TO: Joint Steering Committee for Development of RDA

FROM: John Attig, ALA Representative

SUBJECT: Machine-Actionable Data Elements in RDA Chapter 3: Discussion Paper

Background

Based in part on posts to RDA-L, ALA's Committee on Cataloging: Description and Access voted to create a Task Force on Machine-Actionable Data Elements in RDA Chapter 3. The Task Force was charged to evaluate the structure of data elements in Chapter 3 that contain quantitative information. The Task Force was chaired by Peter Rolla; members included Diane Hillmann, Randy Roeder, Paul Weiss, and Kathy Winzer, John Attig (ex officio), Gordon Dunsire (consultant); Karen Coyle participated in most of the discussions.

After studying the elements for Extent and Dimensions, the Task Force came up with the Aspect–Unit–Quantity model that is described in this paper. The implementation of this model in the RDA element set, not to mention the text of the RDA instructions, would have a significant impact on RDA, and this impact would not be limited to Chapter 3 or to elements containing quantitative information. Therefore, before presenting a formal revision, ALA would like to submit the attached Discussion Paper, and would like to ask for the advice of the JSC on a number of issues.

Questions for Discussion

1. The Task Force recommends that RDA add a treatment of extent that is machine-actionable, using the Aspect–Unit–Quantity model described here as a basis of proposed revisions to the element set and the text of the RDA instructions. Does the JSC agree?
2. The Task Force notes some ambiguity between content and carrier in RDA instructions for extent. We recommend that the FRBR attribute Extent of Expression be added to RDA for recording those aspects of the extent statement that apply to content. Does the JSC agree?
3. The Task Force considered a number of options for integrating the Aspect–Unit–Quantity model into the RDA element set; these are described on pages 10–11 below. Does the JSC prefer:

   Option 1: A single element for Extent (broadly defined as including Dimensions and perhaps Duration), following the Aspect–Unit–Quantity model.

   Option 2: Retain the current elements for recording eye-readable text strings; add a parallel element for machine-actionable Extent (as defined above).

   Option 2a: Same as option 2, but add multiple parallel elements machine-actionable Extent, Dimensions, Duration.

   Option 3: Retain the current elements, but rewrite the instructions to call for recording aspect, unit, and quantity.

To: ALA/ALCTS/CC:DA Committee on Cataloging: Description and Access

From: Peter J. Rolla, Chair, Task Force on Machine-Actionable Data Elements in RDA Chapter 3

Subject: Machine-Actionable Data Elements in RDA Chapter 3: Discussion Paper (May 2012)

Introduction and Charge

The CC:DA Task Force on Machine-Actionable Data Elements in RDA Chapter 3 has been charged with evaluating the structure of data elements in chapter 3 that contain quantitative information in the form of a quantity and a unit of measure (e.g., Extent, Dimensions) and proposing revisions to make these elements more machine-actionable. RDA currently, and as a legacy of AACR2, treats this information as a textual string, but CC:DA decided that, given the nature of the information described by these elements, as well as the general desire to treat bibliographic information where possible as data that can be used and manipulated by computers, that elements like Extent and Dimension are useful test cases to explore a new type of treatment.

The Task Force decided to present a formal Discussion Paper to the JSC and not a revision proposal yet. We have come up with a model for these data elements that we believe will work, but that represents a radical break both with cataloging tradition as well as with how RDA has been structured and written. Because of that we felt that a broader discussion, first among CC:DA and its constituencies and then among the JSC and its member organizations, was necessary before we could propose an actual change to the instructions and text of RDA.

Rationale

One of the most obvious areas where the general intention to move towards library description as “quantitative data” rather than “text” is incomplete occurs in the recording of extent and dimensions. At present, the RDA instructions and examples look very much like they did in AACR2 and its predecessors, emphasizing textual information for human interpretation rather than data that can be parsed and manipulated by machines. In order to change this situation, CC:DA formed this Task Force, whose members have some experience with machine manipulation, to analyze the issues and recommend possibilities.

