String encoding schemes in RDA Toolkit

Gordon Dunsire, RSC Technical Team Liaison Officer

Discussion paper / December 2019

Background

A string encoding scheme (SES) is "A set of string values and an associated set of rules that describe a mapping between that set of strings and a value of an element".

An SES specifies a set of strings and rules for assembling them into a single string that is the value of a structured description of specific kinds of RDA element.

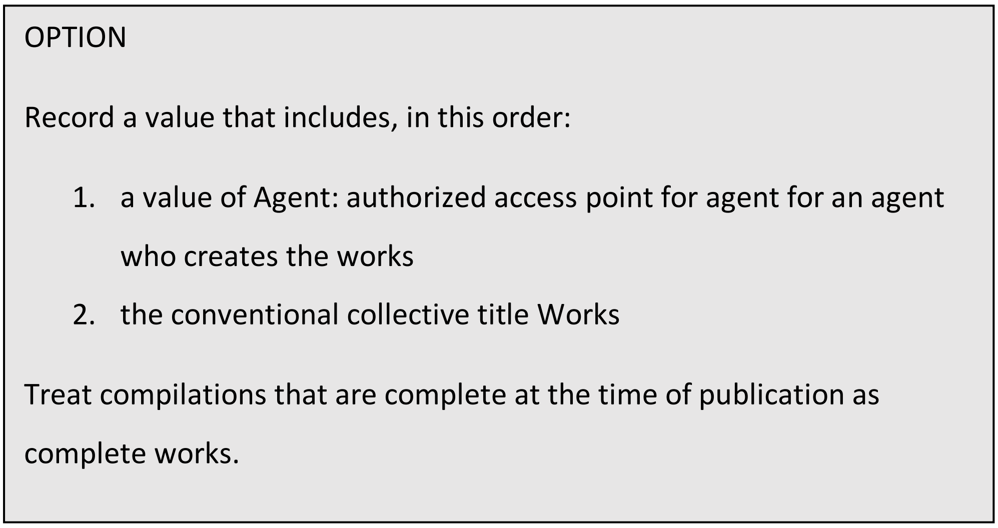
Two distinct kinds of string can be specified:

- The value of another RDA element recorded as an unstructured description, structured description, or identifier.
- A fixed text string or boilerplate value.

The 'rules' for mapping the component strings to the final value string may include:

- The component strings and the order in which they are to be assembled. This may be termed a **value selector**.
- Punctuation or other delimiters of one or more of the component strings. This may termed a **punctuation pattern**.

An example of a value selector from **Work: authorized access point for work** is given in Figure 1.



OPTION

Record a value that includes, in this order:

1. a value of Agent: authorized access point for agent for an agent who creates the works
2. the conventional collective title Works

Treat compilations that are complete at the time of publication as complete works.

Figure 1: Value selector displayed as a standard option in beta Toolkit

The value selector gives the strings and their order of assembly:

1: is a string that is the value of another element.

2: is a fixed string “Works”.

No explicit punctuation pattern is given, but a common pattern is to insert a stop-space between the two component strings. This pattern might be presented as “Value 1. Value2”.

SESs are often recorded and maintained separately from the instructions that use them. For example, ISBD presents a set of “punctuation patterns” that includes the labels of component elements at the beginning of the stipulations for each area of description. The original RDA Toolkit specifies punctuation patterns for specified elements in *Appendix E: Record syntaxes for access point control*, as well as within the content of the instructions.

Beta Toolkit

The original RDA Toolkit contains several value selectors and punctuation patterns.

In the beta Toolkit, value sector specifications have been moved to instructions for authorized and variant access point elements. There are more specifications for Person and Corporate Body than other entities, and none at all for new entities (Agent, Collective Agent, Timespan). RDA Entity is unlikely to require any SESs, and Nomen is out of scope for access point construction.

The original Toolkit has examples of the same value selector used with different punctuation patterns. This results in two separate SESs.

The ultimate number of combinations of value selectors and punctuation patterns is dependent on the communities and applications that use RDA metadata. This presents a challenge to the future development and maintenance of RDA as a tool for constructing access points and other structured descriptions in an international context.

