NLM Tag Set Working Group Minutes
October 2007

Table of Contents

1 Tag Set Working Group Meeting

The Working Group meeting took place by conference call on October 9, 2007 to recommend changes for Version 3.0 of the NLM Tag Sets.

The next meeting will be after there is a version of the tag sets for Working Group test and review.

1.1 Attendees

1.1.1 NLM Staff

- Jeff Beck (Moderator)
- Abraham Becker
- Steve DeRose
- Marilu Hoeppner
- Laura Kelly
- Adeline Manohar
- Kim Tryka

1.1.2 Secretariat (Mulberry Technologies, Inc.)

- Deborah A. Lapeyre

1.1.3 Telephone Participants

- Mark Doyle (American Physical Society)
- Beth Friedman (DCL)
- Linda Good (Cadmus)
- Kathryn Henniss (Highwire)
- Evan Owens (Portico)
- John Meyer (Portico)
- Bruce Rosenblum (Inera)
- B. Tommie Usdin (Mulberry)

1.2 This Document

The following Tag Set icons appears with each suggested change to indicate to which of the Tag Sets a change might apply

 Archiving

 Publishing

 Article Authoring

 Book

A request may be described using the following:

- Request (a statement or restatement of what a user requested)
- Discussion (which may include rationale or use cases for the requested

modification, description or repercussions of the change, expansion of the specific user request into similar elements, etc.)

- Recommendation (suggested course of action)
- Documentation Recommendation (suggested documentation to accompany the request or solve the problem)

To alert the Working Group, the following change-type icons have been used:

 Backwards-incompatible Change

 Make one tag set more like another

2 Model Modification Requests

2.1 Print Versus Online (Item 6.2 from September list)

Scope: 

Request: Add a wrapper element for inline material and a set of attributes for block-level objects such as paragraphs, section, lists, boxed-text, etc to indicate whether the elements are intended for online-only display or print-only display. Any element not so labeled would be assumed to be for both print and online.

Recommendation: Add "specific use" attribute (not just print versus online, but teacher edition versus student version, doctor versus nurse versus patient, etc.) to all block-level elements (paragraphs, sections, and all block display structures). This attribute is already available on the new <styled-content> element.

Minutes: The group liked the idea of the attribute and made specific suggestions as to where it should be allowed. They suggested:

- <ref-list>
- <ref>
- <caption>

(The <ref-list> example given by Bruce Rosenblum was from the journal *Nature*. For some articles, the printed body of the article contains an abbreviated "*Methods*" section. The printed Reference List cites only those references cited in the print. In the online article, there is an extended "*Methods*" section with an extended Reference List to match. The online Reference List does not contain supplemental or additional reading, it contains cited references, foe example the printed list may have references 1 to 28 and the online list adds 29-30.)

Tommie Usdin pointed out that once this attribute exists, it is very easy and backward-compatible to add it to additional elements as use cases are discovered.

2.2 Multiple Versions of a Single Object

Scope: 

Request: There is a need for a structure to hold multiple forms or versions of a single object. The type case is a graphic for which multiple resolutions or multiple formats exist (such as a tif, a gif, a jpeg, and a png). Typically, the user is not aware of these groupings, they are for purposes of processing and for use by production personnel or by software.

Other *potential* use cases include:

- A table multiple ways: as an XML-tagged table, a CSV file for database loading, and an Excel spreadsheet for user experimentation
- A video and a thumbnail representing one image from the video to appear in print
- An equation present as MathML, TeX, and a jpeg image
- Two captions for a figure, one for color and one for black and white
- Two versions of a section, one for print and a significantly longer one for online
- Alternative versions of a <boxed-text>

Discussion: It seems to us that the multiple-forms issue is distinct from the multiple-versions issue, and that the use cases break down into two categories along those lines

- Those that present merely processing alternatives (3 different resolutions of a graphic), and
- Those that present an alternative structure or different textual versions that are used in alternative use situations.

The case of tables straddles these two, seemingly clear-cut situations.

Minutes: The Working Group agreed that there were two subparts to this question: objects where there are processing alternatives (e.g., a single formula as both MathML and a graphic) and a separate case for user-level alternative text versions.

2.2.1 Multiple Versions: Alternative Text

Request: In our opinion, use cases such as the following:

- Two versions of a <sec>,
- Two captions for a <fig>, and
- Two versions of a <boxed-text>

represent a different problem from the multiple processing alternatives. This is where-and-how-is-this-object-used information. To represent this material, we recommend using the new "specific-use" attribute that was created for the <styled-content> element and discussed in 2.1.

```
specific-use    Use distinction that is the reason for the
                differentiation, for example, "web only"
                "print only", "voice only", etc.

                specific-use  CDATA  #IMPLIED
```

To record alternative usage, this new attribute could be placed on selected structures.