The reasons to tackle these issues revolve around the desire for better machine manipulation of the data created. Resolving these issues could provide functionality in the following areas:

- Easier matching for the purposes of determining differing content
- Sorting by size, dimension, or other criteria
- More granular faceting for media materials based on extent
- A better path towards automated determination of extent
- Provision of textual values and labels in a variety of languages
- Ability to compress and itemize more complex extent information for particular users (similar to MARC holdings data)
- Validation of data at the time of input

A reason often given for not providing this level of detail is that at the moment it is not needed for many library functions. While inarguable, this misses the point. If we are to have the data to provide the functionality above, we should start now, rather than later. Another argument against this level of detail is that it would be expensive to create. However, providing the information we have now in a more machine-friendly form should be no more expensive than what we do now in text, particularly with an intelligent user interface in place that supports the use of controlled vocabularies for textual elements.

Although many libraries may have no need for increased functionality around the area of physical description, it is still true that some libraries, particularly those dealing with media and special materials, have many unmet needs with RDA. If those libraries can provide the level of detail and specificity they need using the enhanced extent strategy, extra value can be created within bibliographic descriptions without requiring special efforts for the majority of libraries without those needs. Some creators and users of bibliographic data, such as online bookstores, already make use of detailed information about aspects like size and weight of products to aid in functions like shipping. Because we anticipate sharing data with agencies outside of the immediate library environment, provision for machine-actionable data in the future will make library data more attractive to these users.

If enhanced physical description is implemented, we must consider the display for humans, particularly so that effective summarized or ‘dumbed-down’ information can be provided (preferably via algorithm from more detailed description) for general users who do not need the level of detail that is required for algorithmic processing.

The Task Force’s Work

The Task Force began by going through RDA Chapter 3 and pulling out all the instructions that fit into our charge: i.e., wherever an instruction says to record a numerical quantity and a unit of measurement. The bulk of these examples came from sections 3.4 (Extent) and 3.5 (Dimensions), although we also took into consideration other sections like 3.16 (Sound Characteristics), 3.17 (Projection Characteristics), and 3.19 (Digital File Characteristics). We then analyzed all of these examples to see if they followed a similar pattern and if we could rationalize that pattern to create a new instruction that encompassed all of them. One fact that struck us immediately is that in RDA treats the Extent of different formats of materials as separate elements: i.e., Extent of Cartographic Resource is a different element from Extent of Notated Music which is a different element from Extent of Text, etc. This division into different formats has been carried over into the RDA Element Set, where each of these format-based Extent elements has been registered separately.

After much discussion and analysis of all of these elements and instructions, we did agree that a pattern exists and that it can be turned into a model for recording information like extent and dimensions about resources, whether or not the JSC continues the separation of elements by format as noted. We noticed that all of the various statements about extent, dimensions, etc., have three parts to them: the Aspect being measured, the Unit of measurement, and the numerical Quantity.

We put this model to the test and despite some complications and a few disagreements, decided that it does work, and that using this model to record information about extent, dimensions, etc., would effectively present this information as machine-actionable data. We also envision that the Aspect and Unit pieces of this model would come from open controlled vocabularies, which could facilitate the recording of information by catalogers through the use of drop-down menus or other labor-saving means of data entry.

Issues for Discussion

We want to emphasize that this paper presents a discussion very much still in progress. We have come up with a model which we believe accomplishes our objectives. However, there are aspects of this model and its application that raise important issues that the Task Force is still discussing. Before describing the model and showing examples to illustrate its application, we want to review some of the issues that have arisen and to indicate the current state of our discussion of these issues.

The model described below will not support the encoding of all the detail that is called for in the current RDA instructions. For example, the model does not (at this stage in our thinking) support the distinction in the Extent element between recording units (e.g., 1 score) and recording subunits (e.g., 55 pages); both statements could be recorded, but our model does not as yet indicate the relationship between the two statements. For this reason, it is not clear whether the model should be considered a replacement for the current elements or an alternative structure for recording the same information. Different options for the overall structure of these sections of RDA are explored in the final section of this paper.

Our three-part model works especially well with Dimensions, as the examples that follow will show. However, it becomes a little more complicated when speaking of Extent. Specifically, it is not clear what should be recorded as the Aspect being measured. Our tentative solution for most of the examples below is to treat “extent” or “units/subunits” as the Aspect being measured. The color diagram adds some possible vocabulary for Aspect, but clearly this particular vocabulary-building task requires subject matter expertise, particularly for music and maps.

We have encountered a number of issues relating to the content/carrier distinction. Our charge was specifically to look at chapter 3, which describes carriers, but we think our model also would apply to elements like the Duration of a motion picture. In FRBR and RDA, however, that is an attribute of the expression, not the manifestation, and is therefore officially outside of our charge.

