OCLC Online Computer Library Center, Inc.

FRBR Work-Set Algorithm

Thomas B. Hickey

Office of Research
OCLC Online Computer Library Center, Inc.
6565 Frantz Road
Dublin, Ohio 43017
hickey@oclc.org

Please address all correspondence to this author.

Jenny Toves

Office of Research
OCLC Online Computer Library Center, Inc.
6565 Frantz Road
Dublin, Ohio 43017
tovesj@oclc.org

April 2005

FRBR Work-Set Algorithm

The research work-set algorithm generates an author/title key for each bibliographic record. These keys can then be used to bring work-sets together. The current algorithm ignores format so that the generated work-sets are sometimes at a higher level than a FRBR work.

A work-set is a group of bibliographic records having the same author/title key, generated according to the algorithm in this paper.

Authors and author/titles that match variant headings in the authority file are changed to their established form. This means that the first step in building work-set author/title keys is to build author and author/title mappings from the LC Authority data.

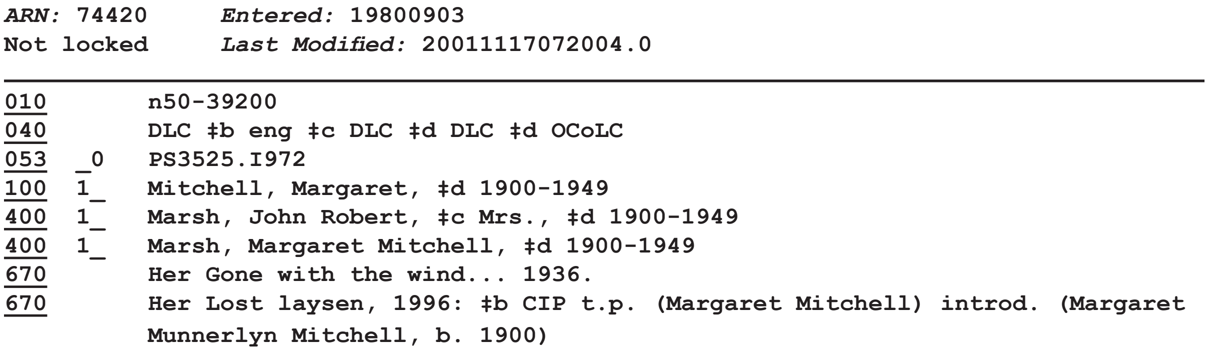
Note: All keys are normalized using the NACO rules (http://www.loc.gov/catdir/pcc/naco/normrule.html). When keys are extracted from fields, it is implied that normalization is also performed. We will only specify normalization occurring in the section dealing with cleaning title keys. OCLC Research has a web page available for experimenting with the NACO normalization rules (http://www.oclc.org/research/researchworks/naco/default.htm).

Note: A list of subfield codes between brackets, like this **'[tn]'**, means to use any or all subfields matching those codes.

Note: We use a slash ('/') as the separator between key parts. The examples show a slash but the choice of exactly how the key appears is an implementation choice.

Authority Mappings

Author Mappings link variant forms of names to the established form of a name. For example, the record for "Mitchell, Margaret, $d 1900-1949" (n50-39200) will link 2 entries in the 400 fields to the established form in the 100 field.

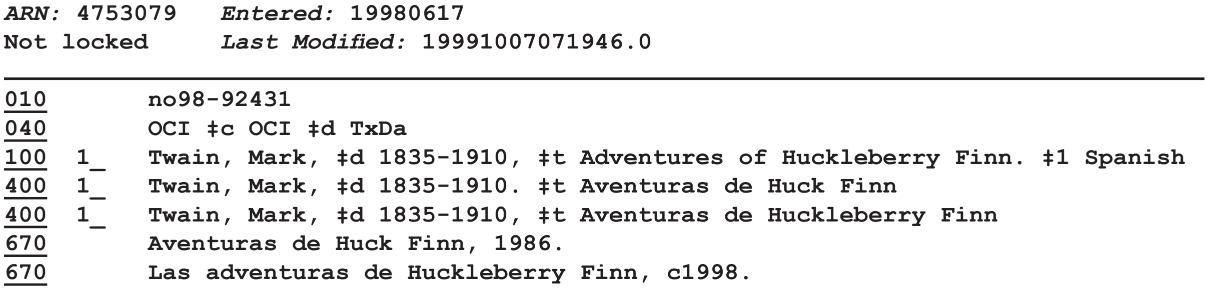


```
ARN: 74420      Entered: 19800903
Not locked      Last Modified: 20011117072004.0

010       n50-39200
040       DLC ‡b eng ‡c DLC ‡d DLC ‡d OCoLC
053  _0   PS3525.I972
100  1_   Mitchell, Margaret, ‡d 1900-1949
400  1_   Marsh, John Robert, ‡c Mrs., ‡d 1900-1949
400  1_   Marsh, Margaret Mitchell, ‡d 1900-1949
670       Her Gone with the wind... 1936.
670       Her Lost laysen, 1996: ‡b CIP t.p. (Margaret Mitchell) introd. (Margaret
          Munnerlyn Mitchell, b. 1900)
```

