The National Library of Medicine Tag Suite for Journal Articles: Taking Over the World of XML Journal Publishing

Deborah Lapeyre
Mulberry Technologies, Inc.
17 West Jefferson Street, Suite 207
Rockville, MD 20850
Phone: 301/315-9631
dalapeyre@mulberrytech.com

Version 1.0 (September 2009)

Mulberry
Technologies, Inc.

The National Library of Medicine Tag Suite for Journal Articles: Taking Over the World of XML Journal Publishing

The National Library of Medicine
Tag Suite for Journal Articles:
Taking Over the World of XML Journal Publishing

The National Library of Medicine Tag Suite for Journal Articles

- What
- Why
- How

Tag Sets for Public Use

- Tag Set of XML elements and attributes to describe
  - journal articles (not just medical)
  - non-article journal material such as editorials, book reviews, letters, features, etc.
- Made available for free use and modification
  - by National Library of Medicine (NLM)
  - in 2003 (latest update Nov 2008)
- By mid-2009, the Tag Set has become the *de facto* standard for journal articles in XML

(Book model is different with different adoption pattern.)

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 own tag set (as a DTD)
  - Aggregators/conversion shops/technical services also wrote tag sets
- The problems
  - numerous mine-to-yours conversions
  - conversion vendors tooled up for 100s of tag sets
  - frustration all around
    - electronic archives struggled
    - many libraries panicked

History of the NLM Tag Sets

- 1999-2001 — NLM's PubMed Central wrote/used an archival tag set
- 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)

Design Informed By Analysis of

The union of current journal practice

- Original PubMed Central (PMC) 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 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
- Two listserves
- Secretariat does update and maintenance

  (Mulberry Technologies, Inc.) (See last slide for URLs)

What and Why of the Tag Sets

Purpose of the Tag Sets is to

- Act as a *transformation target*
- Not set best practice, capture current tagging practice
    - not all elements every used in journals (80/20)
    - built-in escape hatches
    - a few 1-big-publisher-uses-it constructs
- Easy to customize

Purpose of the Tagged XML

- Archiving journal articles
- Interchanging article content
- Publishing articles (print, web, and beyond)
- Creating new journal articles
- Describing NLM Bookshelf books

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)
  - format-specific look and feel elements
  - journal administrative content (TOC, masthead, etc.)
  - display and classified advertising
  - automated testing (CLE, CME, Q&A)

Tag Sets and Tag Suite

- Tag Suite
  - predefined elements and attributes
  - intended for journal articles and some types of books
  - basis for tag sets to be built
- Tag Sets
  - DTDs and schemas made from components of the Tag Suite
  - NLM distributes four Tag Sets
    - Journal article archiving (very loose)
    - Journal Publishing (tighter)
    - Journal Authoring (tight for 1-man authoring)
    - NCBI Bookshelf for books
    - other archives.publishers/service providers have made more

Four 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)
Book (Purple)	Support for the books of the NLM Bookshelf publications

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)
  - prescriptive and enforcing
    - not as many choices
    - more required elements (e.g., `<issn>`)
    - more sequences that were `OR` groups in Green
    - usually one way to do many things (not many, like Green)
    - leans toward best practice

Characteristics of the Authoring Tag Set

Enable authors to create new articles as simply as possible

- The simplest version of the Tag Sets (restricted subset)
- No journal or issue metadata (this is not yet published)
- Almost no control over formatting
  (journal that accepts article will have own style)
- Differences from Publishing
  - even smaller (not as many elements)
  - very prescriptive (not as many choices)
  - enforcing (there is one way to do many things)

Characteristics of the Book Tag Set

Supports the NCBI online libraries

- Not a generic tag set like the journal ones
  (although it gets used as one!)
- Not everything in all books, just what Bookshelf books needed
- Written for
  - converting existing books into XML
  - a particular processing architecture
- Differences from the journal sets
  - top-level is `<book>` not `<article>`
  - no journal metadata tagging
  - some processing-specific tagging

Journal Article Tag Sets Adopted Widely

Unexpected (but welcome) success because

- Was easy to understand/adopt
  - looked a lot like existing journal tag sets
  - cognates for most current patterns
- Hit the sweet spot (85-95% of what you want is already there)
- Provided hooks to add the rest
  - parameter entities everywhere
  - encouraged local customization/differentiation

