IFLA WLIC 2013
SINGAPORE

Submitted on: May 31, 2013

The UNIMARC in RDF project: namespaces and linked data

Mirna Willer
Department of Information Sciences, University of Zadar, Zadar, Croatia.
E-mail address: m.willer@unizd.hr

Gordon Dunsire
Independent Consultant, Edinburgh, Scotland.
E-mail address: gordon@gordondunsire.com

Predrag Perožić
Department of Information Sciences, University of Zadar, Zadar, Croatia.
E-mail address: pperozic@gmail.com

Abstract:

The paper describes the work of a project to represent UNIMARC in Resource Description Framework (RDF), the basis of the Semantic Web and linked data. The current focus is on the UNIMARC Bibliographic format, and the development of an element set for the tags and subfields and of value vocabularies for the coded information block. The paper discusses issues identified by the project for improving the UNIMARC standard, and in particular its alignment with ISBD and other bibliographic standards such as RDA: resource description and access. The paper also gives examples of linked open data derived from UNIMARC records.

Keywords: UNIMARC, Resource Description Framework, ISBD, mappings, interoperability.

1 INTRODUCTION AND BACKGROUND

A paper presented at the World Library and Information Congress: 77th IFLA General Conference And Assembly in San Juan, Puerto Rico, with an updated version subsequently published in the IFLA Journal (Dunsire & Willer, 2011), discussed a basic framework for representing the UNIMARC Bibliographic (UNIMARC/B) and UNIMARC Authorities (UNIMARC/A) encoding formats in Resource Description Framework (RDF), the syntactical basis of the Semantic Web and linked data. The paper identified specific issues, and made a number of recommendations for resolving them and developing namespaces to accommodate UNIMARC element sets and value vocabularies following the pattern already established for other IFLA bibliographic standards, including the Functional Requirements (FR) family of

models and the International Standard Bibliographic Description (ISBD). The Permanent UNIMARC Committee (PUC), responsible for the maintenance of the UNIMARC formats, agreed in 2012 to proceed with the development of such namespaces. Although an application to IFLA's Professional Committee was unsuccessful, the PUC was able to find and allocate enough funds for a project to be initiated in 2013 (PUC, 2012).

The first focus of the project was the development of namespaces for the UNIMARC/B format. However, it was immediately obvious that analysing the format in isolation from the namespaces of other related standards would produce a partial result because the project would not provide the information landscape in which the format functions. Furthermore, the positioning of the format in relationship to IFLA and other relevant standards additionally gives feedback to its developers about gaps and potential development of the format itself. This paper considers the relationship of UNIMARC/B to ISBD, taking into account the impact of the ISBD consolidated edition on the documented alignments, the relationship between the UNIMARC/B and UNIMARC/A formats, and, at a more general level, their alignment with RDA: resource description and access.

In this paper, the terms "element set" and "value vocabulary" conform to the usage recommended by the W3C Library Linked Data Incubator Group (Isaac and others, 2011).

2 BASIC METHODOLOGY FOR NAMESPACE CREATION

The documentation for the UNIMARC formats is only available in machine-readable form as Microsoft Word or Adobe PDF files. These lack the necessary structure for automatic parsing of the data required for element sets and value vocabularies, such as labels, definitions, and scope notes. There is sufficient structure in the layout, however, for human identification of such data. The basic methodology for extracting the data from the files is therefore human-mediated copying, pasting, and subsequent editing.

The finest granularity to be captured, in the case of element sets, is at the level of the UNIMARC subfield code and in the case of value vocabularies, the notation code and corresponding term. A subfield is the smallest unit of encoding for an element. Subfields are often aggregated into fields or tags, encoded according to the ISO 2709 standard by three digits (ISO, 2008). However, the use of one or both indicators for a tag can modify the semantics of the tag's subfields and the tag itself. The full semantics of a UNIMARC subfield may therefore need to include its tag and both indicators. The methodology used by the project assumes this is the default situation, and therefore creates an element for every allowed combination of tag, indicators, and subfield. For example, UNIMARC/B tag 200[1] (Title and statement of responsibility) has a subfield, encoded "$a", for the title proper. Although the second indicator is not used, the first indicator can take one of two values, to distinguish the cases where the title is significant or not significant. The significance of the title, that is the title proper, is captured by creating two elements, one where the title is significant, and the other where it is not.

If both indicators are used in a tag, the number of elements created for each subfield is the multiplication of the number of values for each indicator; if the first indicator can take 3 values, and the second can take 2, then 6 elements are created. This ensures that every difference in meaning is accommodated. In order to reduce the time taken to manually extract

[1] UNIMARC/B, 200 TITLE AND STATEMENT OF RESPONSIBILITY

the required data from the UNIMARC documentation, a spread-sheet is used to store the data for each subfield in a tag. The set of rows for the tag is then duplicated to record the different allowed values for the first indicator, and the whole block of duplicated sets of rows is duplicated for each allowed value for the second indicator. Only the indicator value has to be changed in each duplicated set of subfields, and that can be quickly achieved using the interface for copying cells within the spread-sheet. Each row contains the tag number, indicator values, and subfield encoding, along with the corresponding tag, indicator, and subfield captions. These data are used by a spread-sheet formula to create a somewhat artificial, but human-readable, label for the RDF property that represents the subfield. For example, the two elements for title proper will have the labels "title proper in Title and statement of responsibility (Title is significant)" and "title proper in Title and statement of responsibility (Title is not significant)".

