Creation of File Formats

Tommie Usdin
Mulberry Technologies, Inc.
btusdin@mulberrytech.com

Jeff Beck
National Center for Biotechnology Information,
National Library of Medicine
beck@ncbi.nlm.nih.gov

Creation of File Formats: Journal Articles Now, Books Soon

- What is the Tag Suite?
- NISO Standardized Markup for Journal Articles Working Group
  - Move to NISO
  - Update in Progress
- Models in the Tag Suite
- Future Work: Books

Short Vocabulary Lesson: Standards and Jargon go hand-in-hand

- acronym:
  - **JATS** — Journal Article Tag Suite
- names
  - **Tag Suite** — the complete set of structures and rules used to create all of the published tag sets. Includes a few materials available to creators of customizations that are not used in any of the published tag sets
  - **Tag Set** — a model, created from the Tag Suite, for a particular type of content
- nicknames
  - **Green** — Archiving Tag Set
  - **Blue** — Publishing Tag Set
  - **Pumpkin** — Authoring Tag Set
  - **Purple** — NLM Book Tag Set

What is the Tag Suite — and where did it come from?

History of Journal Tag Sets

- First SGML tag set ever written (AAP) was for journal articles (later ISO 12083)
- Journal publishers were early adopter of SGML, then XML
  - Many used AAP/ISO-12083
  - Big players wrote their own tag sets (as a DTD)
  - Aggregators, conversion shops, and technical services also wrote tag sets
- The problems
  - numerous mine-to-yours conversions
  - conversion vendors tooled up for hundreds of tag sets
  - frustration all around
    - electronic archives struggled
    - many libraries panicked

History of the NLM Tag Sets

- 2000-2002 — NLM's PubMed Central wrote and used a tag set that focused on online display. As PMC's mission changed from simple access and display to take on an archiving function, the pmc-1.dtd was reviewed and revised.
- 2001 Harvard E-Journal Archive DTD Feasibility Study
  - http://www.diglib.org/preserve/hadtdfs.pdf
  - Was it feasible to write one domain-neutral tag set for all journal content?
- Idea behind the Tag Sets: NLM + Harvard/Mellon (as per the Harvard study)
  - Write one tag set
  - Preserve intellectual content of XML / SGML journal material (not the look and feel)
  - Make it easy for publishers / archives to transform documents
    - from their XML or SGML
    - to a standardized XML (for interchange and archiving)
- 2003 — Tag Suite Version 1.0 released with Journal Archiving and Interchange and Journal Publishing Tag Sets

Design Informed By Analysis of

The union of current journal practice

- The revised PubMed Central tag set
- Over 35 existing journal models
  - publishers, archives, aggregators
  - all DTDs but one (an XSD Schema)
- Previous generic journal models (AAP, ISO 12083)
- Hundreds of journals in many disciplines

Keeping the Tag Sets Useable and Relevant

- DTDs and schemas available online
- Tag Set Documentation and FAQ available online
- Online form to solicit user feedback
- International Working Group Created
  - representatives from archives, publishing, software industry
  - recommend changes / additions to tagset

Scope of the Tag Sets

- Journal article content (not just life sciences)
  - research articles
  - review articles
  - editorials, columns, essays, and features
  - book and product reviews
  - letters and errata
- Deliberately out of scope
  - full journals (This is an article model)
  - journal administrative content (TOC, masthead, etc.)
  - format-specific look and feel elements
  - display and classified advertising
  - automated testing (CLE, CME, Q&A)
  - magazine content

Tag Sets and Tag Suite

- Tag Suite
  - defines all elements and attributes
  - intended for journal articles
  - basis for tag sets to be built
- Tag Sets
  - DTDs and schemas made from components of the Tag Suite
  - Three Tag Sets published
    - Journal Article Archiving (very loose)
    - Journal Article Publishing (tighter)
    - Journal Article Authoring (tight for single article authoring so that the author can concentrate on the content and not which structure to use)

Tag Sets Currently Distributed

Tag Set	How to Be Used
Archiving and Interchange (Green)	- Base tag set for XML repositories - Translation target from other tag sets (capture many structures and semantics conveniently) - Common format for interchange of XML between publishers, archives, aggregators, service vendors
Publishing (Blue)	- Common format for the conversion of journal content into XML / Publishing from XML - For archives/publishers that wish to regularize and control their content
Authoring (Pumpkin)	Creation of a single article by an individual (no journal or issue metadata)

Characteristics of Archiving Tag Set

Enable archive to capture structure and semantics of existing material

- Descriptive (tag what is there)
- Non-enforcing
  - almost nothing required
  - inclusive (preserve as much tagging as possible)
  - very little required sequence (metadata in order, little else)
  - many large OR groups (do anything here)
  - capture even very bad practice
- Multiple approaches to common structures supported (there is no “right way”)

Characteristics of the Publishing Tag Set

Enable archive to regularize and control content of full articles or article metadata while retaining semantic information

- Differences from Archiving
  - smaller (not as many elements)
  - concentrates on representing the article content rather than strictly representing existing structures.
  - prescriptive and enforcing
    - not as many choices
    - more required elements (e.g., <issn>)
    - more sequences that were OR groups in Archiving
    - usually one way to do many things (not many, like Archiving)
    - leans toward best practice

NCBI Mulberry Technologies, Inc.

NISO Standardized Markup for Journal Articles Working Group

Move to NISO

- What has moved to NISO?
  - Tag Suite as a whole
  - Three Journal Article models
  - NLM Tag Set Advisory Board became NISO Standardized Markup for Journal Articles Working Group
- What has NOT moved to NISO?
  - NLM Book model
  - NLM Historical Book model
- Why move to NISO?
  - The Tag Sets have become de facto standards but some users were interested in something more formal (notably the Library of Congress and the British Library)
  - NLM is very interested in using and supporting information standards as a general practice
  - The formalized standard process should help reliably maintain the standard in the interests of the journal publishing and archiving communities beyond our personal involvement.

