

PRESENTATION OUTLINE

FRBRization

- 1. Work Level
- 2. Beyond Work Level
 - Expression Level
 - Manifestation Level
- 3. Issues and Challenges

1. FRBRIZATION AT WORK LEVEL

3

WORK-LEVEL FRBRIZATION PROCEDURE

- The FRBRization procedures for identifying and grouping works for our project are based on the OCLC FRBR work-set algorithm (Hickey & Toves, 2005).
 - The original source of the algorithm is:
 - o Hickey, T. B., & Toves, J. (2005). FRBR Work-Set Algorithm. Available at: http://www.oclc.org/research/software/frbr/default.htm
 - Based on several rounds of evaluations, the KSU FRBR team has added to and refined some of the algorithm's steps.

THE COLLECTION

- For this experiment, the collection used was extracted from WorldCat for LoC records at the end of December 2007.
- This collection includes:
 - 13,624,251 bibliographic records
 - **7,283,635** authority records

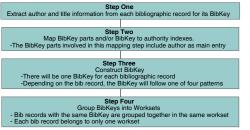
5

OVERALL IMPLEMENTATION PROCEDURES

- The OCLC work-set algorithm essentially covers two major processes:
 - Processing Authority Records



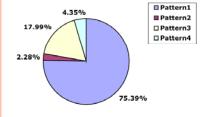
• Processing Bibliographic Records and Creating Work-sets



THE BIBKEY PATTERNS

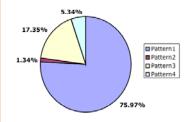
- There are four basic BibKey Patterns:
 - Pattern 1: <author>/<title>
 - Pattern 2: <uniform title>
 - Pattern 3: <title>/<name(s)>
 - Pattern 4: <title>/<control number>

THE LOC COLLECTION: BIBLIOGRAPHIC RECORDS BY PATTERN



Pattern	BibKey Count	Percent
Pattern 1	10,271,186	75.39%
Pattern 2	310,259	2.28%
Pattern 3	2,450,668	17.99%
Pattern 4	592,138	4.35%
Total	13,624,251	100.00%

Note that the KSU LoC BibKey pattern distribution is similar to the one reported in the 2005 OCLC Work-set Algorithm paper, shown below.



Pattern	BibKey Count	Percent
Pattern 1	41,654,567	75.97%
Pattern 2	736,020	1.34%
Pattern 3	9,513,007	17.35%
Pattern 4	2,927,095	5.34%
Total	54,830,689	100.00%

Loc Work-sets by Pattern

	Overall		Sing	e-record workse	t
Pattern	Count	Percent	Count	Percent within pattern	Percent of total worksets
Pattern 1	8,477,369	73.21%	7,424,329	87.58%	64.12%
Pattern 2	242,713	2.10%	221,975	91.46%	1.92%
Pattern 3	2,267,440	19.58%	2,125,564	93.74%	18.36%
Pattern 4	592,138	5.11%	592,138	100.00%	5.11%
Total	11,579,660	100.00%	10,364,006	n/a	89.50%

9

2. FRBRIZATION BEYOND WORK LEVEL

WORK-SET SAMPLE FOR FURTHER FRBRIZATION

- •From the multiple-record work-sets, we selected those that contained at least 10 records to explore FRBRization at the expression and manifestation levels
- Sample size:
 - 12,579 work-sets (1% of the multiple-record work-sets)
 - 273,866 records (2% of the entire LoC collection bib records)

11

PREVIOUS WORK AND BASIS

- Mapping and importance ratings of attributes and MARC fields on user tasks:
 - Functional Analysis of the MARC 21 Bibliographic and Holdings Formats by Network Development and MARC Standards Office, Library of Congress
 - IFLA FRBR Report, Tables 6.1-6.4
 - KSU FRBR project evaluation based on cataloging standards & practice
- MARC records analysis
- FRBR work-set analysis

BACKGROUND PROCESS FOR ALGORITHM DEVELOPMENT

- Mapping FRBR attributes to MARC codes
- o Mapping User Tasks to Attributes & Relationships & MARC codes
 - Analysis is based on FRBR User Tasks (e.g. the LoC "search" was mapped to FRBR "find")
- Evaluation of the importance of each attribute/MARC mapping in finding, identifying or selecting Group 1 entities with a focus on Expression & Manifestation
 - Evaluation & rating scale based on:
 - ${\color{blue} \bullet}$ FRBR evaluation of Attributes & Relationships primary rating
 - ${\color{blue} \bullet}$ LoC analysis of MARC fields, mapping to FRBR attributes & relationships and value of importance secondary rating
 - ${\color{blue} \bullet}$ KSU FRBR project evaluation based on standards & practice