Minutes: The Working Group felt that the "specific-use" attribute met most of the need, but should be used and documented carefully. It was felt that it should be used on as low a structural level as possible. For example, if Figure 3 has one set of captions in English and one set in French, this is not two Figures; this is one Figure with a single ID and which contains two captions, with different specific-use and language attributes. The situation we are trying to avoid is having two Figure 3s (which would have the same unique identifier for reference.)

The group discussed the fact that such a figure might have two titles, two captions, or alternative images (color versus black and white). Bruce Rosenblum described an existing case of the same article in English and Japanese, that was coded by the vendor as parallel sections, first in one language and then the other. Several suggestions were preferred:

- Describe two subarticles and enter the text in each. (In the examples given, 1) one subarticle would be in English and one in Japanese and 2) one subarticle would contain version A of Figure 3 and the other would contain version B of Figure 3.
- For the two title or two captions case, the Archiving tag set has a repeatable <caption> element. We recommend that Publishing does not need one.
- Multiple figures could be enclosed in a Figure Group (which can be used anywhere Figure can be used) but a vocal minority felt that this was semantic overload at best and tag abuse at worst, since there it would interfere with the interpretation of real Figure Groups.
- Multiple version of table and math can be handled as described in 2.2.2.1 below.

While this attribute does not take care of all use cases, it is adequate to the moment. Specifically, multiple sibling structures such as Figures and Sections (with their need for the same ID attribute) are not handled. Paragraphs could be handled as <named-content> with "specific-use" attributes for text and alternatives inside a structure for structures. We can revisit alternatives for paragraph. boxed-text, and section when new use cases arrive.

2.2.2 Multiple forms: Alternative Processing

2.2.2.1 <alternatives>

Request: There is a need to record multiple processing alternatives, such as a graphic in three formats or in several different resolutions.

Recommendation: We define a new noun for multiple versions of one object, named <alternatives>

- <alternatives> is defined as containing multiple logically equivalent (substitutable) versions of the same information object.
- The type case for <alternatives> is a graphic which is shipped with the article in multiple versions (such as a tif, jpeg, and SVG image) or an inline equation that is available as a tif, a MathML, and a TeX.
- <alternatives> is neither a block object nor an inline object, those are the properties of the element that contains it.

- <alternatives> can be used everywhere that graphic can be used now, for example:
  - Inside <fig> as part of the large OR group in the middle of the <fig> model that defines the figure content, that is, after the <caption> and before the <permissions>:

```
(id*, label*, caption*, ...,
   (disp-formula | chem-struct-wrap | disp-quote |
    speech | statement | verse-group | table | para |
    list | def-list | graphic | media | preformat)*,
(attrib | permissions)* )
```

  - Similarly inside the table-level content of <table-wrap>, as an alternative to

```
(disp-quote | speech | statement | verse-group| list |
 def-list | chem-struct | graphic | media | preformat  |
 table)*
```

  - Similarly inside the <chem-struct> content of <chem-struct-wrap>, as an alternative to:

```
chem-struct | graphic | media | preformat
```

  - Loose inside paragraph to handle graphics that are loose inside a paragraph rather than wrapped within a figure or similar container.
- The model for <alternatives> is an OR group containing at least:

```
graphic | media | mml:math | preformat | tex:math |
table | supplementary-material
```

- Here is a tagged example, using figure:

```
<fig>
   <caption><title>Big Dogs</title>
   are really very cool. </caption>
   <alternatives>
     <graphic...a tif .../>
     <graphic ... jpg .../>
     <media ... a CSV file .../>
     <preformat...>...</preformat>
   </alternatives>
   <permissions>...</permission>
</fig>
```

The Group felt that this solution worked well for alternatives with element content.

2.2.2.2 Mixed-content Models

The solution for <alternatives> just outlined contains one real problem, which occurs within the elements that need to contain <alternatives> but that have mixed content rather than element content models:

<disp-formula>

<chem-struct>

<inline-formula> [Added by the Group during discussion.]

