos. 0: **r**

- **Zeitschrift** -> Aleph-Feld **LDR**, Pos. 23: **h**; Aleph-Feld **052**, Pos. 0: **p**

- **Zeitung** -> Aleph-Feld **LDR**, Pos. 23: **h**; Aleph-Feld **052**, Pos. 0: **z**

- **Band** (RDA: Teil mit abhängigem Titel) -> Aleph-Feld **LDR**, Pos. 23: **u**

- **Unselbstständiges Werk** (RDA: Teilwerk, analytisch erfasst) -> Aleph-Feld **LDR**, Pos. 23: **h**; Aleph-Feld **051**, Pos. 0: **a Unverknüpfte Stücktitel**

3. Match&Merge-Verfahren im BVB

3.1. Vergabe von Kennbuchstaben

Zur weiteren Verarbeitung der Titeldaten werden je nach Aleph-Satzkennung (Aleph-Feld LDR, Pos. 23) und Aleph-Satztyp (Aleph-Feld 051/052, Pos. 0) verschiedene Kennbuchstaben vergeben. Teilweise werden auch mehrere Aleph-Satztypen unter einem Kennbuchstaben zusammengefasst.

| Satztyp | Kennbuchstabe | Kriterium |
|--|---------------|---|
| Monographie | M | Feld LDR, Pos. 23: h <u>und</u> Feld 051, Pos. 0: m |
| Stücktitel (RDA: Teil mit unabhängigem Titel) | S | Feld LDR, Pos. 23: h <u>und</u> Feld 051, Pos. 0: s |
| Mehrbändiges begrenztes Werk (RDA: mehrteilige Monographie) ohne/mit Zählung innerhalb einer Reihe | G | Feld LDR, Pos. 23: h <u>und</u> Feld 051, Pos. 0: t |
| Serie/Zeitschrift/Zeitung | E | Feld LDR, Pos. 23: h <u>und</u> Feld 052, Pos. 0: r, p, z |
| Band (RDA: Teil mit abhängigem Titel) | B | Feld LDR, Pos. 23: u |
| Unselbstständiges Werk | U | Feld LDR, Pos. 23: h <u>und</u> Feld 051, Pos. 0: a |

3.2. Abgleich nur innerhalb des gleichen Kennbuchstabens

Im weiteren Verlauf des Match&Merge-Verfahrens werden nur Datensätze der gleichen Kennbuchstaben miteinander verglichen.

Beispiele:

- Monografien der Lieferdatei (Kennbuchstabe M) werden gegen Monografien im B3Kat geprüft (Kennbuchstabe M)
- Serien der Lieferdatei (Kennbuchstabe E) werden gegen Serien, Zeitschriften und Zeitungen im B3Kat geprüft (Kennbuchstabe E)

Aber

- Stücktitel der Lieferdatei (Kennbuchstabe S) werden **nicht** gegen Monografien im B3Kat geprüft (Kennbuchstabe M)
- Bandsätze der Lieferdatei (Kennbuchstabe B) werden **nicht** gegen Monografien im B3Kat geprüft (Kennbuchstabe M)

Überordnungen und verknüpfte Stücktitel/Bände

Gelieferte Überordnungen (mehrbändige begrenzte Werke/Serien/Zeitschriften/Zeitungen) werden – getrennt nach Kennbuchstaben – gegen Überordnungen im B3Kat dublettengepüft.

Bei erfolgreicher Dublettensammenführung der Überordnungen werden die gelieferten verknüpften Stücktitel/Bände unter den B3Kat-Überordnungen zusammengefasst.

Die gelieferten verknüpften Stücktitel werden anschließend mit den Verbund-Stücktiteln innerhalb der gleichen B3Kat-Überordnungen abgeglichen.

Die gelieferten verknüpften Bände werden **nicht** weiter gegen die Verbund-Bände innerhalb der gleichen B3Kat-Überordnungen dublettengepüft.

Wird zu einer gelieferten Überordnung **keine** Dublette im B3Kat gefunden, so werden sowohl die gelieferte Überordnung als auch sämtliche damit verknüpfte Stücktitel/Bände zu Neuaufnahmen.

Durch dieses Verfahren soll gewährleistet werden, dass gelieferte verknüpfte Stücktitel/Bände unter einer Überordnung zusammenbleiben.

