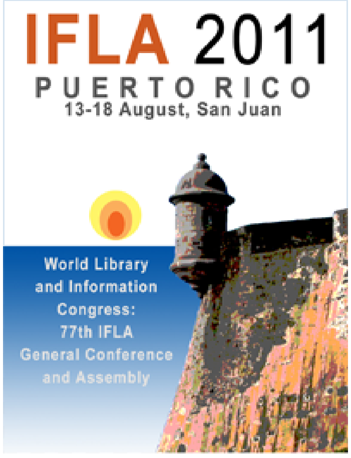
http://conference.ifla.org/ifla77 Date submitted: July 2, 2011


IFLA 2011
PUERTO RICO
13-18 August, San Juan
World Library and Information Congress: 77th IFLA General Conference and Assembly

UNIMARC and linked data

Gordon Dunsire
Freelance consultant
Edinburgh, UK
E-mail: gordon@gordondunsire.com

Mirna Willer
University of Zadar
Department of Library and Information Science
Zadar, Croatia
E-mail: mwiller@unizd.hr

Meeting: **187 — *Advancing UNIMARC: alignment and innovation* — IFLA UNIMARC Programme (UNIMARC)**

Abstract:

The main objective of this paper is to present arguments for and recommendations to representing UNIMARC formats for bibliographic and authority data in RDF (Resource Description Framework), the W3C standard for structuring data in Semantic Web and Linked Data environment. This is a continuation of the work already started by IFLA's respective groups in representing ISBD, and conceptual models FRBR, FRAD and FRSAD. The authors highly recommend that the PUC propose to IFLA the funding of the development of UNIMARC representation in RDF as a research and development project.

Introduction and background

"The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web."[1] In this approach, data is expressed as simple statements using Resource Description Framework (RDF), and connected using machine-readable identifiers conforming to the syntaxes of the Uniform Resource Identifier (URI). RDF statements take the form of a three-part subject-predicate-object structure, with the subject identifying what the statement is about, the predicate identifying the specific aspect of the subject being described, and the object identifying or presenting the value of that aspect. An RDF statement is therefore commonly known as a "triple". The basis of a triple is its predicate, which is represented as an RDF property, and the specific subject and object of a triple are represented as members of RDF classes. Classes describe things, and properties describe the relationship between those things; classes and properties are the basic types of element in RDF. The thing described as a class can be any type of resource or entity we want to make a statement about; it is used as the subject of a triple. Controlled terminologies used as the

objects of triples can be represented as "value vocabularies" using Simple Knowledge Organization System (SKOS)[2], which is a special set of RDF elements designed for simple thesauri and taxonomies. The object can also be represented as a literal string of data, such as a personal name, edition statement, etc., not supported by a vocabulary or controlled terminology.
A triple is essentially metadata, or data about data; in this case data about the subject of the triple. Linked data should therefore be of particular interest to the library community which has evolved sophisticated user-centred approaches to bibliographic metadata in the form of catalogues governed by internationally-agreed standards. A feature of linked data is its Web-scale, the Semantic Web, allowing the sharing of data at a global level between multiple heterogeneous sources. Again, this should be a significant utility for libraries which have been exchanging machine-readable cataloguing (MARC) records since the 1960s.

Library linked data derived from existing records based on international standards will be of high quality and high quantity, covering many of the available information resources in which users of the Semantic Web are likely to be interested. OCLC's WorldCat alone contains over 230 million bibliographic records[3]. An analysis of MARC content[4] found over 13 million subfields in approximately 420 thousand records; assuming that each subfield can generate one triple, this suggests an average of 31 potential triples per record. This figure is not reduced by the effects of duplication within WorldCat, as it is easily offset by records not aggregated in WorldCat, indicating that there must be at least billions of triples locked in legacy records. Of equal importance are the data created by libraries for authority control, covering persons, organizations, places, subject topics, and other things likely to be of interest to a wider audience than traditional users of libraries.

Using current library standards as the bases of new triples and the extraction of triples from legacy records requires the representation of such standards in RDF, either by creating appropriate RDF elements or mapping to existing elements. This will not just allow the Semantic Web to benefit from library metadata; it should also improve interoperability between bibliographic entities, attributes, and relationships described by different, but related, standards. RDF properties can be chosen from different standards and mixed within a single application to meet its functional requirements, using a Dublin Core Application Profile[5] or ontology expressed in RDF/OWL[6].

IFLA, as a standardizing body, should be particularly interested in entering the Linked Data and Semantic Web environment because of its mandate to develop and maintain bibliographic models and standards, and thus enable the library community to better serve its users in technologically changing conditions. Besides, by supporting developments leading towards the presentation of its internationally-agreed upon standards in RDF, IFLA provides authenticity and trustworthiness in library-produced metadata which is of exceptional importance in an environment which lets “anyone say anything about any resource”, while at the same time promoting its brand beyond library community boundaries. Using explicitly defined relationships “it is possible to computationally create a web of trust [Godlbeck and Parsia]. Establishing a system of trust in the Semantic Web will make it easier for computers to determine which information comes from an authoritative source and which does not”[7].

