Ein neuer Verbundkatalog entsteht

Praxisbericht zur Dedoublierung und Migration der bibliografischen Daten für SLSP

Silvia Witzig

Eine von vielen großen Herausforderungen im Projekt SLSP war die Zusammenführung der bibliografischen Daten aus den sechs Verbünden, die zu SLSP migrierten.[1] Einerseits mussten Aufnahmen für identische Ressourcen erkannt und ohne Datenverlust zusammengeführt werden und andererseits sollte die Migration zur zentralen Bereinigung von Altdaten und Harmonisierung im Hinblick auf die gemeinsame Katalogisierung in SLSP genutzt werden. Diese Aufgaben wurden im Auftrag von SLSP von der Universitätsbibliothek Basel übernommen.

Die UB Basel betrieb von 2009 bis 2021 swissbib als Metakatalog und Datenhub aller Schweizer Hochschulbibliotheken, der Nationalbibliothek, zahlreicher Kantonsbibliotheken und weiterer Institutionen. In swissbib wurden bibliografische Daten aus 28 Quellen[2] zusammengeführt. Entsprechend waren die Datennormalisierung und insbesondere die Dedoublierung für swissbib von Projektbeginn an ein zentrales Thema. Der Datenhub von swissbib basierte mit CBS[3] von OCLC auf einer Softwarekomponente, die sehr ausgereifte und differenzierte Möglichkeiten zum Clustering von Daten bietet. Für swissbib wurden diese Algorithmen auf die Daten der Schweizer Bibliotheken angewendet und deren Konfiguration kontinuierlich verbessert. Dadurch ist im swissbib-Team eine sehr gute Kenntnis dieser Daten und viel Erfahrung in der Arbeit damit vorhanden. Die Zusammenarbeit im Migrationsprojekt ermöglichte es, dieses Wissen in SLSP einfließen zu lassen.

Migration der bibliografischen Daten

SLSP betreibt swisscovery, ein Alma-System mit einer zentralen Network Zone (NZ) sowie 29 Institution Zones (IZ).[4] Die NZ enthält den Verbundkatalog von SLSP, in dem die bibliografischen Metadaten aller SLSP-Bibliotheken verwaltet werden.[5] Für das Migrationsprojekt bedeutete diese Topologie, dass Daten aus den verschiedenen Verbundsystemen einerseits in die geteilte NZ und andererseits in die einzelnen IZs migriert werden mussten.

Die bibliografischen Daten wurden über bereits für die Services von swissbib bestehende Workflows aus den Verbundsystemen exportiert, von der UB Basel verarbeitet und anschließend Ex Libris zum Import in die NZ zur Verfügung gestellt. Gleichzeitig wurden die bibliografischen Daten zusammen mit allen lokalen Daten[6] über die üblichen Migrationspfade von Ex Libris in die jeweilige IZ migriert. Dabei wurden Daten aus dem gleichen Verbund in verschiedene IZs aufgeteilt, gleichzeitig wurden gewisse IZs aus mehr als einem Verbundsystem gespeist. Anschließend wurden die bibliografischen Daten in der IZ mit den dedoublierten Daten in der NZ über die Systemnummer aus dem ehemaligen Verbundsystem, die sowohl in der IZ als auch in der NZ in Feld 035 (MARC21) enthalten war, verknüpft.