42

2.1. FRBRIZATION AT EXPRESSION LEVEL

EXPRESSION LEVEL FRBRIZATION – ATTRIBUTES (1)

o EA1: Form of expression

Defined based on the values in Leader/06 and Leader/07:

- 1. Book
- 2. Collection
- 3. Continually updated resource
- 4. Electronic resource
- 5. Kit
- 6. Manuscript
- Map
- 8. Mixed materials
- 9. Music score
- 10. Musical sound recording
- 11. Nonmusical sound recording
- 12. Object
- 3. Serial
- 14. Visual material

15

EXPRESSION LEVEL FRBRIZATION – ATTRIBUTES (2)

o EA2: Language of expression

- \bullet 008/35-37 for language code
- 041 \$a for language code
- → Form a language array

• EA3: Music – Arranged statement for music

- 130 \$o
- 240 \$o

If data contains "arr" "arranged" or "arrangement", the attribute value is "Arranged"

• EA4: Music – Type of score

For music scores, find type of score in 008/20 in coded data:

- a Full score
- b Full score, miniature or study size
- \boldsymbol{c} Accompaniment reduced for keyboard
- d Voice score
- ${\bf e}$ Condensed score or piano-conductor score
- g Close score

- h Chorus score
- i Condensed score
- j Performer-conductor part
- $\ensuremath{\mathsf{m}}$ Multiple score formats
- n Not applicable
- u Unknown

EXPRESSION LEVEL FRBRIZATION – ATTRIBUTES (3)

• EA5: Serials - Extended frequency of issue

For serials, extract frequency from 008/18, and 008/19

• EA6: Cartographic – Scale

For maps, extract data in 034 \$b

• EA7: Cartographic - Projection

For maps, use 255 \$b for display description, not for grouping.

18 - Frequency	19 - Regularity
# - No determinable frequency a - Annual b - Bimonthly c - Semiweekly d - Daily e - Biweekly f - Semiannual g - Biennial h - Triennial i - Three times a week j - Three times a month k - Continuously updated m - Monthly q - Quarterly s - Semimonthly t - Three times a year u - Unknown w - Unknown w - Weekly z - Other - No attempt to code	n - Normalized irregular r - Regular u - Unknown x - Completely irregular - No attempt to code

17

EXPRESSIONS PER WORK-SET DISTRIBUTION

- o Expression
 - The merged expression string (EA1-6) for each record serves as a basis for evaluating how many different expressions are in a work.
 - Individual forms of expression and language elements serve as a basis for display.

Expressions-per-workset	Number of worksets	%	Cumulative %
1	3,794	30.2%	30.2%
2	2,918	23.2%	53.4%
3	1,480	11.8%	65.1%
4	1,227	9.8%	74.9%
5	927	7.4%	82.2%
6	610	4.8%	87.1%
7	423	3.4%	90.5%
8	302	2.4%	92.9%
9	225	1.8%	94.6%
10	154	1.2%	95.9%
11	103	0.8%	96.7%
12	74	0.6%	97.3%
13	56	0.4%	97.7%
14	48	0.4%	98.1%
15	35	0.3%	98.4%
16	37	0.3%	98.7%
17	31	0.2%	98.9%
18	21	0.2%	99.1%
19	20	0.2%	99.3%
20	17	0.1%	99.4%
>=21	77	0.6%	100.0%
Total	12,579	100.0%	

Overall distribution in the sample

Minimum	Maximum	Average
1	276	3 64

FORMS OF EXPRESSION PER WORK-SET DISTRIBUTION

Forms-per-work-set	Number of work-sets	%
1	7,972	63.4%
2	3,915	31.1%
3	652	5.2%
4	35	0.3%
5	4	0.0%
6	1	0.0%
Total	12,579	100%