The first initiative to start reviewing IFLA’s standards in the context of Web technologies and services can be traced back to 2006 when the IFLA Cataloguing Section’s ISBD Review Group decided to act upon its Material Designations Study Group recommendation to develop an XML Schema for the ISBD. The ISBD/XML Study Group[8] was set up in 2008 with such an objective; however, as the work by the FRBR Review Group[9] on FRBR[10] related to RDF had started in the previous year, the Study Group decided to bypass a general XML mark-up, and consider representing ISBD itself in RDF. The Study Group's three year project is now in its final phase, to be finished by December 2011. [11,12] The FRBR Review Group continued its work on representing IFLA models, extending it to the models for authority data (FRAD[13]), and subject authority data (FRSAD[14]). All three models as well as ISBD have been created using the Open Metadata Registry (OMR)[15].

It should be mentioned, however, that all these activities have been liaising with similar research in the field by other interested parties, feeding back into the development of IFLA standards representations in RDF[i]. It should also be noted that research was done to test the potential for applying RDA: Resource Description and Access as a content standard for UNIMARC, in addition to and in alignment with ISBD, in the context of the Semantic Web.[16]

The latest, third editions of UNIMARC formats for bibliographic and authority data were published in 2007[17] and 2009[18] respectively, with subsequent updates in preparation by the Permanent UNIMARC Committee (PUC). UNIMARC for authority data has already, in its 3rd edition, implemented specific features of the FRAD model in order to be aligned more closely to that model[19], while its alignment with FRSAD is still pending. The alignment of UNIMARC for bibliographic data with FRBR, and also the new, consolidated edition of ISBD[20], is in the process of approval. It goes without saying that UNIMARC formats follow closely other IFLA standards; in the case of bibliographic data this is ISBD in particular. Thus, the work on representing UNIMARC in RDF specifically for bibliographic data is the extension of the ISBD/XML Study Group's work which forms the basis of this paper.

UNIMARC namespaces

RDF requires classes and properties to be given machine-processable identifiers conforming to the syntax of the Uniform Resource Identifier (URI)[21]. A set of URIs with basic information about corresponding classes and properties, and published and managed in a single context is known as a namespace. All URIs in a namespace will usually be constructed from a common string of characters, known as a base domain, to which is added a distinguishing string, known as a local part. One advantage of this approach is that the base domain can be abbreviated to shorten the URI for display to humans; it is expanded automatically for machine-processing. UNIMARC RDF elements and vocabularies will initially be created and maintained using the OMR, following the same approach used for ISBD and the Functional Requirements family of metadata models (FRBR, FRAD, and FRSAD). In particular, the OMR supports multilingual labels and other annotations. This is an important requirement for IFLA standards which are intended for application in an international environment and designed for multilingual interoperability. UNIMARC has been translated from English into Chinese, Croatian, French, Italian, Lithuanian, Portuguese, Russian, etc. RDF is essentially language-neutral because it is intended for machine-processing, but allows labels, definitions, scope notes, and other annotations in multiple languages to be assigned to the same element. The longer-term infrastructure required to manage IFLA namespaces will be investigated and developed by the IFLA Namespaces Technical Group, following the recommendations made in the report by that Group[22].

[i] For more information, see: Dunsire, G.; M. Willer. Ibid.

Identifying UNIMARC elements by tag/subfield/character position

In this paper, abbreviated codes are used to identify UNIMARC elements specified by tags (fields), subfields, indicators, and character positions in the 1-- Coded information block, using the pattern:

Tag + 1st indicator + 2nd indicator + subfield code; using "b" to indicate a blank indicator or hash (#) value for a space. The subfield identifier ($) is not required because each abbreviated code pertains to a single subfield, with the sixth position indicating the subfield code.

e.g. 010bba = Number (ISBN)
e.g. 2001ba = Title Proper (title is significant)
e.g. 2000ba = Title Proper (title is insignificant)

For the Coded information block, the character position is added to form the abbreviated code.

e.g. 100bba8 = Type of publication date code
e.g. 100bba17-19 = Target audience code
e.g. 100bba34-35 = Script of title code

Namespace domains

One or more namespaces with corresponding base domains will be required to represent UNIMARC elements and vocabularies in RDF.

Re-use of existing namespaces

It is good practice to re-use RDF elements and vocabularies from existing namespaces, where appropriate: it saves time and effort in developing the elements and vocabularies, and in maintaining them; it is simpler to develop metadata applications and services; it encourages a mix-and-match approach to applications; it fosters the web of connected elements and linked data. It is most important, however, to ensure that re-used elements are tightly coupled to the standard using them, so that any change in their direct or indirect meaning (semantic neighbourhood) is immediately reflected in the related namespace in order to prevent semantic "drift" between the two namespaces.

UNIMARC Bibliographic is aligned with ISBD, which already has a published namespace for its element set and vocabularies (Area 0 Content form and media type area). So there is a choice for UNIMARC:

1. Re-use ISBD classes and properties where appropriate, instead of creating separate ones in the UNIMARC Bibliographic namespace. This option is only appropriate if proposed changes to either standard involve consideration of the impact on the other standard; that is, if both standards are managed and maintained "as one".

