To: Joint Steering Committee for Development of RDA

From: Gordon Dunsire, Chair, JSC Technical Working Group

Subject: RDA accommodation of relationship data

Abstract

This paper discusses the general approaches used by RDA to accommodate data about entities related to the entity being described. The paper makes some general recommendations for developing RDA to improve its accommodation of relationship data.

Introduction

A task (#8) for the JSC Technical Working Group in 2015 is to investigate and prepare proposals or discussion papers on how RDA accommodates relationship data as unstructured data, structured textual notes, authorized access points, and identifiers (including linked data).

Scope

Relationship data refers to data identifying an entity that is related to the resource being described, and the nature of the relationship.

This paper is primarily concerned with the data identifying the related entity. RDA accommodates the identification of relationships in specific elements and relationship designators, discussed in 6JSC/CILIP Rep/3. There are few technical issues with this approach, so the paper discusses data about the relationship only when it is entangled with data about the related entity.

To clarify, the task is focussed on how related entities are identified, not how relationships are identified.

The Working Group took into consideration the discussion on 6JSC/ALA/Discussion/3 Instructions for Recording Relationships: Discussion Paper.[1]

Models

RDA attempts to accommodate the range of entity identification data that is typically used for cultural heritage and information resources.

[1] http://www.rda-jsc.org/6JSC/ALA/Discussion/3

RDA 24.4 refers to four basic methods that can be used:

1. Unstructured description
2. Structured description
3. Authorized access point
4. Identifier

This is nicknamed the "four-fold path".

RDA 24.4, covering *Works*, *Expressions*, *Manifestations*, and *Items*, allows all four methods to be used.

RDA 29.4 and RDA 18.4.1.1, covering *Persons*, *Families*, and *Corporate Bodies*, only allow two of those methods to be used:

1. Authorized access point

2. Identifier

Instructions for recording subject relationships between *Works* and all RDA entities are lacking.

Path 1: Unstructured description

RDA Glossary description: “A full or partial description of a resource written as a sentence, paragraph, etc.”

RDA gives an explicit instruction for providing an unstructured description of a related entity at RDA 24.4.3 (Description of the Related Work, Expression, Manifestation, or Item).

Example:

"Reprint of the revised and updated edition published in 1971 by Farrar, Straus & Giroux"

The data are recorded as a literal text string.

The data describe aspects of one or more related entities and relationships.

The nature of the relationship is described in the data itself.

The values are not recorded or aggregated according to any specified schema. Generally, the values cannot be automatically parsed from the string.

Path 2: Structured description

RDA Glossary definition: “A full or partial description of a resource using the same structure (i.e., the same order of elements) that is used for the resource being described.”

RDA gives an explicit instruction for providing a structured description of a related entity at RDA 24.4.3.

Example: "Reprint of: Venice / by Cecil Roth. — Philadelphia : The Jewish Publication Society of America, 1930. — (Jewish communities series)"

The data are recorded as a literal text string.

The data describe aspects of one or more related entities and relationships.

The nature of the relationship is described in the data.

The values are recorded, aggregated, ordered, and delimited according to a specified schema (ISBD in this example). The string can be automatically assembled from its component values, and, generally, the values can be automatically parsed from the string. This process may be lossy if the schema and string syntax are not completely aligned with RDA.

The RDA element *preferred citation* is a structured description using a non-RDA schema. It can also resemble an access point and include identifier data.

Path 3: Authorized access point

RDA Glossary definition: “The standardized access point representing an entity.”

RDA provides explicit instructions for providing an Authorized Access Point of a related entity at 24.4.2 (Work or Expression) and 29.4.2 (Agent).

A similar instruction for providing an AAP in a subject relationship to a Work, Expression or Agent is missing from 23.4.1.2.2.

Example: "Shakespeare, William, 1564–1616. Taming of the shrew"

The identification data is a text string.

The data contains one or more values describing aspects of the related entity. The relationship is not described in the data.

The values are recorded, aggregated, ordered, and delimited according to a schema partially specified by RDA, but the choice of options effectively renders this as a local schema. The string can be automatically assembled from its component values. It is very useful if, conversely, the values can be automatically parsed from the string.

Path 4: Identifier

RDA Glossary definition: “A character string uniquely associated with … or with a surrogate for … The identifier serves to differentiate that … from other …”

RDA provides an explicit instruction for providing an identifier of a related entity at RDA 24.4.1 (*Work* or *Expression*) and RDA 29.4.1 (*Person, Family, or Corporate Body*).