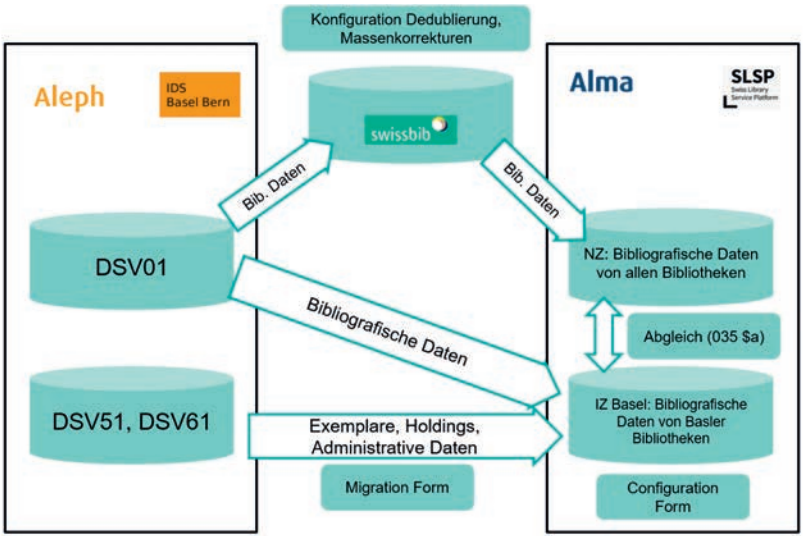


Illustration zur Migration der Daten aus dem IDS Basel Bern zu SLSP

1 Dabei handelt es sich um die im Informationsverbund Deutschschweiz (IDS) lose organisierten Verbünde IDS Basel Bern, IDS Luzern, IDS St. Gallen und NEBIS sowie um jeweils einen Teil der Bibliotheken aus RERO (Réseau des bibliothèques de Suisse occidentale) und aus Sbt (Sistema bibliotecario ticinese). Die Verbünde betrieben vor der Migration zu SLSP separate Bibliotheksysteme auf der Basis von Aleph (IDS-Verbünde und Sbt) bzw. Virtua (RERO).

2 Bibliotheksverbünde und Repositories von Hochschulen, vgl. https://ub-basel.atlassian.net/wiki/spaces/SWISSBIB/pages/1923186689/Participating+networks+and+repositories (eingesehen am 21.10.2021).

3 Central Bibliographic System

4 Stand Oktober 2021, vgl. https://registration.slsp.ch/libraries/?lang=de (eingesehen am 28.10.2021).

5 https://slsp.ch/de/news/20181207-2 (eingesehen am 14.10.2021).

6 Holdings, Exemplare und Erwerbungsdaten.

Die Migration zu Alma wurde in drei Testmigrationen ausführlich getestet und produktiv im November/Dezember 2020 durchgeführt. Die Testmigrationen wurden einerseits zur Prüfung der migrierten Daten, aber auch für Funktionstests von Alma genutzt. Die Datenverarbeitung der UB Basel wurde ebenfalls als Teil der Testmigrationen getestet. Ergänzt wurden diese durch separate Testmöglichkeiten außerhalb von Alma für das Testen der Dedoublierung.

Kriterien für die Dedoublierung

Bei der produktiven Migration wurden aus den sechs Verbundsystemen insgesamt 21.203.655 Aufnahmen an swissbib geliefert. Nach der Dedoublierung wurden 16.144.919 Aufnahmen an Ex Libris abgeliefert und in die NZ migriert. Dabei handelt es sich um 3.282.111 Aufnahmen, die aus 8.340.847 Aufnahmen zusammengeführt wurden, sowie um 12.862.808 Aufnahmen, die nicht dedoubliert werden konnten.

Die Algorithmen von CBS sehen ein Clustering unter Verwendung unterschiedlicher Kriterien mit wählbarer Gewichtung vor. Die für swissbib erarbeiteten Kriterien und die damit gemachten Erfahrungen dienten als Basis und wurden im Verlauf des Projekts auf die zu migrierenden Daten abgestimmt und optimiert. Die Vorgabe von SLSP war, möglichst viele Aufnahmen korrekt zu dedoublieren, aber im Zweifelsfall eher auf eine Zusammenführung zu verzichten, als falsch dedoublierte Aufnahmen in Kauf zu nehmen.

Bei den zu migrierenden Aufnahmen handelt es sich um bibliografische Daten für Publikationen von Mitte des 15. Jahrhunderts bis heute, die in zwei Generationen von Vorsystemen nach unterschiedlichen Regelwerken erfasst wurden. Grundsätzlich kamen alle gelieferten Aufnahmen für die Dedoublierung in Frage. Aktiv von der Dedoublierung ausgeschlossen wurden nur exemplarspezifische Aufnahmen sowie Aufnahmen ohne gültige Publikationsjahre. Zudem bestand die Möglichkeit, Aufnahmen im Quellsystem mit einem Code zu kennzeichnen und damit von der Dedoublierung auszuschließen. Von dieser Möglichkeit wurde bei problematischen Fällen, die bei den Tests aufgefallen sind, Gebrauch gemacht.

