Tag Set Working Group Recommendations
November 2006

Table of Contents

1.   Tag Set Working Group Meeting

1.1   Introduction

The Working Group meeting took place by conference call on November 14, 2006.

This document contains a record of the advice of the Working Group when the group was able to reach consensus.   Those items where the group was not able to make a recommendation, and those that were not addressed because of time limitations are not included here. They will be included in the list of discussion topics for the next Working Group conference call, scheduled for December 20th 2006, at 11:00 am EST.

1.2   Attendees

1.2.1   NLM Staff
*   Jeff Beck (Moderator)
*   Steve DeRose
*   Marilu Hoeppner
*   Laura Kelly
*   Adeline Manohar
*   Kim Tryka

1.2.2   Secretariat (Mulberry Technologies, Inc.)
*   Deborah A. Lapeyre (Mulberry)
*   B. Tommie Usdin (Mulberry)

1.2.3   Telephone Participants
*   Mark Doyle (American Physical Society)
*   Beth Friedman (DCL)
*   Kathryn Henniss (Highwire)
*   Sharon McCamant (Cadmus)
*   John Meyer (Portico)
*   Jules Milner-Brage (Highwire)
*   Evan Owens (Portico)
*   Bruce Rosenblum (Inera)

1.3   Work Group Recommendations

*   Requests for additional changes and leftover business items will be posted on the mailing lists before the next Working Group meeting.

*   As soon as possible after the December meeting:

    *   Version 2.3 of the DTDs will be created, documented, frozen, and distributed. This version will be fully backward compatible.

    *   A list of proposed non-backward-compatible changes will be distributed.

    *   The following release will be a major release 3.0, which will for the first time be non-backward-compatible.

1.4 Change Applicability

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

A Archiving

P Publishing

A Article Authoring

B Book

2. Tag Set Changes

2.1 New Features

2.1.2 Make <title> optional in <app>

Scope: **A** **P**

Request:

We commonly find appendices without any titles as the formatting engine generates the word Appendix from the <app> element in Archiving, please make the appendix title optional.

Working Group Advice:

For Archiving: make appendix title optional.

For Publishing: make either <title> or <label> required for <app> (as the titles/labels for <sec> are modeled in Publishing).

2.1.4 Add <kwd-group> to <sec-meta>

Scope: **A** **P** **A** **B**

Request:

Requestor asks for an optional, repeatable <kwd-group> to be added to section metadata.

Working Group Advice:

Add <kwd-group> to the section metadata model:

```
sec-meta  (contrib-group*, kwd-group*, permissions?)
```

2.1.5   Add Elements to Citation

2.1.5.1 Add Identifier to Citation

Scope: A P A B

Request:

There has been some confusion about the use of object-id to record DOIs in citations. We suggest that the documented definition of object-id be observed strictly: *used to record an identifier, such as a DOI, for elements other than the top-level element <article>, for example, for the elements or <fig>*. Following that logic, a object-id inside a citation would be the identifier for that component of and location in the article (e.g., the first citation in the ref list, or the third citation, etc.) should a publisher choose to identify the components of the article to that granularity and <pub-id> should be used to point to entire publications.

Working Group Advice:

The internal versus external distinction between identifiers inside citations should be clearly documented and examples added.

2.1.5.2 Add <date> and <string-date> to citation

Scope: A

Request:

The elements <date> and <string-date> are needed inside <citation> because we encounter input has text elements marked as dates but not fully fielded as day, month, year.

Working Group Advice:

Change the model of %references.class; to the following to add <date> and <string-date>"

```
<!ENTITY % references.class

            "access-date | annotation | article-title |
            collab | comment |
            conf-date | conf-loc | conf-name | date |
            day | edition | email | elocation-id |
            ext-link | fpage |  gov | isbn |
            issn | issue | issue-id | issue-title |
            lpage | month | name | object-id |
            page-count | page-range |
            patent | person-group | pub-id |
            publisher-loc | publisher-name |
            role | season |
            series | source | std | string-date |
            supplement | time-stamp | trans-source |
            trans-title | uri |
            volume | volume-id | year"            >
```

2.1.6   Add <label> to <address> and <product>

Scope: **A**

Request:

It would be helpful to allow <label> inside <address> and <product> inside <citation> as we find that some publishers include text labels in these contexts. Examples include descriptive labels such as "Affiliation:" or sometimes footnote marker such as "*" that lack proper cross referencing markup.