Furthermore, the Task Force realized that in at least some circumstances the Extent element is actually recording the extent of the content (i.e., Extent of Expression), and that a distinction may need to be made between Extent of Manifestation and Extent of Expression. One example of this is the instruction to recording the extent of a cartographic resource both in terms of its content (e.g., “1 map”) and in terms of the carrier (e.g., “on 6 sheets”). We think that perhaps it might be appropriate to introduce the FRBR element Extent of Expression to deal with the “content” component of such an extent statement.

Such discussions have brought up issues relating to the RDA Extent property and its use of terms for both content and carrier. Resolving these issues is not part of this group’s charge, although they do complicate our work.

The Task Force also had some difficulty in coming up with the proper vocabulary to use in our model. We finally agreed to call the three pieces of our model Aspect, Unit, and Quantity. These work well for Dimensions and possibly for Extent (note our caveats above), but we are still not certain about the words to use for the larger discussion. Some members of the Task Force feel that we still need to use the current RDA terms for Extent, Dimensions, Duration, etc., to refer to separate elements, although others have been thinking in terms of a new Extent element that contains sub-elements for these same categories. Neither approach, however, gives us a good term to refer to the larger concept we are discussing: “Things that can be described using a unit of measurement and a numeric quantity.”[1] The questions about vocabulary extend to specific examples of our Aspect–Unit–Quantity model: in some cases we do not have agreement on what to best call the specific Aspect. This is where we will need guidance and feedback from the various constituencies.

The last major complication, hinted at in the discussion of vocabulary above, is how to fit what we are proposing into RDA. At the end of this paper we will present a few different options of how we think it could happen. Here, though, are some of the questions we are not yet able to answer:

- Do we want to propose a radical revision of RDA that sweeps out the bulk of chapter 3, or do we want to fit our model into RDA as it is currently written, even if that will require a pretty big shoehorn?
- Are we proposing a separate and parallel element – i.e., the textual strings that RDA currently calls for will not change and we will just add a new, parallel, machine-actionable Extent element to live alongside the current instructions?
- Or do we want to get rid of current instructions, replace them with this model that allows for computer manipulation, and then let our systems turn the Aspect–Unit-Quantity pieces of data into a human-readable text string?

These are important questions for RDA development as a whole, and will be relevant too as we consider integrating legacy data with what we intend to build prospectively.

The Aspect–Unit–Quantity model

Information about the physical properties of resources can be expressed through descriptions based on three individual pieces: the Aspect being measured, the Unit of measurement, and the numerical Quantity.

The flexibility of this model, and its applicability to a wide variety of resources, will allow a greater simplification of the instructions, and will also present the data in a machine-actionable way. In this model, the cataloger determines what Aspect is being measured, and the Unit of measurement — and both of those elements can be represented by controlled vocabularies — and then records the Quantity.

[1] The term “measurand,” meaning a physical quantity, property, or condition which is measured, has been suggested. It entered our discussion very late in the process, however, and we have not had enough time to determine if we find it a useful and appropriate term or not.

Currently, and using ISBD punctuation, RDA would have us describe a printed volume as:

> 245 pages ; 23 cm

The Aspect–Unit–Quantity model would break up that statement into its separate parts:

Aspect:	extent: units/subunits
Unit:	pages
Quantity:	245

Aspect:	height
Unit:	centimeters
Quantity:	23

With this approach we can have multiple descriptions for more complicated situations or to include more information when desirable. There could be parallel descriptions as, for example a score made up of 38 leaves:

Aspect:	extent: units/subunits
Unit:	score
Quantity:	1

Aspect:	extent: units/subunits
Unit:	leaves
Quantity:	38

Multiple descriptions can also be hierarchical, for example to describe both a map and the sheet of paper it is printed on:

- *Textual statement:* 20 × 30 cm, on sheet 25 × 35 cm

content/carrier: ***map***

Aspect:	width
Unit:	cm
Quantity:	20

Aspect:	height
Unit:	cm
Quantity:	30

carrier: ***sheet***

Aspect:	width
Unit:	cm
Quantity:	25

Aspect:	height
Unit:	cm
Quantity:	35

Currently there are vocabularies for units of extent in RDA (and below we have an argument for merging them into one vocabulary), but RDA does not currently specify a vocabulary for units of measurement for dimensions. We believe that RDA should have a vocabulary for dimensions, and if the Task Force's work moves beyond this discussion paper and into an actual revision proposal we will want to propose at the same time a vocabulary for units of measurement for dimensions. Also, as stated above, the element set needs to be expanded to include Extent of Content.

