Archiving and Interchange Tagset
Working Group Meeting

18 August, 2003
NLM Board of Regents Room
National Library of Medicine

Attendees

Archiving and Interchange Tagset Working Group

- Adriana Ardeleanu (Microsoft)
- Matthew Cockerill (BioMed Central)
- Mark Doyle (American Physical Society)
- Beth Friedman (Data Conversion Laboratories, Inc.)
- Gerry Grenier (IEEE)
- Evan Owens (JSTOR Electronic Archiving Initiative)
- Diana Robinson (HighWire Press)
- Nick Twyman (Public Library of Science)
- Tommie Usdin (Mulberry Technologies, Inc.)
- Roy Williams (Cal Tech)

NCBI

- Jeff Beck
- Laura Kelly
- Sergey Krasnov
- David Lipman
- Jim Ostell
- Ed Sequeira

Archiving and Interchange DTD Secretariat

- Deborah Lapeyre (Mulberry Technologies, Inc.)

1.0 Agenda (distributed at the meeting)

1. Call to Order and Opening Remarks
   housekeeping details
2. Discussion of Scope for Working Group
3. Open Discussion on Suggested Additions/Changes to Tagset
4. Review of Notable Suggestions Made by Users
5. Summary Of Conclusions And Identification Of Areas For Discussion
6. Closing remarks

NOTE: There will be no confidential information discussed at the Archiving and Interchange Tagset Working Group meeting.

2.0 Administrative Material

2.1 Scope of the Working Group

The Archiving and Interchange Tagset Working Group has been established to provide NCBI with industry feedback concerning the Archiving and Interchange DTD Suite as well as tools and utilities to store, retrieve, convert, display, and archive data tagged according to this suite. The hope is that, with guidance from the Working Group, the Suite and related products can be made useful, not only to NCBI, but to the larger electronic publishing and archiving communities.

2.2 Working Group Meetings

Subsequent meetings of the Working Group will take place as conference calls, with much of the work accomplished through email discussions. The first remote meeting of this Working Group will be during the week of November 3-7, 2003, not to be held on US election day. The numbers for the conference call and agenda will be distributed in advance. The date will be chosen soon and emailed to all participants.

2.4 Discussion Groups Established During the Meeting

Authoring Tool Discussion Group

- Adriana Ardeleanu
- Matthew Cockerill
- Mark Doyle
- Nick Twyman
- Jeff Beck (to take lead in forming group)

Discussion Group for External XML Structures

- Jim Ostell
- Mathew Cockerill
- Nick Twyman
- Tommie Usdin
- Roy Williams
- Jeff Beck (to take lead in forming group)

3.0 Issues Discussed

- The scope, purpose, and relationships among the Archiving and Interchange DTD (hereafter Green DTD), the Publishing DTD (hereafter the Blue DTD), and the original and revised original PMC DTDs.
- What went wrong with ISO 12083 and how not to have this DTD suffer the same fate
- How to avoid the negative baggage that NLM projects sometimes arouse in the publishing community.
- What tools should be developed to enable use and facilitate adoption
- The process by which additions, corrections, and tools should be considered and implemented.
- The specific list of Tagset and DTD Suite changes brought to the meeting by the Archiving and Interchange Secretariat. These suggestions came from the open discussion list and from the "Phase II" comments in the original DTD.
- What is necessary to get this group to consider a modification/addition to the Tagset?
- How does the DTD Suite handle "Islands" of foreign tagging, such as CML or other valid (to some DTD or schema) XML Tagset?
- Print version of the Tag Library

4.0 Recommendations

4.1 Develop and Promote DTD Tools and Utilities

The world at large will not adopt a DTD suite, however meritorious, unless there is ample tool support. Therefore, to increase adoption, NCBI should develop, solicit, promote, and make centrally available tools to work with this suite and links to such tools on other sites.