2. Represent all UNIMARC Bibliographic elements in a specified UNIMARC namespace, and link to equivalent classes and properties in the ISBD namespace. This option should be chosen if ISBD and UNIMAR continue to be developed separately, even if there is close liaison between them.

This choice must be made before there is any substantive development of a namespace for UNIMARC Bibliographic Format.

Table 1 shows an example of mapping potential UNIMARC/B properties to existing ISBD properties. Note that the namespace domain of each of the UNIMARC and ISBD property URIs is not included for the sake of brevity.

UNIMARC	English label	ISBD property	English label
P205bba	has edition statement	P1008	has edition statement
P205bbb	has issue statement*	P1011	has additional edition statement
P205bbd	has parallel edition statement	P1009	has parallel edition statement
P205bbf	has statement of responsibility relating to edition	P1010	has statement of responsibility relating to edition
P205bbg	has subsequent statement of responsibility**	P1010	has statement of responsibility relating to edition

Table 1. Example of mapping UNIMARC properties to existing ISBD RDF properties: UNIMARC Bibliographic 205 (Edition statement)

* Differences in labels, e.g. P205bbb, can be accommodated using the SKOS property for an alternate label. This has been used for some FRBR properties which have different labels in FRAD.

** This is an example where the UNIMARC element is more specific than the ISBD one. To preserve the UNIMARC element and finer granularity, a UNIMARC property is required in the UNIMARC Bibliographic namespace.

UNIMARC Authorities, as already mentioned, "takes into account attributes of the entities and entity relationships as specified in the *Functional Requirements for Authority Data: Conceptual Model* (FRAD)"[ii], in the following aspects: "change of terminology, definition of fields, and control subfield $5, Relationship Control […]. The blocks are renamed to 2-- Authorized Access Point, 4-- Variant Access Point, 5-- Related Access Point, and 7-- Authorized Access Point in Other Language and/or Script, while tags designate names of the entities which the controlled access points represent, such as Personal Name, Corporate Body Name, Title"[iii]. However, FRAD is a model, while UNIMARC/A is a content carrier at the level of application, so the equivalence of FRAD and UNIMARC/A definitions will need to be checked. Also, unlike ISBD, FRAD is an extension of FRBR, and itself re-uses appropriate FRBR namespace elements. It should be noted that the alignment between UNIMARC Bibliographic and Authorities is an intrinsic one, which means that UNIMARC/A develops following changes and additions to the bibliographic format. The correspondence between UNIMARC/A and UNIMARC/B is specified by subfield $3 in access point fields in UNIMARC/B, and specifically in Guidelines for Use in UNIMARC/A. Therefore, it will be necessary to investigate how to position UNIMARC/A in relation to UNIMARC/B 7-- Responsibility Block on the one side and FRAD on the other to determine element relationships for specific instances of linked data. For example, if a specific access point/entry element in a UNIMARC/B record has a corresponding UNIMARC/A record which has also been published in RDF, then a linked data chain can be established between them; otherwise the UNIMARC/B access point, for example 700 Personal Name – Primary Responsibility,

[ii] UNIMARC Manual: Authorities Format. Ibid., p. 14.
[iii] Willer, M. Foreword to the third edition. Ibid., p. 8.

used without the reference to the UNIMARC/A field by subfield $3 can be represented as a literal string.

Specific namespaces for UNIMARC

ISBD does not cover access points or headings and their corresponding authority records, so UNIMARC Authorities will not have corresponding ISBD classes or properties. A namespace for UNIMARC Authorities is definitely required.

The example in Table 1 shows that the ISBD elements do not extend to the same level of granularity as UNIMARC Bibliographic, so a namespace for UNIMARC Bibliographic is required for those elements not covered by ISBD, irrespective of the re-use of ISBD elements. UNIMARC Bibliographic and UNIMARC Authorities should have separate namespaces, to reflect the separate publication of the texts and distinguish between similar tag/subfield encodings.

It is proposed that the UNIMARC namespaces should follow the pattern already established by the IFLA Namespaces Task Group for ISBD and FRBR, FRAD, and FRSAD.

For UNIMARC Authorities format elements, the base namespace domain is:

http://iflastandards.info/ns/unimarc/unimarca/elements/

This can be abbreviated to "unimarca" for display purposes. Note that this is not a URL; it is a "cool" URI.

For UNIMARC Bibliographic format elements, the base namespace domain is:

http://iflastandards.info/ns/unimarc/unimarcb/elements/

This can be abbreviated to "unimarcb".

For UNIMARC vocabularies, a separate base domain is used for each vocabulary, following ISBD practice. The base domain consists of a vocabulary-specific identification string added to an overall base domain for UNIMARC vocabularies:

http://iflastandards.info/ns/unimarc/terms/ + identification string

The specific identification string cannot be based on the tag/indicators/subfield/character position code because some vocabularies, such as Script, are assigned to more than one tag within UNIMARC/B and UNIMARC/A:

UNIMARC/A: 100bba21-22
UNIMARC/B: 100bba34-35

Instead, an abbreviation of the vocabulary title can be used:

e.g. http://iflastandards.info/ns/unimarc/terms/graphicssmd as the namespace for the vocabulary for the specific material designation for graphics used in 116bba1.