Other examples of the Aspect–Unit–Quantity model:

- *Textual statement:* 1 map on 4 sheets

Aspect:	extent: units/subunits
Unit:	map
Quantity:	1

Aspect:	extent: units/subunits
Unit:	sheets
Quantity:	4

- *Textual statement:* 1 atlas (xvii, 37 pages, 74 leaves of plates)

Aspect:	extent: units/subunits
Unit:	atlas
Quantity:	1

Aspect:	extent: units/subunits
Unit:	pages
Quantity:	xvii

Aspect:	extent: units/subunits
Unit:	pages
Quantity:	37

Aspect:	plates
Unit:	leaves
Quantity:	74

- *Textual statement:* 1 score (viii, 278 pages) and 24 parts

Aspect:	extent: units/subunits
Unit:	score
Quantity:	1

Aspect:	extent: units/subunits
Unit:	pages
Quantity:	viii

Aspect:	extent: units/subunits
Unit:	pages
Quantity:	278

Aspect:	extent: units/subunits
Unit:	parts
Quantity:	24

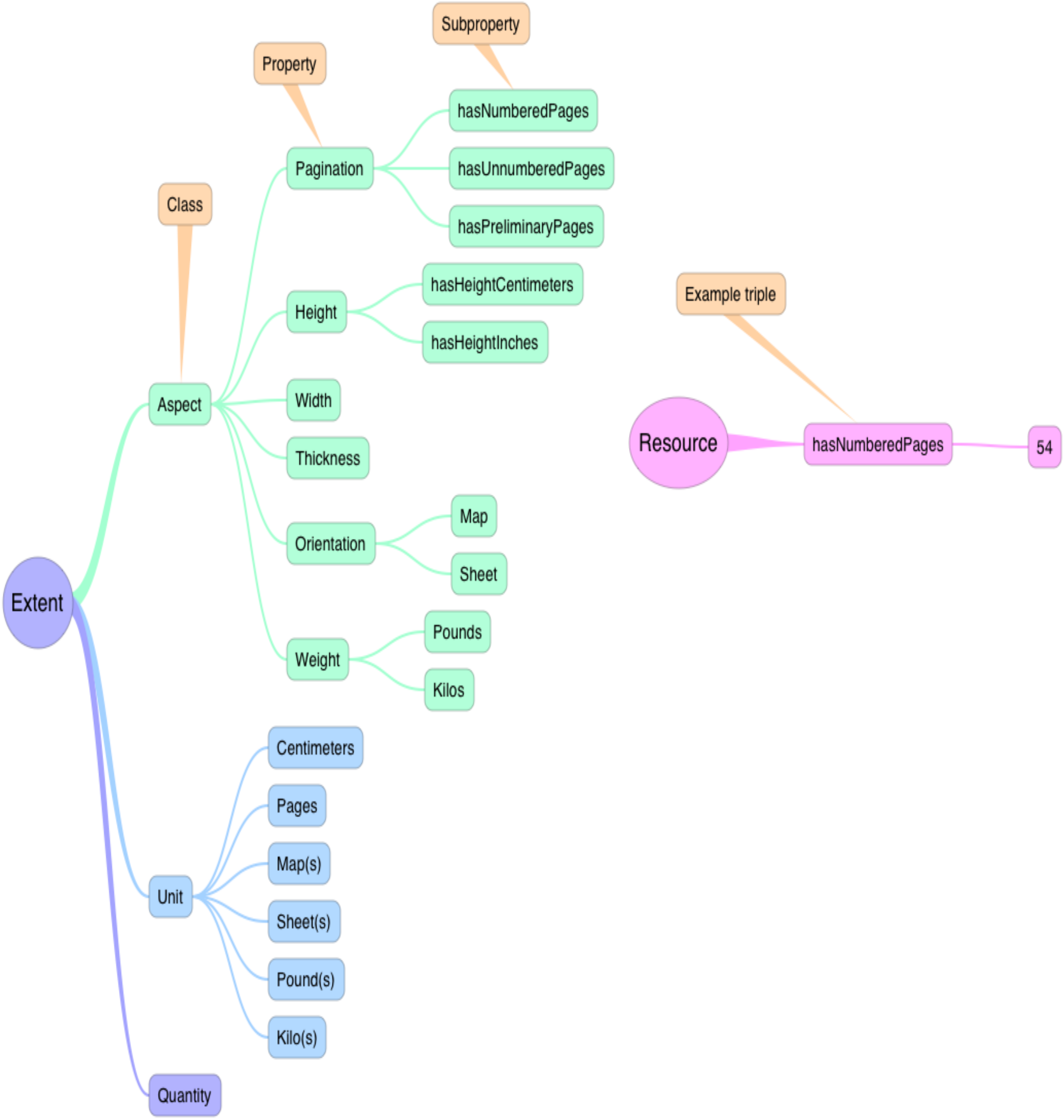
Changes to the RDA Element Set

As mentioned above, under our Issues for Discussion, there are several decisions that will need to be made if the cataloging community wants to move forward with the changes the Task Force is proposing here. Our proposal, it is worth pointing out, would affect the RDA Element Set as well as the instructions. The changes to the instructions would follow the changes that are made to the Element Set. Here, then, are three possible options that could be made to the Element Set in order to incorporate the Aspect–Unit–Quantity model into physical description in RDA.

Option 1: Single "Extent" class

Collapse the various format-based RDA elements into a single element (as a 'class'). Properties within this class could follow the Aspect–Unit–Quantity model, with subproperty vocabularies relevant for Aspect and Unit specified under each (the blue graph, below). If this approach is favored, more study should be undertaken to understand its impact on RDA in RDF, as the best method of doing this is not obvious. However, since RDA is avowedly format neutral, that question should be secondary to the questions of desired functionality for users.

Also important to note that in the example below, an attempt was made to suggest a more complete Aspect vocabulary, which is ripe for comment by specialized cataloging communities, particularly music and maps, where differences in point of view come most immediately into focus.



Either approach could be encoded in XML, with the ‘nesting’ aspect expressed in the instance data and/or the schema instead of ‘behind’ the data as in RDF.

Option 2: Retain current elements; add single new "Extent" element

Leave the current RDA elements defined as they are at present, for the purpose of recording textual statements. Add a single additional element for Machine-actionable Extent, as in Option 1, which would exist alongside the textual statements (analogous to the 006/007/008 fields in the MARC format, which replicate, in coded form, information that exists in textual strings in the variable fields of the MARC record).

Option 2a: Retain current elements; add multiple new "Extent" elements

