Mapping and Transforming MARC21 Bibliographic Metadata to RDA/LRM/RDF

SWIB22 : Linked Library Data I
28 November 2022

Theodore Gerontakos

Crystal Yragui

Zhuo Pan

Project foundations

“Presenters [will] describe the theoretical foundation of the project”

“Foundation”

Original motivation not so theoretical

Theory: continue and expanding the project

Now: project is an international, collaborative

- Foundations not common to all participants

Project foundations (or motivations)

Origin: University of Washington, ca. 2014

RDA and BIBFRAME ontologies were being developed

- RDA Registry, 2014
- BIBFRAME 1.0, 2012; BIBFRAME 2.0, 2016

We asked, “Is BIBFRAME fit to handle RDA data?”

No

2016 proof -of-concept mapping

RDA-to-BIBFRAME (2016)

- RDA:BF
  - Not an ontology-to-ontology alignment
  - Demonstrate how well BIBFRAME accommodates the PCC RDA BIBCO Standard Record
  - “Not too well,” we decided
  - N:1 (properties), especially entity-to-entity relationships
  - Entity mismatch (no Expression entity on BIBFRAME)

Reaffirmed in RDA-to-BIBFRAME (2020) ontology-to-ontology alignment

Assumptions (Motivational assumptions)

RDA ontology *does* accommodate the PCC RDA BIBCO Standard Record

RDA data will be most accurate and complete as RDF using the RDA ontology

Institutions implementing RDA can exchange RDA data with each other

Accurate/complete RDA is well-suited to produce other exchange formats

...years passed...

We remained motivated by our assumptions

We launched projects to continue RDA exploration

BIBFRAME has become much more widely adopted that RDA /LRM/RDF

Continuing the work

Not destiny: lack of RDA/LRM/RDF adoption

- “historical accident”

We continue to help increase RDA/LRM/RDF adoption

- construct an extensive RDA/LRM/RDF graph
- better understand the benefits/drawbacks
- see how it compares to other models
- put-into-play the RDA Registry as we adopt “new RDA” maybe in 2023

Theoretical *justifications*

Strengthen metadata **interoperability**

- Interoperability; alignments/mappings have an important role

Develop core tools (selected application profiles, ontologies, mappings, etc.) in detail using committed **human intellectual engagement**

- Mapping ontologies manually, in detail; useful for our core ontologies/vocabularies

Develop a **shared RDA/LRM/RDF graph**

- For testing and reviewing
- Derive metadata using alternate data models and assess the results

Assemble mapped ontologies in machine-readable formats for additional processing

- For example, to assist automated ontology matching

Develop **fluency across multiple data models** among metadata professionals as a core proficiency

We’re open to more participation!

We meet on Wednesdays

The discussions are in depth, MARC subfield-by-subfield

Tasks are adopted by volunteers from a task board as-time-permits

The work is

- Interesting
- Engaging
- Mind-boggling
- Well documented (we think)

And now for more detail...

Crystal Yragui, the project manager

Mapping Structure & Format

Iterative Development

- Ideal format:
  - User friendly for MARC21 and RDA experts (not developers)
  - Machine readable
- Ended up with spreadsheets, prioritizing accessibility over machine readability because of:
  - Need for volunteers to complete complex and specialized work
  - Lack of budget
  - Lack of existing, well-documented, machine-readable format we could use

Spreadsheet Documentation

- Initial format:
  - Used Python to create initial spreadsheets
  - Based on entire MARC21 bibliographic format & RDA Registry RDA-to-MARC map
  - Human-readable rules for expressing MARC tags and conditions
  - Granular notes categories
  - Mapping itself is not structured enough for computer readability
- Working documents:
  - Split up by MARC tags
  - Located in Google Sheets for interoperability and real-time collaboration without requiring volunteers to push/pull using Git
- GitHub repository:
  - Instructions and description
  - Project management and discussion
  - Updated .csv version of mapping through semi-automated script process