Unverknüpfte Stücktitel

Die auf Wunsch der Bibliothek gesondert aufbereiteten und in einer separaten Datei gelieferten unverknüpften Stücktitel werden mit jedweden anderen Stücktiteln im B3Kat abgeglichen. Als Ergebnis können Stücktitel verschiedener Überordnungen als Treffer ermittelt werden. Der gelieferte Bestand zu einer Überordnung wird somit u.U. nach dem Match&Merge-Lauf auf verschiedene B3Kat-Überordnungen verteilt.

Unverknüpfte Stücktitel, die nicht gesondert aufbereitet und nicht in einer separaten Datei geliefert wurden, werden in jedem Fall zu Neuaufnahmen.

Bandsätze (Kennbuchstabe B) und unselbstständige Werke (Kennbuchstabe U)

Bandsätze und unselbstständige Werke sind mangels zuverlässiger Vergleichskriterien generell von der Dublettenprüfung ausgenommen und werden Neuaufnahmen.

Ausnahme:

Enthalten Bandsätze und unselbstständige Werke in den gelieferten Daten eine gültige BVB-Identifikationsnummer (BV-Nummer) von B3Kat-Dubletten des gleichen Kennbuchstabens, so werden sie auf die entsprechenden B3Kat-Datensätze zusammengeführt. (Zusammenführung über BV-Nummer s.u.)

3.3. Match-Kriterien

Feld-für-Feld-Abprüfung

Für die Dublettenprüfung werden pro Kennbuchstaben die in der folgenden Übersicht aufgeführten ASEQ-Kategorien herangezogen, die nach individuell festgelegten Kriterien verglichen werden. Die Match-Kriterien wurden in den vergangenen Jahren immer weiter verfeinert, um einerseits auch bei nicht zeichengleicher Erfassung gleicher Inhalte diese als dublett zu erkennen, andererseits aber falsche Zusammenführungen zu verhindern.

| Kategorie | Beschreibung | Kennbuchstaben |
|------------------|---|-----------------------|
| 025z | Überregionale Identifikationsnummer der ZDB | E |
| 026_ | Überregionale Identifikationsnummer | M, G, S, U, B |
| 089 | Bandangabe der Bandaufführung in Vorlageform | B |
| 100/104 | Ansetzungsform der 1./2. Person | M, E, G, S, U, B |
| 200 | Ansetzungsform der 1. Körperschaft | M, E, G, S, U, B |
| 310 | Ansetzungssachtitel | M, E, G, S, U, B |
| 331 | Hauptsachtitel | M, E, G, S, U, B |
| 425 | Erscheinungsjahr | M, E, G, S, U, B |
| 451 | 1. Gesamttitel in Vorlageform | S |
| 540 | ISBN | E, G, S, U |
| 542 | ISSN | E, G, S |
| 599 | Identifikationsnummer des selbstständigen Werks | U |

Inhaltlich als nicht dublett erkannte Felder erhalten eine Wertung. Die Summe der Werte aller abgeprüften Felder einer Titelaufnahme entscheidet über das Match-Ergebnis: „zusammenführen“ oder „nicht zusammenführen“.

In die Gesamtbewertung eines Dublettenabgleiches fließt auch ein, ob ein abzuprüfendes Feld in einer der Vergleichsaufnahmen oder in beiden Vergleichsaufnahmen nicht vorhanden ist.

Zusammenführung über die BVB-Identifikationsnummer (Aleph-Feld 001, 026d) bzw. über die ZDB-Identifikationsnummer (Aleph-Feld 025z)

Bereits in der Verbunddatenbank erfasste Titelaufnahmen der Bibliothek, die Teil der Datenerlieferung sind und somit beim Match&Merge gegen sich selbst dublettengprüft werden, werden über die BV-Nummer im Aleph-Feld 001 auf sich selbst zusammengeführt.

