

February 2007
NLM Tag Set Working Group
Discussion and Recommendations

Table of Contents

| | | |
|-------|--|---|
| 1. | Tag Set Working Group Meeting | 1 |
| 1.1 | Attendees..... | 1 |
| 1.1.1 | NLM Staff..... | 1 |
| 1.1.2 | Secretariat (Mulberry Technologies, Inc.)..... | 1 |
| 1.1.3 | Telephone Participants..... | 1 |
| 1.2 | Change Applicability | 1 |
| 2. | Final Recommendations for the V2.3 Release..... | 2 |
| 2.2.1 | <i>Format/Display Oct/Nov Recommendations for Reconsideration</i> | 2 |
| 2.2.5 | <i>Use of <x> inside <ref></i> | 2 |
| 3. | New Dawn Discussion..... | 2 |
| 4. | Meeting Action Items | 4 |
| 5. | Potential User Questions..... | 4 |
| 6. | The Layout Issue Discussion | 5 |
| 6.1 | <i> element (previously #2.2.1)</i> | 6 |
| 6.2 | <i>Layout-hint attribute <table-wrap> et al. (previously #2.2.2)</i> | 6 |
| 6.3 |  <i>Prohibit Face Markup in <label> (previously #3.6)</i> | 7 |
| 7. |  <i>MathML Namespaces (previously #3.9)</i> | 7 |
| 8. | New Book DTD | 8 |
| 9. | Pending Discussions for Version 3.0..... | 8 |

1. Tag Set Working Group Meeting

The Working Group meeting took place by conference call on February 6, 2007. This document contains a record of the discussion of the Working Group. The agenda items from the February meeting that were not addressed because of time limitations are not included here, but will be included in the list of discussion topics for following Working Group meetings.

1.1 Attendees

1.1.1 NLM Staff

- Jeff Beck (Moderator)
- Steve DeRose
- Marilu Hoepfner
- Laura Kelly
- Adeline Manohar
- Kim Tryka

1.1.2 Secretariat (Mulberry Technologies, Inc.)

- Deborah A. Lapeyre

1.1.3 Telephone Participants

- Alex Brown
- Mark Doyle (American Physical Society)
- Beth Friedman (DCL)
- Sharon McCamant (Cadmus)
- Evan Owens (Portico)
- Bruce Rosenblum (Inera)
- B. Tommie Usdin (Mulberry)

1.2 Change Applicability

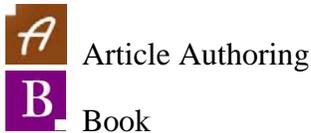
The following Tag Set icons appear with each suggested change to indicate to which of the Tag Sets a change applies:



Archiving



Publishing



2. Final Recommendations for the V2.3 Release

2.2.1 Format/Display Oct/Nov Recommendations for Reconsideration

Scope: A graphic showing four colored squares: green with 'A', blue with 'P', brown with 'A', and purple with 'B'.

Inline-graphic attribute — Request that a CDATA #IMPLIED attribute be added to <inline-graphic> to hold the relationship between the inline element and the baseline. This attribute may contain numeric offsets from the baseline (as Elsevier does) or keyword values like “top | bottom | middle” (as other DTDs have used).

Advice:

Add the attribute to preserve the semantic distinctions sometimes indicated by a baseline shift. Look at the values for the “baseline-shift” attribute in XSL-FO and the “vertical-align” attribute in CSS for documenting values. Since the value will be unconstrained, the documentation should make best practice recommendations.

2.2.5 Use of <x> inside <ref>

Scope: A graphic showing a green square with a white letter 'A'.

Currently <x> is allowed inside <label> and <citation> but not inside <ref>. Requesting an <x> for use inside <ref> as well as inside <label> and <citation> so that <x> can be used in all three places.

Advice:

Add <x> to the model of <ref>. Best practices documentation should describe that the archive that wishes to preserve everything (including periods and spaces) within a reference should tag

- 23. Jones, Jill
as follows:
- <ref><label>23<x>. </x></label><citation>...</citation></ref>

3. New Dawn Discussion

The NLM Tag Sets have been in testing and use since December 2002. So far, all versions of the Suite and the Archiving, Publishing, Authoring, Book, Collection, and Historical Tag Sets have been document-backward-compatible, meaning that a document valid to 1.0 will still be valid to 2.3. But non-backward-compatible changes have been requested, and are being seriously considered.

If backwards compatibility is no longer a limiting requirement, perhaps this is the time to step back, review the Tag Sets in the light of users' experiences, and ask the user community what modifications and additions would be useful. What really works? What is less than optimal? In what direction should the next version go?

Evan pointed out that many major publishers are switching to the Tag Sets or can at least provide a clean export from their native XML to NLM. Now is an excellent opportunity to get their feedback. It would also be productive to ask those who reviewed the Tag Sets but chose not to use them, to determine what issues they had. Tommie pointed out many publishers have used the customization mechanisms, sometimes extensively, and it could be productive to see such customizations were aligning. Bruce reminded us that some DTDs have been written as NLM supersets customized and some new DTDs have been "informed by" the Tag Sets and we should talk to both groups.