The RDA Steering Committee discussed the treatment and accommodation of SESs during its 2019 meeting in Santiago, Chile. It was agreed to test an approach that moves the content of SESs out of the element pages to an area of the Toolkit where they can be better maintained. The content remains available in an element page using a link or an expandable box. A similar approach is already in used for maintaining and displaying examples in the new Toolkit.

This paper presents the results of the test so far and makes recommendations for the next stages of development.

Scaling issues

There are no global SESs that are used by all libraries and cultural heritage organizations for all resources meeting specified conditions. The specification of an SES is dependent on the local application and is generally considered to be in the scope of a policy statement or application profile.

It is difficult to estimate the number of different value selectors that might be required to meet the needs of the wider international and cultural heritage communities. A difference in the selection of component strings or in their order requires a separate SES. A translation of a fixed component string does not require a separate SES.

Table 1 gives the numbers of selectors extracted for each entity.

Table 1: Number of value selectors used for access points for an entity

Entity	Value selectors
Corporate Body	65
Expression	5
Family	1
Manifestation	3
Person	10
Place	5
Timespan	10
Work	90

The numbers for Corporate Body and Work are inflated because the value selectors specify which instance of an entity is the source of a component string. For example, a value of **Corporate Body: preferred name of corporate body** may be filtered as pertaining to a broader/parent instance, an intermediary instance, or the instance being described.

It is also difficult to ascertain the number of different punctuation patterns that might be required. As well as the original Toolkit Appendix E, specific punctuation patterns may conform to international standards such as ISBD, national and language standards as recorded by IFLA, or specific application profiles and policies.

Table 2 shows the initial results from extracting punctuation patterns from the element instructions. Patterns are given explicitly and implicitly in the instructions.

Table 2: Number of punctuation patterns for each number of selected strings

Number of selectors	Number of patterns
1	2
2	4
2 +	2
3	4
3 +	2
4	1
5	1

The plus (+) indicates that there is an indeterminate repeat of a selector string in the pattern. For example, if a corporate body hierarchy spans more than two levels, there is an unspecified number of intermediary bodies between the corporate body being described and its top-level parent body.

The number of distinct SESs moved from the RDA instructions is well over 100. It is not easy to count them at this stage.

Toolkit element page

SESs mainly apply to elements for authorized and variant access points, in the context of instructions for constructing specific access points.

The use of any SES is optional.

There is no more than one RDA SES in each Toolkit option box.

Examples are tied to a specific (RDA) SES.

Figure 2 gives the generic components and their sequence in a new beta Toolkit option box.

OPTION

SES

Examples

Navigation, etc.

[Citation number]

Figure 2: Components of an Option box

Potential display of SES

An SES can be displayed from the option block:

1. As a preview in the right rail. There should be sufficient depth in the preview pane to display the whole content of most SESs, but not all.
2. As an expandable box. This would be the same as for examples and provides maximum flexibility for display.

1. has the potential for screen clutter if policy statements are open.

2. uses the same approach as for examples, which have similar scaling issues. Only minor development of existing software and scripts is required to implement this option.

Recommendation 1: Configure the display of an RDA SES as an expandable box

Layout of access point elements

Access point, authorized access point, and variant access point elements have a standard layout that separates out instructions for the base of an access point from instructions for qualifying the base.

There are issues with the clarity of instructions that reference the output of applying ‘related’ instructions, for example, a variant access point that depends on the option that was chosen for an authorized access point. Many of these instructions appear in variant access point elements.

Variant access point elements use the name or title of the entity as the base. This blurs any distinction between preferred and variant names and titles.

There are two distinct categories of variant access point:

1. Access points that are variants for the same base as an authorized access point.
2. Access points that use a different base as an authorized access point.

Category 2 is not accommodated in the new Toolkit.

Note that category 2 should use both the instructions for authorized access point, and the instructions in category 1 for variant access point.

To avoid proliferation of value selectors, they should refer to the broader name/title or access point element where possible and leave selection of preferred/authorized to preliminary instructions (for the selection of base and qualifiers).