Existing and already planned tools include:

- Preview/Converter (XML by Green DTD to HTML)

- Preview/Converter (XML by Blue DTD to HTML)
- Converter(XML by pmc-1 DTD to Green DTD)
- Converter (DocBook XML to Green DTD)
- Other Converters (from XML according to a few of the widely used Publisher DTDs such as Keton and HighWire)
- Print (Blue instance to PDF)
- W3C XML Schema version of Blue DTD

Other Tools Suggested by the Working Group

- Authoring tool for Microsoft Word
- Online Validator to check Blue and Green instances
- Additional converters from major publishing DTDs

4.2 User Contribution Area

- Need an area of the website where individuals and organizations may contribute tools such as stylesheets, authoring customization packages, and converters.
- NCBI should establish a tool catalog with descriptions of tools and URIs (partly a list of links, partly an archive for downloading)
- Authors of such products will be encouraged to put their source into SourceForge (or other open access mechanism) and provide a description, metadata, and links to NCBI for the tool catalog

4.3 Open Change Process

- The Working Group should include one or more members from active, traditional scientific and medical publishers.
- Suggestion for changes, additions to the tagset will continue to be taken from the open, public list.
- Changes (except error corrections) should be undertaken on a regular basis (quarterly, twice-annually, and annually were mentioned)
- The Working Group will vet suggestions and set priorities.
- A mechanism should be established to let users, easily and in one place, review current issues under consideration.

5.0 Further Discussion Needed

There were several semantic tagging or enhanced structural tagging suggestions where the Working Group felt that further information was needed. Any parties who desire the following changes should present use cases and scenarios, tagging utility or rationales, and a description of any structural-only alternative to the tagset proposed. How is the equivalent structure being handled by DTDs and schemas now?

Such unresolved issues include:

- Questions and Answers (in the formal sense)
- Taxonomic Key
- SVG
- Island of foreignly-tagged material within the Green XML stream

6.0 Out of Scope

Potential additions to the Tagset that the Working Group felt were out of scope for this group:

- Packaging or delivery wrapper standards;
- Digital rights management (rights metadata and acceptable usage metadata should be in another layer and not part of the DTD Suite);
- Metadata complete enough for content harvesting;
- Forms and fill-in-the-blank formatting tagging; and
- Non-article material such as: ads, job ads, classified advertising, display advertising; calendars, meeting schedules, and announcements; and journal material such as author guidelines, policy and scope statements, editorial and advisory boards, and detailed indicia, except as such material could be translated into ordinary paragraphs, lists, sections, etc.

Appendix A. Detailed Tagset Suggestions

A.1 How to Submit Tagset Addition Requests

The following informal change process was recommended until NCBI writes more formal procedures. Error reporting should be on the list or by email, as succinctly as possible. For additional structural or new semantic material, a petitioner should provide at least the following:

- Rationale for change; assertion of benefit;
- Description of need and community;
- Coded examples; and
- If appropriate, display examples in a readily read format like HTML or PDF.

A.2 Element/Structure Changes and Additions

- **Tables**—#1)Add the CALS DTD fragment to the Green DTD #2) Add the CALS fragment to the Blue DTD and restrict the Green DTD to whatever the PubMed Central archive wants
- **Alternative versions**—Allow more than one version of tables, graphics, etc. e.g, a table and a gif for a tabular structure; MathML , TeX, gif combination for an equation, etc.
- **Media Object**—yes, give us a tag for it, either by adding a new `<media>` element (or equivalent) or be redefining what a `<graphic>` is.
- **Named content**—expand its content model to allow more inside it
- **Labels**—in Green, allow them on almost anything, but treat this as an override and document it as such
- **URL in DTD**—Consider at least a list of the most recent URLs where the material could be found, or point to the top level
- **Full journal title**—Add an element in the metadata

A.3 Attribute Additions and Changes

- Add a `"type"` attribute or equivalent to `<target>`
- Add an `"ID"` attribute or equivalent to `<paragraph>`
- Add the ability to say that a footnote concerns Conflict of Interest
- Add an `"information-class"` and/or `"role"` attribute or equivalent to (at least) `<paragraph>` and consider it for most large structures such as def lists, sections, etc. For green this is CDATA, for other this may be an explicit list with an "other" escape hatch and CDATA attribute for recoding the other.

A.4 Consider Incorporation of Other Standards

- OAI (Open Archives Initiative Protocol for Metadata Harvesting)

A.5 Postponed for Now

- XLink extended links

Appendix B. Suggestions for DTD and Documentation Changes

This list was provided to the attendees at the meeting. Some of the items were discussed briefly, if at all. All conclusions and recommendations from those discussions are listed in the minutes.

Suggestions for DTD & Documentation Changes

August 2003

These comments and suggestions have been collected from comments made on the DTD discussion lists and email from people working with the DTDs.