Instructions and Description:
https://github.com/uwlib-cams/MARC2RDA/tree/main/Instructions

Current version of mapping as .csv:
https://github.com/uwlib-cams/MARC2RDA/tree/main/Working%20Documents/Draft%20Field%20By%20Field%20Spreadsheets/csv

Spreadsheet Structure: Rows

- Correspond to single MARC tag value or combination of values in some cases
- Specific circumstances or layers of circumstances (conditions) resulting in a single LRM/RDA/RDF mapping
- Started with approximately 70,000

Status	MARC Field	MARCFieldLabel	MARCInd1Label	MARCInd1Value	MARCInd1ValueLabel	MARCInd2Label	MARCInd2Value	MARCInd2ValueLabel	CharacterPosition	CharacterPositionLabel	MARCSubfield	CharacterPositionLabel	MARCSubfield	MARCSubfieldLabel	RDA Registry URI	RDA Registry Label	Recording Method	Justification for Mapping	X	Y	Z
																			Transformation Notes	Problems with Mapping	Notes (Uncategorized)
reviewed	490	SERIES STATEMENT (R)	Series tracing policy	*		Undefined	*				a, x, v		a, x, v		http://rdaregistry.info/Elements/m/datatype/P30106	has series statement	structured description	Chose not to use sub-properties because number of conditions is not sustainable for transformation. MARC does not have separate subfields for each element. Chose to retain MARC subfields when ISBD punctuation is absent in order to maintain structure of description.	when LDR 18 = a or i, remove marc subfields and rely on ISBD punctuation. When LDR 18=c, retain marc subfield codes to separate pieces of information coded --TG, 2022-11-13		
reviewed	490	SERIES STATEMENT (R)	Series tracing policy	*		Undefined	#	Undefined			3		3	Materials specified (NR)	http://rdaregistry.info/Elements/m/datatype/P30137	nas note on manifestation	unstructured description	$3 becomes a note	coded --TG, 2022-11-13		See: https://github.com/uwlib-cams/MARC2RDA/discussions/353
reviewed	490	SERIES STATEMENT (R)	Series tracing policy	*		Undefined	#	Undefined			6		6	Linkage (NR)					nothing coded --TG, 2022-11-		See: https://github.com/uwlib-cams/MARC2RDA/wiki/Decisions-Index#h-6
not mapped	490	SERIES STATEMENT (R)	Series tracing policy	0	Series not traced	Undefined	#	Undefined			8		8	Field link and sequence number (R)					not coded --TG, 2022-11-13		https://github.com/uwlib-cams/MARC2RDA/issues/343
not mapped	490	SERIES STATEMENT (R)	Series tracing policy	1	Series traced [REDEFINED]	Undefined	#	Undefined			3		3	Materials specified (NR)				Series tracing not relevant for mapping	nothing coded --TG, 2022-11-		
not mapped	490	SERIES STATEMENT (R)	Series tracing policy	1	Series traced [REDEFINED]	Undefined	#	Undefined			6		6	Linkage (NR)				Series tracing not relevant for mapping	nothing coded --TG, 2022-11-		
not mapped	490	SERIES STATEMENT (R)	Series tracing policy	1	Series traced [REDEFINED]	Undefined	#	Undefined			8		8	Field link and sequence number (R)					not coded --TG, 2022-11-13		https://github.com/uwlib-cams/MARC2RDA/issues/343
not mapped	490	SERIES STATEMENT (R)	Series tracing policy	1	Series traced differently	Undefined	#	Undefined			8		8	Field link and sequence number (R)					not coded --TG, 2022-11-13		https://github.com/uwlib-cams/MARC2RDA/issues/343