As in option 2, leave the current RDA elements as they are at present for recording textual statements. Add multiple elements for machine-actionable counterparts of the current elements (i.e., Machine-actionable Extent, and Machine-actionable Dimensions).

Option 3: Retain Extent, modify description

In this option, no additional elements are added to RDA. The addition of elements would be an application-level decision. The RDA text needs to say that there is an aspect, a unit and a quantity, and this is true for textual statements as well as for structured data. How the data is recorded in a machine-readable record is left to an application. There could be applications that ask for separate Aspects, Units, and Quantities. There could be situations where the application already "knows," for example, that for a book there is only height and it is in centimeters, and therefore simply asks the user for a number. This solution separates the instructions for decision-making from the data format, which follows the RDA stated goal of being format neutral. There would, however, need to be changes in the examples to illustrate the three parts of the quantitative statement of extent. These may need to show both the Aspect, Unit, and Quantity as well as possible user displays.

Here is an example of how the text of RDA might change under this scenario:

Currently 3.4.1.3 says, "Record the extent of the resource by giving the number of units and an appropriate term for the type of carrier as listed under 3.3.1.2. Record the term in the singular or plural, as applicable. (For instructions on using other terms to designate the type of unit see 3.4.1.5.) If the resource consists of more than one type of carrier, record the number of each applicable type. Specify the number of subunits, if applicable, as instructed under 3.4.1.7–3.4.1.9."

This instruction could be re-worded to look something like: "Record the extent of the resource by giving the aspect, the unit and the quantity that are appropriate to the carrier being described. If the resource consists of more than one type of carrier, record the aspect, unit and quantity of each carrier. Specify the number of subunits, if applicable, as instructed under 3.4.1.7–3.4.1.9."

Appendix I: Simplifying the current RDA element set

In addition to the RDA elements for Extent and Dimensions, RDA currently defines a number of element sub-types for:

- Extent of cartographic resource
- Extent of notated music
- Extent of still image
- Extent of text
- Extent of three-dimensional form
- Dimensions of map, etc.
- Dimensions of still image

The Editor made a one-to-one correspondence between the outline of the text of RDA and the structure of the element set; all sections identified by two-part numbers (e.g., 3.4) were elements, and all sections identified by three-part numbers (e.g., 3.4.2) were either sub-elements or element sub-types. However, it could be argued that these should not be interpreted in this way, that some are simply identifying groups of instructions that deal with the application of the element in particular situations.