Category: FAQ

Change which? **Both** Change what? **Documentation**

Request No: **17**
Requestor: **Nick Twyman**

I was hoping to get a better understanding of the status of the publishing-dtd and it's relationship to the existing pmc-1.dtd document.

We're in the process of setting up our publishing systems and want to make sure we are in-line with the NCBI/PMC standards right from the start. Question is, when will NCBI start using the Journal Publishing DTD for submissions into PMC?

I am assuming that if we adopt the Journal Publishing DTD from the start this will not be a problem.

Category: FAQ

Change which? **Archiving** Change what? **Documentation**

Request No: **26**
Requestor: **"Hollinger, Blaine"**

Expand the description of the reason for these tag sets and the expected uses.

Why:

Users who send email to the list such as:
I have no idea what this is about and you don't do a good job of telling me what benefit it is or why I should want to use it?

Category: Major

Change which? **Both** Change what? **Tag Set**

Request No: **5**
Requestor: **DTD comment via BTU**

Add structures for encoding Questions and Answers. (We can currently do this with paragraphs, but that doesn't capture the real content of the Q&A pairs.)

Why:

This would be useful in instructional material and interviews. Also in CME materials (see suggestion on CME)

Category: Major

Change which? **Both** Change what? **Tag Set**

Request No: **6**
Requestor: **DTD comment via BTU**

Add structures for Continuing Medical Education materials

Why:

These materials are sometimes included in journals. At present they can be encoded as paragraphs, but that is not a satisfactory way to capture the true intent of the material.

Category: Major

Change which? **Both** Change what? **Tag Set**

Request No: **7**
Requestor: **DTD comment via BTU**

Forms and fill-in-the-blank material. We could add structures for capturing the content of forms, including the blanks to fill in, etc.

Why:

These materials are sometimes included in journals. At present they can be captures as graphics, which are not generally searchable.

Category: Major

Change which? **Both** Change what? **Tag Set**

Request No: **10**
Requestor: **Original DTD comment**

Advertising included in the journal. This may include: job ads, classified advertising, display advertising.

Why:

Some readers may feel that the advertising included in the journal is as interesting as the article content. We currently have no way to capture anything in addition to article content from journals.

Category: Major

Change which? **Both** Change what? **Tag Set**

Request No: **11**
Requestor: **Original DTD comment**

Calendars, meeting schedules, and announcements.

Why:

These can be handled as articles, but the difference in type of material may be important to some users.

Category: Major

Change which? **Both** Change what? **Tag Set**

Request No: **12**
Requestor: **Original DTD comment**

Add structures to preserve about the journal materials, including author guidelines, policy and scope statements, editorial and advisory boards, detailed indicia, etc.

Why:

These can be handled as articles, but the difference in type of material may be important to some users.

Category: Major

Change which? **Archiving** Change what? **Tag Set**

Request No: **19**
Requestor: **Wendell Piez**

Consider an information class ("class" or "type" perhaps) attribute on most generic structures: paragraph, sec, speech, stmt, def-list, etc. for capturing more specific tagging in source documents.

Request No: **43**
Requestor: **Wendell Piez**

Elsevier also adds a "role" attribute to the <p> element. This is documented as "The attribute role allows one to categorize paragraphs, and attach a special meaning to them. For instance, it makes it possible to mark a paragraph as a "motto", and handle it in a different way than an ordinary paragraph." This would essentially be like the HTML class attribute in that it gives an additional back door to semantically qualify domain-specific material without adding new elements.

Category: Major

Change which? **Archiving** Change what? **Tag Set**

Request No: **36**
Requestor: **Bruce D. Rosenblum**

add structures for Taxonomic Keys

For the moment, it looks like it's best to handle with head-less sections, named-content, and list-content. However I would like to request that the review board consider adding a taxonomic model to the DTD at some point in the near future to accommodate the needs of the bioone DTD along with several other publishers that have this requirement.

named-content is working really well for us, too. We have directed several DTD users towards it with very nice results. We also have a group that's using the list-content attribute for a taxonomic key section! It was exactly what they needed to avoid a set of new elements.

Category: Major

Change which? **Archiving** Change what? **Tag Set**

Request No: **37**
Requestor: **Bruce D. Rosenblum**

handle borders on table cells.

