Recommendations of the Primo FRBR / Dedupe advisory group

December 2019

At the request of the ELUNA and IGELU Primo working groups, this group was formed with the following goals:

- Define the optimal default OTB configuration for FRBR and Dedup. This group will also discuss needs for additional opt-in options, including strict / loose dedup, or additional match points that can be enabled or disabled.
- Define the elements for a "super record." To improve consistency in user experience, the preferred record should not change dynamically. Consider selection and display of a VSR i.e. “Virtual Super Record”, as for Primo Central Index.
- For VE, discuss deduplication with external data sources, to achieve feature parity with Primo BO.

Scope: Both Primo VE and Primo BO. Unless otherwise specified, “Primo” refers to the discovery interface regardless of the deployment model (VE vs BO).

NOTE: deduplication / FRBR with CDI is not the focus of this group at this time. Deduplication / FRBR with CDI should be revisited by relevant PWGs after CDI has been rolled out widely.

Please see the appendix for a list of report contributors.

⇒ This report is intended to provide guidance to Ex Libris. Any specific enhancement requests should be made in NERS. Any requests already entered into NERS for the 2020 ballot have been noted below (as of December 3rd, 2019). REMINDER: Deadline for 2020 requests is January 10, 2020. For issues important to your institution, you are encouraged to add requests to NERS for any issues noted below that are not already in NERS for the 2020 ballot.

FRBR

UI behavior

- Add *static* preferred record presentation
    - That is, it should be possible to configure which record is used as preferred regardless of search query / ranking score
    - The current *dynamic* preferred record presentation option in Primo is problematic for 2 reasons:
        - Not all users see the same preferred records which causes confusion. It is especially challenging for research librarians to work with patrons when they each see something different
        - It's a serious problem that use of the preferred record presentation does not calculate facets for non-preferred records in the cluster
    - Algorithm for preference of *static* record presentation should be (in this order):
        - Prefer latest record (based on pub date in 008)
            - Development for this is already committed via NERS 2019 request # 6230: *Enable option to favour latest edition of bibliographic work to top the FRBR group*
        - Prefer specific data source based on configuration option
            - For example, customer should be able to prefer Alma IZ, then Alma CZ, then specific external source, etc.
        - Prefer online, then print, then microform
            - Currently there is no way to distinguish print from microform because they are both calculated as Physical
    - The search result entry for the FRBR group should have these elements, based on the preferred record:
        - Resource type icon
        - Cover image
        - Data elements based on View configuration for Brief Display
- When the search result contains a group, **all records in the FRBR group** should be accounted for in the facets. (This is not happening in the current UI behavior).

- Physical inventory should continue to display in the FRBR group brief record. Furthermore, if a member of the group has online availability, the online link should also display in the brief record, or there should be an indicator that the group contains online content.
    - If this is not possible, it would be preferable to not display any inventory in the brief record.
- After group expansion, record order in the list should use same algorithm as above

⇒ NERS 2020 ballot request # 6583: FRBR group static preferred record presentation

FRBR OTB settings for new customers

- Note: for existing customers, their active settings should be retained. The recommendations below are opt-in for existing customers
- FRBR is off by default. If switched on, defaults will be as follows
- Clustering will apply to books and serials only
    - That is, books can cluster with other books, and serials can cluster with other serials. Books and serials do not cluster with each other
    - Other formats are often too complex to include as OTB defaults, example: TV programs
- Customer should be offered 2 profiles that control the level of clustering most appropriate for their institution
    - **Strict cluster**: include all formats (print/electronic/microform/digital) of a single resource based on identifiers
        - Different editions will not cluster in this model
        - Book data elements (normalized): 020a, 776z, 035a(OCoLC), 776w(OCoLC)
        - Serial data elements (normalized): 022a, 022l, 775x, 776x, 035a(OCoLC), 776w(OCoLC)

    - **Broad cluster (default)**: include all editions and formats (print/electronic/microform/digital) of a book or serial. Consider all identifiers above, in addition to:
        - Books: 1XX, 240, 245
        - Serials: 1XX, 222, 245
    - When a record is too minimal to successfully group, it should not be considered for any FRBR group. In this case minimal is defined as having a 245 and 008 Date and no other data elements.
    - *Use case:* a CZ bib has 008 Date of 1999 and a 245: *Best loved poems*. This minimal bib should not group with the print and online version of: Lyons, Liam, and Thomas F. Walsh. *Best Loved Poems : Favourite Poems from the West of Ireland*. Currach Press, 2013.
    - Records which fail to dedup merge with print or online equivalent, should still be considered for the same FRBR cluster if it meets a match point. (E.g. case 00711091)