This is most easily seen with regard to Dimensions. There is nothing distinctive about the data that is recorded in the case of maps or still images that make it structurally different from the data that is recorded for any other type of resource. In all cases, the data consists of a quantity and a unit of measurement, along with an identification of what is being measured. The only thing that is distinctive in the cases of maps and still images is that the instructions regarding *what to measure* are different and are more complex because of the inclusion of both content and carrier in a single statement.

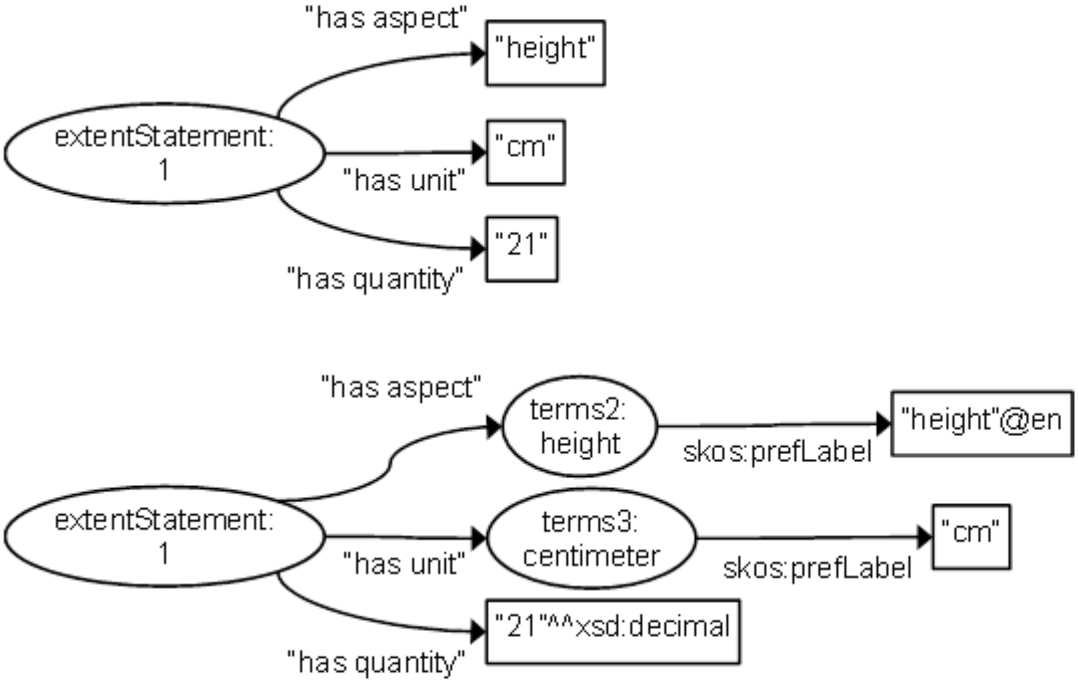
The case with Extent is a bit different, because in this case the vocabularies for recording the units being counted are different. On the other hand, these vocabularies are not completely disjoint. One example is the use of “volume” both in applying the general instructions and in applying the instructions for Extent of Text. It makes sense to consider that there is a single (open-ended) vocabulary for recording units of extent, but that there are sections devoted to giving detailed instructions for recording Extent data for some specific types of resources.

The major advantage of looking at things in this way is that it simplifies the structure of these elements. This will be important if we adopt option 1 above, which would add significant layers of complexity to the RDA element set, albeit complexity primarily for machines, but likely hidden from human users. This approach would also allow us to define a single vocabulary for units of extent, and a single vocabulary for units of measurement for dimensions (and perhaps also units of time)

Appendix II: Machine-actionable elements: RDF graphs

The following RDF graphs show how the proposed elements might fit structurally.

