FRBR Work-Set Algorithm

Version 2.0

Thomas B. Hickey
Chief Scientist

Jenny Toves
Software Architect

A publication of OCLC Research

FRBR Work-Set Algorithm, v. 2.0
Thomas B. Hickey and Jenny Toves, for OCLC Research
August 2009

OCLC (WorldCat): 645543531

Please direct correspondence to:
Thomas B. Hickey
Chief Scientist
hickey@oclc.org

Suggested citation:
Hickey, Thomas B. and Jenny Toves. 2009. FRBR Work-Set Algorithm. Version 2.0. Dublin, OH: OCLC Online Computer Library Center, Inc. (Research division). Published online at: http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf.

The research work-set algorithm generates a key for each bibliographic record. These FRBR 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 FRBR key, generated according to the algorithm in this paper.

Authors and titles that match variant headings in the mapping files are changed to their preferred form. This means that building the mapping files is a prerequisite for building FRBR keys.

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. OCLC Research has a web page available for experimenting with the NACO normalization rules (http://www.oclc.org/research/researchworks/naco/default.htm). In 2007, PCC expanded the NACO normalization rules to include Unicode. OCLC Research implemented these changes for FRBR processing but has not changed the NACO web page.

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.

Note: Subfields within a key portion are separated by a '\'. The choice of character here is an implementation detail. The subfield delimiter between subfields 'a' and 'b' in a title field is replaced with a space.

The Keys

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. Records that only have a 24X field (no 1XX or 7XX fields exist in the record) get combined with their Worldcat number to force unique keys. We cannot combine those matching titles since we don't have enough information to reliably group the items.

If an author exists, combine it with a title

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 oclc number to the title to make the key unique

This gives four possible key patterns. The key parts are separated by a slash although that is an implementation detail. A recent run of 138,649,513 bibliographic records created the following counts for work set keys:

Key Type	Occurrences	Example
<author>/<title>	97,961,220 (70.65%)	bjorling, jussi\1911 1960/opera arias and duets
<uniform title>	1,569,352 (1.13%)	10 commandments
/<title>/[one or more <name>]	26,559,404 (19.16%)	/bergler/bergler, friedrich
/<title>/<oclc number>	12,559,537 (9.06%)	/britain and antarctica/289903387

Building Author Portion

1. The author portion of a FRBR key is built from subfields [abcdq] from tag 100, 110, 111 field or an 880 field linked to a 100, 110, or 111
2. The normalized author is looked up in the mapping file and substitutions are made

Author Type	MARC Field	Normalized String
Mapped Personal Name	100 1 Twain, Mark, $d 1839-1918	twain, mark\1835 1910
Unmapped Personal Name	100 1 Twain, Mark, $d 1835-1910	twain, mark\1835 1910
Corporate Name	110 2 Friendship Village of Dublin (Dublin, Ohio)	friendship village of dublin dublin, ohio
Conference Name	111 2 Conference on a Century of Russian Foreign Policy $d (1961 : $c Yale University)	conference on a century of russian foreign policy\1961\yale university

Building Name Portions

1. The name portion(s) of a FRBR key is(are) built from subfields [abcdq] from all tags 700, 710, 711 or 880 linked to a 700, 710, 711 as long as the field does not have a [tk]

2. If no name or author has been identified in a record and a 720 without a [tk] exists then use the name in the 720

3. The normalized author is looked up in the mapping file and substitutions are made

Grouping Title/Name Keys

Title/Name keys must be grouped. The records are first grouped by title. Then records with intersecting sets of names are grouped within title. If record 1 contains names A and B and record 2 contains names B and C then records 1 and 2 will be grouped because of the overlapping names. The actual value assigned to the name portion of a title/name key could be any or all of the overlapping names as long as the groups are calculated on the total membership of all names.

Record Number	Names	Key
Record 1	"smith" and "doe"	/day in the park/smith
Record 2	"doe" and "jones"	/day in the park/smith
Record 3	"jones"	/day in the park/smith
Record 4	"harvey"	/day in the park/harvey

Building Uniform Title Portion

Uniform titles are built from 130 [amnpr]. The title is cleaned by skipping any characters indicated by a non-filing indicator, by processing bracketed text and by applying NACO normalization. If $a is entirely enclosed in brackets then the brackets are ignored. If the entire field is entirely enclosed in brackets then the brackets are ignored. Otherwise, text within brackets is deleted. If the cleaned title is in the list of noise titles then no uniform title is generated.

Building Title Portion

Titles are built from a list of possible title fields. The list is in order of preference and the first field to yield a usable title is the one used. Each field is cleaned and then checked against a list of noise titles. If the title cleans to nothing or is a noise title then the search continues with the next title field. If all titles are noise titles then the first noise title is used. If all titles are empty then the record number is used as the title.

Usable fields must contain at least one of the subfields in the “Must have” column. The used subfields are in the “Short title” and “Full title” columns. Where indicated, both a short title and a full title are constructed. If the short title matches a mapping then the mapping is used. Otherwise the full title is used for look ups in the mappings and is used as the title portion of the FRBR key.

Cleaning titles involves skipping any characters indicated by a non-filing indicator, by processing bracketed text and by applying NACO normalization. If $a is entirely enclosed in brackets then the brackets are ignored. If the entire field is entirely enclosed in brackets then the brackets are ignored. Otherwise, text within brackets is deleted.