Similarly, the Uniform Resource Indicator (URI) of each element is derived using a spread-sheet formula to concatenate the tag number, indicator values, and subfield encoding letter to form a partial URI which is unique within the UNIMARC encoding scheme. The final form of the rest of the URI, the base of the namespace, will not be finalised until the data entry is completed. For example, the URIs of the title proper RDF properties are likely to be *http://iflastandards.info/ns/unimarc/unimarcb/elements/2XX/U2001_a* and *http://iflastandards.info/ns/unimarc/unimarcb/elements/2XX/U2000_a* respectively. An underscore (_) is used to mark the place of the second indicator, which is not used in this field. Although it is not strictly necessary for this example, it allows the automatic derivation of the URI from the actual encoding used in a UNIMARC bibliographic record. This approach allows any UNIMARC record to be transformed into data triples using a simple computer program.

The two different properties for title proper can be "combined" by adding a third property with URI *http://iflastandards.info/ns/unimarc/unimarcb/elements/2XX/U200__a* and the label "title proper in Title and statement of responsibility"; that is, ignoring the value of the first indicator. This third property can be declared a super-property of the other two. This allows automatically generated data triples which use the first two properties to be "dumbed-down" to the super-property by losing the distinction of significance caused by the indicator values. An application which makes the distinction can use the original data triples; an application which does not need the distinction can use the dumbed-down triple.

This "sub-property ladder" technique can be extended to other bibliographic formats with RDF namespaces. Figure 1 shows a potential map for the title proper element from UNIMARC, ISBD, and RDA using only the RDF syntax (RDFS) property "subPropertyOf".

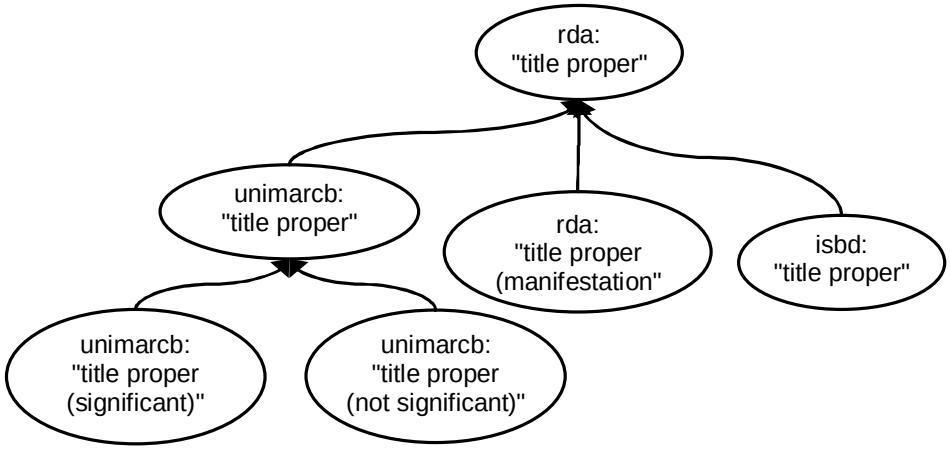


Figure 1: RDF graph of "title proper" property from UNIMARC, ISBD, and RDA element sets. All predicates are the RDFS "sub-property of" property and URIs are replaced with abbreviated labels for clarity.

For clarity and brevity, the URIs in Figure 1 are abbreviated into a "base" part, for example "unimarcb" standing for "http://iflastandards.info/ns/unimarc/unimarcb/elements/2XX/" and a "local" part, for example ISBD's "P1004", with a human-readable label such as "title proper" enclosed in quotes or brackets. The rest of this paper will use a similar convention.

The graph in Figure 1 is based on the use of unbounded or "unconstrained" properties in the proposed mappings of the ISBD and RDA element sets (Dunsire, 2011; Dunsire & ISBD Review Group, 2012). These mappings also use elements for aggregated statements, so the project will include data for field-level aggregations in the spread-sheet, as well as finer aggregations found with common patterns in their subfield codes.

The URIs and labels derived from the spread-sheet formulas will be uploaded into the OMR on completion. The upload will also include definitions and scope notes for the properties. Determining these has not proved to be straightforward, and requires significant editorial intervention.

2.1 Transcription issues

Like the rest of the namespace content, definitions and scope notes can be copied and pasted into the spread-sheet, or any RDF vocabulary management system, but doing so rightly takes the text out of the context of the documentation. Humans can readily discern a definition from embedded scope notes, text formatting, or usage instructions when it is presented in conjunction with the rest of the vocabulary, but in isolation this embedded text becomes mostly noise, obscuring or over-complicating the definition of the element. Definitions and scope notes need to be readable, understandable, unambiguous, and separate.

The project has identified several types of situation affecting the transcription of UNIMARC documentation to the namespaces:

- The definition contains phrases introduced by "including", "excluding", "for example", "e.g." and so on. An example is *unimarcb:U2000_a* with the definition "The chief title of the item, includes alternative title but excludes other title information (e.g. subtitles) and parallel titles". These phrases have been treated as parts of the element's scope notes in transcriptions of other IFLA standards.
- The definition references other elements directly. An example is *unimarcb:U2000_a* with the definition "The title proper in another language and/or script relating to a title proper appearing in a $a or $c subfield". The references are syntactical, using field and subfield codes and references to structure, and need to be separated from the semantics of the definition. They are properly represented in an application profile.
- The definition contains usage phrases. An example is *unimarcb:U2000_v* with the definition "Used to indicate a particular part of an item that is related to another item." The phrases make the definition more difficult to read and understand outside of the context of the manual.
- The definition contains formatting information. An example is *unimarcb:U2000_z* with the definition "Coded identification of the language of a parallel title that appears in a $d subfield." How the content is represented is not part of the semantics of the element. This is a similar issue as that of embedded syntactical information. This information is best represented in domain and range constraints and an application profile.

