

5JSC/Editor/3
31 May 2007

To: Joint Steering Committee for Development of RDA
From: Tom Delsey, RDA Editor
Subject: Encoding RDA Data

The attached document was discussed at the April 2007 JSC meeting. It has not been updated as a result of that meeting.

No constituency responses to this document are required.

Encoding RDA data

RDA as a metadata element set

RDA can be viewed as a metadata element set (similar to the Dublin Core Metadata Element Set) insofar as it:

- a) specifies a set of elements, element sub-types, and sub-elements that reflect the properties of a resource
- b) defines each element, element sub-type, and sub-element
- c) establishes parameters for the value representations recorded for each element, element sub-type, and sub-element.

As a metadata element set, RDA incorporates the following additional features (similar to those in the Dublin Core Library Application Profile):

- a) it establishes requirements (required, required if applicable, and optional) for each element, element sub-type, and sub-element
- b) it incorporates (either directly or by reference) encoding schemes for certain elements, element sub-types, and sub-elements (e.g., the ISO encoding schemes for standard identifiers such as ISBNs and ISSNs).

RDA has not, at this stage, been formally registered as a metadata element set, nor have Uniform Resource Identifiers (URIs) been assigned to RDA data elements, element sub-types, and sub-elements.

Encoding syntax

In the absence of a formally registered RDA metadata element set, data created using RDA must be encoded using a "proxy" syntax. That is to say that RDA elements, element sub-types, and sub-elements must be encoded using a syntax that can be mapped to RDA specifications. The candidate proxies include MARC 21, MODS, and XML implementations of Dublin Core.

Using a proxy encoding syntax means that the data is encoded using the element set defined for that syntax rather than the RDA element set *per se*. RDA elements, element sub-types, and sub-elements are mapped to the nearest corresponding element, element sub-type, or sub-element defined in the proxy syntax, and encoded accordingly.

Once the initial phase of RDA development has been completed, it should be possible to formally register an RDA metadata element set. At that stage, RDA data could be encoded directly in XML using the registered RDA terms to identify the elements, element sub-types, and sub-elements. The mappings of RDA to MARC 21, MODS, and Dublin Core could then be used as the basis for conversion to and from the RDA encoding syntax when importing or exporting data in MARC 21, MODS, or Dublin Core in XML.

Encoding schemes for RDA value representations

As noted above, for certain elements, element sub-types, and sub-elements defined in the RDA element set, the RDA instructions reference “external” encoding schemes (e.g., the ISO encoding schemes for standard identifiers such as ISBNs and ISSNs). For certain other elements, element sub-types, and sub-elements, RDA provides instructions on recording value representations in a structured form that function, in effect, as “internal” encoding schemes. For example, the controlled lists of values for elements such as Media type, Carrier type, and Content type function as vocabulary encoding schemes for those elements. At this stage, however, none of the controlled lists of values specified in RDA have been formally registered as vocabulary encoding schemes.

When RDA data is encoded in a proxy encoding syntax, the values for which RDA provides an “internal” vocabulary encoding scheme must be recorded in the corresponding element, element sub-type, or sub-element defined in the proxy encoding syntax. In some instances, the corresponding element, element sub-type, or sub-element in the proxy encoding syntax will have defined its own “internal” vocabulary encoding scheme (e.g., the coded data in the 007 and 008 fields in MARC 21). In those cases, the vocabulary specified for the RDA element, element sub-type, or sub-element must be mapped to the vocabulary specified in the proxy encoding syntax. In other instances, there will be no vocabulary encoding scheme defined in the proxy encoding syntax (e.g., the variable data for “other technical details” recorded in field 300, subfield b in MARC 21), and the RDA value representation must be recorded simply as an unstructured value string.

Once the initial phase of RDA development has been completed, it should be possible to formally register the controlled lists of terms specified for various RDA elements, element sub-types, and sub-elements as vocabulary encoding schemes, with assigned URIs. At that stage, the URI could be used to identify such a term recorded in the corresponding element, element sub-type, or sub-element in a proxy encoding syntax as a term derived from a specific RDA vocabulary encoding scheme (assuming the proxy syntax allows the use of “external” vocabulary encoding schemes and makes provision for identifying the scheme from which the recorded term was derived).

Alternatives for recording RDA values

The Joint Steering Committee (JSC) is currently assessing the implications of incorporating into RDA alternatives for recording values for three specific types of data: (1) elements, element sub-types, and sub-elements for which RDA specifies a controlled list of values; (2) designations used to indicate roles and relationships; and (3) elements, element sub-types, and sub-elements for which there is an international standard encoding scheme.

Controlled lists of values

The current draft of RDA specifies controlled lists of values for a number of elements reflecting characteristics of both the carrier and the content of the resource being described. The values specified in those controlled lists are all represented as terms (i.e., text strings). Consideration is being given to incorporating a general guideline into RDA that would allow the recording of a coded value as a substitute for a term specified in a controlled list.