Title Subfields	Normalized Title
$a [Let’s visit the school]	lets visit the school
$a Let’s visit the school $b [today]	lets visit the school
$a [Let’s visit the school $b today]	lets visit the school today
$a [Let’s visit the school $c today]	lets visit the school\today

Special Condition	Tag	Must have	Short Title	Full Title	Has Skip Indicator?
	240	[amnpr]		[amnpr]	i2
Language != ‘eng’	246	[abnp]	[a]	[abfgnp]	
	24[2567]	[abnp]	[a]	[abfgnp]	242 i2 245 i2
	740	[anp]		[anp]	i1
	245	[kfg]		[kfg]	i2
Linked to tag 240	880	[amnpr]		[amnpr]	i2
Language != ‘eng’ and linked to tag 246	880	[abnp]	[a]	[abfgnp]	
Linked to tag 24[2567]	880	[abnp]	[a]	[abfgnp]	242 i2 245 i2
Linked to tag 740	880	[anp]		[anp]	i1
Linked to tag 245	880	[kfg]		[kfg]	i2

Noise titles are: 'speeches', 'cantatas', 'choral music', 'constitution', 'chamber music', 'essays', 'operas', 'annual report', 'organ music', 'vocal music', 'plays', 'orchestra music', 'correspondence', 'instrumental music', 'short stories', 'piano music', 'treaties etc', 'songs', 'poems', 'laws etc', 'works', 'selections'.

Mapping Keys:

Various mappings are done on the parts of the FRBR key. It is possible for multiple mappings to apply to a single key as in the case where the author is mapped and then the author/title is mapped. The mapping files are derived from LCNAF and from data mining within Worldcat.

Author – If the normalized author or name from a 7xx is not found in the mapping file then the name is examined to see if it appears to end in a date that is missing a subfield delimiter. If the name ends in a blank plus digits then a subfield indicator is inserted before the digits. If the digits were preceded by ‘b’ or ‘d’ then that is included with the date. This edit is not remembered if no match was made in the mapping file.

Author/Title - > Author/Title combinations can map to another author/title or to a uniform title. If the first lookup fails, then a series of edits are attempted with look ups at each stage. If any edited look up matches, then the match is used. If all edits fail, then the title portion of the key is the final edited version of the full title.

Step	Key	Edit
1	Short title	Cleaned title
2	Short title	Strip leading articles
3	Short title	Surname pattern
4	Short title	Strip leading articles
5	Full title	Cleaned title
6	Full title	Strip leading articles
7	Full title	Surname pattern
8	Full title	Strip leading articles

Articles that are stripped from the front of a title are: 'the ', 'die ', 'la ', 'an ', 'der ', 'le ', 'das ', 'les ', 'de ', 'el '.

The surname pattern edit uses the first word in a 100 field as the surname. If surname ends in a comma then the next word is the forename. If there is only a surname then this rule attempts to remove “<surname> “ or “<surname>s “ from the front of the title. If there is a forename then the rule attempts to remove “<surname> <forename> “ or “<surname> <forename>s “ from the front of the title.

Given these fields:

100 1 $aSmith, Johan$d1962-

245 10 $aJohan Smith's The Ultimate Guide to Fall : $b cleaning the gutters

Here are the attempts to make a mapping match:

Step	String
Step 1	johan smiths the ultimate guide to fall
Step 2	(*no leading article so nothing to do*)
Step 3	the ultimate guide to fall
Step 4	ultimate guide to fall
Step 5	johan smiths the ultimate guide to fall cleaning the gutters
Step 6	(no leading articles so nothing to do)
Step 7	the ultimate guide to fall cleaning the gutters
Step 8	ultimate guide to fall cleaning the gutters

Uniform Title – Uniform titles are looked up and may be mapped to another uniform title or to an author/title.

Uniform Title	Maps To
acta de casa mata	plan de casa mata
wvmp radio	smith ready, jeri/wvmp radio

Building Mappings

The mappings that are derived from the LCNAF are built from personal name, name title and uniform title records. Other mappings that are derived from Worldcat data mining and map to personal names, titles and uniform titles in LCNAF are beyond the scope of this paper. Usable LCNAF records are established headings (008/9 is 'a' or 'f'), based on LCSH if they are appropriate for use as a subject and they are appropriate for use as a main heading (008/14 is 'a').

Author Mappings – These map names in a 400 to the name in the 100. Also if stripping one or both dates from a name does not result in ambiguous mappings then the name minus one and/or both dates is also saved as a mapping. This allows us to match the common case where a death date is added to an authority record but the associated bibliographic date is not updated. If the same normalized 400[abcdq] maps to multiple 100 fields then weights are calculated for each 100 and the 100 with the highest weight is used in the mappings.

Author Weights are calculated based on the number of records in Worldcat with that name plus the square root of the holdings attached to those records.

Author/Title Mappings – These map names and titles in a 400 field to the corresponding name/title in the 100 field or to the uniform title in the 130.

Uniform Title Mappings – These map uniform titles in a 430 field to the corresponding name/title in a 100 field or to the corresponding uniform title in the 130.