Working Group Advice:

Add <label> to %address-model;:

```
<!ENTITY % address-model
        "(%address.class; | %address-link.class; | %label.class;)*">
```

Add <label> to %product-elements;:

```
<!ENTITY % product-element
                        "|%article-link.class; |
                        %appearance.class; | %break.class; |
                        %emphasis.class; |
                        %inline-display.class; |
                        %inline-math.class; | %math.class; |
                        %phrase.class; |  %price.class; |
                        %references.class; |
                        %simple-link.class; | %subsup.class; |
                        %x.class; | %label.class;"              >
```

2.1.7   Add ID Attribute in Various Contexts

Scope: **A P B**

Request:

The "id" attribute should be allowed in many more elements than it is currently. Any element that is a substantial component of the article should be allowed to have an id because a publisher may have a cross-reference to that object.

Working Group Advice:

Add an "id" attribute of type ID to the following elements: abstract, app-group, attrib, author-comment, contrib-group, custom-meta, fn-group, gloss-group, person-group, product, ref-list, sig-block, title.

```
        id      ID      #IMPLIED
```

2.1.8   Add Translated Journal Title and Subtitle tags

Scope: **A P**

Request:

Requester has asked that <trans-title> and <trans-subtitle> be added to <journal-meta>

Working Group Advice:

The elements <trans-title> and <trans-subtitle> will be added to <journal-meta>.

Whether that addition will be as:

```
journal-title, journal-subtitle*,
 trans-title*, trans-subtitle*
```

or as the more correctly paired::

```
(journal-title, journal-subtitle*)*,
(trans-title, trans-subtitle*)*
```

Will be decided when non-backward-compatible changes are discussed in December.

2.1.10 Align <related-article> Content Model and Attributes

Scope: A P A B

Request:

It would be helpful to all <journal-id> in the content of <related-article>, so that all possible attributes would also be possible as textual elements.

Working Group Advice:

Add <journal-id> to the content of <related-article>.

2.1.12 Allow <page-range> without <fpage> in <article-meta>

Scope: A

Request:

As currently defined, it is impossible to have a <page-range> element without also having an <fpage> element; this seems over-prescriptive.

Working Group Advice:

Change the article metadata model to include the following:

```
((fpage, lpage?)?, page-range?) |
  elocation-id )?,
```

2.1.13 Sequence Information

Scope: A P A B

Request:

The treatment of sequence information in the Tag Set is incomplete. There are at least three different ways that publishers encode sequence information that we have observed:
- Sequence on the page; e.g., page 100, sequence 1; page 100, sequence 2; page 100, sequence 3'
- Sequence in the issue; e.g., issue 1, page 1, sequence 1; page 10, sequence 2; page 10 sequence 3; page 20, sequence 4; issue 2, page 30, sequence 1; etc.; and
- Sequence in the volume.

Working Group Advice:

Add "seq" as an IMPLIED CDATA attribute to volume and issue as well as to fpage and document the three separate meanings carefully.

2.1.14 <break> in Emphasized-text (<bold>, <italic>, etc.)

Scope: **A**

Request:

We have encountered publishers who use the break element in create explicit line breaks in table cells (e.g. in header cells) and in article titles. It would be helpful to allow break inside the emphasized-text elements.

Working Group Advice:

Change the parameter entity emphasized text to add line breaks:

```
<!ENTITY % emphasized-text
                        "| %all-phrase; | %break.class;"   >
```

2.1.15 <x> inside <contrib-group> and <contrib>

Scope: **A**

Request:

For publishers who do not use generated text but rather attempt to capture all the text of their publications in their XML files, it is necessary for the tag set to allowing connecting text between the elements of <contrib-group> and <contrib.> using <x>:

Working Group Advice:

Make <contrib-group> into a large unordered OR group as though it were modeled:

```
(contrib | %contrib-info.class; | %x.class;)+                 >
```

2.1.16 X in Additional Contexts

Scope: **A**

Request:

We have found it necessary to add the x element to many elements to reflect structures possible in other models and a variety of possible contexts for generated text.

Working Group Advice:

Use the %x.class; to add <x> to all of the following: address-model, conference-model, copyright-statement-elements, etal-model, kwd-elements, label-elements, rest-of-para.class, role-elements, title-elements, and xref-elements.

2.1.17 X Content Model Change

Scope: **A**

Request:

The contents of X is currently restricted to PC data with no markup at all. We propose expanding that to allow emphasis and the sub and sup elements as these are commonly used in expanding out generated text, one of the stated purposes of the element.

