Tag Set Working Group Meeting
May 2007

Table of Contents

1. Tag Set Working Group Meeting

The Working Group meeting took place by conference call on May 3, 2007. This meeting was spent in discussion of what we had learned in interviews with NLM tag set users, producers of tag sets that were “informed by” the NLM tag sets, and informed non-users.

The next meeting will be Thursday 14 June, 2007 at 11:00 am Eastern Daylight Time.

1.1 Attendees

1.1.1 NLM Staff

- Jeff Beck (Moderator)
- 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 Press)
- Evan Owens (Portico)
- John Meyer (Portico)
- Bruce Rosenblum (Inera)
- Parke Shissler (Cadmus)
- B. Tommie Usdin (Mulberry)

2. Action Items

- For All WG Members:
  - Advisory Group — The group seems to have dwindled since the early days. We need more quality members. All think of appropriate names and send email to Jeff, who will issue invitations.
  - CITATION — Send Jeff any thoughts on citation models new and old and what the tag sets might be doing.
  - SAMPLES — Everyone agrees that the documentation needs more samples. But the Secretariat does not publish or archive journal articles. We have no samples to share. Semi-politely put: You want samples, you send samples. We agree to tag them, but we need material we have permission to share. We would also like examples of things that are difficult to tag using the current Tag Sets.

- For Portico:
    - Supply example of Parts of a journal issue and what metadata needs to be recorded to call an article part of a Part.
    - Produce a list of what is “missing” from Portico’s point of view in 2.3 and send it to the Secretariat.
    - Keywords. Supply samples of Elsevier’s tiered keywords, string keywords, and major-heading/minor-heading keywords.
    - Supply an example of the use of and need for <roman> emphasis.
    - Bring us an epigraph and tell us what about it is not just a <disp-quote>.
    - Make a list of things for which you have had to use <named-content>.
    - Supply examples of some of Elsevier’s (or other publisher’s) nesting that we cannot handle with <named-content>.
- Jeff:
    - NLM CITATION — Determine if the order is important in the NLM citation, or if a large element OR group would work with current processing. Take a stab at writing up the distinction we need to make between our citation types.
    - TABLE – Capturing the whole table in a comment. Track back and find out who and why. What problem were they trying to solve?
- NCBI:
    - Public Posting — Take a more active role in making tag set changes and news known to the public. As part of the new NLM DTD website, post the minutes of the Working Group meetings. Post information concerning upcoming tag set changes. In particular, when the 3.0 changes are decided, post them and ask for comment emails to be sent to the Secretariat. (The request was also made that the change items being considered for 3.0 also be posted. No decision was made on this.)
    - Validation (suggestion) — Write (and post on the website as a user tool) examples of validation outside the DTD or schema, such as Schematron rules. For example, make a series of validations for different citation types that people could copy and modify for their own systems.
- Secretariat:
    - Get back to the person with the keyword request for a sample and details.
    - Get permission from the coded keyword person to use a their sample of their keywords.
    - Interview Evan for the survey.
    - Get back to the user with the volume number issue and get details.

- Produce a list of 3.0 discussion requests from the ones we tabled until after the survey, the ones in this document, and any that have come in on the comment form.