Element layout sections:

Access point

Select base: name/title element

Format base

General conditions/options

Qualifiers: list elements. Do not list specific designators (fixed text).

Data provenance

Authorized access point

Select base: preferred name/title element

Format base

Cross-reference: access point Format base

Conditions/options for authorized access point format

Qualify base

Conditions/options for authorized access point qualifiers

Data provenance

Variant access point

Select base 1: variant name/title element

Format base 1

Cross-reference: authorized access point Format base

Cross-reference: variant access point Format base 2

Qualify base 1

Cross-reference: authorized access point Qualify base

Select base 2: preferred name/title element

Format base 2

Conditions/options for variant access point format

Qualify base 2

Conditions/options for variant access point qualifiers

Data provenance

This layout has been applied to the authorized access point and variant access point elements in the **Toolkit Development** site.

Recommendation 2: Apply a regular layout reflecting the base+qualifiers workflows for the construction of access points.

Terminology

The LRM discusses access points in the context of the Nomen entity and does not describe the construction of access points in any detail. The proposed layout of options in RDA access point elements follows FRAD: select a base for the access point; format the base if necessary; add qualifiers for various purposes.

The element page menu headings for the standard layout use the phrase 'basis of access point'. This is slightly ambiguous, as it may imply that a qualifier must be added to form the access point. The FRAD terminology of 'base access point' is better at indicating that it may be a full access point without adding qualifiers.

Recommendation 3: Use 'base access point' instead of 'basis of access point' in RDA Toolkit headings and instructions.

Categorization of access point qualifiers

Some options for adding qualifiers to access points for some elements are assigned one of the categories:

- to distinguish the access point from a value of an access point for another entity
- to assist in the identification of the entity
- to conform to a string encoding scheme

The categories are derived from the wording of the current Toolkit; the last is a catch-all and is a tautology because all access points conform to an SES.

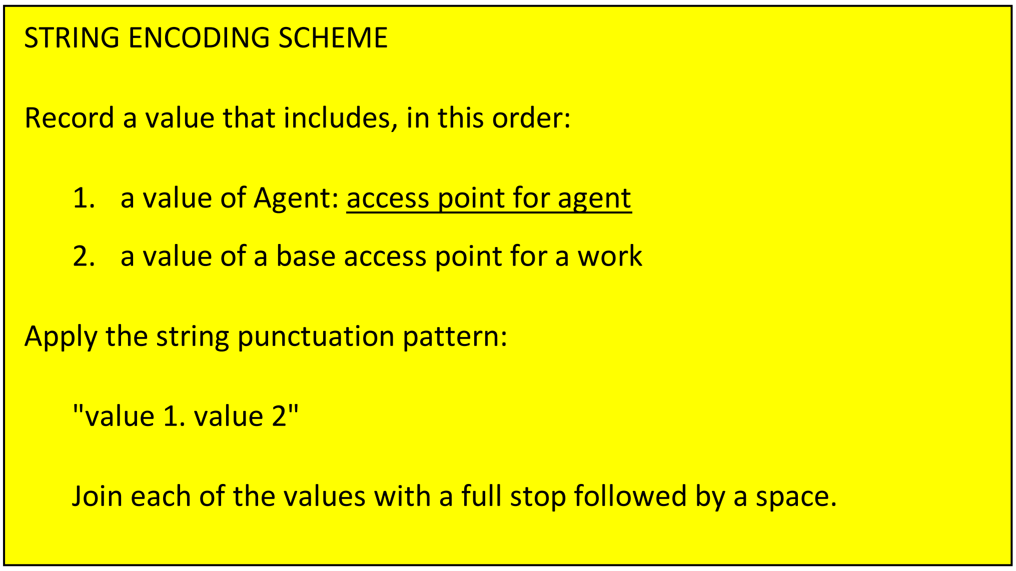
The categories are even less useful in the treatment of SESs proposed in this document.

A qualifier can only be assigned to one category, but local applications may want to use the option under another category. It is better to devolve all categorization of qualifiers to policy statements and application profiles.