Within such models, there may be an alternative that is just #PCDATA characters, for example, there may be an equation in MathML, as well as a jpeg image and *a plain text version* for search. The plain text version should logically be inside <alternatives>. This could be addressed by adding a <text> element, so that <alternatives> would contain:

```
graphic | media | mml:math | preformat | tex:math |
```

```
table | supplementary-material | plain-text-version
```

Original recommendation: We recommended that a new element <text> would be used *only* inside <alternatives> to allow

```
<text>a + b = c</text>
```

to be an alternative to <graphic> inside, for example, a <disp-formula>.

Minutes Recommendation: The Working Group liked the idea of an element to contain the different taggings of the same object. In addition, the Working Group recommended:

- Delete the current "alternate-form-of" attribute, since the <alternatives> element solves the same problem more neatly.
- Extend the <text> concept even more by getting rid of the mixed content model (also never an unqualified success) for <disp-formula>. The element <text> will be a mixed content model including elements such as MathML, not limited to ASCII text, so it will need a new name (<mixed-content>?). This element will handle both the use case shown above (<text>a + b = c</text>) and also the use case of equation in two alternative forms: 1) a graphic image, and 2) interspersed data characters and MathML fragments.

The Group could envision a display formula with alternatives (all versions of the same object) within it that includes 1) MathML, 2) other computer algebra, 3) interspersed math and text, 4) other media, and 5) supplementary material.

2.2.2.3 <table-wrap> as a Special Case

Recommendation: We believe that the <alternatives> solution also works for the straddle case of <table-wrap>. The parameter entity inside-table-wrapper, which contains the guts of the table, currently includes the following elements:

```
chem-struct |  def-list | disp-quote |  graphic |
list |  media | preformat  | speech | statement |
table | verse-group
```

<alternatives> could be added to this list, allowing a table wrapper to contain:

- a table (XHTML model rows and columns)
- a graphic (jpeg of the formatted table)
- a spreadsheet (as a supplementary-material)
- a CSV file (as a supplementary material)

for example:

```
<table-wrap>
  <id> ... a DOI</id>
  <label>Table 6.</label>
  <alternatives>
     <table...>...</table>
     <preformat>...</preformat>
     <media>...</media> )
  </alternatives>
  <table-wrap-foot>...</table-wrap-foot>
  <attrib>...</attrib>
  <permissions>
     <copyright-statement>It's ours</copyright-statement>
```

```
  </permissions>
</table-wrap>
```

Minutes: The Group agreed that this solution handled <table-wrap> nicely.

The tagged example (which used <supplementary material> where <media> is now, led to a follow-up discussion concerning the nature of Supplementary Material. Evan Owens posited the following situation: There is a table (tagged in XHTML XML) that summarizes the results of a study. There is also a spreadsheet and a tag-delimited-text file of the material behind the calculations. Should these be considered multiple manifestations of the same work (all one table with alternatives) or are they different in significant ways. This is a philosophical issue, but even if they are different, can't we use <alternatives> (fudging slightly) to cover the case? Another fudge-able example: a short version of a table and a longer version (more rows or columns) of the same table could be considered alternatives. The "specific-use" attribute could report which was which.
In other words, our <alternatives> need not be (by design) 100% equivalent. A common example would be alternatives of a video and one frame of it as the print and still web alternative. These are not completely equivalent: they are *processing* alternatives, not *semantic* alternatives.

Minutes: Documentation Recommendation: Describe the processing/semantic distinction and then show examples of how <alternates> might be used.

2.3 Equation Group

Scope: 

Request: There are groups of equations that are processed together, in the same way that groups of tables or groups of figures are processed. The current tag sets provide no structure for this.

Minutes Recommendation: We recommend adding a <disp-formula-group> element modeled on figure groups that can be used anywhere <disp-formula> can be used.

The model in default, Archiving, Publishing, and Book:

```
(label?, caption?, (%access.class; | %address-link.class;)*,
 (disp-formula )*   )
```

The content model in Authoring might be:

```
(caption?, (%access.class; | %address-link.class;)*,
 (disp-formula )*   )
```

Minutes Discussion: The situation with numbered equations is very complex. Published examples we have seen include names or numbers for an entire equation group and then individual numbers for each display equation in the group.

There are several problematic cases:

- The major problem may be alignment, for equation numbers, where lists, etc. MathML alignment only works within one set of math tags. If the multiple math tags are all in one formula, you can get alignment. If there are several formulas in a formula group, no. But identifiers need to be placed on each individual formula as well as on the group.

- We have seen cases where images were used for the equations, but the label/number is text and other cases where the label/number is built into the image.
- Evan Owens pointed out that the MathML spec recommends that labels/numbers be built using a labeled table row, but that current MathML systems do not render this properly.
- Bruce Rosenblum expressed the concern that "You may not be able to auto-number equations, even for Authoring".