This example can be abbreviated to "unimarcgsmd" for display purposes.

Note that there is no need to use the base domain to indicate if the vocabulary is from Bibliographic or Authorities. Vocabularies can be used as the object of any appropriate property used in a triple, and usage of the vocabulary will be represented in an application profile as a Vocabulary encoding scheme linked to the relevant RDF properties. This approach decouples a vocabulary from its specific use within UNIMARC, and will make it easier for other communities to re-use it for non-UNIMARC applications.

Application profiles

One or more DC Application Profiles will be required for UNIMARC, to represent the re-use, if any, of ISBD and FRAD classes and properties, the use of aggregated statements composed of two or more properties (as in ISBD), the use of specified vocabularies as Vocabulary Encoding Schemes, and any other constraints on the use of elements in a well-formed UNIMARC Bibliographic or Authorities record, such as mandatory and repeatability status.

Both UNIMARC/B and UNIMARC/A specify mandatory elements, and repeatable and non-repeatable elements, in UNIMARC records. Mandatory fields in both formats are 001 Record Identifier, 100 General Processing Data (certain data elements only, identical in both formats), and 801 Originating Source, while specific to a format are 200$a Title proper in UNIMARC/B (apart from some fields specific to the type of resource), and 2-- Authorized Access Point in UNIMARC/A. Both formats specify repeatable and non-repeatable elements at the level of fields and subfield identifiers, such as 010 ISBN is repeatable, while 010$a Number is not in UNIMARC/B. UNIMARC/A field 220 Authorized Access Point – Family Name is repeatable, but only for alternative script forms, while 220 $a Entry element is not. The order of subfield identifiers in a UNIMARC record is not specified, as order is determined by the data.

Meta-metadata

Data about a specific UNIMARC record is held in the Record label and 1-- Coded information block. This is meta-metadata, or data about metadata. In RDF there are a number of techniques that can be used to represent such data, such as the language qualifier that can be added to a literal string, for example a title, used as the object of a triple; for example "@en" indicates that the string is in English, "@fr" for French, etc. These techniques do not require specific UNIMARC elements, and are excluded from further discussion in this paper.

There are also meta-metadata specific to a UNIMARC record as an instance of the ISO 2709 structured exchange record format, such as Record length, Implementation codes, Indicator length, etc. These elements are not relevant when metadata is represented in RDF as triples.

Coded information block vocabularies

The codes and corresponding values and definitions used in the Coded information block to describe a resource (rather than the UNIMARC record) are best represented as a SKOS vocabulary, in the same way as the ISBD Area 0 vocabularies.

The UNIMARC code for a vocabulary term can be used as the local part of its URI.

e.g. http://iflastandards.info/ns/unimarc/terms/graphicssmd#a = "collage"

Where a term code is a number, it should be prefixed with a letter such as "T" (for term) to avoid XML problems with local parts starting with a numeric character; this follows ISBD practice.

Using the term code in this way retains the language-independence of the URI, avoids overloading the URI with semantics, and avoids confusion if the (English) term is changed in the future (say from "collage" to "mixed-media two-dimensional sculpture").

The UNIMARC code itself can be explicitly represented using the skos:notation property. The following example triples using this property have the URI for a term as the subject and the term code as the value of the object:

<http://iflastandards.info/ns/unimarc/terms/graphicssmd#a> skos:notation "a".
(or, using a namespace abbreviation: unimarcgsmd#a skos:notation "a".)
<http://iflastandards.info/ns/unimarc/terms/publicationdatetype#f> skos:notation "f".
<http://iflastandards.info/ns/unimarc/terms/titlescript#ca> skos:notation "ca".

Similarly, the term itself can be represented using the skos:prefLabel property with a language qualifier.

e.g. <http://iflastandards.info/ns/unimarc/terms/graphicssmd#a> skos:prefLabel "collage"@en.
e.g. <http://iflastandards.info/ns/unimarc/terms/publicationdatetype#f> skos:prefLabel "monograph, date of publication uncertain"@en.
e.g. <http://iflastandards.info/ns/unimarc/terms/titlescript#ca> skos:prefLabel "Cyrillic"@en.

Table 2 gives a full example of a Coded information block vocabulary with labels in English, Italian and Portuguese taken from official translations of UNIMARC.