These have necessitated close human scrutiny of the text of the format manuals followed by editing to create the RDF property definitions and notes. This is time-consuming, but will ultimately inform any future development of the manuals themselves.

3 ALIGNING UNIMARC/B AND ISBD

In the 2XX section or block[2] covering the descriptive fields, with tags ranging from 200 to 225, the UNIMARC Bibliographic format is constructed according to the provisions of ISBD. The analysis of the alignment between UNIMARC/B and ISBD is relevant to the methods and practices by which data, published as linked data following one or the other IFLA standards, can be put into interoperable relationship. It should be noted, though, that there is an issue because the UNIMARC/B format does not provide rules concerning the content of the record: it specifically provides an indication whether descriptive data elements are in accordance with the provisions of ISBD or not, thus allowing data created by non-ISBD practices to be accommodated in the format. For example, the UNIMARC/B Record label character position 18 (Descriptive Cataloguing Form)[3] has values for full, partial, or no conformance with ISBD. Also, although it provides definitions of fields/subfields for ISBD data elements, it refers to the ISBD documents for those definitions.[4]

This is particularly important in interpreting the set of alignments given in Table 1 in which the alignment relationships are based on the semantic coherency of respective definitions, categorized as equal to (“=”), broader than (“>”) or narrower than (“<”) in meaning. There are two other issues that should be taken into account: the first is the nature of the format itself and its method of content designation; the second is the nature of the descriptive data with regard to their order of appearance on the resource, and therefore stipulations for their

[2] UNIMARC/B, 2-- DESCRIPTIVE INFORMATION BLOCK
[3] UNIMARC/B, RECORD LABEL, 18 Descriptive cataloguing form
[4] UNIMARC/B, 1.3 Definitions

transcription in the record. Although the latter issue is considered to be the focus for the development and implementation of an application profile (Willer, Dunsire & Bosančić, 2010), it will be shown that it also has relevance to the definition of namespace elements and the analysis of their alignment.

UNIMARC			**ISBD**	
Property	**Label**	**A**	**Property**	**Label**
200 2000_ 2001_	Title and statement of responsibility	=	P1159	has title and statement of responsibility area
			P1170	has title statement
			P1012	has title
200__a	Title proper	= <>	P1004	has title proper
			P1117	has title of individual work by same author
			P1137	has common title of title proper
200__b	General material designation			
200__c	Title proper by another author	=	P1118	has title of individual work by different author
200__d	Parallel title proper	=	P1005	has parallel title
			P1182	has common title of parallel title
			P1183	has dependent title of parallel title
			P1184	has dependent title designation of parallel title
200__e	Other title information	=	P1006	has other title information
			P1140	has parallel other title information
200__f	First statement of responsibility	>	P1007	has statement of responsibility relating to title
200__g	Subsequent statement of responsibility	<	P1007	has statement of responsibility relating to title
			P1141	has parallel statement of responsibility relating to title
200__h	Number of part	=	P1139	has dependent title designation of title proper
200__i	Name of part	=	P1138	has dependent title of title proper
200__j	Inclusive dates			
200__k	Bulk dates			
200__r	Title page information following the title proper (for older monographic publications)			
200__v	Volume designation			
200__z	Language of parallel title proper			
200__5	Institution to which the field applies			

Table 1: UNIMARC Bibliographic format to ISBD alignment, tag 200 (draft).

The very first subfield in the block, for the element *unimarcb:U200__a* (title proper) is a good example for all three issues mentioned. This is the first data element in the field, and its definition corresponds to the one in ISBD for the same element; therefore they are “equal” in meaning. But the fact is that it can be considered at the same time to be both broader and narrower in meaning than the ISBD element. Namely, the economy of the content

designation of the UNIMARC/B format designates this first data element to be the one that comes first in the context of the resource being described: not only “title proper” and “common title of title proper”, but also “title of individual work by same author”. The “common title” can be aligned with its ISBD equivalent and computed as such using an algorithm that detects the presence of subfield $h (number of part) or subfield $i (name of part) in the same tag, although the method is not 100% accurate. However, in the linked data environment each data triple from a record is potentially de-linked from every other triple and such a computational method is not valid after the triples have been published. The same is true for the second case, where the element “title of individual work by same author” is encoded in repeats of the $a subfield. We can conclude from this example that informational value is being lost in both directions of the alignment: in the UNIMARC/B to ISBD mapping, the UNIMARC/B property with URI *unimarcb:U200__a* subsumes the meaning of the *isbd:P1004*, *isbd:P1137* and *isbd:P1117* properties, while in the inverse ISBD to UNIMARC mapping the discrete ISBD properties lose their specificity in one UNIMARC property. In the context of an implementation of an application profile, it would be possible to align repeats of the $a subfield to *isbd:P1117*, but this is not the case with *isbd:P1137* for “common title”.