Spreadsheet Structure: Columns

- Status (in progress/done/etc.)
- MARCField
- MARCFieldLabel
- MARCInd1Label
- MARCInd1Value
- MARCInd1ValueLabel
- MARCInd2Label
- MARCInd2Value
- MARCInd2ValueLabel
- CharacterPosition
- CharacterPositionLabel
- MARCSubfield
- MARCSubfieldLabel
- CodeValue
- CodeValueLabel
- MARCTagCondition1
- Condition1Value
- MARCTagCondition2
- Condition2Value
- RDA Registry URI
- RDA Registry Label
- Recording Method
- Justification for Mapping
- Transformation Notes
- Problems with Mapping
- Notes (Uncategorized)

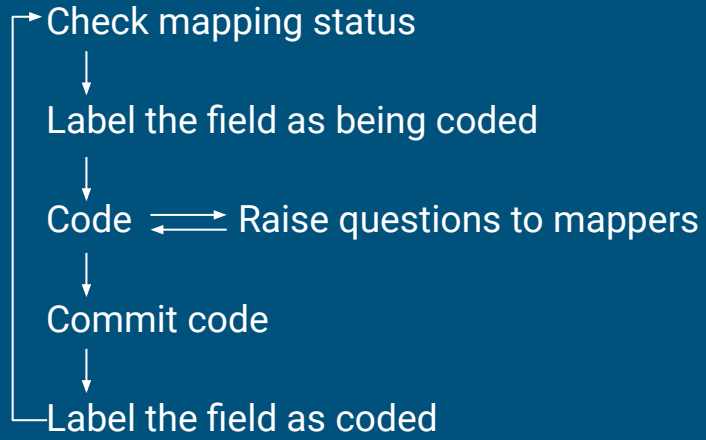
https://github.com/uwlib-cams/MARC2RDA/tree/main/Instructions

Conversion Tool

Overview

- **Goals**
  - Faithful representation of the mapping
  - Well-formed, error-free RDA data
  - Readability for people who are not developers over code economy
- **Language**
  - MARCXML → RDA/RDF/XML
  - XSLT 3.0

- **Workflow (field by field)**

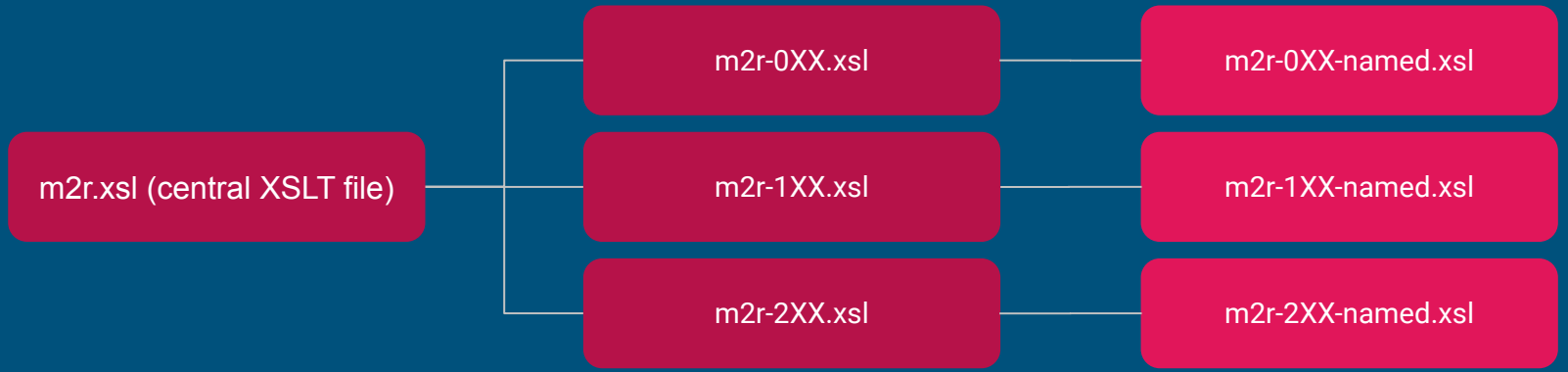


Template Design

Match marc:record
Mint IRIs for WEM
Create relationships between WEM
Non-aggregate only

Templates for each MARC field

Named templates for conditions, subfields, etc.