N	PrefLabel@en	PrefLabel@it	PrefLabel@pt	Definition@en	ScopeNote@en
a	collage	collage	colagem	An original work created by affixing various materials (paper, wood, newspaper, cloth, etc.) to a surface.	
b	drawing	disegno	desenho	An original visual representation (other than a print or painting) made with pencil, pen, chalk, or other writing instrument on paper or similar non-rigid support.	

c	painting	pittura	pintura	An original visual representation produced by applying paint to a surface.	
d	photomechanical reproduction	riproduzione fotomeccanica	reprodução fotomecânica	Any picture produced in imitation of another picture through the use of a photographic process to transfer the image to a printing surface.	~~Hence,~~ a snapshot made to document a painting or a Xerox copy of a print are considered photomechanical reproductions. Art reproductions, postcards, posters, and study prints are included here.
e	photonegative	fotonegativo	negativo fotográfico	A piece of film, a glass plate, or paper on which appears a "negative" image, i.e. directly opposite to a "positive" image (photoprint), slide, or transparency. Used to produce a positive print.	Does not include negative photoprints, photoprints that are a combination of negative and positive images, photographs or solarized prints, all of which are considered to be techniques used when making photoprints.
f	photoprint	riproduzione fotografica	positivo fotográfico	A positive image made either directly or indirectly on a sensitised surface by the action of light or other radiant energy.	The term "photoprint" is used here as a more precise term than "photograph", which technically can cover both the print and the negative. Radiographs and opaque stereographs are included here.

h	picture	immagine	imagem	A two-dimensional visual representation accessible to the naked eye and generally on an opaque backing.	This term is used when a more specific designation is unknown or not desired.
i	print	stampa	gravura	A design or picture transferred from an engraved plate, wood block, lithographic stone, or other medium.	Generally, there are four types: planographic print, relief print, intaglio print, and stencil print.
k	technical drawing	disegno tecnico	desenho técnico	A cross section, detail, diagram, elevation, perspective, plan, working plan, etc., made for use in an engineering or other technical context.	
m	master	master	matriz	Any plate, mould, matrix, die etc. which allows the reproduction of the same impression.	
z	other non-projected graphic type	altro tipo di documento grafico non proiettabile	outro material gráfico não-projectável	~~Other types not included in the above.~~ [Types other than collage, drawing, painting, photomechanical reproduction, photonegative, photoprint, picture, print, technical drawing, master.]	Includes mixed media productions made by a combination of freehand and printing techniques when one or the other does not predominate. In some cases, where mixed media are applied, one must decide

					whether the creator intends the item to be a photoprint (even though it is painted over the photographic image). Hand colouring is considered a technique applied to a printing process; this aspect is covered by a character position 3. Computer-produced graphics and the various duplication masters (including spirit masters and transparency masters) are included here.

Table 2. Full example of a Coded Information Block vocabulary: Vocabulary of 116bba0 = Coded data for graphics: Specific material designation (column N = Notation, which is also the local part of the URI)

Terms, definitions, and scope notes are taken from the following source texts:

@en: http://archive.ifla.org/VI/8/unimarc-concise-bibliographic-format-2008.pdf
@it: http://unimarc-it.wikidot.com/116
@pt: http://www.ifla.org/files/uca/Unimarc_bib_3%C2%AAed_abrev.pdf

Definitions and scope notes include mark-up to show their derivation from the English text.

Using external SKOS vocabularies

Some UNIMARC coded value sets are explicitly based on an external vocabulary or terminology. For example, the language of incipit in 036bbaz and the language of the item in various places in tag 101 use a 3-letter code taken from Appendix A, which is the same as the MARC List for Languages. This list is available as a SKOS vocabulary[23], which can be used directly in any UNIMARC triples. Similarly, the geographic area codes for 660bba are also available in SKOS.[24]

The country of publication in 102bba uses a 2-letter code from Appendix B, which is ISO 3166-1. This is not the same as the MARC List for Countries which the Library of Congress has also published as a

SKOS vocabulary[25]. However, an RDF representation of ISO 3166-1 is available[26], although further investigation of its suitability for UNIMARC is required.

The availability of SKOS representations of other external vocabularies used in UNIMARC needs to be checked and verified. If no SKOS representation can be found, the Permanent UNIMARC Committee would have to contact the owner of the vocabulary to discuss the development of an appropriate representation in RDF.

Some internal UNIMARC vocabularies may already have suitable SKOS representations, even though they are not explicitly based on an external vocabulary or terminology. A possible example is the character set codes used in 100bba26-29. That is, another community may have developed a SKOS representation of a similar vocabulary which contains all of the terms and codes used by UNIMARC; if so, the SKOS URIs can be re-used by UNIMARC. This also requires further investigation and verification, for example by using an ontology search engine such as Swoogle.

Date values

Coded elements which are dates in a specified format, such as year, can be represented using the rdfs:range property. For example, Publication dates 1 and 2 in the UNIMARC Bibliographic General processing data can use the triples (following the URI convention given below):

unimarcb:100bba09-12 rdfs:range xsd:gYear.
unimarcb:100bba13-16 rdfs:range xsd:gYear.

UNIMARC classes

As with ISBD, there is only one RDF class to consider for UNIMARC Bibliographic, excluding classes for Syntax encoding schemes required for aggregated elements, as noted below. This is the ISBD class Resource, which can be used as the domain for all UNIMARC Bibliographic RDF properties; there is no need to create a UNIMARC class for Resource.