The lack of alignment is obvious also in the case of parallel title data elements. There is only one UNIMARC property for that type of content: *unimarcb:U200__d* (Parallel title proper). According to its definition, it is “The title proper in another language and/or script relating to a title proper appearing in a $a or $c subfield”, meaning that its semantics are related to *unimarcb:U200__a* and all repeats of the subfield, and *unimarcb:U200__c* (Title proper by another author). Only the sequence or order of data elements, transcribed from the resource and processed by an application profile, can “say” to which of these two or more titles the property *unimarcb:U200__d* corresponds. Although the definition of *isbd:P1005* (has parallel title) does not specify to which kind of title proper it relates, the examples show that the meaning is equivalent. All other parallel data in the UNIMARC/B format are indicated within the relevant subfield: UNIMARC/B states that “If '= ' is required by ISBD rules at the start of any other subfield, it must be entered explicitly.”[5] In other words, the UNIMARC/B element is semantically refined (the opposite of "dumbed-down") by the syntax of its content, so the MARC encoding by itself is insufficient to delineate the semantics of the formatted data. This is shown by the following example:[6]

> 200 1#$aBibliographica belgica$fCommission belge de bibliographie$f= Belgische Commissie voor bibliografie

Here, the subfield $f data "Commission belge de bibliographie" has the same semantics as *isbd:P1007* (has statement of responsibility relating to title) while the subfield $f data "Belgische Commissie voor bibliografie" has the semantics of *isbd:P1141* (has parallel statement of responsibility relating to title). That provision means that the UNIMARC/B namespace should contain a property for the "200__f=" case, literally a "parallel property", and so on for all other cases except for *unimarcb:U200__a* and *unimarcb:U200__c*. These should be created in order to support the alignment with ISBD properties for parallel data. If this is not done, the ISBD to UNIMARC alignment would cause loss of the parallel information value of the data. It must be noted that the same provision for the treatment of parallel data is given in all other 2XX fields.

[5] UNIMARC/B, 200 TITLE AND STATEMENT OF RESPONSIBILITY, Parallel data
[6] UNIMARC/B, 200 TITLE AND STATEMENT OF RESPONSIBILITY, EX 6

We should also consider here the treatment of definitions in UNIMARC/B: as already mentioned, the Manual does not define ISBD data elements, but refers to the specific ISBD for definitions. The first edition of UNIMARC Bibliographic Format, in the same style as the concise versions of current editions, named only fields and subfield data elements without providing definitions. It was the following edition entitled UNIMARC Handbook: Bibliographic Format, and subsequently UNIMARC Manual: Bibliographic Format that included definitions based on the contemporary ISBD editions. The status of UNIMARC/B field/subfield definitions should be viewed therefore from the aspect of the maintenance of the format in relation to the changes of ISBD data element definitions resulting from the replacement of ISBD(G) and seven specialized ISBDs by the consolidated edition, and also of its general intentions in referring the user to another document. The alignment column of the UNIMARC/ISBD Table 1 shows, in fact, that only *unimarcb:U200__a*, and *unimarcb:U200__f* and *unimarcb:U200__g* are not considered to be equal in semantics to corresponding ISBD properties. The case of *unimarcb:U200__a* is discussed earlier in this paper, while in the other cases, it was necessary to distinguish between first and subsequent statements of responsibility due to the different ISBD punctuation required for these two data elements. The different categorization of meaning, however, blurs the situation in which the first statement of responsibility relates not only to the title proper but to other elements which UNIMARC/B specifies as: "The first statement of responsibility for a title appearing in subfield $a, $c or $d, or for a numbered or named part of a work appearing in subfields $h or $i." ISBD examples demonstrate the same treatment; or, better to say, UNIMARC follows the ISBD provisions "in the best possible way". Furthermore, the ISBD considers "The difference between the first and subsequent statements of responsibility is merely a matter of order" and treats it as a single repeatable element, and thus there is no separate RDF property to distinguish the first occurrence. If it were represented as a separate element, it would consequently be aligned with *unimarcb:U200__g*, and the information value of data would be equal in both directions of the alignment.

The decision was taken at the outset of the UNIMARC namespace development to not re-use elements from the ISBD namespace. The context was recognized to be different from the case of the FR family of models where the RDF property for the same element is used for all three of models, for example the classes for "work" or "person". The fundamental argument was that each set of elements should be coherent and complete in itself, which would therefore enable their independent maintenance and further development. This decision has been justified, as the alignments of other UNIMARC/B 2-- fields show the same issues as presented in the case of tag 200. It is not realistic to expect the UNIMARC/B format to update definitions of existing fields and subfields following ISBD changes in the consolidated edition, but only in those that are new, such as Area 0 Content form and media type, primarily because the format is basically a container for data. In any case, element *unimarcb:U200__a*, for example, can be used for other purposes when the record contains data following non-ISBD descriptive cataloguing practices.

Certain changes in the format related to the change of field or subfield name should be mentioned here because they can directly influence the definition of the domain of the UNIMARC/B RDF properties. Specifically, the change of terminology in ISBD(ER) from "publication/item" to "resource" impacted on the renaming of tag 207[7] and tag 206[8]. The change in the name of the fields is from "material" to "resources". The field and subfield

[7] UNIMARC/B 207 MATERIAL SPECIFIC AREA: NUMBERING OF CONTINUING RESOURCES
[8] UNIMARC/B 206 MATERIAL SPECIFIC AREA: CARTOGRAPHIC MATERIALS - MATHEMATICAL DATA

definitions in 206 were changed to replace "item" with "resource", while the subfield definitions of 207 retain the use of the term "item". It should be noted though that, except for 206 field, UNIMARC consistently follows the change of term only in the case of electronic, and continuing and integrating resources, and thus the term "resource" should be considered a *terminus technicus* for describing specific types of material.