NCBI Mulberry Technologies, Inc.

Tag Suite Update in Progress: Why Update in 2010?

- Previous release (version 3.0) was in November 2008
- Many comments and suggestions have been received since then
- Wanted to start NISO process with best product possible

NCBI Mulberry Technologies, Inc.

What to expect in this update

- Fully backward compatible with Version 3.0 of November 2008
- Improved models for:
  - multi-lingual documents
  - creating accessible documents
  - customizing the tag sets
  - improvements and extensions to accommodate user requests
- Improved documentation:
  - creating accessible documents, including Section 508 compliance
  - clarification in response to user requests

NCBI Mulberry Technologies, Inc.

Structure of the upcoming standard

Warning: This is Prediction - this is uncertain

- Normative
- Non-normative
- Supporting resources

NCBI Mulberry Technologies, Inc.

Normative

Some information currently available will become normative in the new standard:

- Element identifiers (tags) and names
- Element definitions
- Element models
- Attribute identifiers
- Attribute definitions
- Attribute values, restrictions, and types
- DTD versions of the models

Non-normative parts of the Standard

- Recommended uses and common tagging practices
- Remarks and Related elements
- Examples of tagged documents
- Suggested uses of attributes
- W3C XML Schema (XSD) versions of the models
- RELAX NG (RNG) versions of the models

Supporting Resources

- Tag Libraries in current form will be maintained at NLM
- Starter stylesheets to create HTML and PDF via XSL-FO will be maintained
- Existing Tag Set discussion lists will be merged and maintained at NLM
- JATS-Con: The Journal Article Tag Suite Conference, November 1-2, 2010

Future Work: Books

Current Book Model

- specific to NLM Bookshelf
- not moving to NISO

Warning: Personal Opinions Follow

- This is what these authors think and/or hope.
- This is not NISO policy
- This is not Working Group concensus
- This is not based on rigorous study or extensive surveys

Tag Suite users need a Book model

- many also publish books
- want books in same databases/systems as articles
- same staff and vendors create books as articles

Tag Suite Book model should be

- identical to article when possible
- compatible with article when not identical
- familiar and comfortable to users of article models

Books differ from Journal Articles

- more varied structures (especially above the “chapter” level)
- different metadata
- larger
- often more complex

Limited scope for Tag Suite Book Model

- Journal Article models don’t model all periodicals
  - *Does model*: scientific, technical, scholarly journal articles and related documents
  - *Does NOT model*: magazines, catalogs, TV guide, theater programs
- Book model will succeed if limited scope
  - *Should model*: scientific, technical, scholarly books, technical reports and related documents
  - *Should NOT model*: cook books, trade books, poetry, graphic novels, grade school texts, and many other important (but different) types of books

Where to Look for Information

- at National Library of Medicine (current and all previous versions)
  - The home page for the Tagset (the Suite) and the Archiving DTD: http://dtd.nlm.nih.gov
  - Journal Publishing DTD: http://dtd.nlm.nih.gov/publishing/
  - The FAQ: http://dtd.nlm.nih.gov/faq.html
  - Online form for Comments/Suggestions: http://www.mulberrytech.com/DTD-Comment/CommentForm.html
  - Archiving DTD Documentation: http://dtd.nlm.nih.gov/archiving/tag-library/3.0/index.html
  - Publishing DTD Documentation: http://dtd.nlm.nih.gov/publishing/tag-library/3.0/index.html
  - Authoring DTD Documentation: http://dtd.nlm.nih.gov/articleauthoring/tag-library/3.0/index.html
  - PubMed Central: http://www.ncbi.nlm.nih.gov/pmc/
  - Discussion list for Archiving DTD: http://www.ncbi.nlm.nih.gov/mailman/listinfo/archive-dtd
  - Discussion list for Publishing DTD: http://www.ncbi.nlm.nih.gov/mailman/listinfo/publishing-dtd
  - In the Future: all supplementary and non-normative and user support.
- at NISO
  - Working Group Charter and materials
  - In the Future: the standard, including all normative materials

Where to Look for Information

- at National Library of Medicine (current and all previous versions)
  - The home page for the Tagset (the Suite) and the Archiving DTD: http://dtd.nlm.nih.gov
  - Journal Publishing DTD: http://dtd.nlm.nih.gov/publishing/
  - The FAQ: http://dtd.nlm.nih.gov/faq.html
  - Online form for Comments/Suggestions: http://www.mulberrytech.com/DTD-Comment/CommentForm.html
  - Archiving DTD Documentation: http://dtd.nlm.nih.gov/archiving/tag-library/3.0/index.html
  - Publishing DTD Documentation: http://dtd.nlm.nih.gov/publishing/tag-library/3.0/index.html
  - Authoring DTD Documentation: http://dtd.nlm.nih.gov/articleauthoring/tag-library/3.0/index.html
  - PubMed Central: http://www.ncbi.nlm.nih.gov/pmc/
  - Discussion list for Archiving DTD: http://www.ncbi.nlm.nih.gov/mailman/listinfo/archive-dtd
  - Discussion list for Publishing DTD: http://www.ncbi.nlm.nih.gov/mailman/listinfo/publishing-dtd
  - In the Future: all supplementary and non-normative and user support.
- at NISO
  - Working Group Charter and materials
  - In the Future: the standard, including all normative materials

NCBI

Mulberry Technologies, Inc.