Classes for UNIMARC Authorities require further investigation, especially in relation to FRAD/FRSAD. The FRAD namespace has classes for all FRAD entities such as Bibliographic Entity, Name, Identifier, Controlled Access Point, Rules and Agency, and subclasses for Name of a Person, Name of a Corporate Body, Name of a Family and Name of a Work; it also has a sub-class for Corporate Body because its definition is modified from the one in FRBR, and a class for Family which is not defined an entity in FRBR's published document. The FRAD namespace does not, however, include classes for other entities such as Person, Work, Expression, Manifestation, etc., because they are already published in the FRBR namespace. On analysis, it becomes obvious that FRBR classes do not accommodate all types of possible UNIMARC/A classes: Person, Corporate Body, Work could in general be considered as aligned, but there are examples where some UNIMARC/A types of entities or candidates for classes require special analysis. Such an example is Place which in FRBR is defined simply as "A location", which can be aligned with UNIMARC/A Territorial or geographical name, but not really to Place Access – the access point/field which was originally designed to record the place (country and town) of printing for older publications, but was subsequently extended to cover Place and date of publication, performance, provenance, etc.; additionally, as UNIMARC/A is an integrated format for name and subject authorities, the "place" class should be considered also in the context of FRSAD. Another example is Work/Expression: UNIMARC/A defines among its types of entities the following: title, collective title, name/title and name/collective title, but in its 3rd edition does not distinguish whether the type of entity is Work or Expression[iv]. FRAD has a subclass for Name of a Work, but if UNIMARC defines a new type of entity Expression, it should add to its namespace a UNIMARC subclass for Name of an Expression. Possible classes outside the scope of FRBR and FRAD are the UNIMARC/A types of entities Trademark, and Form, Genre and Physical Characteristics, which FRSAD intentionally excluded from its consideration.

UNIMARC properties

Generally, UNIMARC tag/indicator/subfield elements will be represented as RDF properties, following ISBD practice.

Not all UNIMARC elements are suitable or appropriate for representation as RDF properties. These include the meta-metadata elements discussed above. However, other data in the record label, such as type of record (in both formats) are actually resource metadata; although much of this information should be present in the body of the record, this is not always the case and some of the record label elements will require RDF representation. These elements are identified as, for example, bibliographic and hierarchical levels in UNIMARC/B, and type of entity in UNIMARC/A. This requires further investigation.

The tag/indicator/subfield abbreviated codes used in this paper can form the local part of the URI. "Slash" URIs are recommended rather than "hash" URIs where there are large numbers of properties. The local part should be prefixed with a letter to avoid XML problems with local parts starting with a numeric character; ISBD uses "P" (for property) and UNIMARC can follow this convention. For example, the URI for the Edition statement might be:

http://iflastandards.info/ns/unimarc/unimarcb/elements/P205bba

(or, using the namespace abbreviation: unimarcb:P205bba)

This approach is language-independent, avoids semantic loading of the URI, and avoids confusion if the "caption" associated with the tag, indicator, or subfield changes.

The caption itself can be represented using the rdfs:label property and RDF language qualifier. The caption may require synthesis from the separate tag, indicator and/or subfield captions. Following ISBD practice, the RDF property labels can be made verbal by prefixing the caption with "has ".

e.g. unimarcb:P205bba rdfs:label "has edition statement"@en.

The OMR requires a separate registry name using the reg:name property; this can be constructed in the usual way, as a CamelCase version of the rdfs:label.

e.g. unimarcb:P205bba reg:name "hasEditionStatement".

[iv] PUC has a specific mechanism to distinguish between the Work and Expression in the process of approval, to be published in the next update of UNIMARC formats.

Table 3 gives a full example of the RDF properties derived from a single tag with no indicators.

URI	Label@en	Name
P205bba	has edition statement	hasEditionStatement
P205bbb	has issue statement	hasIssueStatement
P205bbd	has parallel edition statement	hasParallelEditionStatement
P205bbf	has statement of responsibility relating to edition	hasStatementOfResponsibilityRelatingToEdition
P205bbg	has subsequent statement of responsibility	hasSubsequentStatementOfResponsibility

Table 3. Full example of RDF properties representing a UNIMARC field with no indicators: Edition statement

As already discussed, each unique combination of indicators and a subfield within a tag potentially constitutes a separate RDF property, with a suitable distinct label. A method for achieving this is to qualify the subfield caption with the indicator "caption", as shown in Table 4.

URI	Label@en	Name
P206bba	has mathematical data statement (unstructured)	hasMathematicalDataStatementUnstructured
P206bbb	has statement of scale (unstructured)	hasStatementOfScaleUnstructured
P206bbc	has statement of projection (unstructured)	hasStatementOfProjectionUnstructured
P206bbd	has statement of coordinates (unstructured)	hasStatementOfCoordinatesUnstructured
P206bbe	has statement of zone (unstructured)	hasStatementOfZoneUnstructured
P206bbf	has statement of equinox (unstructured)	hasStatementOfEquinoxUnstructured
P2060ba	has mathematical data statement (structured)	hasMathematicalDataStatementStructured
P2060bb	has statement of scale (structured)	hasStatementOfScaleStructured
P2060bc	has statement of projection (structured)	hasStatementOfProjectionStructured
P2060bd	has statement of coordinates (structured)	hasStatementOfCoordinatesStructured
P2060be	has statement of zone (structured)	hasStatementOfZoneStructured
P2060bf	has statement of equinox (structured)	hasStatementOfEquinoxStructured

Table 4. Full example of RDF properties representing a UNIMARC field with a single, binary indicator: Material Specific Area: Cartographic materials – mathematical data

When both indicators are used in a tag, with multiple values for each indicator, the number of potential RDF properties is affected by a combinatorial explosion, as demonstrated in Table 5.