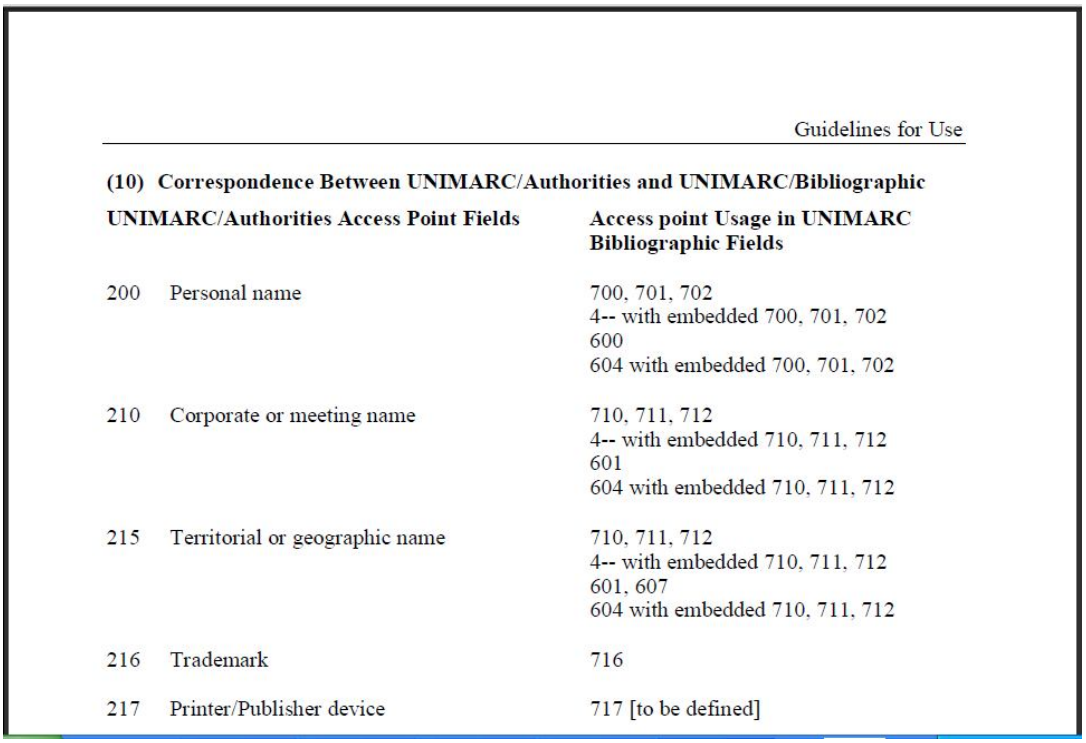
The 2XX block shows another issue relevant to the development and maintenance of the respective documents. The "General material designation" (GMD) element was removed from ISBD element 1.2 and replaced by the new area ISBD 0 Content form and media type area in the consolidated edition. The new edition renumbered the elements so the 1.2 element was changed from being the GMD to the "Parallel title" element, and so on. The UNIMARC 2012 Update does not reflect this renumbering, and retains subfield $b of tag 200 for the GMD. The format cannot delete a defined element in the expectation that it is used in legacy records; it can only make the subfield obsolete, which, however, the PUC has not yet done. At the same time, it should be noted that the ISBD Review Group did not take into consideration the option to deprecate the element, and not reuse the position number because the situation is the same as in the case of UNIMARC/B format; there are legacy data to be considered, as well as those data that will be produced by continuing practices that for one reason or the another will not follow the newly defined area. The same is the case with the UNIMARC/B tag 230[9] which is defined as equivalent to the ISBD(ER) Type and Extent of Resource (Area 3), but was deleted from the same area in the consolidated edition. The PUC has not made this field obsolete yet, either.

4 TWO FORMATS, TWO NAMESPACES?

The correspondence between the UNIMARC/Authorities to UNIMARC/Bibliographic formats is built on structural compatibility, the primary reason being that the two types of records are intended to be used together in integrated library systems. This means that data elements for the same access points appear in the same subfields in both formats, while the tags and field names differ because of the different functions of bibliographic and authorities records.

[9] UNIMARC/B 230 MATERIAL SPECIFIC AREA: ELECTRONIC RESOURCE CHARACTERISTICS

This correspondence is documented in UNIMARC/A, Guidelines for Use (10); part of the correspondence table from the 3rd edition is shown in Figure 2.



Guidelines for Use

(10) Correspondence Between UNIMARC/Authorities and UNIMARC/Bibliographic

UNIMARC/Authorities Access Point Fields		Access point Usage in UNIMARC Bibliographic Fields
200	Personal name	700, 701, 702 4-- with embedded 700, 701, 702 600 604 with embedded 700, 701, 702
210	Corporate or meeting name	710, 711, 712 4-- with embedded 710, 711, 712 601 604 with embedded 710, 711, 712
215	Territorial or geographic name	710, 711, 712 4-- with embedded 710, 711, 712 601, 607 604 with embedded 710, 711, 712
216	Trademark	716
217	Printer/Publisher device	717 [to be defined]

Figure 2: UNIMARC Authorities and UNIMARC Bibliographic correspondence table (part)

At the level of fields, the two formats differ in how the relationship for corporate body access point is to be categorized; that is, the three separate UNIMARC/B tags for Corporate body name[10] correspond to the UNIMARC/A tags 210[11] and 215[12]. The subfields names and definitions in tags 71X and 210 are equal in meaning, while tag 215 is defined to "contain a territorial or geographic access point [...] Territorial names alone or only with subject subdivisions as additions are considered territorial names (field 215); territorial names followed by corporate body subdivision as considered corporate body names (field 210)." Therefore, the category of relationship between UNIMARC/B tags 71X and UNIMARC/A tag 215 is debatable.