Recommendation 4: Remove the explicit categorization of access point qualifiers by removing the headings in access point elements, but retain the list of reasons for qualification to support context and cataloguer's judgement.

Layout of string encoding scheme box

Figure 3 shows a draft layout and boilerplate for an SES for an authorized access point for Work.



STRING ENCODING SCHEME

Record a value that includes, in this order:

1. a value of Agent: access point for agent
2. a value of a base access point for a work

Apply the string punctuation pattern:

"value 1. value 2"

Join each of the values with a full stop followed by a space.

Figure 3: Draft layout and boilerplate for an SES box

This SES is the first one given in the example in the Appendix.

Recommendation 5: Apply a standard layout and boilerplate for the value selector and punctuation pattern components of an SES.

Other issues from the test

The example in the Appendix illustrates several issues arising from this approach.

Both of the Condition boxes shown have only SES options, some with and some without accompanying examples. This occupies a significant amount of screen real estate.

There is no means of distinguishing the Option boxes. The only way of doing this is to expand the SES box. It is not feasible to label either the Option box or the SES box.

The overall length of the display is significantly reduced. The seven lines of Figure 3 are collapsed to a single line with an icon/button.

The retention of an Option box for each SES allows a boilerplate or simple Policy Statement to be applied: Use this option = apply this string encoding scheme.

Policy Statement writers and Translators have been informed of the element pages that are likely to change in layout and have been asked to avoid processing these files for the time being.

RSC needs to decide on the strategy for future development of SESs in the context of Policy Statements and application profiles to allow early release of the affected content to writers and translators.

The test offers three options for further development:

Option 1: Completely remove the SES boxes, together with associated Condition boxes and Examples boxes, from the element instructions.

The SES boxes would retain an individual URL so that each can be referenced from within the content of a Policy Statement or from the SES column of an RDA application profile.

This option removes the major part of the screen display for access point elements, as indicated by Table 1.

This option requires significant development of the CMS data and management infrastructure but provides the best future-proof path.

Option 2: Develop the beta Toolkit with the approach used in the test.

The impact on Toolkit users can be tested while the Toolkit is in beta. It is essential to include a facility to turn on or off the expansion of an SES box for a session, as with Example boxes.

This option requires minor development of the CMS software, and further development will be required if new RDA communities require additional SESs.

This option allows quick addition of Policy Statements for SESs.

Variation: develop a Tooltip display for the SES box.

Option 3: Retain the current approach in the beta Toolkit.

The immediate issue with punctuation pattern displays can be resolved by replacing the SES box button in an Option box with the hard-coded information given in Figure 3.

Other inconsistencies and gaps found in the test can be applied to the instructions in the same way. This will result in a significant increase in the length of some element pages.

This option allows quick addition of Policy Statements for SESs.

Options and recommendations

Option 1: Completely remove the SES boxes, together with associated Condition boxes and Examples boxes, from the element instructions.

Option 2: Develop the beta Toolkit with the approach used in the test.

Option 3: Retain the current approach in the beta Toolkit.

Recommendation 1: Configure the display of an RDA SES as an expandable box

Recommendation 2: Apply a regular layout reflecting the base+qualifiers workflows for the construction of access points.

Recommendation 3: Use 'base access point' instead of 'basis of access point' in RDA Toolkit headings and instructions.

Recommendation 4: Remove the explicit categorization of access point qualifiers by removing the headings in access point elements but retain the list of reasons for qualification to support context and cataloguer's judgement.

Recommendation 5: Apply a standard layout and boilerplate for the value selector and punctuation pattern components of an SES.

Appendix: Examples of proposed Toolkit layout

Element: Work: authorized access point for work

Single works

CONDITION

A work is a *single work*.

A work is created by one agent.

OPTION

String encoding scheme

Example

CONDITION

A work is a *single work*.

A work is created by two or more agents in collaboration.

OPTION

String encoding scheme

Example

OPTION

String encoding scheme

Example

OPTION

String encoding scheme

OPTION

String encoding scheme

Example

OPTION

String encoding scheme