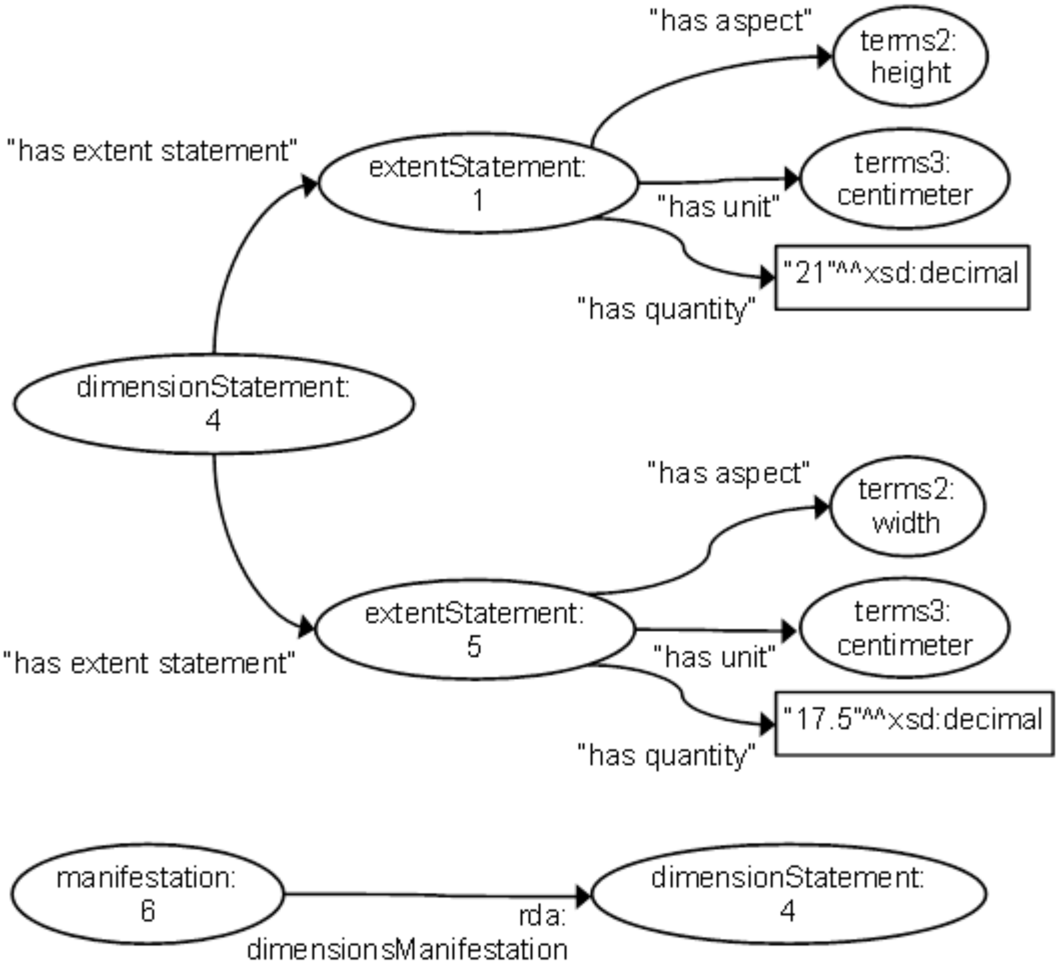
Fig.1: RDF graphs of the basic extent sub-elements



The first graph is the simplest, with untyped literals as the object values of the sub-elements. The second graph is a linked-data version, where the object values are URIs from controlled vocabularies or typed literals.

The sub-elements are linked to a standard extent statement as an instance URI. The statement would usually be associated with a syntax encoding scheme in an application profile. Different profiles can have different encoding schemes, so the sub-element data can be displayed flexibly.