It was proposed that the NLM Tag Set Secretariat (Mulberry) undertake a series of guided telephone conversations (rather than a more static questionnaire) to take a deeper look at:

- Who is using the Tag Sets;
- How they are using them (and at what point in the journal lifecycle);
- What alterations did they make and why; as well as
- Who is not using the Tag Sets and why they chose another route.
-

A subcommittee of the Working Group will draft a fairly detailed "what are we looking for?" document and that can be used to guide the one-on-one conversations. The same individual will conduct the conversations to keep the bias as constant as possible. The material will be recorded by a note-taker, but if anyone is uneasy at having their name recorded, we will not record the specific source.

The difficult part of such conversations is getting the correct level of person in each organization on the phone. We need a starter list of specific people to ask, and then we need to go beyond those we know.

Evan suggested that we frame the "why" of this very carefully, emphasizing benefit to the community; properly done this is a community service. Bruce pointed out that there are business benefits, such as legal deposit in the UK that can also be mentioned. We can promise anonymity of the sort "a North American journal publisher" and agree not to identify sources explicitly.

Suggestions were made that we troll for feedback from the lists such as SSP, OmniMark Users List, CrossRef Publishing Technology blog, etc., asking both technical and non-technical publishers as well as archives and vendors. We agreed that the Working Group does not know everyone who is using the Tag Sets. Evan suggested that we take a lesson from the METS registry, where people register their METS use. Alex Brown volunteered

coordination with the online validation service his company runs for the Journal of Korean Medical Science.

NLM and Mulberry will draft solicitation statements, which the Working Group can approve. Once done, these can be posted to the NLM list serve and on the NLM site for cross-linking.

4. Meeting Action Items

- All Working Group Members:
 - To help put together a list of users for the guided conversations, send people's names (providing brief rationale for inclusion and contact information) to Tommie Usdin
 - To help the subcommittee write up the base document for the conversations, send potential questions and areas for questions to Tommie Usdin
 - To help publicize the call for feedback, send names of list serves to which you can post the notice to Tommie as well, so she can coordinate.
- New Dawn Subcommittee will write:
 - Known-party introductory outline to question conversation (for Working Group members to send to their clients and contacts);
 - Blanket solicitation to question conversation;
 - List of questions for customizers of NLM Tag Sets;
 - List of questions for Tag Sets that were “informed by” NLM Tag Sets;
 - List of questions for those who chose not to use NLM Tag Sets; and
 - List of dissemination assignments so that we do not post to the same lists multiple times or bug people by duplication.

5. Potential User Questions

This is a preliminary question list created by brain-storming during the phone conference and should not be taken as final:

- Top ten things you hate about the Tag Sets
- What is awkward, what things you have most trouble with?
- What have you modified (light)?
- From people whose Tag Sets were “informed by” the NLM Suite, who therefore made heavy modifications, may we see your models?
- Open ended: What should the NLM Suite be for? Full text retrieval? Archival storage? Typesetting and production?
- Are there any areas that seem ambiguous or confusing, where you were at a loss

- for which of several ways to encode?
- For publishers/archives that use the Tag Set for some of their material but not for others, what was the deciding factor in which to use where? What did you want to do in some of the documents?
 - Are you using any Processing Instructions? For what purpose?
 - Are you tagging fonts or other formatting? Are the current tagging capabilities sufficient to meet your needs? Have you added structures for tagging fonts or formatting?
 - What are the major omissions? Which Tag Set needs to be expanded to deal with this?
 -

Some questions should be asked early, as context, to determine the extent of the organization's involvement with the Tag Sets and because the answers may influence all other answers:

- Which Tag Set have you used? What color(s)? What version(s)? What constraint language(s)?
- What kind of material are you tagging? Journal articles? Conference Proceedings? Book (STM, Historical, etc.)? Legal Code?
- Are you tagging full text? Just metadata? References? Supplementary material? Some combination less than all?

Some questions need to be asked late in the interview, and some at the very end:

- Where in the document lifecycle/your workflow are you tagging with NLM? Prior to page creation? After page creation? (We sort of expect answers like Microsoft Word to InDesign to offshore XML conversion.)
- Are you working in or exporting to NLM? Is NLM only for interchange or is it your primary format? Do you archive NLM?
- We'd like to ask software and tools questions, but need to be very careful. Maybe do not ask about tools explicitly. Maybe ask: What software products do you use routinely with the Tag Sets? Into what tools does your XML go after creation? Maybe merely ask: Is your choice of Tag Set version constrained by tool or other requirements?
- If you are not using the most recent version, why not, what are the constraints?
- What resources can we provide you to make what you are doing easier? (We expect answers like tool customizations or a PDF to XML converter.)
- Is there anyone else of whom we should be asking these questions?