Overall distribution in the sample

Minimum	Maximum	Average
1	6	1 42

19

LANGUAGE EXPRESSIONS PER WORK-SET DISTRIBUTION

Varied Language Es			
per-work-set	Number of work-sets	%	Cumulative %
1	6,202	49.3%	49.3%
2	2,101	16.7%	66.0%
3	1,272	10.1%	76.1%
4	886	7.0%	83.2%
5	592	4.7%	87.9%
6	399	3.2%	91.0%
7	273	2.2%	93.2%
8	219	1.7%	95.0%
9	141	1.1%	96.1%
10	103	0.8%	96.9%
11	80	0.6%	97.5%
12	59	0.5%	98.0%
13	43	0.3%	98.3%
14	31	0.2%	98.6%
15	30	0.2%	98.8%
16	20	0.2%	99.0%
17	27	0.2%	99.2%
18	13	0.1%	99.3%
19	21	0.2%	99.5%
>=20	67	0.5%	100.0%
Total	12,579	100.0%	

Overall distribution in the sample

Minimum Maximum Average

2.2. FRBRIZATION AT MANIFESTATION LEVEL

21

MANIFESTATION LEVEL FRBRIZATION - ATTRIBUTES

• Manifestation:

- o Identifier
- o Manifestation title
- Form of carrier
- Series statements
- Statement of reasonability
- Serials Numbering
- o Edition/issue designation
- Date of publication/distribution
- Publisher/distributor
- ${\color{red} \bullet} \ Visual\ material-Presentation \\ format \\$

- o Microform Reduction ration
- Microform/projection Generation
- Microform/projection Polarity
- Extent of carrier
- \circ Image Color
- Physical medium
- o Dimensions of carrier
- o Capture mode
- Electronic resources System requirements
- Electronic resources File characteristics
- Electronic resources Mode of access

Manifestations per Expression Distribution

Number of Ms in an E	Frequency	%	Cumulative %
1	20,496	44.8%	44.8%
2	5,915	12.9%	57.7%
3	2,890	6.3%	64.0%
4	1,865	4.1%	68.1%
5	1,337	2.9%	71.0%
6	1,111	2.4%	73.4%
7	962	2.1%	75.5%
8	996	2.2%	77.7%
9	1,090	2.4%	80.1%
10	1,872	4.1%	84.2%
11	1,326	2.9%	87.1%
12	1,007	2.2%	89.3%
13	717	1.6%	90.8%
14	556	1.2%	92.1%
15	439	1.0%	93.0%
16	359	0.8%	93.8%
17	307	0.7%	94.5%
18	249	0.5%	95.0%
>=19	2,281	5.0%	100.0%
Total	45,775	100%	

Overall distribution in the sample $\,$

Minimum	Maximum	Average
1	3,395	6.0

- - -

3. FRBRIZATION ISSUES AND CHALLENGES

- 1. Algorithm related
- 2. Legacy data, current cataloging practices and standards related
- 3. Display related

3.1 ALGORITHM RELATED CHALLENGES

- Existing work-set algorithm decisions & modifications undocumented
- Relationships have not been explored; limited to attributes
- o Algorithm basis: each bib record can belong to only one work-set: Bib record ← Work-set
- Source of data
 - Coded data vs. free text. E.g., 008 vs. 041 vs. 240 \$1 008 vs. 260 \$c
- Processing titles for works
 - · Trailing "English", "Tragedy of" and "Comedy of" in titles
 - Partial matching
- o Other titles not used in mappings of works
 - E.g. 730, 740, 7xx \$t
- Work identifiers

25

3.1 ALGORITHM RELATED CHALLENGES

- Non-mapped and mis-mapped FRBR elements to existing data. E.g.,
 - form and context of work, title of expression
 - Edition (250 field)
 - Edition of manifestation
 - Other distinguishing characteristic of expression
- Additional FRBR elements applying only to special types of materials – one complex algorithm or multiple algorithms
- Definition of an expression
 - Form or language
 - Form + language
 - Form + language + special material characteristics (scale, frequency, etc.)