FRBR Advanced config options

- Ability for a customer to define clustering rules using resource type as an element that is taken into consideration
    - *Example*: customer wants musical scores to cluster together (and only scores, not sound recordings, etc.) based on uniform title + author
- Ability for customer to *ignore* resource types during clustering and look at other data elements only
    - *Example*: customer wants a thesis to cluster with its corresponding CD-ROM, based on title + author data.
    - *Example*: customer wants to be able to cluster a newspaper archive database with records for related newspapers e.g. The Economist Historical Archive (1843-2014) with The Economist (1845-)
- Continued ability to prevent clustering for certain records based on specific criteria
    - (Already possible in BO and VE)
    - *Example*: records with "Bible" in uniform title

- *Example*: all portfolios from a certain collection because the portfolios are so similar they cluster even though they shouldn't
- *Example: all records with physical inventory linked to a specific Library or holdings location (i.e. Special Collections)*

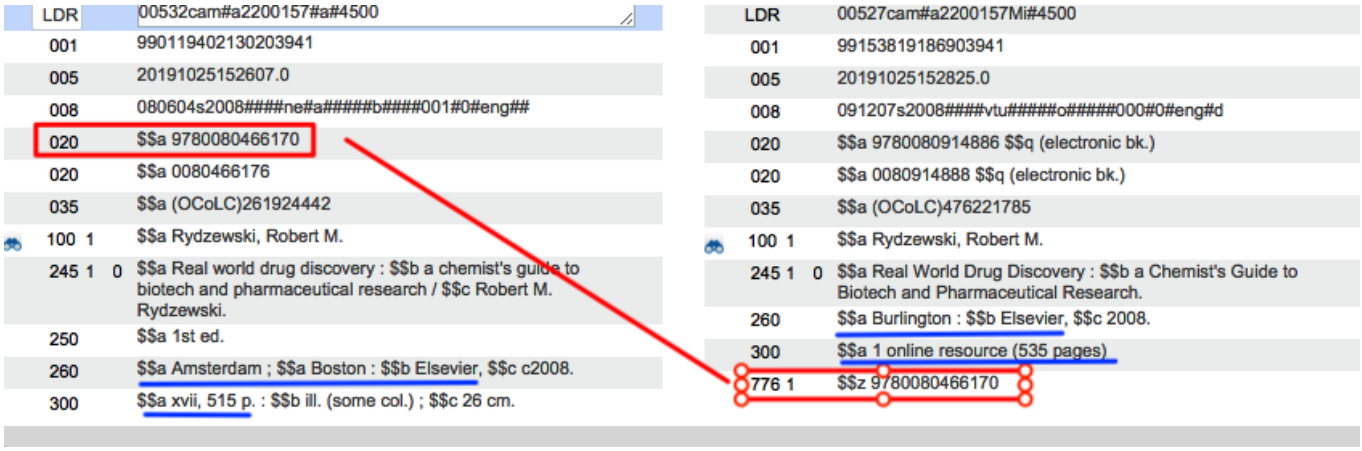
⇒ No NERS request needed?

Dedupe

OTB Dedupe

- Not making a recommendation about whether it should be on/off for new implementations. Case dependent. Should be part of implementation questionnaire.
- If turned on, existing OTB dedupe rules should also dedupe *Print , electronic, digital, microform* based on identifiers in linking fields ~~775/~~776 should be used to identify these corresponding records
    - Use subfield w (OCoLC) to match to 035 (OCoLC)
        - National libraries use different prefixes here so the system must support other codes as well. What is important is that the matching be code specific -- that is, an OCLC number should not be compared to a number from a national library. Note WorldCat usage of $w: "In WorldCat, subfield ǂw may be used to cite record control numbers for the Library of Congress, OCLC, national libraries, and other national level bibliographic networks. Subfield ǂw may be repeated when multiple record control numbers apply."
    - Use ISBN in 776 $z and ISSN in 776 $x for matching, in addition to 020 $a, 022 $a $l.
    - Matches on these identifiers in linking fields should outweigh minor differences in subtitles and publication.
        - Ability for system to ignore *1 online resource (...)* for ebook item to create a pagination match with print equivalent in the 300 field (example below)