The correspondence between subfield names and definitions is categorized as equal following the aforementioned structure and function of the formats. The one difference is found in the field name of UNIMARC/B tag 500 "Preferred access point" and UNIMARC/A tag 230 "Authorized access point". In general, UNIMARC/A terminology follows that of FRAD, while UNIMARC/B is closer to RDA. The synchronous development and maintenance of the two UNIMARC formats has been attained by the 2012 updates to the 3rd edition of each format.

[10] UNIMARC/B 710 CORPORATE BODY NAME - PRIMARY RESPONSIBILITY, UNIMARC/B 711 CORPORATE BODY NAME - ALTERNATIVE RESPONSIBILITY, and UNIMARC/B 712 CORPORATE BODY NAME - SECONDARY RESPONSIBILITY
[11] UNIMARC/A 210 AUTHORIZED ACCESS POINT - CORPORATE BODY NAME
[12] UNIMARC/A 215 AUTHORIZED ACCESS POINT - TERRITORIAL OR GEOGRAPHICAL NAME

The structural compatibility of the two formats is reflected in their encoding, which drives syntactical correspondence between them. Semantic correspondence at the subfield and tag levels is only partial, justifying the decision to create separate namespaces for each format. However, for elements where the syntactic correspondence is also a semantic correspondence, the relevant parts of the UNIMARC/B spread-sheets can be re-cycled for UNIMARC/A, saving significant time in developing its namespaces.

5 EXAMPLES OF UNIMARC DATA AS RDF LINKED DATA

The UNIMARC/B 1XX block uses codes as data values, with the codes themselves having captions or labels and sometimes definitions. Each set of codes can be represented in RDF as a value vocabulary, assigning a URI to each concept associated with a code. An example of such a vocabulary is the frequency of issue of continuing resources. Each value of frequency has its own code which is stored in a UNIMARC data record as character position 1 of tag 110.[13] The coded data fields use fixed character positions rather than subfields to delimit the elements, so the URI pattern is slightly different. Thus the "frequency of issue" element has the URI *unimarcb:U110__a1*. The value vocabulary has been represented in the OMR as part of the project (UNIMARC Frequency of issue, 2013).

This allows the publication of data triples linked by the URI of a code value. For example, a UNIMARC record for a daily newspaper will contain a 110 tag with the character position code values given in Table 2.

Character position	Value	Notes
0	c	newspaper
l	a	daily
2	a	regular
3	#	n/a
4-6	###	n/a
7	0	Not conference proceedings
8	x	n/a
9	x	n/a
10	0	No cumulative index, etc.

[13] UNIMARC/B 110 CODED DATA FIELD: CONTINUING RESOURCES $a/1 Frequency of issue

Table 2: Tag 110 code values for a daily newspaper.

Figure 3 shows the linked data graph for the first three characters. The URI for the "daily" frequency is linked to the code or notation "a" and the preferred label in English, Italian, and Portuguese.

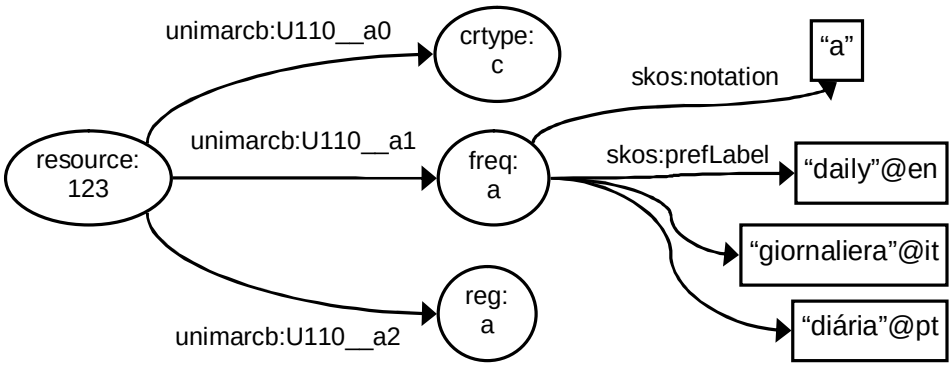


Figure 3: RDF graph of part of the coded information block of a UNIMARC record for a daily newspaper.

The graph in Figure 3 can be augmented with similar links to the notation and preferred labels of the continuing resource type code "c" and the regularity code "a" when their value vocabularies are published at a later stage of the project. The use of coded information in UNIMARC is thus more compatible with the syntax of RDF than the use of free-text labels. This has immediate benefit in a multilingual environment, as Figure 3 shows, by using RDF's in-built language identification system.

6 MAPPING VALUES

In addition to the UNIMARC value vocabulary, the OMR contains three other vocabularies for the frequency of continuing resources such as serials and collections, from Dublin Core Collection Level Description, MARC 21 and RDA: resource description and access. A search for the term "daily" gives results from all four vocabularies, as shown in Figure 4.