URI	Label@en
P210bba	has place of publication, distribution, etc. (sequence of publication data not applicable or earliest available publisher; produced in multiple copies, usually published or publically distributed)
P210b1a	has place of publication, distribution, etc. (sequence of publication data not applicable or earliest available publisher; not published or publically distributed)
P2100ba	has place of publication, distribution, etc. (intervening publisher; produced in multiple copies, usually published or publically distributed)
P21001a	has place of publication, distribution, etc. (intervening publisher; not published or publically distributed)
P2101ba	has place of publication, distribution, etc. (current or latest publisher; produced in multiple copies, usually published or publically distributed)
P21011a	has place of publication, distribution, etc. (current or latest publisher; not published or publically distributed)

Table 5. Partial example of RDF properties representing a UNIMARC field with two multiple-valued indicators: Publication, distribution, etc.

The total number of potential properties for this example is 3 (values for 1st indicator) times 2 (values for 2nd indicator) times 8 (subfields): 48.

There are UNIMARC tags with much larger combination numbers:

327 Contents note: 4 x 2 x 12 = 96
620 Place and date of publication, performance, etc.: 7 x 3 x 15 = 315
621 Place and date of provenance: 7 x 3 x 16 = 336
852 Location and call number: 7 x 4 x 16 = 448

These require further investigation to determine if some combinations are invalid and do not require a separate property.

Aggregated statements

All repeatable tags with more than one subfield form an aggregated statement. It is necessary to keep the subfields together for each repeat, so that they do not get mixed up.

e.g. 010bba International Standard Book Number Number + 010bbb: International Standard Book Number Qualification

Aggregated statements are represented in RDF using Syntax encoding schemes (SES), and ISBD practice should be followed. Re-use of ISBD elements which are themselves aggregated statements will avoid the need for developing UNIMARC equivalents.

Conclusion and recommendations

The involvement of IFLA in the activity of publishing its internationally-agreed models and standards in RDF, as the first step to mark-up library metadata as authoritative and trustworthy in the Semantic Web, has already been done. However, although these first steps involve all three conceptual models, FRBR, FRAD and FRSAD, and the bibliographic standard ISBD, further work is necessary. This paper, by presenting some solutions and raising questions for further analysis, argues for the need to represent IFLA's UNIMARC formats – bibliographic and authorities - in the same way. The authors also argue that the coordination of the work on representing IFLA standards documentation should be brought more closely together because in practice they are considered and used in unison, and also because their further development would be more efficient and economical. Another aspect of the work in representing standards in RDF is that it offers feedback on the standards themselves, their structure, precision in wording concepts and definitions, consistency, interoperability with other related library and different communities metadata standards, etc., which is required in the new technological paradigm of the Semantic Web.

Recommendations to the Permanent UNIMARC Committee for further discussion and approval are:

- Approve the method of identifying UNIMARC elements and vocabularies.
- Decide on initial creation and maintenance of UNIMARC elements and vocabularies in the Open Metadata Registry (OMR).
- Support and promote the translation of UNIMARC classes and properties in national languages.
- Decide between re-use of existing ISBD namespaces for UNIMARC/B or representing all UNIMARC/B elements and link to existing ISBD classes and properties as appropriate.
- Investigate further the re-use of existing FRAD/FRBR and FRSAD namespaces or representing all UNIMARC/A elements and link to existing FRAD/FRBR/FRSAD classes/subclasses and properties as appropriate.
- Approve the pattern for namespaces for UNIMARC/B and /A elements and vocabularies.
- Discuss and consider the requirements for Application Profiles for UNIMARC.
- Check and verify the availability of SKOS representations of other external vocabularies used in UNIMARC.
- Investigate and verify internal UNIMARC vocabularies for suitable SKOS representations; consider approaching the owners of external vocabularies to liaise on developing SKOS representations.
- Investigate further the appropriate classes for UNIMARC/A in relation to UNIMARC/B, FRAD/FRBR and FRSAD.
- Investigate further the "combinatorial explosion" of UNIMARC properties; determine if some combinations are invalid and do not require a separate property.
- Consider and approve the re-use of aggregated ISBD elements which are represented in RDF using Syntax encoding schemes (SES), which will avoid the need for developing UNIMARC equivalents.
- Monitor relevant MARC21 developments, especially the Bibliographic Framework Transition Initiative recently announcement by the Library of Congress[27].

The authors of this paper highly recommend that the PUC propose to IFLA the funding of the development of UNIMARC representation in RDF as a research and development project.