Demo

Test dataset of 54 records

Coded 13 fields

Output:

- 162 RDA entities
- 1,000+ RDA properties

```
1690      <rdf:Description rdf:about="http://fakeIRI2.edu/1145068737man">
1691         <rdf:type rdf:resource="http://rdaregistry.info/Elements/c/C10007"/>
1692         <rdamo:P30139 rdf:resource="http://fakeIRI2.edu/1145068737exp"/>
1693         <rdamd:P30134>Native American, American Indian, and Alaska Native LGB
1694         <rdamd:P30156>Native American, American Indian, and Alaska Native LGB
1695         <rdamd:P30105>by Adrian D. Zongrone, M.P.H., Nhan L. Truong, Ph.D., J
1696         <rdamd:P30111>New York, NY : GLSEN, [2020]</rdamd:P30111>
1697         <rdamd:P30088>New York, NY</rdamd:P30088>
1698         <rdamd:P30176>GLSEN</rdamd:P30176>
1699         <rdamd:P30011>[2020]</rdamd:P30011>
1700         <rdamd:P30280>©2020</rdamd:P30280>
1701         <rdamd:P30002>computer</rdamd:P30002>
1702         <rdamd:P30002>c</rdamd:P30002>
1703         <rdamd:P30001>online resource</rdamd:P30001>
1704         <rdamd:P30001>cr</rdamd:P30001>
1705         <rdamd:P30106>Erasure and resilience : the experiences of LGBTQ stude
1706         <rdamd:P30137>"A report from GLSEN and the Center for Native American
1707         <rdamd:P30455>Includes bibliographical references (pages 43-49).</rda
1708      </rdf:Description>
1709   </rdf:RDF>
```

Code:
https://github.com/uwlib-cams/MARC2RDA/tree/main/Working%20Documents/transformationCode

```
ex:1268154196wor  a       <http://rdaregistry.info/Elements/c/C10001> ;
        fake:rdawP10065   "Engel, Michael S., author." ;
        rdawd:P10002      "1268154196wor" ; # has identifier for work
        rdawo:P10078      ex:1268154196exp . # has expression of work

ex:1268154196exp  a    <http://rdaregistry.info/Elements/c/C10006> ;
        rdaed:P20001    "txt" , "text" ; # has content type
        rdaed:P20002    "1268154196exp" ; # has identifier for expression
        rdaeo:P20059    ex:1268154196man ; # has manifestation of expression
        rdaeo:P20231    ex:1268154196wor . # has work expressed

ex:1268154196man  a    <http://rdaregistry.info/Elements/c/C10007> ;
        rdamd:P30001    "nc" , "volume" ; # has carrier type
        rdamd:P30002    "n" , "unmediated" ; # has media type
        rdamd:P30011    "[2020]" ; # has date of publication
        rdamd:P30088    "Emporia, Kansas" ; # has place of publication
        rdamd:P30105    "Michael S. Engel" ; # has statement of responsibility relating to title proper
        rdamd:P30106    "The Kansas school naturalist, 0022-877X ; volume 64, no. 2 (December 2020)" ; # has series statement
        rdamd:P30111    "Emporia, Kansas : Emporia State University, Department of Biological Sciences, [2020]" ; # has publication statement
        rdamd:P30134    "Bees in Kansas" ; # has title of manifestation
        rdamd:P30137    "Cover title." , "\"Checklist of bees recorded from Kansas\": pages 5-13." ; # has note on manifestation
        rdamd:P30156    "Bees in Kansas" ; # has title proper
        rdamd:P30176    "Emporia State University, Department of Biological Sciences" ; # has name of publisher
        rdamd:P30455    "Includes bibliographical references (page 13)." ; # has supplementary content
        rdamd:P30456    "polychrome" , "monochrome" , "color" , "black and white" ; # has colour content
        rdamo:P30139    ex:1268154196exp . # has expression manifested
```