Entsprechend musste die Dedoublierung mit einem heterogenen Datenset umgehen. Dazu war ein großes Set von Kriterien zum Vergleich der Aufnahmen sinnvoll. Eine nur auf Identifikatoren basierende Dedoublierung war keine Option, da nur in rund 40% der Aufnahmen eine ISBN oder ISSN vorhanden war. Wie auch in swissbib wurde deshalb mit diversen Kriterien gearbeitet. Neben den zu erwartenden Kriterien wie ISBN, ISSN und weiteren Identifikatoren sowie Titel, Autoren und Erscheinungsjahr wurden weitere Kriterien wie z.B. Seitenzahlen, Bandnummern und Verlagsangaben hinzugezogen. Die Kriterien wurden zum Teil stark normalisiert, um eine bessere Vergleichbarkeit zu erreichen. Neben den auf alle Publikationsformen anwendbaren Kriterien galt es z.B. für Karten und Musikdrucke auch weitere Angaben zu berücksichtigen. Um die Dedoublierung für diese Materialien zu verbessern, waren u.a. auch Koordinaten und die musikalische Ausgabeform Teil der Kriterien.

Besondere Beachtung wurde den sehr rudimentären Aufnahmen, die nur wenige verwendbare Informationen enthalten, geschenkt. Diese mussten in den vorhandenen Elementen gesamthaft eine höhere Übereinstimmung haben, damit sie mit anderen Aufnahmen dedoublieren konnten.

Im Verlauf des Projekts wurden diverse Kriterien geprüft, einige davon wurden integriert, andere mussten aufgrund der heterogenen Datenqualität und entsprechend negativer Auswirkungen auf die Dedoublierung wieder verworfen werden.

Nachdem zwei oder mehr Aufnahmen als Dubletten identifiziert wurden, wurden diese zusammengeführt. Ziel war dabei einerseits die Dopplung von Angaben zu vermeiden und andererseits keine Informationen zu verlieren. Nur teilweise lösbar war dies in Fällen, wo semantisch gleiche Informationen auf unterschiedliche Art erfasst waren, sei dies mit unterschiedlichen Formulierungen, z.B. in Fußnoten, oder in einer anderen Sprache.

Herausforderung Mehrsprachigkeit

Im Migrationsprojekt wurden die Daten von Bibliotheken aus sechs Verbünden und aus drei Sprachregionen der Schweiz migriert. Die Mehrsprachigkeit und heterogene Anwendung von Katalogisierungsregeln in den Regionen war und ist auf vielen Ebenen eine Herausforderung für SLSP.[7]

Zielformat für SLSP als auch Ausgangsformat für die Migration war MARC21. Allerdings gab es in den Quellsystemen unterschiedliche Formatvarianten, insbesondere durch das in den IDS-Verbünden und Sbt verwendete Format IDSMARC[8]. Bei der Migration der bibliografischen Daten sollten die heterogenen Daten soweit möglich aneinander angeglichen und auf die Erschließungspraxis in SLSP vorbereitet werden.

7 Marty, Thomas und Küssow, Jürgen. «swisscovery – eine neue nationale Plattform vernetzt die wissenschaftlichen Bibliotheken der Schweiz» b.i.t. online, 3, 2021, pp. 305-307.

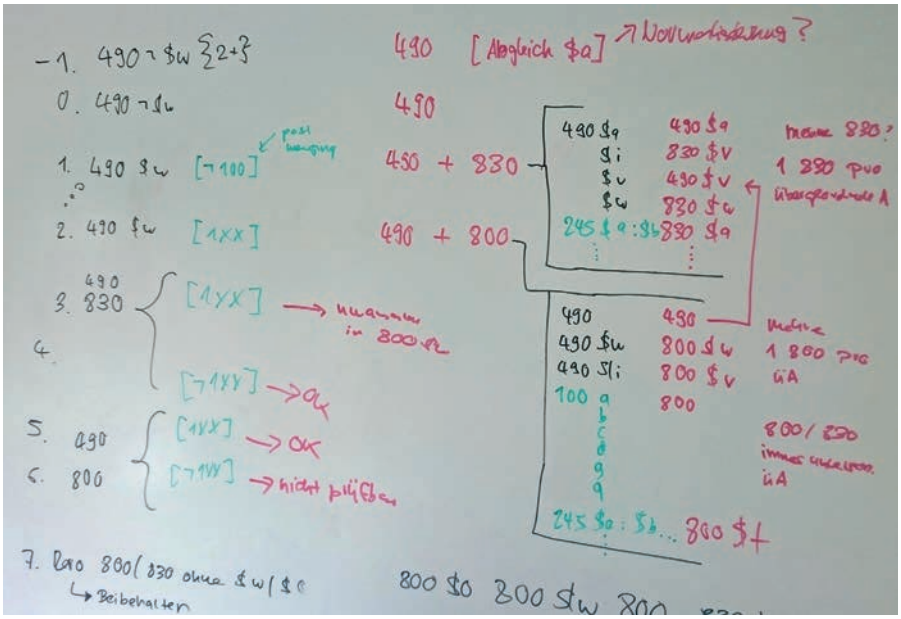
8 IDSMARC wurde in den IDS-Verbünden bereits 2016 mit der Einführung von RDA zum größten Teil an MARC21 angeglichen. Einige Spezifika blieben aber bis zur Migration zu SLSP erhalten.