References

[1] Bizer, Christian; Tom Heath, Tim Berners-Lee. Linked data – the story so far. // International Journal on Semantic Web and Information Systems (IJSWIS), vol. 5, issue 3. (2009). Pre-print available at: http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

[2] W3C. SKOS Simple Knowledge Organization System - home page. 2010. Available at: http://www.w3.org/2004/02/skos/

[3] OCLC. WorldCat facts and statistics. 2011. Available at: http://www.oclc.org/worldcat/statistics/default.htm

[4] Moen, William E.; Penelope Benardino. Assessing metadata utilization: an analysis of MARC content designation use. // International Conference on Dublin Core and Metadata Applications, DC-2003--Seattle Proceedings. Available at: http://dcpapers.dublincore.org/ojs/pubs/article/view/745/741

[5] Coyle, Karen; Thomas Baker. Guidelines for Dublin Core application profiles. 2009. Available at: http://dublincore.org/documents/profile-guidelines/index.shtml

[6] W3C OWL Working Group. OWL 2 Web Ontology Language: document overview. 2009. Available at: http://www.w3.org/TR/owl2-overview/

[7] Graves, Mike; Adam Constabaris, Dan Brickley. FOAF: connecting people on the Semantic Web. // Knitting the Semantic Web / Jane Greenberg, Eva Méndez, editors. Binghampton, NY: The Howarth Information Press, 2007. P. 196.

[8] IFLA. Cataloguing Section. ISBD Review Group, ISBD/XML Study Group. 2008. Available at: www.ifla.org/en/node/1795

[9] IFLA. Cataloguing Section. FRBR Review Group. Meeting Report Durban, August 21, 2007. Available at: www.ifla.org/files/cataloguing/frbrrg/meeting_2007.pdf

[10] Functional requirements for bibliographic records: final report / IFLA Study Group on the Functional Requirements for Bibliographic Records. München: Saur, 1998. Amended 2009, available at: www.ifla.org/en/publications/functional-requirements-for-bibliographic-records

[11] Dunsire, Gordon; Mirna Willer. Standard library metadata models and structures for the Semantic Web. // Library hi tech news, vol. 28, no. 3. (2011), pp. 1-12. Available at: http://dx.doi.org/10.1108/07419051111145118

[12] Willer, Mirna; Gordon Dunsire, and Boris Bosančić. 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://dx.doi.org/10.4403/jlis.it-4536

[13] Functional requirements for authority data : a conceptual model / edited by Glenn E. Patton ; IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). Final report, December 2008 / approved by the Standing Committees of the IFLA Cataloguing Section and IFLA Classification and Indexing Section, March 2009. München: K. G. Saur, 2009. Also available at: www.ifla.org/publications/functional-requirements-for-authority-data

[14] Functional requirements for subject authority data (FRSAD) : a conceptual model / edited by Marcia Lei Zeng, Maja Žumer and Athena Salaba ; IFLA Working Group on Functional Requirements for Subject Authority Records (FRSAR). Berlin/München: De Gruyter Saur, 2011. Final report available at: www.ifla.org/files/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf

[15] Open Metadata Registry. No date. Available at: http://metadataregistry.org/

[16] Dunsire, Gordon. UNIMARC, RDA and the Semantic Web. In: International Cataloguing and Bibliographic Control (ICBC), vol. 39, no. 2 (April/June 2010). Based on a paper presented to the World Library and Information Congress: 75th IFLA General Conference and Assembly, 23-27 August 2009, Milan, Italy; available at: www.ifla.org/files/hq/papers/ifla75/135-dunsire-en.pdf

[17] UNIMARC Manual: Bibliographic Format / edited by Alan Hopkinson. 3rd edition. München: Saur, 2007.

[18] UNIMARC Manual: Authorities Format / edited by Mirna Willer. 3rd edition. München: Saur, 2009.
[19] Willer, Mirna. Foreword to the third edition. // UNIMARC Manual: Authorities Format / edited by Mirna Willer. 3rd edition. München: Saur, 2009. Pp. 7-9.
[20] International standard bibliographic description (ISBD). Consolidated edition. Draft as of 2010-05-10. Available at: www.ifla.org/files/cataloguing/isbd/isbd_wwr_20100510_clean.pdf. Standard edition is expected to be published by IFLA 2011.
[21] Semantic Web Education and Outreach (SWEO) Interest Group. Cool URIs for the Semantic Web. 2008. Available at: http://www.w3.org/TR/cooluris/
[22] IFLA Namespaces Task Group. IFLA namespaces - requirements and options. 2010. Available at: http://www.ifla.org/files/classification-and-indexing/ifla-namespaces-requirements-options-report_corrected.pdf
[23] Library of Congress. No date. MARC list for languages. Available at: http://id.loc.gov/vocabulary/languages.html
[24] Library of Congress. No date. MARC list for geographic areas. Available at: http://id.loc.gov/vocabulary/geographicAreas.html
[25] Library of Congress. No date. MARC list for countries. Available at: http://id.loc.gov/vocabulary/countries.html
[26] Martin, Earle. ISO 3166 RDF representation. 2005. Available at: http://downlode.org/Code/RDF/ISO-3166/
[27] Library of Congress. Bibliographic Framework Transition Initiative. 2011. Available at: http://www.loc.gov/marc/transition/