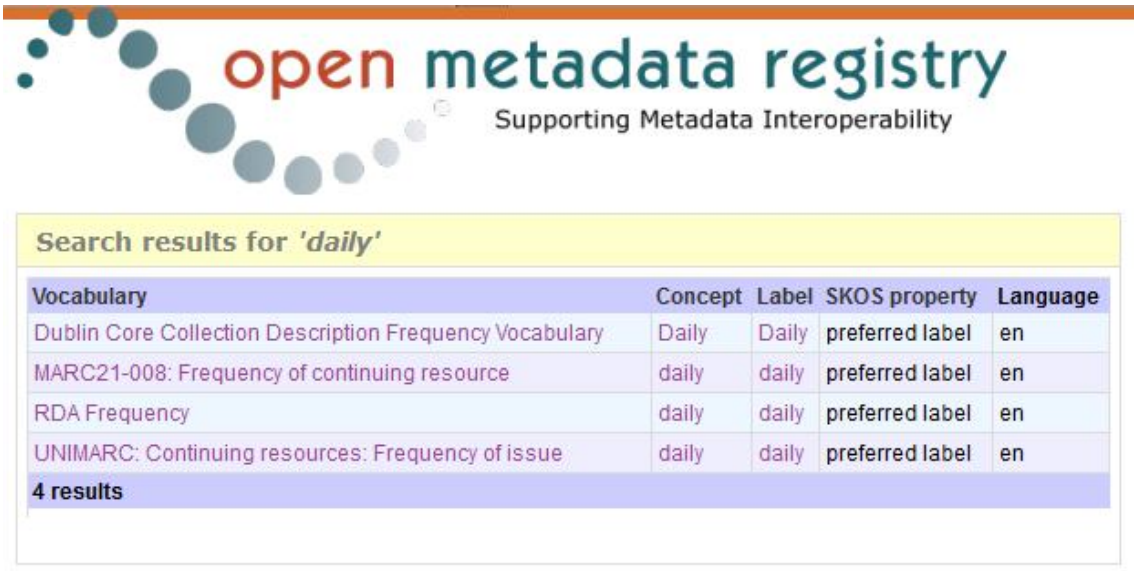


Figure 4: Screenshot of the results of a search for the term "daily" in value vocabularies registered in the Open Metadata Registry: http://metadataregistry.org/conceptprop/search?concept_term=daily.

Interoperability of linked data based on these four standards is improved if similar concepts can be related. To determine the relationship between the concepts of "daily" in each of the vocabularies, it is necessary to examine their definitions and scope notes:

- The UNIMARC term has no definition or scope note.
- The Dublin Core term has the definition "The event occurs once a day". The term is derived from MARC 21 Holdings, 853-855 - Captions and Pattern-General Information, subfield $w. Note that this is a fifth vocabulary, although not yet represented in RDF, because the subfield has a different context to that of MARC 21 Bibliographic.
- The MARC 21 term has the definition "Once a day" and a scope note "Includes Saturday and Sunday"; this information is derived from the manual for MARC 21 Bibliographic (MARC 21 2010).
- The RDA term has the definition "Frequency for a resource issued or updated once every day, usually exclusive of nonworking days". That is, the RDA concept excludes Saturday and Sunday.

In the absence of a definition, vernacular use of the term "daily" suggests that the UNIMARC concept covers all 7 days of the week, but this should not be taken as certain and it is unsafe to assume an exact match with the MARC 21 concept. The Dublin Core and MARC 21 concepts are broader than the RDA concept, which covers only 5 days of the week. Mapping relationship properties from the Simple Knowledge Organization System (SKOS) namespace (SKOS 2009) can be used to create an RDF graph relating the three concepts. In terse triple language (Beckett & Berners-Lee, 2011), this is serialized as:

@prefix cld: <http://purl.org/dc/cld/freq/> .
@prefix marc21: <http://marc21rdf.info/terms/continuingfre#> .
@prefix rda: <http://rdvocab.info/termList/frequency/>.
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix unimarc: <http://iflastandards.info/ns/unimarc/terms/continuingfreq#> .
daily
unimarc:a skos:closeMatch cld:daily .
unimarc:a skos:closeMatch marc21:d .
cld:daily skos:closeMatch marc21:d .
rda:1001 skos:broadMatch cld:daily .
rda:1001 skos:broadMatch marc21:d .
rda:1001 skos:broadMatch unimarc:a .

The inverse relationships can be automatically inferred to be:

cld:daily skos:closeMatch unimarc:a .
marc21:d skos:closeMatch unimarc:a .
marc21:d skos:closeMatch cld:daily .
cld:daily skos:narrowMatch rda:1001 .
marc21:d skos:narrowMatch rda:1001 .
unimarc:a skos:narrowMatch rda:1001 .

The complete RDF graph is shown in Figure 5.

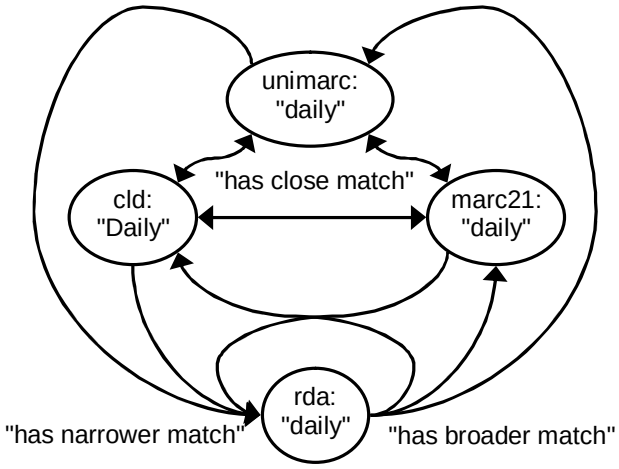


Figure 5: RDF graph of "daily" frequency of issue from MARC 21, RDA, and UNIMARC value vocabularies. URIs are replaced with labels for clarity.