3.1 ALGORITHM RELATED CHALLENGES

- Processing form of expression
 - Leader 06 & 07 combinations; groups vs. all. E.g.,

Leader06	Leader07	Leader-06.meaning	Leader-07.meaning
0	a	Kit	Monographic component part
0	b	Kit	Serial component part
0	С	Kit	Collection
0	d	Kit	Subunit
0	i	Kit	Integrating resource
0	m	Kit	Monograph/Item
0	s	Kit	Serial

- Processing *language* of expression
 - Source: 008 vs. 041 vs. 240 \$l
 - Multiple languages: treat each as a separate expression or as one?

bibsID	worksetID	leader06	leader07	language
3751493	13952	a	m	lat heb grc syr sam ara arc gez pal

27

3.1 ALGORITHM RELATED CHALLENGES (CONT.)

- Processing Manifestations
 - The work-set algorithm cannot recognize duplicate records, and therefore duplicates are considered separate manifestations
 - Assumption that each current bib record describes one manifestation
 - No further processing for grouping by "manifestation" set
 - Current record fields are all used as manifestation information

3.2 LEGACY DATA RELATED CHALLENGES

- Typos and coding problems
- o Multiple identifiers
- Missing important elements
 - 240s in AT pattern
 - 130s in bibliographic records
 - Incorrect or missing language indication
 - · Missing, incomplete, incorrect 041s for translated works
 - · Relator codes or other "relationships"
- Current cataloging practice
 - Collocating uniform titles
 - o Musical works and other Genre headings
 - o Selections, Works, etc.
 - Inconsistent use of uniform titles, e.g., TV and radio programs

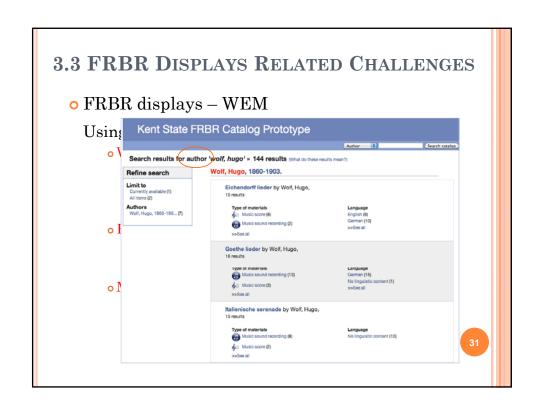
29

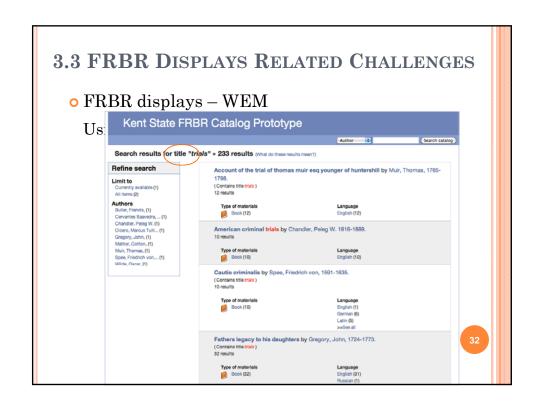
3.3 FRBR DISPLAYS RELATED CHALLENGES

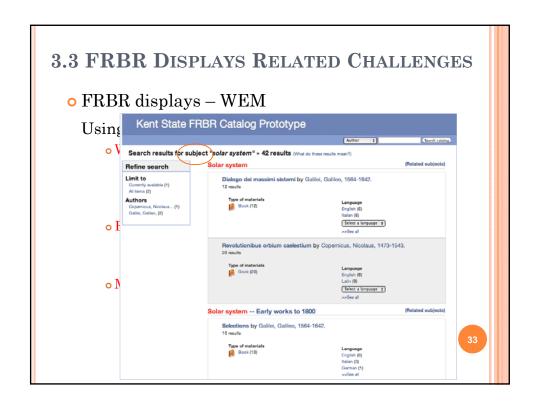
• FRBR displays – WEM

Using the hierarchy of Group 1 as basis:

- Work level display of results and refinement
 - Author search
 - Title search
 - Subject search
- ${\color{red} \circ}$ Expression level display of results and refinement
 - · Form as primary, language as secondary
 - · Language as primary, form as secondary
- o Manifestation display and refinement