It is assumed that when data formulated according to RDA specifications is recorded using a proxy encoding syntax such as MARC 21, a term derived from a controlled list for a specific RDA element, element sub-type, or sub-element would normally be recorded as an unstructured value string in the corresponding variable field of the encoding syntax.

If the alternative under consideration is introduced, a coded value defined in the proxy encoding syntax (e.g., in field 007 or 008 in MARC 21) that corresponds to the term specified in RDA could be recorded as a substitute for recording the RDA term in the corresponding variable field. If there is a coded value defined in the proxy encoding syntax that is the direct equivalent of the RDA term, that coded value would be recorded. If, however, there is no coded value defined in the proxy encoding syntax that is the direct equivalent of the RDA term, a coded value for "other" would have to be recorded in the coded data field. In that case, the term specified in RDA (or a more appropriate or specific term, as provided for with all "open" lists in RDA) could be recorded in the corresponding variable field in the proxy encoding syntax.

The key factors relating to the introduction of such an alternative that JSC needs to consider are the following:

- a) the degree of correspondence (and/or conflict) between the RDA values and the coded values defined in the encoding syntaxes that are most likely to be used as proxies for encoding RDA data
- b) the potential advantages to be gained by recording values in a coded form that could be used to facilitate automated processing of the data (e.g., in record matching, as search limiters, or to generate the display of an equivalent term in a form suited to the community that the database is designed to serve)
- c) the implications for user displays of recording certain of the specified values for an element, element sub-type, or sub-element as coded values (e.g., in field 007 or 008 in MARC 21) while coding other specified values (i.e., those that could only be coded as "other" in a coded data field in the proxy encoding syntax) as text strings (e.g., in the corresponding variable field in MARC 21)

The other alternative that could be considered is simply to acknowledge the possibility of recording an RDA value in an equivalent coded form in a proxy encoding syntax not as a substitute for, but in addition to, recording the specified RDA value as a term (i.e., a text string) in the corresponding variable field. That alternative would not necessarily require "redundant" data entry if the data entry software provided a menu of specified terms to be recorded, and the selection of a term triggered both the recording of the term in the corresponding variable field and the generation of the equivalent coded value for the corresponding coded data element.

Ultimately, JSC's assessment of the alternatives will likely hinge on what can realistically be expected by way of support for both data entry and data display from the developers of the software that will manage the data created using RDA.

Designations of roles and relationships

The current draft of RDA provides for the optional use of a designation of role in conjunction with an access point representing a person, family, or corporate body

associated with the resource being described, to indicate the role played by that person, family, or corporate body in relation to the resource. Similar provisions are made for the use of a designation to indicate the nature of the relationship between the resource being described and a related resource. JSC is currently assessing the implications of allowing the use of either a coded value or a term for such designations.

For designations of role, the current draft does not specify the designations, but recommends the use of “standard” lists. The draft does, however, define element sub-types for associated persons, families, and corporate bodies that reflect the role played by that person, family, or corporate body in relation to the resource. Similarly, for designations of relationship, the current draft does not specify the designations (nor, in this case, does it explicitly recommend the use of “standard lists”). It does, however, define element sub-types that reflect the nature of the relationship between the resource being described and the related resource.

To the extent that the element sub-types defined in RDA for persons, families, and corporate bodies associated with the resource and for related resources can be mapped to an equivalent element sub-type in the proxy encoding syntax, the definition of the element sub-type itself will serve as a means of designating the role or relationship. However, where a more specific designation of role or relationship is needed, a term or code derived from a “standard” list would have to be used in conjunction with the appropriate element sub-type.

The key factors relating to allowing the use of designations of role and relationship derived from “standard lists” that JSC needs to consider are the following:

- a) the degree of correspondence (and/or conflict) between the RDA element sub-types that reflect roles and relationships and the values defined in the “standard lists” defined for use with the proxy encoding syntaxes that are most likely to be used for encoding RDA data
- b) the implications for user displays of recording designations derived from a variety of “standard” lists that are independent of RDA specifications.

The alternative that has been suggested for consideration is to specify RDA values for designations of role and relationship (i.e., at a more specific level than that identified by the element sub-type).

International standard encoding schemes

As previously noted, there is at least one instance in the current draft of RDA where the instructions specify (by reference) the use of an international standard encoding scheme when recording the value for an element (i.e., resource identifiers for which there is an international standard). However, when considering the suggested use of the same approach for the Duration element, JSC chose not to make specific reference to the relevant ISO encoding scheme in the instructions, but rather to word the instruction in neutral terms and to add an example illustrating duration formatted according to the ISO standard.

The key factors relating to referencing international standard encoding schemes (either as the prescribed form for recording an RDA element, element sub-type, or sub-element, or as an alternative form) that JSC needs to consider are the following:

- a) the potential advantages to be gained by recording values in accordance with a standard encoding scheme that could be used to facilitate automated processing of the data (e.g., for retrieval or matching)
- b) the implications for user displays of recording data in a form that is designed primarily for automated processing and may not be in a form that is readily recognized or understood by a user.

Again, JSC's assessment of the alternatives will likely hinge on what can realistically be expected by way of support for data display from the developers of the software that will manage the data created using RDA.