Working Group Advice:

Change %x-elements; to so that <sub> and<sup> will be added to <x>:

```
"|%emphasis.class;|%subsup.class;" >
```

2.1.18 Style Attribute
Scope: A P A B

Request:

Publisher DTDs containing varying amounts of format and presentation information, depending on the style of the DTD and its relationship to typesetting or page layout system. We have found a number of instances where there is significant meaning encoded in presentation information that there is no way to encode in the Interchange Tag Set. To fix that problem, Portico has resorted to using CSS values stored in a style attribute on the following elements:
- th (for cell borders and shading)
- td (for cell borders and shading)
- font (for width, font family, etc.)
- graphic (for vertical alignment)
- inline-graphic (for vertical alignment)

Working Group Advice:

A number of suggestions grew from this discussion:
- Add the HTML "style" attribute back in the XHTML table model for <th> and <td>
- Add a CDATA #IMPLIED attribute to <inline-graphic> that states relationship of the graphic to the baseline. Legitimate values might be numeric offsets from the baseline (as Elsevier does) or "top|bottom|middle" as other DTDs have done.
- Add several new emphasis elements: <roman>, <san-serif>
- As a first step, add "font-name" and "font-size" attributes to <font>, as well as a "class" attribute bucket to hold additional formatting parameters.

2.1.19 Loosen values of "mimetype" & "mime-subtype" attributes of <graphic>
Scope: A P A B

Request:

There is currently a list of mimetypes and subtypes; the attributes for mimetype and mime-subtype should not be tightly constrained in the Interchange DTD.

Working Group Advice:

Change the values of the "mimetype" and "mime-subtype" attributes in the %graphic-atts; parameter entity from an explicit list to CDATA for <media> and <graphic>

Since this change would make <graphic> the same as <media> for Archiving, the documentation must stress the differences. Graphic is used for still images. Media is used for video, audio, all moving images and time-dependent graphic types.

2.1.20 Improve documentation of @citation-type

Scope: A P A B

Request:

Document should provide guidance on how to tag citations that seem to be of multiple types, such as a citation to a book published by a government.

Working Group Advice:

The section type attribute allows for multiple values by means of providing two values separated by an or-bar, e.g. "materials | methods". A similar arrangement should be adopted and documented for citation types, for example, "book | gov".

2.1.21 Add a <float> at end of <body> to contain all floating structures

Scope: A P

Request:

Many users prefer to have all floating structures (<table-wrapper>, <boxed-text>, <chem-struct-wrapper>, <fig>) in a known location at the end of the <body>. Some have been creating an "extra" <sec> after all regular <body> content to contain all floating objects, which is clearly tag abuse, since such a "section" is not part of the narrative of the article and should not, for example, appear in the Table of Contents.

Working Group Advice:

Add a special floating objects wrapper elements to the end of article (before any subarticles), subarticle, and response.

2.1.22 Make Tagging Recommendations Individually Addressable

Scope: A P A B

Request:

That each page in the Tag Library be made uniquely addressable by giving it a unique number.

Working Group Advice:

Giving each page a unique address that an end user can see might lead to abuse of the navigation system. But Mulberry can build a table of identifiers to each page of the tag library that can be used by applications.

3. New Business for Future Discussion

3.1 Country attribute on <patent>

We currently have no way to identify the country associated with a <patent>. The suggestion is that we should add country as an attribute.

3.2 Flag to Suppress DOI of Citation for Display

In the conversation about et al and ibid also discuss flagging DOIs and other portions of citations that are not to be displayed to the end user.

3.3 Change <long-desc> model to %sec-model;

The current model for <long-desc> limits the content to #PCDATA. NCBI has found this limitation problematic and would like to make the model for long-desc the same as %sec-model.

4. Working Group Action Items

- **All:**

    - **DTD Rationale** — Provide feedback on the differences between, the appropriate audience for, and the rationales behind the Archiving, Publishing, and Authoring DTDs.

    - **Formatting as a Goal** — Consider whether or not "typesetting" (i.e., computer pagination) should be among the stated goals of the Publishing DTD.

    - **Font Attributes** — Find examples and use cases for attributes for the <font> element. (Do we consider XSL-FO values as guides?)

- **Bruce Rosenblum:**

    - **Layout Hint Attributes** (tables, figures, graphics) — Investigate the issue with an outside expert. Report back on the various strengths and weaknesses of DTD changes, processing instructions, namespaced-DTD changes, or the prevalence of layout information in industry DTDs.

- **Evan Owens:**

    - **Layout Hint Attributes** (tables, figures, graphics) — Investigate the extent of layout information in a few well-known industry DTDs.

    - **Chemical Structures in Citations** — Find one or more examples, so that we can see where inside the citation a chemical structure was being used.

    - **Graphical Alignment Attribute** — Find one or more examples or explain use cases

- **Jeff Beck/Laura Kelly:**

    - **Layout Hint Attributes** (tables, figures, graphics) — Investigate the extent of layout information by BIOMed Central.

- **Bruce/Debbie/Jeff**

    - **Type Attributes** — Make a recommendation for additional "type" (information classing) attributes for those elements (listed in the agenda) that do not have type attributes.