6. The Layout Issue Discussion

The Tag Sets do not preserve article look and feel by design, but some markup indicates semantic information even if the specific semantics are unknown (why this word is italic). Jeff asks where is the line and how far do we want to push it? The Tag Sets need a rule of thumb. Alex points out that if we can derive the formatting from context, we can definitely ignore it, but some things are not derivable at rendering time without tagging hints. Tommie reminds that just because we cannot derive it does not necessarily mean that we want to tag it in the tag set. This is the debate.

Bruce: the key goal of Green was to preserve intellectual content. Elements such as font family and font size do not directly preserve content, they are for the convenience of the users [processing].

Evan: Did not agree. For certain pieces of a journal article such as tables and display math, the content is in the presentation. All vectors are bold.

There is general agreement that math, tables, and running text are separate cases. Alex pointed out that in table there is an intellectual value in rendition. As an example, Bruce mentioned red shading on a table cell that is mentioned in the caption. Laura reminds us that our imperative is to understand the content, so the rule is summed up this way:

What we need to capture is style that is imperative to maintaining the meaning. Style that can be used to carry meaning (and that we would not consider ridiculous) should be preserved.

6.1 `` element (previously #2.2.1)

Scope: 

Font — The initial request had been to add the attributes “font-name”, “font-size”, and “class” to the element ``. These were rejected for version 2.3. The further suggestion, as a result of the layout discussion summarized above, was that the `` element be removed entirely, since its presence in the DTD is an artifact of inheriting it from the XHTML DTD.

Advice:

Leave `` alone for Version 2.3; do nothing. We do not even have enough information to deprecate its use at this time. Consider asking user as part of the survey about `` and elements like it and possibly delete it (or not) as part of 3.0, based on user feedback.

6.2 Layout-hint attribute `<table-wrap>` et al. (previously #2.2.2)

Scope: 

The current tag set already incorporates a “position” attribute for some block-type objects that provides information/hints regarding presentation. The attribute has these values:

- | | |
|--------|--|
| anchor | The object must remain in its exact location in the text flow. |
| float | The object is not anchored and may be moved to a new column, a new window, a new page, the end of the document, etc. |
| margin | In print, the object should be placed in the margin or gutter; |

online, the object should remain closely associated with the text.

The attribute is available on the following elements:

<boxed-text> Boxed Text; <chem-struct-wrapper> Chemical Structure Wrapper;
<fig-group> Figure Group; <fig> Figure; <graphic> Graphic; <media> Media Object;
<preformat> Preformatted Text; <supplementary-material> Supplementary Material;
<table-wrap-group> Table Wrapper Group; <table-wrap> Table Wrapper

- a) For tables and figures, another very useful piece of presentation information is whether they are most suitable for presentation as single column, double column, or as full page. Values seen in current DTDs include words such as “pagewide”, “col-wide”, “full-page”, “fold-over”, etc. Should a “width” be available for all elements that can have the “position” attribute? For any elements?
- b) Another useful piece of information is whether a table should be shown as portrait or landscape. Book tag sets use values like “portrait” and “landscape” or the angle of rotation in positive and negative degrees. Should a “rotation” attribute to tables, figures, and similar block objects?
- c) If they are to be added to the Suite, do (a) and (b) represent one piece of information or two?

Advice:

Leave <table-wrap> attribute alone for Version 2.3; do nothing. Ask user as part of the survey about these and attributes like them and possibly add them (or not) as part of 3.0, based on user feedback.

6.3 **Prohibit Face Markup in <label> (previously #3.6)**

Scope: 

Face markup (<bold>, <italic>, <monospace>, etc.) in a <label> is generally used to represent the format of the label in the context. For example, all of the <label>s in a list might be tagged <bold>. This is inappropriate; if all the <label>s are bold that is the format of that object and should not be tagged in the XML. We suggest prohibiting face markup inside <label>.

Advice:

In Archiving and Publishing, leave as is since they are not meant to be enforcing DTDs. Similarly for Book. However, for Authoring only, remove face markup from <label>. This will be a version 3.0 change.

7. **MathML Namespaces (previously #3.9)**

Scope: 

For reasons of backwards compatibility, the MathML prefix for the Suite has continued to be “mml”, although the latest MathML DTD defaults to a prefix of “m”. It

could/should be changed to “m” to stay compatible with common practice in the MathML community

Advice:

This only matters in the DTD world, since with schemas, prefixes are immaterial. Make no change for 2.3, but do not make this into a survey question. For 3.0, it would be very ugly in the DTD world to try to use both. The more up-to-date users and vendors are using "m", the majority "mml", thus there is no perfect solution. We decide that since the prefix is not supposed to matter, we will leave it alone for now and document how to customize the change.

8. New Book DTD

Advice:

Wait until after the whence-whither-what-now discussions for the current Tag Sets. Some of the material learned may be relevant to writing a new book DTD.

9. Pending Discussions for Version 3.0

Pending discussion items for version 3.0 will be addressed in a separate document for discussion in the March and subsequent meetings.