Minutes Documentation Recommendation: Documentation must talk about alignment, labeling, and identifiers and provide examples.

Action Item: Mark Doyle described how his group handles the alignment problems and promised us email describing this in detail.

2.4 Supplementary Material

Scope: 

Request: There is not a specific request, more a general query: Are the current tag sets adequate to handle the cases of supplementary material that we as a group are seeing in the wild?

Discussion: Supplementary Material is the spreadsheets, datasets, audio files, and other material that is more and more frequently associated with the "text" of an article. Publisher's practice, even the definition of supplementary material, varies widely. Our working definition is "the stuff that publishers want to accompany an article that does not print or display as part of the ordinary narrative text". Thus figures, tables, and sidebars are all part of the article, but the database of genomic material is not.

The Working Group separated supplementary material from the problem of multiple forms of material (the jpeg, the gif, and the png for a graphic). That may still leave several use cases for such material:

- Material that was used to support the conclusions of the narrative, e.g., a dataset and a survey for a paper that presents the highlights learned by the survey and recorded in the dataset.
- "Extra" tables that do not print or show directly in the HTML but that record measurement on which the article is based.
- Material added for enhancement purposes, such as a video of a reaction that is also described with text and shown in still images.
- Material that is logically part of the published narrative material but that is not printed or displayed, e.g., 10 columns of a 45-column table are displayed. (Is this really just an alternate form, as a video might be an alternate form of a still image?)

Minutes: The Group stated that this request was closely related to alternatives proposal, and this was examined in the light of that discussion.

The real problem is that, while we provisionally defined supplementary material as extra or enhancements, "material that may be essential to the scientific research behind and

article but not essential to the article itself", there are publishers who define it differently. Some publishers define it as "Figures 6 through 10". Is an object supplementary because it does not do into the print copy? How does supplementary compare with the Physics Society's "essential non-text elements".

To the Group, a spreadsheet is not necessarily supplementary material just because of its form, "if it has 12 extra pages of data, then it would really be supplementary".

Minutes Recommendation: We recommend no further action at this time. The <alternatives> solves some of this problem.

2.5 <person-group>

Scope: 

Request: Even in the loose <citation> model, it is not possible to reorder the person's name or to punctuate it, except inside <string-name>. Some publishers would like to reorder the name to match their presentation order.

Discussion: This is exactly what <string-name> is intended to be used for, to render the punctuation or change the order of the name components. There seems to be an odd reluctance to use <string-name> instead of <name>, as though one were better than the other.

Documentation Recommendation: Document more uses of <string-name> or redefining it to explain circumstances like this use case will help.

Minutes Recommendation:  Currently, <string-name> can only be used inside <person-group> in Archiving and Book. We recommend adding inside <person-group> and inside <citation> in Publishing for purposes of punctuation consistency in the loose citation model. We realize that this will allow <string-name> inside <product>, <related-article>, and other places where it is not a problem. Jeff Beck will check out <nlm-citation>

2.6 <source> in <citation> Attributes

Scope: 

Request: Add an attribute to <source> to make it possible to distinguish whether a source in multiple formats was web-accessed or the print version was used.

Minutes: When this material is available it is in the text of the citation; only a human reader could extract it for an attribute. We reject this suggestion, but would be willing to re-open the discussion if presented with a few strong use cases.

3 Minutes: The Next Steps

- The very next step is to produce and distribute minutes of this meeting.
- Outstanding action items will be emailed to potential actors as a reminder.
- NLM will then go though all the changes lists with the Secretariat (Mulberry) so that Mulberry can finish the Version 3.0 tag set changes.

- It is the intent to float a version of the tag sets to the Working Group for testing, evaluation, and experimentation before making the full version public. Ideally, a Public version would be available after January of next year.

Evan Owens suggested that Working Group members should get specific homework assignments. Jeff Beck remarked that the outcomes from such assignments would be really useful to beef up the documentation. Bruce Rosenblum volunteered Inera to see how much work it is to shift someone from a typical 2.3 installation to Version 3.0. In general, the Group feels that Version 3.0 will be a great improvement, but will require some work to convert.

The documentation changes such as live URLs in the DTD versions and a new comprehensive index will take even longer than the tag sets to develop. The DTD and schema versions of the tag sets are likely to be released to the committee for experimentation before the full documentation is ready.

We will attempt to stay in touch during this development with postings to the list and "virtual meeting" (voted recommendations) rather than by phone

---------- document end ----------