- Ensure that system can use both 260 and 264 to look for publisher value. (Older cataloging uses 260, newer cataloging uess 264)
- Example of two records that should dedupe despite minor differences in 260 and 300, based on linking identifiers



⇒ NERS request not yet added but encouraged for VE (if this turns out to be a parity issue no points may be needed)

Ability to exclude certain resource types from Dedupe

- *Problem:* Media formats represent different challenges for deduping
- *Recommendation:* ability for a site to choose to let books and serials dedupe, while excluding sound recordings, videos, etc. from deduping because of their different characteristics.
- *Use case:* a library catalogs each season of a TV show on separate records. Dedupe rules evaluate field 300 (pagination) to determine eligibility for dedupe, and seasons that happen to have the same run time (in field 300 for videos) are deduped while seasons with different run times are not. For example, seasons 2 and 3 dedupe, but not seasons 1 and 4. In this case, it would be preferable to not attempt to dedupe videos at all, but the library still wants to dedupe books/serials.

⇒ NERS request not yet added but encouraged for VE (if this turns out to be a parity issue no points may be needed)

Ability to Dedupe Alma records with External data (not including CDI which is outside the scope of this group)

- *Problem*: In VE it is not possible to dedupe Alma records with external data. (This is possible in BO)
- *Recommendation*: in VE, support ability to dedupe with external data.
- *Use case*: Library participates in a shared print collection with outside institutions. Library wishes to load external data for this shared print collection into Primo, and dedupe with local data to prevent redundant search results for users, as well as redundant search results in a FRBR cluster.
- *Use case:* Library has an external data source for finding aids. Each one corresponds to an Alma record. Library wants finding aid to be searchable as full text, and to dedupe with its corresponding MARC record. In Back Office, this is possible because the external data source has the MMS ID in an attribute of the XML header, which can be used as a data element for matching.

⇒ NERS request not yet added but encouraged for VE (not needed for BO) (if this turns out to be a parity issue no points may be needed)

Choosing preferred record during Dedupe

- *Problem*: need ability to choose preferred record
- *Recommendation*: Add ability to identify a preferred record based on format as well as data source (including external data)
    - For format, need the ability to distinguish between online, print, and microform. Currently print and microform are both calculated as Physical and there is no way to prefer print over microform.
- *Use case*: Library participates in a shared print collection with outside institutions. Library wishes to load external data for this shared print collection into Primo, and dedupe with

local data to prevent redundant search results for users, as well as redundant search results in a FRBR cluster. The library wants to always prefer its own print record.

- *Use case:* Library uses Primo as discovery interface for data from Alma and other external systems (archive system, digital repository), but always wants the Alma record to be the preferred record (if available).
- *Use case:* Library has separate records in Alma for print and microform. When these are for the same publication, they want to dedupe the records in Primo and prefer the print record.

⇒ NERS 2020 ballot request # 6584 : Dedupe preferred record and data elements

Data elements from Deduped-Merged record

- *Problem*: duplicate fields not being removed.
    - The system does not normalize the data before comparing for likeness, so you end up with duplicate elements where the only difference is punctuation or capitalization, e.g.:
        - <lds34>Everyman</lds34>
        - <lds34>Everyman.</lds34>
    - In other cases, the data is only slightly different, but the library would not wish to retain both fields, e.g.:
        - <lds14>By A. F. Sperry.</lds14>
        - <lds14>by A.F. Sperry.</lds14>
- *Recommendation*:
    - System should normalize data when doing field comparison
    - Library should be able to configure which local display fields from the non-preferred record will be retained

⇒ Combined with NERS 2020 ballot request # 6584 above

APPENDIX

Report contributors:

Name	Institution
Corinna Baksik (Lead)	Harvard University
Diana Jiang	CDC
Emma Booth	The University of Manchester Library
Helen Garner	Sheffield Hallam University
Jeff Peterson	University of Minnesota
Jen Froetschel	George Washington University
Leslie Engelson	Murray State University
Lukas Koster	Library of the University of Amsterdam
Michael Norman	University of Illinois at Urbana-Champaign
Peta Hopkins	Bond University
Richard Lee Guinn	Texas Medical Center Library
Sandra McKenzie	National Library of New Zealand
Sian Thomas	National Library of Wales
Stacey Beach	Southern Methodist University
Teressa Keenan	University of Montana