Fig. 2: RDF graph of dimensions statement

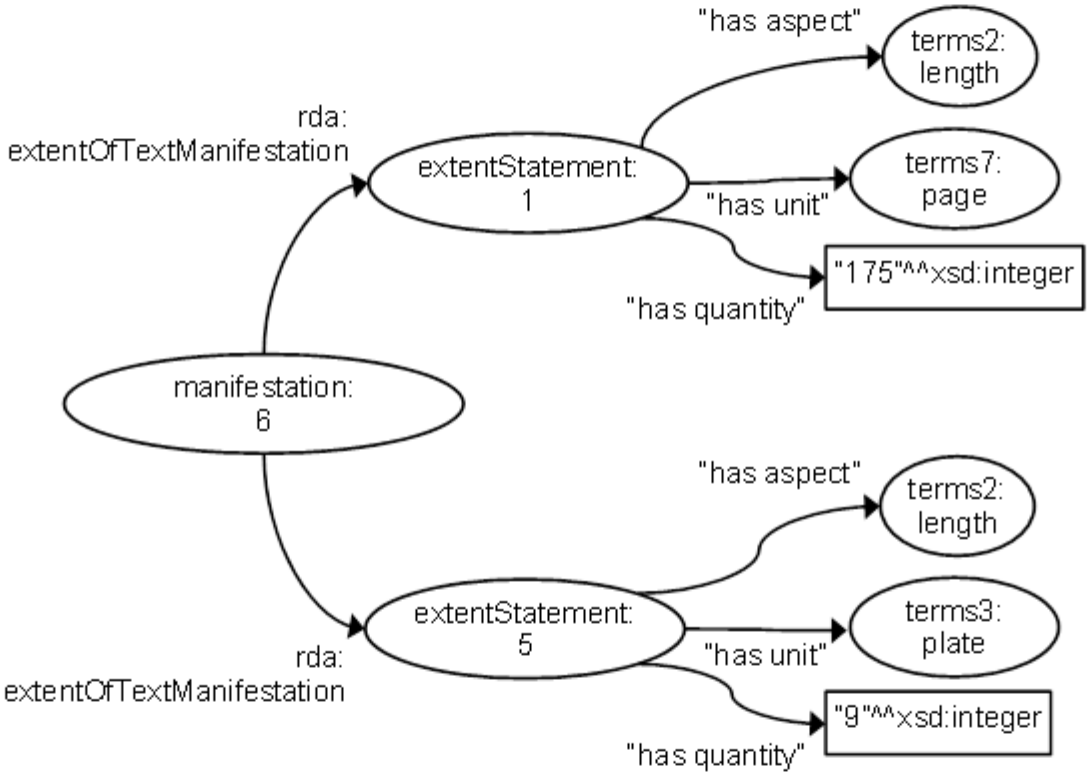


This graph merges two basic extent sub-graphs to form a dimension statement for a two-dimensional carrier.

The dimension statement is, again, an instance URI. It would also usually be associated with its own syntax encoding scheme in an application profile, this time governing how the basic extent statements are to be displayed (e.g. width x height).

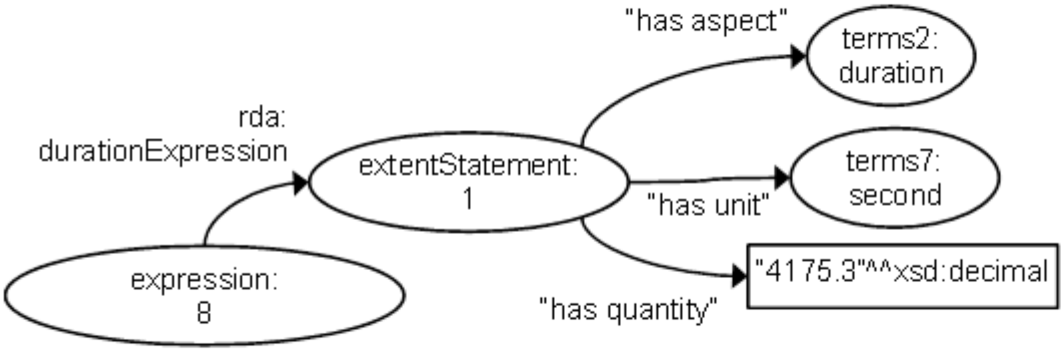
The sub-graph at the bottom shows the relationship between a dimension statement and the existing RDA property.

Fig. 3: RDF graph of extent statements directly linked to a manifestation



This graph shows a high-level "extent" structure composed of two extent statements for a pagination and plate count. There is no compound extent statement, which will probably be required for display purposes via an application profile as outlined above.

Fig. 4: RDF graph of a duration



This graph shows a possible way of recording a duration using the basic sub-element structure. The extent statement is directly linked to an expression using the existing RDA property.

General remarks: The current approach hinges on the “has aspect” property and a set of associated controlled terminologies for the specific aspect value. This needs more thought, as there may be scaling issues, vocabulary overlaps, etc. We are effectively using these vocabularies in place of specific sub-elements/properties such as “has duration” - this is not the same property as the existing rda:durationExpression, which needs a statement compounding the quantity and unit if a different approach is used. Although the current approach is very neat and simple, it is possible that vocabulary-determined pseudo-property graphs may not be recommended by RDF and ontology specialists, except for quite specific circumstances.