Figure 1 - A Margaret Mitchell Record (Screen shot from Connexion)

Author/Title mappings link established authors paired with variant forms of titles to the established form of a title. Bibliographic data must match the author/title pair exactly in order to map to an established title. For example, the record for "Twain, Mark, $d 1835-1910. $t Adventures of Huckleberry Finn. $l Spanish" (no98-92431) will link 2 entries in the 400 fields to the established form in the 100 field.



```
ARN: 4753079   Entered: 19980617
Not locked     Last Modified: 19991007071946.0
____________________________________________________________________________________
010        no98-92431
040        OCI ‡c OCI ‡d TxDa
100  1_    Twain, Mark, ‡d 1835-1910, ‡t Adventures of Huckleberry Finn. ‡1 Spanish
400  1_    Twain, Mark, ‡d 1835-1910. ‡t Aventuras de Huck Finn
400  1_    Twain, Mark, ‡d 1835-1910, ‡t Aventuras de Huckleberry Finn
670        Aventuras de Huck Finn, 1986.
670        Las adventuras de Huckleberry Finn, c1998.
```

Figure 2 - A Mark Twain Record (Screen shot from Connexion)

Constructing Authority Indexes

Author

1. Construct the established form
    a. Skip records with no tag 100
    b. Skip records with a tag 100 with subfields [tnmpr] (title subfields)
    c. Skip records where field 008, byte 9 is not 'a' or 'f' (use only established headings)
    d. Skip records where field 008, byte 15 is 'a' and byte 11 is not 'a' (if the field is appropriate for use as a subject heading then it must specify LCSH)
    e. Build the key from tag 100, subfields [abcdq]

2. Construct the variant forms
    a. Skip records where field 008, byte 14 is not 'a' (records that are not an established form)
    b. Build the key from tag 400 (field is repeatable), subfields [abcdq]
    c. Save the mapping if the key does not match the established form

Author/Title

1. Construct the established form
    a. Skip records missing either an author or a title
    b. Skip records where field 008, byte 9 is not 'a' or 'f' (use only established headings)
    c. Skip records where field 008, byte 15 is 'a' and byte 11 is not 'a' (if the field is appropriate for use as a subject heading then it must specify LCSH)
    d. Construct the title from tag 100, subfields [tmnpr], or tag 130, subfield [amnpr]
    e. Construct the author from tag 100, subfields [abcdq]
    f. Build the key as ‹author›/‹title›

2. Construct the variant forms
    a. Skip fields with tag 400 and no subfields [tmnpr]
    b. Build the title portion of the key from tag 400 (field is repeatable), subfields [tmnpr]
    c. Build the author portion of the key from the same tag 400, subfields [abcdq]
    d. Save the mapping of ‹author›/‹title› if the key does not match the established form

Matching Authority Indexes

Author

There are multiple possible ways to match a name from a bibliographic record to a variant name in the authority file author mapping.

1. The first check is for an exact match.

   Example: "Marsh, John Robert $c Mrs $d 1900-1949" is an exact match and links to the established form of "Mitchell, Margaret $d 1900-1949".

2. If multiple exact matches are found then pick the established form that is used most frequently. We define "used" as the number of records in WorldCat with that author plus the square root of the number of holdings attached to those records. "Most used" is the author with the highest "used" score.

3. Last, if the name from the bibliographic record looks like it might contain a date outside of subfield 'd' then add a subfield marker in front of the date in the name from the bibliographic record and repeat previous steps.

   Example: "Marsh, John Robert $c Mrs 1900-1949" can be changed to "Marsh, John Robert $c Mrs $d 1900-1949".

Author/Title

Titles undergo up to five manipulations in an attempt to find an exact match in the authority indexes. The bibliographic title key generation produces two possible titles. The short title and the full title are used in an attempt to match an authority index entry. The manipulations are performed in the given order and stop when a match is successful.

1. The author combined with the short title is looked up in the authority author/title index.

2. If the short title starts with:

   <author's last name><a blank or slash>
   <author's last name>s<a blank or slash>
   <author's first name><blank><author's last name><blank or slash>
   <author's first name><blank><author's last name>s<blank or slash>

   and the author's name is not the entire title, then remove the author's name(s). Check if the author combined with the short name with the surname pattern applied is a match.

3. The author combined with the full title is checked for a match.

4. The author combined with the full title with the surname patterns applied is checked for a match. If the full title with the surname patterns applied starts with "tragedy of", "tragedie of", "tragedia de", "comedy of", or "single plays", then delete the matched phrase plus any newly leading articles ("a", "an", or "the") and check for a match.