3. Add to the 3.0 Discussion List (if not already there)

- AFFILIATIONS — By design can be in several places, grouped with individual authors or with groups of authors, etc. Do we want to indicate a “best practices” preference? Is this a documentation problem where we need to show both styles and explain the differences?
- CITATIONS — Are obviously a problem. Many users are trying to use the “NLM citation” model because they perceive it to be “better”, with NLM’s approval upon it.
    1. We need to counter that perception.
    2. We may need to rename the NLM citation model.
    3. The suggestion has been made that we have at least two citation models: a) the loose #PCDATA one that we already have, b) an all-element all the time one that has all the reference elements but NO #PCDATA, c) the current NCBI one to provide an order for those who have no preset opinion.
    4. We need to document the distinctions between our models and give multiple examples (at least a journal and a book) in each format.
- EDITION — Does <edition> need face markup? Why? Anything else?
- FORMATTING AND VISUAL CONTROL — Now is the time. We have been ducking the issue. How much? How little? Customizations or not? We need to come up with a general principle so that we have an answer when we are getting nickel-and-dimed to death.
- FUNDING AGENCY — We need a real model for.
- ISSN — <issn> is required in Blue, a fact that drives some to Green who might like the restrictions in Blue. Should we make it optional?
- ISBN — <isbn> needs to be added to the metadata for articles. Should be multiple with an type attribute to identify the reason for the ISBN
- JOURNAL PARTS — Issues of journals are divided into parts. Need to be able to record that.
- KEYWORDS — Keywords are not always the simple #PCDATA elements that we have modeled. Some keywords some with an attached code:
  863   Icelandic sagas
  Some come as a major heading paired with a minor heading. (Portico to provide sample.) Elsevier supports nested keywords, not compound but two tiers. (Portico to provide sample.) Evan (and others) have also seen “string keywords” where many keywords are placed in to a single wrapper that would have to be parsed out to assemble individual keywords.

- MULTIPLE VERSIONS
    - GRAPHICS — There has always been grumbling about the inadequacy of the “alternate-form-of” attribute. A cleaner solution might be to make a graphics wrapper that can contain multiple resolutions or other forms of a graphic, potentially with attributes for describing them.
    - EQUATIONS — Like multiple forms of the same graphic, there has always been grumbling about the connection between an equation and a picture of the same equation. Similar solution?
    - WRAPPER — The multiple graphic and equations variations inside a single wrapper may be part of a larger question of how to map the text of an article to a manifest of its components. We did not include a manifest DTD, although other tag sets have done so.
- NAMED CONTENT

    1) Works well with single-level objects that can be tagged as name-value pairs. It would not work for RDF-style triples (subject/object/relationship) or for things with a more complex hierarchical structure. Is this a problem we want to solve? can solve?

    2) We agree that <named-content> is useful to describe semantic distinctions (gene, peptide, genus-species). Should it also be used to describe production artifact distinctions? (print versus online, landscape tables, etc.)
- PUB-DATE — Publication Date — is kept separate from <history>, which is a series of <date>s that record received, accepted, etc. Is that still a good idea? If several of the dates in history are print publication and others are online publication dates, is that a problem? How does cover date fit into the pub-date/history issue?
- ROMAN — The emphasis element that would not die. Wanted again for Green (Portico to provide example). Should it be added to Blue as well? What problem is it solving? Inside math or elsewhere?
- UNIT OF MEASURE — Do we need one as a floating element?
- SIGNATURE BLOCK — What is it? Why is it there? What do we expect from it? After we decide that, we should say what it needs inside it.
- SUPERTITLE — (aka eyebrow) An example is the regular column title (“From the Castle” in the Smithsonian magazine) that prints above the title of the article (“Washington, Go Fly a Kite!”). Some extension DTDs have modeled this as a new element and some have use the article categories to accomplish this. Some have built elaborate article category hierarchies, where there is a part of the publication, which then as sections, which then have supertitles). Can this be solved in the documentation with more complex examples (review, reviews). At the very least, do we need to document the difference between the “article-type” attribute and subject categories or series titles?
- X-tags — Are really not allowed in enough places for Green. How about

everywhere text is allowed (%all.phrase;) and inside <given-names> as a start. More?

4. Add to the Documentation

- ISSUE — More than one issue number (issue 1-2) is still a single <issue> element. Provide an example.
- MULTI-PART SURNAMES — Provide a few examples of multi-part surnames (without hyphens): Andrew Lloyd Webber, David Ben Gurion, one Spanish example.
- PREFIX — Fix the “Prince Charles” example to use <prefix>. His last name is “Wales”, do we care? Find another ambiguous example.

5. Observations Worth Noting

- DTDs are still the thing. Nearly everybody is still using the DTD, even the few who have software that needs an XSD or RELAX NG for some purpose. PubMed Central has had their first submission from someone who used the XSD schema.
- Most tagging books are using the Publishing DTD, not the Book DTD, implying a need for a generic book DTD.