A similar instruction for providing an Identifier in a subject relationship to a *Work*, *Expression* or *Person, Family or Corporate Body* is missing from RDA 23.4.1.2.2.

Example: "ISBN 978-1-74146-163-3"

The identification data is a text string.

The data contains a single value identifying the related entity.

The value is recorded specifically as an entity identifier, and carries no additional descriptive data. The value may include meta-metadata in the form of qualifiers.

The nature of the relationship is accommodated separately as a relationship designator.

URIs

There are no specific references to recording URIs in the RDA instructions. There is a single example of a URI at RDA 17.4.2.1, repeated at RDA 17.7.1.3.

Example: http://larvatusprodeo.net

The identification data is a URI.

Discussion

Links and relationships

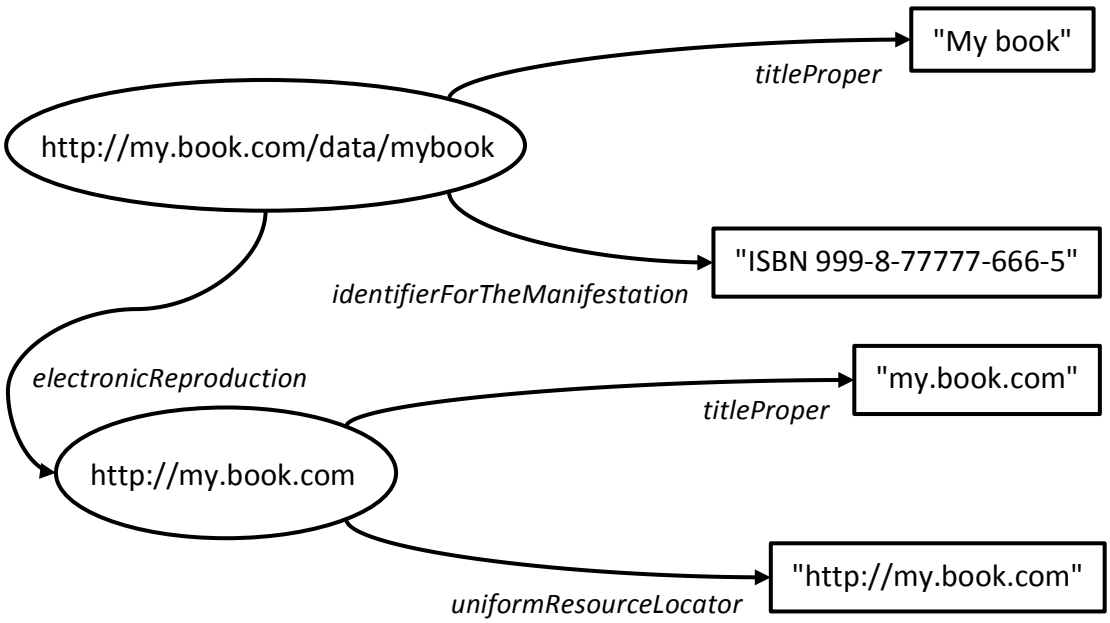
The treatment of data consisting of a direct link to the related resource itself, for example the URL of an online resource, needs to be clarified.

Example: An online document reproducing a printed text. This is a relationship between two manifestations.

Title of the printed text: "My book"
ISBN of the printed text: "999-8-77777-666-5"
Title of the website: "my.book.com"
URL of the website: "http://my.book.com"
URI of the website: <http://my.book.com>
URI of the printed text: <http://my.book.com/data/mybook>

This can be expressed in linked data terms:

<http://my.book.com/data/mybook> rdam:titleProper.en "My book" .
<http://my.book.com/data/mybook> rdam:identifierForTheManifestation.en "ISBN 999-8-77777-666-5" .
<http://my.book.com/data/mybook> rdam:electronicReproduction.en <http://my.book.com> .
<http://my.book.com> rdam:titleProper.en "my.book.com" .
<http://my.book.com> rdam:uniformResourceLocator.en "http://my.book.com" .



The URL is treated as a type of identifier for the manifestation.

Note: This is referenced in RDA 2.15.1.1, but not referenced back from RDA 4.6. The implied sub-property relationship is not stated in the RDA Registry. RDA 4.6.1.3 references RDA 27.1.

Recommendation 1: Clarify and make explicit the relationship between the elements *identifier for the manifestation* and *uniform resource locator* in RDA Toolkit and the RDA Registry.

The URL is recorded as a literal string because it is a type of identifier.