Item, Metadata Work & Reification (in progress)

561 Ownership and Custodial History

First Indicator:

0 - Private

Field contains private information

561 0# $a From the collection of L. McGarry, 1948-1957.

```
ex:123456789Aited12e90
        a               <http://rdaregistry.info/Elements/c/C10003> ;
        rdaid:P40001    "ited12e90" ; # has identifier for item
        rdaid:P40026    "From the collection of L. McGarry, 1948-1957." ;
        # has custodial history of item
        rdaio:P40049    ex:123456789Aman . # has manifestation exemplified

<http://marc2rda.edu/fake/MetaWor/d12e901>
        a               rdf:Statement , <http://rdaregistry.info/Elements/c/C10001> ;
        rdf:object      "From the collection of L. McGarry, 1948-1957." ;
        rdf:predicate   rdaid:P40026 ;
        rdf:subject     ex:123456789Aited12e90 ;
        rdawd:P10002    "MetaWor/d12e901" ; # has identifier for work
        rdawd:P10004    "Private" ; # has category of work
        rdawo:P10616    ex:123456789Aited12e90 .
        # is metadata description of item
```

Source of MARC:
https://www.loc.gov/marc/bibliographic/bd561.html

Vision for Ongoing Project

Future of Conversion Tool

- Implement RDA data models
  - Aggregates
  - Collections
- Address issues that lack clear guidance from RDA
  - Metadata work
  - Nomens & Non-Latin scripts
  - Non-RDA entities
- XSLT 3.0
- Test on large-scale datasets

Looking Ahead:
MARC21 to LRM/RDA/RDF Mapping Project

- Current Milestones:
  - PCC BSR (BIBCO Standard Record) Core MARC21 Fields
  - Mapping Review (concurrent with mapping)
  - Transformation (concurrent with mapping)
- Next:
  - PCC CSR (CONSER Standard Record) Core MARC21 Fields
  - Remaining Non-Obsolete MARC Fields
  - Publication
- Publication and Implementation:
  - Network Development and MARC Standards Office (NDMSO)
  - RDA Steering Committee, RDA Registry
  - Committed to keeping mapping and transformation freely available for adoption by library metadata creators and vendors

Regarding Publication

Project documents will remain public (probably on Github)

There will be standalone representations of the mapping

- Spreadsheets for human consumption
- What for machine consumption?

No well-known standard to represent mappings

Some well-known standards/specifications used to accommodate mappings:

- OWL
- SKOS
- SPARQL
- RDFS
- DITA

How represent a mapping/alignment

Some lesser-known standards/specifications used to accommodate mappings:

- MAFRA Semantic Bridge Ontology (2002)
- Semantic Web Rule Language (SWRL) (2004)
- RDFS Plus

How represent a mapping/alignment

Some standards/specifications are extended to accommodate mappings

- Context OWL (C-OWL) (2003?)
- Something homemade
  - "Extended MARC-XML" (extended to describe a mapping)(does not exist!)

How represent a mapping/alignment

Some standards/specifications were created for ontology matching:

- SEKT-ML (2004)
- Alignment Format (2004)
- XeOML (2004)
- Expressive Alignment Format (2006?)
- Expressive and Declarative Ontology Alignment Language (EDOAL) (2007? 2011?)

Current preference

RDF Mapping Language (RML)

- Best for our purposes?
- Support for XPath
- Navigates the MARC XML to express complex conditions
- Our special situation:
    - we are not matching ontology-to-ontology
    - MARC is not an ontology

Thank You!

Questions?

Theodore Gerontakos, tgis@uw.edu

Crystal Yragui, cec23@uw.edu

Zhuo Pan, panzhuo@uw.edu

GitHub Repository:
https://github.com/uwlib-cams/MARC2RDA