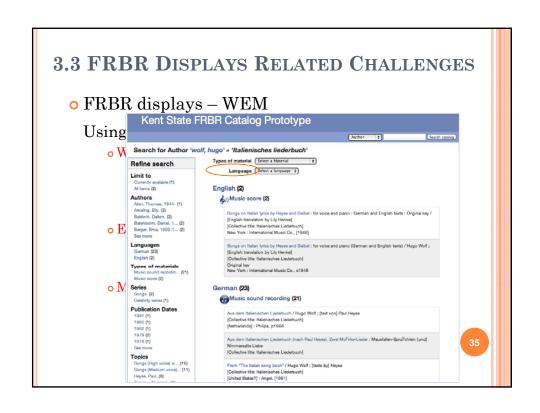


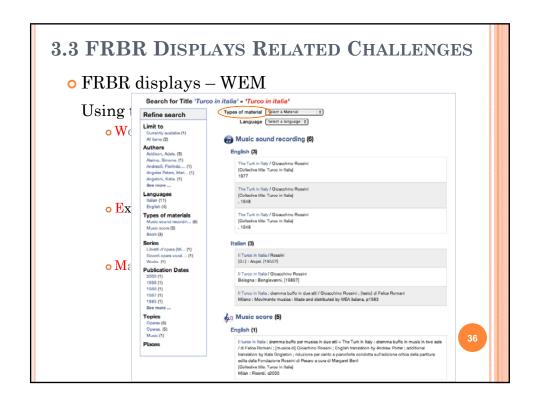
3.3 FRBR DISPLAYS RELATED CHALLENGES

• FRBR displays – WEM

Using the hierarchy of Group 1 as basis:

- Work level display of results and refinement
 - · Author search
 - Title search
 - Subject search
- Expression level display of results and refinement
 - · Form as primary, language as secondary
 - · Language as primary, form as secondary
- o Manifestation display and refinement





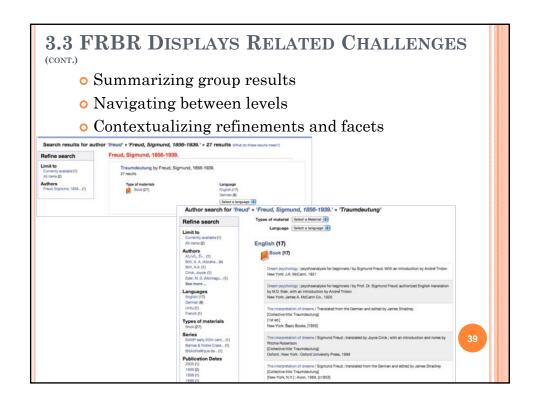
3.3 FRBR DISPLAYS RELATED CHALLENGES

• FRBR displays – WEM

Using the hierarchy of Group 1 as basis:

- Work level display of results and refinement
 - Author search
 - Title search
 - Subject search
- Expression level display of results and refinement
 - · Form as primary, language as secondary
 - · Language as primary, form as secondary
- o Manifestation display and refinement

37



FUTURE WORK AND CHALLENGES

- Authority data in OPAC
- Algorithms to process legacy data & integrate with new FRBR and/or RDA data
- RDA elements FRBR attribute relationship
- RDA testing results
- Legacy AACR2/MARC data RDA data relationship

ACKNOWLEDGEMENTS

We would like to thank the following people and groups:

- Thom Hickey and Jenny Toves, for their clarification of the FRBRization process and algorithm;
- OCLC, for providing the LC records that were used;
- · The project advisory team,
 - ${\color{red} \circ}$ Dr. Barbara Tillett, Library of Congress
 - o Dr. Edward T. O'Neill, OCLC
 - o Dr. Diane Vizine-Goetz, OCLC
 - o Dr. Thomas B. Hickey, OCLC
- Research team:
 - Lei Xie (Programmer)
 - Jake Schaub (Graduate Assistant)
 - o Vicki Ceci (Graduate Assistant)
 - Theda Schwing (Graduate Assistant)
 - Matt Shreffler (Graduate Assistant)
 - o Melissa Higey (Graduate Assistant)
 - Alison McCarty (Graduate Assistant)
 - o Joanna Samuelson (Graduate Assistant)
 - o Grace Brinker (Graduate Assistant)

41

MORE INFORMATION

Project website:

http://frbr.slis.kent.edu

Contact information:

Yin Zhang (<u>yzhang4@kent.edu</u>) Athena Salaba (<u>asalaba@kent.edu</u>)

Thanks!!

Questions and Suggestions?



Implementing FRBR in Libraries: Key Issues and Future Directions by Yin Zhang; Athena Salaba Publisher: New York: Neal-Schuman ©2009.

ISBN: 9781555706616 1555706614