This is the basic distinction between a URL and URI: a URL is a string; a URI is a thing.

Both are types of identifier: the URL is a "local" identifier that can be changed arbitrarily to identify a different document and therefore cannot be used safely as the subject of a global metadata statement; the URI is a global identifier built-in to the syntax of RDF and is intended to be used as the subject of a global metadata statement.

In operational terms, a URI can act like a URL. This happens if the URI is able to be de-referenced using http content negotiation. A normal browser treats it as a URL, and the

server is set-up to send back a human-readable HTML document. A semantic application treats it as a URI, and the server is set-up to send back a machine-readable RDF document.

It is reasonable to assign a URI to an online resource that is different from its URL. In the example, the URI assigned to the website could be changed to <http://my.book.com/data/website/1>:

<http://my.book.com/data/website/1> rdam:uniformResourceLocator.en "http://my.book.com" .

This might avoid some of the ambiguity between strings and things, but there is no further advantage because of de-referencing mechanism maps the URI to the URL.

Surrogates

The definition of identifier in the RDA Glossary allows the identifier to refer to a surrogate for the related entity. This requires review and clarification.

A surrogate such as a metadata "record" is treated as a *Manifestation* of a separate, descriptive *Expression* and *Work* in the FRBR model. The conflation of using the same identifier for an entity and a description the entity is a common cause of confusion in discussions on linked data representations of catalogue records. The identifier of a surrogate can be included in an unstructured description, in the same way as a link to a Wikipedia article, etc., but it should not be accommodated in RDA as the identifier for the entity described by the surrogate. For example, an ISBN identifies a specific *Manifestation* but does not identify a catalogue record for that *Manifestation*. The identifiers for the catalogue record, or RDA WEMI record, etc., are likely to be local, and will not include ISBNs.

The recent development of the subject relationship allows RDA to accommodate surrogates correctly, by using a subject relationship designator:

> Entity – (is) described in (manifestation) – Manifestation [surrogate]

The identifier for the surrogate of a related entity is accommodated in a chain, but not directly attached to the entity in focus:

> Expression1 – (is) abridgement of (expression) – Expression2 - (is) described in (manifestation) – Manifestation1 – (has) identifier – "ISBN"

Data from the surrogate is used to construct an AAP for the related entity.

Recommendation 2: Remove references to surrogates in the definition of identifier, and develop instructions for accommodating current use cases using other approaches.

Entities and unstructured descriptions

Unstructured data is assumed to include transcribed data, albeit with augmentation to make it a sentence, etc.

RDA does not provide an instruction for an unstructured description of a related Person, Family, or Corporate Body. The generalization of the RDA statement of responsibility element in the April 2015 release of RDA Toolkit is assumed to accommodate similar functionality.

Examples:

“Contributor: Starring, in alphabetical order: Josie Bissett, Thomas Calabro, Doug Savant, Grant Show, Andrew Shue, Courtney Thorne Smith, Daphne Zuniga”

“Edited and special effects by You Oughta Be in Pixels; production design by Paula Dal Santo; director of photography, Luis Molina Robinson; music by Mark Oates”

Use of the generalized RDA element *statement of responsibility* stretches the traditional concept of "relationship", but clearly a statement of responsibility usually contains data that relates an *Expression* associated with the *Manifestation* to one or more *Agent* entities.

In a similar way, data relating the *Manifestation* to *Agents* can be found in the RDA elements *manufacture statement*, *production statement*, and *publication statement*.

Entities and structured descriptions

Traditionally, a "structured description" has always been a citation for a *Manifestation*; all of the examples at RDA 24.4.3 are of this type, although two of the labels used in those examples are for either *Works* or *Expressions*.

This raises questions about the citation of a *Work*, *Expression*, or *Item* in a structured description.

The specification of what elements to include in a structured description is discussed in 6JSC/ALA/Discussion/3. A basic question is whether those elements should be confined to the related entity, or include elements from entities related to the related entity.

For example, a structured description for a common occurrence of a related *Work* is "Container of (Work): Emma, everyday matters / William Deresiewicz -- Pride and Predjudice, growing up / William Deresiewicz".

However, “Emma, everyday matters / William Deresiewic” and “Pride and Predjudice, growing up / William Deresiewicz” appear to be structured descriptions for *Manifestations*, because they use the same structure and order for the *Manifestation* elements *title proper* and *statement of responsibility* that is used for the resource being described.