Much to my dismay, I just discovered that XTHML has NO table cell borders. How are we supposed to put lines under table headers, or under individual cells in table headers. Please let me know if I'm missing something, since I suspect this will cause significant problems when converting many CALS tables to the NLM DTD.

Category: Major

Change which? **Both** Change what? **Documentation**

Request No: **48**
Requestor: **DAL**

Beef up the examples. Provide at least one rich example of each element. For many that can be used in several ways, provide examples of several contexts in which the element can be used.

Category: Minor

Change which? **Both** Change what? **Tag Set**

Request No: **1**
Requestor: **Jeff Beck**

We need to be able to allow an "alternate version" for <display-formula> and <inline-formula> or at least allow for multiple encodings or representations of the same piece of math.

Why:

We need to be able to allow an "alternate version" for <display-formula> and <inline-formula>. or at least allow for multiple encodings or representations of the same piece of math.

Category: Minor

Change which? **Both** Change what? **Tag Set**

Request No: **2**
Requestor: **Jeff Beck**

need to add a way to tag that an article is "open access" not just leave off the copyright line

Request No: **9**
Requestor: **Jeff Beck**

Electronic and digital rights management information

Category: Minor

Change which? **Archiving** Change what? **Tag Set**

Request No: **3**
Requestor: **Jeff Beck**

We need to allow movies as a display object (like figures) and not just as supplemental data.

Category: Minor

Change which? **Both** Change what? **Tag Set**

Request No: **8**
Requestor: **DTD comment via BTU**

Conflict of interest statements and financial disclosures.

Why:

This content is now encoded as paragraphs, generally in a footnote. Since this content is important to many readers it might be useful to specifically identify it.

Example:

<ack> The meeting was sponsored by the Medical Research Council, UK, with financial support from the Wellcome Trust, and Carl Zeiss Ltd. We thank Jim Duffy for drawing <figr rid="f2">Fig. 2</figr>, and the meeting's speakers for agreeing to the included citations. S.M.G. is supported by the Swiss National Science Foundation and D.L.S. by the National Institute of General Medical Sciences/NIH, USA. </ack>

<ack> This work was supported by the Imperial Cancer Research Fund and the Human Frontiers Science Program. A.C. was supported by a European Science Exchange Fellowship from the Royal Society and the Swiss National Foundation, M.T. was supported by an EMBO Fellowship and J.K.K. is a Boehringer Ingelheim Fonds scholar. </ack>

Category: Minor

Change which? **Both** Change what? **Tag Set**

Request No: **13**
Requestor: **BTU**

Add a place to store the full journal title. We can currently capture a wide variety of journal title abbreviations, but not the full title.

Example:

<title type="journal">Epilepsy Currents</title>
<title type="journal">Neurology</title>

Category: Minor

Change which? **Both** Change what? **Tag Set**

Request No: **14**
Requestor: **B Rosenblum**

Add extended links to xlink. I just took a look at common.ent and noticed that in both cases, xlink:type is fixed as "simple". Did you chose to avoid extended linking because, in reality, no one's really doing it yet? Any comments on the topic, or the specific questions raised above would be appreciated.

Why:

Debbie noted: Yep, all we gave them was HTML href-style links. IF anybody is actually doing extended multi-headed links in journal publishing, that can be one of the first things they ask the advisory board for, I'll believe it when I see it.

I have seen good use of out-of-line linking, but not in STM journal publishing and not in the kind of material that Pub Med Central and Harvard/Mellon/Yale are archiving, even in the book material.

Tell them that, putting something into the DTD is an implicit promise that it will exist, and therefore a burden on every implementor. We did not go with fancy linking, by design, to cut the implementation and conversion burden, UNTIL SUCH TIME AS IT PROVES USEFUL.

Category: Minor

Change which? **Both** Change what? **Documentation**

Request No: **15**
Requestor: **B Rosenblum**

In the link.ent section, use of XLink global attributes is mentioned. I would appreciate clarification on the scope of the XLink functionality. Useful features of XLink are extended linking and out-of-line linking. If possible, could you please provide further explanation or example of how this might work in practice.

Category: Minor

Change which? **Archiving** Change what? **Tag Set**

Request No: **18**
Requestor: **Wendell Piez**

Named-content was envisioned as an inline, phrase-level object. As such, it should contain at least the %inline-display.class inline-graphic private-char and probably the %simple-intable-display.class as well as chem-struct, graphic, and preformat

But there is also the question of whether named content should be a larger structure, allowed to contain anything a paragraph could contain.