Bei „Altaufnahmen“, die noch nicht in der Verbunddatenbank erfasst sind, kann die Bibliothek im Vorfeld der Datenmigration selbst intellektuell in der Verbunddatenbank nach Dubletten suchen und bei erfolgreicher Suche in ihre Titelaufnahmen in einem gesonderten Feld (Aleph-Feld 026d) die Identifikationsnummern der passenden B3Kat-Titel erfassen.¹

In diesem Fall – und bei gleichzeitig passendem Kennbuchstaben - wird beim Match&Merge ausschließlich, d.h. ohne weitere Prüfung, über dieses Kriterium zusammengeführt. Dabei wird vorausgesetzt, dass die eingetragenen IDs korrekt und aktuell sind. Stimmen die Kennbuchstaben nicht überein, wird trotz eingetragener BV-Nummern nicht zusammengeführt. Enthält z.B. ein gelieferter Bandsatz (Kennbuchstabe B) eine BV-Nummer, die sich auf einen Stücker Titel im B3Kat (Kennbuchstabe S) bezieht, so findet keine Zusammenführung statt.

Erfasst die Bibliothek ihre Zeitschriften bereits in der ZDB, enthalten die Zeitschriften-Titelaufnahmen die ZDB-ID und können über diese beim Match&Merge-Lauf direkt auf die B3Kat-ZDB-Aufnahmen zusammengeführt werden.

¹ Hinweis: Findet die gelieferte Identifikationsnummer keine Entsprechung im B3Kat wird sie nach dem Match&Merge-Durchlauf gelöscht.

Hat die Bibliothek bisher noch keine Zeitschriften in der ZDB erfasst, kann sie im Vorfeld der Datenmigration die Zeitschriften in der ZDB erfassen oder ihren Bestand zu bestehenden ZDB-Aufnahmen an die ZDB melden und in ihre lokal erfassten Zeitschriftenaufnahmen in einem gesonderten Feld (Aleph-Feld 025z) die ZDB-IDs eintragen. Auch in diesem Fall werden die gelieferten Titelaufnahmen beim Match&Merge - gleicher Kennbuchstabe vorausgesetzt - ausschließlich über Feld 025z auf die zugehörigen B3Kat-ZDB-Datensätze zusammengeführt.

Kennt das Lokalsystem der Bibliothek keine dem vom BVB betriebenen Aleph-System entsprechenden Satztypen-Einteilungen, muss die Bibliothek – ggf. in Absprache mit der beauftragten Firma – Merkmale aus dem Satzformat festlegen, anhand derer die in Punkt 3.1 genannten Satztypen unterschieden werden können.

3.4. Match-Filter

Bei der Dublettenprüfung werden folgende Aufnahmen aus dem B3Kat generell nicht berücksichtigt:

- Titelaufnahmen mit ausschließlich Großschreibung der Felder 331 und 410
- ZDB-Serien (Serien werden primär im B3Kat erfasst, ZDB-Serien werden für die Katalogisierung nicht berücksichtigt.)

Folgende Filter können außerdem bei Bedarf verwendet werden:

- Titelaufnahmen eines bestimmten Zeitraumes (Intervall für das Erscheinungsjahr in Aleph-Feld 425a)

3.5. Gewinner-Aufnahme

Eine gelieferte Titelaufnahme wird bei erkannter B3Kat-Dublette immer vollständig auf diese zusammengeführt. Die Feldinhalte werden dabei nicht gemergt. D.h. auch einzelne Felder der gelieferten Titelaufnahme, die in der B3Kat-Dublette noch nicht enthalten waren, können in der Regel nicht erhalten bleiben. (In Ausnahmefällen können zusätzliche Felder aus der gelieferten Titelaufnahme im Nachgang zur Datenmigration in die Gewinneraufnahme nachgeführt werden).

Gibt es mehrere Treffer-Titelaufnahmen im B3Kat, so wird auf die Verbund-Titelaufnahme zusammengeführt, mit der die meisten Lokalsätze verknüpft sind. Wahlweise kann auch auf die Prüfung der verknüpften Lokalsätze verzichtet werden. In diesem Fall wird auf die Titelaufnahme mit der kleinsten BV-Nummer zusammengeführt.

3.6. Lokale Identifikationsnummern nach Match&Merge

Lokale Identifikationsnummern werden bei der Dublettenzusammenführung gegen BVB-Identifikationsnummern ausgetauscht. Sie bleiben jedoch in einem anderen Feld (Aleph-Feld 027) für den Datenaustausch beim Import der Daten ins Lokalsystem erhalten.