Beispiel für die Verknüpfung von Sucheinstiegen mit 880 (vollständige Aufnahme: https://swisscovery.slsp.ch/discovery/sourceRecord?vid=41SLSP_NETWORK:VU1_UNION&docId=alma991050216219705501)

```
700    1#$aSchib, Karl $d1898-1984 $0(DE-588)118754858 $6880-02
880    1#$6700-02 $aSchib, Karl $d1898-... $0(IDREF)07608339X
700    1#$aAmmann, Hektor $d1894-1967 $0(DE-588)118648810 $6880-03
880    1#$6700-03 $aAmmann, Hektor $d1894-1967 $0(IDREF)075045222
```

Ziel war, so viele Korrekturen wie möglich zentral bei der Datenverarbeitung durch die UB Basel durchzuführen, in den Verbünden fanden nur verbundspezifische Korrekturen statt.

Da die Katalogisierung in SLSP mehrsprachig und unter Verwendung von sprachspezifischen Normdateien erfolgt, war der Umgang mit den Sucheinstiegen ein wichtiges Thema. SLSP entschied, Feld 880 (MARC21) für die Verknüpfung einer bereits erfassten Entität mit weiteren Normdateien zu verwenden. Zur Vorbereitung dafür wurde bei der Migration ein Abgleich der Sucheinstige implementiert, der bei Übereinstimmung ein Feld 880 generierte.



Konzeptskizze für die Migration und Massenkorrektur der Reihenverknüpfungen

Weitere Massenkorrekturen betrafen die Umstellung von IDSMARC auf MARC21, wofür bestimmte Felder anders strukturiert oder in ein anderes Feld übertragen werden mussten. Ein komplexes Beispiel dafür ist das Feld für Reihenverknüpfungen. Im IDS wurde nur Feld 490 für die Verknüpfung mit übergeordneten Aufnahmen verwendet und nicht die nach MARC21 vorgesehenen Felder 800, 810, 811 und 830. Hier galt es das fehlende Feld basierend auf der übergeordneten Aufnahme zu ergänzen und mit den existierenden Feldern aus RERO zu kombinieren.[9]

Fazit

Mit dem Entscheid, die Dedoublierung und die Massenkorrekturen für die Migration der bibliografischen Daten mit den Mitteln von swissbib durchzuführen, war es möglich, auf die speziellen Anforderungen und Bedürfnisse der Schweizer Bibliotheken einzugehen und die Datenverarbeitung in Zusammenarbeit mit den Expertinnen und Experten bei SLSP und in den Bibliotheken abzustimmen.

Durch die regelmäßigen Tests konnten Problembereiche im Projektverlauf identifiziert und mittels Verbesserungen bei der Verarbeitung oder Maßnahmen in den Verbünden adressiert werden. Im Austausch entstand ein Verständnis für die Datenverarbeitungsprozesse und Wissen über Vor- und Nachteile sowie über Probleme, die händisch zu bearbeiten sind.

Auch wenn ein maschinelles Verfahren nicht alle Dubletten erkennen kann und Daten weiter zu bereinigen sind, konnte mit der intensiven Arbeit an den bibliografischen Daten bei der Migration eine gute Basis für den Start von SLSP und die gemeinsame Katalogisierung geschaffen werden. ❙

Silvia Witzig
Metadatenspezialistin,
Universitätsbibliothek,
Universität Basel
silvia.witzig@unibas.ch

9 Dazu kam nach der Zusammenführung der Aufnahmen ein auf Apache Flink basierender Workflow (vgl. https://gitlab.com/swissbib/slsp/series-transformation/volumes-series-enrichment-flink) zum Einsatz.