Figure 5 is a map of the concept "daily" from the three value vocabularies. It can be extended to other vocabularies such as collection accrual periodicities developed by Dublin Core (Dublin Core Collection Description Task Group, 2007). This graph can be combined with the graph of Figure 3, suggesting new possibilities for the global interoperability of bibliographic linked data.

7 FURTHER WORK

The project is on-going and important work from the initial plan is scheduled for the rest of 2013 and 2014. This includes the development of value vocabularies for all of the code sets in the coded information block, internal mappings with the UNIMARC namespaces, and mappings from UNIMARC to external element sets and value vocabularies, based on the examples discussed in this paper. The project has identified additional significant areas requiring further work within the project and in the UNIMARC environment.

Within UNIMARC, the treatment of syntactical and semantic metadata embedded in record data needs to be modelled in RDF. An example of embedded syntactical metadata is the non-filing characters indicator, while an example of embedded semantic metadata is the use of the "=" sign for parallel data. This will have to be considered along with the representation of aggregations of subfields in a planned future phase of the project.

Considering UNIMARC/B and ISBD, the draft UNIMARC/B to ISBD alignment table should be extended to ISBD 7 Note area and ISBD 8 Resource identifier and terms of availability area, and a ISBD to UNIMARC/B alignment table should be developed. The process will have to take into account the necessity of developing unconstrained UNIMARC properties, that is, with no RDF domain or range. The alignment exercise disclosed various problems in both UNIMARC/B and ISBD which should be considered in close cooperation between the maintaining bodies. It is expected that other issues will be disclosed during the work on the ISBD to UNIMARC/B alignment. Also, the UNIMARC/B correspondence table between tag 2XX subfields and ISBD should be updated to correspond to the numbering of elements in the ISBD consolidated edition.

The complexity of the structure of the two UNIMARC formats which has to meet their goal of being compatible, and at the same time taking into account their respective functions, will further be shown by the process of publishing RDF properties for UNIMARC/B tag 500 and 7XX subfields, and UNIMARC/A itself. It is recommended that the work is conducted in parallel. Additional issues will be disclosed in the case of subject data elements in UNIMARC/B 6XX tags which are treated at the level of subfields in UNIMARC/A, because UNIMARC/A is an integrated name/title and subject authorities format.

8 REFERENCES

Beckett, David, and Tim Berners-Lee (2011). Turtle - Terse RDF triple language. Available at: http://www.w3.org/TeamSubmission/turtle/

Dublin Core Collection Description Task Group (2007). Dublin Core collection description frequency vocabulary. Available at: http://dublincore.org/groups/collections/frequency/2007-03-09/

Dunsire, Gordon (2011). Mapping ISBD and RDA element sets: briefing/discussion paper. Available at: http://www.rda-jsc.org/docs/6JSC-Chair-4.pdf

Dunsire, Gordon, and Mirna Willer (2011). UNIMARC and Linked Data. IFLA Journal, 37, 4 (December 2011) (pp314-326). Available at: http://www.ifla.org/files/hq/publications/ifla-journal/ifla-journal-37-4_2011.pdf

Dunsire, Gordon, and ISBD Review Group (2012). Alignment of the ISBD: International Standard Bibliographic Description element set with RDA: Resource Description & Access element set. Version 1.1. Available at: http://www.rda-jsc.org/docs/6JSC-ISBD-Discussion-1-Alignment.pdf

Hopkinson, Alan (ed.) (2008). UNIMARC Manual: Bibliographic Format. 3rd. ed. München: K. G. Saur.

IFLA, UNIMARC Core Activity, Permanent UNIMARC Committee Available at: www.ifla.org/unimarc/puc

IFLA, PUC (2012). UNIMARC Core Activity, Permanent UNIMARC Committee, Minutes of the Informal Meetings of the Permanent UNIMARC Committee, 13, 14 and 16 August 2012, IFLA Congress, Helsinki, Finland, Draft: 2012 October 31.

Isaac, Antoine, William Waites, Jeff Young, and Marcia Zeng (2011). Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets. W3C Incubator Group Report 25 October 2011. Available at: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

ISBD (2011). ISBD : International standard bibliographic description. Consolidated ed. Berlin ; München : De Gruyter Saur.

ISO (2008). ISO 2709:2008: Information and documentation -- Format for information exchange.

MARC 21 (2010). 008 - Continuing Resources (NR). Available at: http://www.loc.gov/marc/bibliographic/bd008s.html

SKOS (2009). SKOS Simple Knowledge Organization System: Reference. W3C Recommendation 18 August 2009. Available at: http://www.w3.org/TR/skos-reference/

UNIMARC Frequency of issue (2013). UNIMARC: Continuing resources: Frequency of issue. Available at: http://metadataregistry.org/vocabulary/show/id/324.html

Willer, Mirna (ed.) (2009). UNIMARC Manual: Authorities Format. 3rd. ed. München: K. G. Saur.

Willer, Mirna, Gordon Dunsire, and Boris Bosančić (2010). ISBD and the Semantic Web. JLIS.it Journal of Library and Information Science. Italy, vol. 1, no. 2, (2010), pp. 213-236. Available at: http://leo.cilea.it/index.php/jlis/article/view/4536