This calls into question the mapping of MARC21 tag 505 as a Related Work relationship.[2]

It is simpler, and therefore likely to more efficient in applications, if the data values specified in the syntax encoding scheme for the structured description are immediately "local" and linked directly to the entity in focus. This avoids addition processing, for example in de-referencing another entity to obtain the values, and semantic incoherency.

In practical terms, the values for the elements can be quickly derived from how the entity in focus describes itself. Values for related entities require additional research, for example in looking-up the recorded data or references about the related entity (the human equivalent of de-referencing).

It is therefore better if a structured description only specifies elements assigned to the related entity being described.

The RDA guidelines assume that a description of the related entity exists (or is being created at the same time as the description of the entity in focus), and that it contains data for elements from which a structured description can be constructed and added to the description of the entity in focus.

The requirement is to identify what data already exists in the description of the related resource. However, there will be cases in which there is no description of the related resource and no intention to create one. The only source of data for the structured description is the entity in focus. It can be supplemented by reference sources, but that is the procedure for creating access points rather than structured descriptions, and may be seen as the distinction between the two approaches: a structured description is derived from the entity in hand, while an access point or citation also uses external sources of information.

The data source for a structured description is therefore the entity in focus: what it says about the related resource.

This confines the data source for a structured description to the *Manifestation* entity.

Recommendation 3: RDA should specify the source of data for a structured description as the manifestation being described, and confine the elements to be used to the related entity. For example, the structured description of a related work should include only Work elements with data values derived from the manifestation in hand.

Convergence of structured descriptions and AAPs

If this recommendation is accepted, then Paths 2 and 3 are seen to converge.

[2] http://access.rdatoolkit.org/jscmap2_jscmap2-6646.html

In both paths, the elements describing the related entity are confined to the entity. The structured description for a *Work* is confined to *Work* elements; the AAP for a *Person* is confined to *Person* elements.

Path 3 assumes the source of data values is a separate description of the related entity, whereas Path 2 is used when the source of data values is the manifestation in focus.

This is useful for accommodating differences in data creation workflows. Path 2 seems to work better with *Manifestation* attributes; path 3 seems to work best with *Work*/*Expression* attributes.

There is no real difference at the data model level. A structured description and an AAP both require a specification of data elements with syntax to allow their values to be assembled into a literal text string. The specified elements will vary between entities according to the element set of the entity. The specification of syntax is a matter for application profiles and not the RDA instructions.

A structured description and an AAP are equivalent in data terms. The “AAP of a manifestation” is another term for “structured description of a manifestation”.

Recommendation 4: RDA should conflate the instructions for constructing structured descriptions and authorized access points.

A new four-fold path

Converging Paths 2 and 3, and making the distinction between identifiers and URIs, suggests that RDA should accommodate four methods of recording relationship data for any entity:

1. An unstructured description, the equivalent of a note.
2. A structured citation or AAP for the related entity.
3. An identifier for the related entity.
4. A link using the URI of the related entity

New Paths 2 and 3, covering the existing structured descriptions, AAPs, and identifiers, are modelled using the Nomen entity. Related issues are discussed in 6JSC/TechnicalWG/4.

- Data from Path 1 are intended for *manual* search for the related entity.
- Date from Paths 2 and 3 are intended for *manual and programmatic* search for the related entity.
- Data from Path 4 are intended for *programmatic* linking to, and therefore search for, the related entity.

Recommendation 5: RDA should provide guidelines and instructions covering each Path explicitly, whatever approach is developed.

Beyond relationships

The four-fold path is related to the different database implementation scenarios supported by RDA (5JSC/Editor/2/Rev). There are also relationships with workflow scenarios, alluded to above, covering data taken from the item in hand, data taken from reference sources, data taken from related metadata, and direct links to related metadata.

This suggests that the approaches to accommodating relationship data in RDA could be generalized to cover other types of data.

Recommendations

Recommendation 1: Clarify and make explicit the relationship between the elements *identifier for the manifestation* and *uniform resource locator* in RDA Toolkit and the RDA Registry.

Recommendation 2: Remove references to surrogates in the definition of identifier, and develop instructions for accommodating current use cases using other approaches.

Recommendation 3: RDA should specify the source of data for a structured description as the manifestation being described, and confine the elements to be used to the related entity. For example, the structured description of a related work should include only Work elements with data values derived from the manifestation in hand.

Recommendation 4: RDA should conflate the instructions for constructing structured descriptions and authorized access points.

Recommendation 5: RDA should provide guidelines and instructions covering each Path explicitly, whatever approach is developed.