Why:

We actually wanted to put a graphic in a named-structure.

Category: Minor

Change which? **Archiving** Change what? **Tag Set**

Request No: **20**
Requestor: **Wendell Piez**

Consider adding non-hierarchical heading at the paragraph level.

Why:

Capture headings that are not part of the structure of the article, such as in Nature.

Category: Minor

Change which? **Archiving** Change what? **Tag Set**

Request No: **22**
Requestor: **B Rosenblum**

Add <no> to a whole lot of elements. While we're cross-checking, Elsevier has the <no> element (their equivalent of <label>) in the following places:
<cor>
<aff>
<sec>*
<bib>*
<bb>*
<other-ref>*
<app>*
<fn> (did I really leave it at the symbol attribute here?)
<enun>* (our <statement>)
<l>
<dl>
<list>*
<list-item> (I think we agreed to forego it here)
<tbl>*
<tblfn>
<fig>*
<textbox>
<upi>
<fd>
On a quick look, I think we got the ones with the *, but it looks like we may have missed a bunch of others

Why:

People converting content from Elsevier journals.

Category: Minor

Change which? **Archiving** Change what? **Documentation**

Request No: **29**
Requestor: **Robert C. Leif**

By the way all DTDs, Schema, and other XML documents that can be downloaded should include the URL from whence they came. The W3C even includes the URL of the latest version.

Category: Minor

Change which? **Archiving** Change what? **Tag Set**

Request No: **34**
Requestor: **Bruce D. Rosenblum**

Now that we're going into production with the DTD for one of our customers, we keep running into little things. Today I noticed that somehow we didn't add <label> to <list-item>. Can you please add this to the errata list. It's essential to correctly handle University of Chicago and Elsevier 5.0.

Request No: **21**
Requestor: **Bruce D. Rosenblum**

I just was setting something up and noticed we don't have <label> in <aff>.

We need it to replicate the Elsevier model of:

<!ELEMENT aff - o (no?, %data;, %cty-cny;)>
<!ATTLIST aff
id ID #IMPLIED>

Request No: **41**
Requestor: **Bruce D. Rosenblum**

In addition to the <p> changes I sent earlier today, <aff> needs <label>, and actually so does <corresp> (see Elsevier DTD 4.3 and 5.0 - in 4.3 and later, their QCTool requires <label> in a cor). I think you talked me out of label in corresp before by arguing that footnote could be used instead, but for ease of conversion we really do need it. I'm more than happy to have you bury it in a parameter entity so you can easily remove it from publishing DTDs :)

Category: Minor

Change which? **Archiving** Change what? **Tag Set**

Request No: **40**
Requestor: **Wendell Piez**

Remove <attribution> from <verse-group> as it is a) susceptible to tag abuse and b) semantically incorrect modeling.

The attribution is part of the larger unit (such as a display-quote) (or epigraph if we had them) that should be able to contain a title, verse-group, attr, etc. Display quote has such a model.

So, either add <poem> or other wrapper or delete <attribution> from versegroup.

Request No: **45**
Requestor: **Wendell Piez**

Describe/demonstrate how to encode the attribution of verse.

Why:

People who won't know how to use the DTD in this case.

Category: Minor

Change which? **Archiving** Change what? **Tag Set**

Request No: **42**
Requestor: **Bruce D. Rosenblum**

We should add an ID attribute to <p>. I'm starting to see DTDs (e.g. Elsevier 5.0) in which this is being done so that linkages can be made to arbitrary paragraphs.

Category: New Service or Facility

Change which? **Archiving** Change what? **Documentation**

Request No: **23**
Requestor: **David Holden**

Do you have a printable version of your archive-dtd documentation?

Category: New Service or Facility

Change which? **Archiving** Change what? **Tag Set**

Request No: **24**
Requestor: **Robert C. Leif**

Please translate your DTDs into XML schemas. XMLSpy can be used for this purpose.

Why:

The schemas will then work with Microsoft Word 2003 and other modern software.

Category: New Service or Facility

Change which? **Archiving** Change what? **Tag Set**

Request No: **27**
Requestor: **Alain Boisvert**

publish stylesheets to use with the DTD.

I would like to use this DTD but we need XSL style sheets to publish

Are there already developed style sheets?