Work-Set Keys

There are three steps in building our FRBR Work-Set identifiers.

1. Construct the author portion
2. Construct the title portions
3. Combine the key parts

Constructing Key Parts

Author

1. Extract subfields [abcdq] from tag 100, 110 or 111
2. Look for the name in the Authority Author Mappings using the rules described in the section Matching Authority Indexes
3. If an established form was found then substitute the established form

Cleaning Title Keys

1. Use the skip indicator when it is available
2. If the field is a 130 or 240, convert to lowercase and strip trailing "english" or "english."
3. Delete all bracketed text from the title unless it would result in an empty key in which case only the brackets are removed

   Example: "March of the Bees $h [video] $b a long movie" becomes "March of the Bees $h $b a long movie". The empty subfield will be removed during normalization.

   "[March of the Bees] $h [video] $b [a long movie]" becomes "March of the Bees $h video $b a long movie" because removing all bracketed text would result in an empty key.

4. Normalize using NACO
5. Delete leading articles ("an", "the")
6. Delete leading & trailing blanks

Short Title

1. Extract a title from the first field in the following list
   - Field 130, subfields [amnpr]
   - Field 240, subfields [amnpr]
   - Field 242, 245, 246, or 247
     - If the language of the record (008 bytes 35-37) is not 'eng' then 246 is the preferred title field
     - If the record contains a 110 or no author (100 or 111) use subfields [abfgnp]
     - Else use subfield [a]
2. Clean the short title

Full Title

1. If the short title came from a 242, 245, 246 or 247 and only consists of subfield [a] then extract subfields [abfgnp] for the full title
2. Else the full title is the same as the short title
3. If the full title is empty then try to get one from 740 subfields [anp]
4. If the full title is still empty then try to get one from 245 subfields [k]
5. Clean the full title

Combining Key Parts

The goal is to create a key that can uniquely and confidently identify a work-set. The best cases occur when we have an author with a title or a solitary uniform title. If we don't have an author or a uniform title then we try to find name fields (7XX tags) to help identify related items. The processing to group related items is discussed in the section Post-Processing for Sets with No Author. Records that only have a 24X field (no 1XX or 7XX fields exist in the record) get combined with their control number to force unique keys. We cannot combine those matching titles since we don't have enough information to reliably group the items.

This gives us four possible key patterns. As with the Authority keys, the key parts are separated by a slash although that is an implementation detail. A recent run of 54,830,689 bibliographic records created the following counts for work set keys:

‹author›/‹title›: 41,654,567 (75.97%)
‹uniform title›: 736,020 (1.34 %)
/‹title›/[one or more ‹name›] 9,513,007 (17.35%)
/‹title›/‹control number› 2,927,095 (5.34%)

- If an author exists
    1. Look for the author and title in the Authority Author/Title Mappings as described in the Author/Title section of Matching Authority Indexes
    2. If a mapping was not found, use author + full title with the surname patterns applied as described in the Author/Title section of Matching Authority/Indexes
- Else if a 130 exists then the title alone is a sufficient key
- Else if 7XX (700, 710, 711) fields exist then add the names to the title. Skip 7XX fields with subfields [tk]. Use subfields [abcdq] as the name
- Else add the control number to force the title to be a unique key

Post-Processing for Sets with No Author

For each group of records with:

- a matching title portion and
- associated names instead of authors (has 7XX fields but no 1XX)

build sets of records that have overlapping name elements. Given the following items:

1. Title/Name A
2. Title/Name A/Name B
3. Title/Name B/Name C

4. Title/Name C
5. Title/Name D/Name E

We would create 2 sets. A most frequently occurring name can be used to create the work-set identifier for the set.

The first set would contain items 1, 2, 3 and 4. It could have an identifier of “title/name a”. Note that items 1 and 4 have no overlapping elements if considered by themselves but including items 2 & 3 allows us to pull the entire group together.

The second set contains item 5. It could have an identifier of “title/name d”.

Appendix A - Notes on Subfield Selections

The subfields and their meanings are taken from http://www.loc.gov/marc/authority and http://www.loc.gov/marc/bibliographic/ecbdhome.html.

Authority subfields used as a name are: 100/abcdq and 400/abcdq.

Authority subfields in a 100 that indicate a name-title are: [efhklmnoprst].

Authority subfields used for a title key: 130/amnpr, 100/tmnpr, 400/tmnr.

a=title, m=medium, n=number of a part, p=name of section, r=key, s=version

Bibliographic subfields used for a title are: 130/amnpr, 240/amnpr, 740/anp (740s do not have ‘m’ or ‘r’), [242|245|246|247]/abfgnp (242 does not have ‘f’ or ‘g’).

a=title, b=remainder of title, f=inclusive dates, g=bulk dates, n=number of a part, p=name of a part, k=form