Design Features That Aided Adoption Include

- Modeled *current* documents and process
- Descriptive not prescriptive
  - very little required, but much possible
  - your current order usually works
- Enabled inclusion of standard vocabularies (MathML, HTML tables, CALS tables)
- Was a DTD (not a schema)
- *Was documented!*

Reasons Users Give for Adoption

- We will *not* write our own tag set
  (on time, under budget)
- Familiarity of conversion, hosts, aggregators, archives is key
- We need
  - rich metadata
  - multiple taxonomy support
- Vendor-independence is good
- Flexibility is nice, too
- We want to submit to PubMed Central

(Nice talk summarizing why IMechE adopted NLM at
`www.alpsp.org/ForceDownload.asp?id=608`

Pattern of Adoption

- Repositories and Aggregators standardize on conversion target
- Web host, fulfillment service, or compositor supports and/or requires
- Publishers
  - replace proprietary tag sets
  - replace ISO 12083 and AAP
  - wanted XML but did not have a tag set
- Recent XML adopters prefer public
- SGML publishers take the opportunity of moving to XML to get out of the tag set business
- Conversion vendors push it, makes their processing easier
- Buzz on the street
  - better than TEI for production
  - better than DocBook for journals (references in particular)
  - better than DITA in journal focus

Who *is* Using the Tag Sets

- National Library Archives
- Commercial Archives
- Aggregators/Technical Service Providers
- Commercial Publishers
- Society Publishers
- Conversion vendors and tagging services know it and work with
- Software companies provide support (Styllus Studios, Inera, Microsoft, et al.)

Archives

- PubMed Central (PMC) of National Library of Medicine (NLM)
- JSTOR (Ithaka/JSTOR/Portico)
- Library of Congress
- British National Library

PubMed Central (PMC)

- NLM Archives for biomedical and life sciences journals
- Mandated for government-sponsored biomedical research
- More than 2,650, 000 Publications (March 2009)
- Hundreds of journals, such as
  - *American Journal of Human Genetics*
  - *Molecular Biology of the Cell*
  - *Proceedings of the National Academy of Sciences of the U.S.A.*
  - *The EMBO Journal*
  - All of the *BioMed Central (BMC)* Journals
  - American Society of Microbiology Journals

Ithaka / JSTOR The Scholarly Journal Archive

(Ithaka/JSTOR/Portico)

- JSTOR is an independent not-for-profit organization building a scholarly digital archive
- Archiving full journal issues
- Archive has more content than ... anybody
  - over 750 million articles (as of Jan 2009)
  - over 7300 academic journals
    - humanities
    - social sciences
    - sciences
  - More than 5,200 academic and other institutions
  - Over 600 learned societies, university presses, and other content contributor

Commercial Publishers (Large and Small)

including...

- Commonwealth Scientific and Industrial Research Organization (CSIRO)*
- Haworth Press
- Lippincott Williams and Wilkins
- IGI Global
- Oxford University Press Oxford Journals
- Thieme
- The University of Chicago Press
- Blackwell Publishing
- St. James Publishing (*Journal of Burns & Surgical Wound Care*)

(Some publishers with custom tag sets can deliver in NLM)

*First outside adopter

Society Publishers (Large and Small)

including...

- Public Library of Science (*PLoS Biology* & *PLoS Medicine*)
- American Institute of Biological Sciences (BioOne)
- American Association for the Advancement of Science
- National Academy of Sciences
- Association of Computing Machinery
- American Mathematical Society
- Optical Society of America
- CFA Institute (The Global Association of Investment Professionals)
- Institute of Mechanical Engineers
- Society for Microbiology (UK)
- National Athletic Trainers' Association (*Journal of Athletic Training*)

Other "Publishers"

including...

- World Health Organization
- World Bank
- Textbook publishers
- Medical publishers
- Equipment manufacturers
- Government agencies

Aggregators/Technical Service Providers

including...

- Ingenta, a division of Publishing Technology plc
- Highwire Press Library of the Sciences and Medicine
- Jouve (FR)
- Atypon

Most Big Compositors and Conversion Shops

(and many small ones)

- Claim NLM conversion experience (typically from Microsoft Word files)
- Can deliver tagged content in
- Can provide references from happy clients
- Know the tag sets, so they promote them (sometimes inappropriately)
- Offer InDesign or Quark pages made from NLM XML

Future of the Tag Sets

What Resources are Available on the Web?

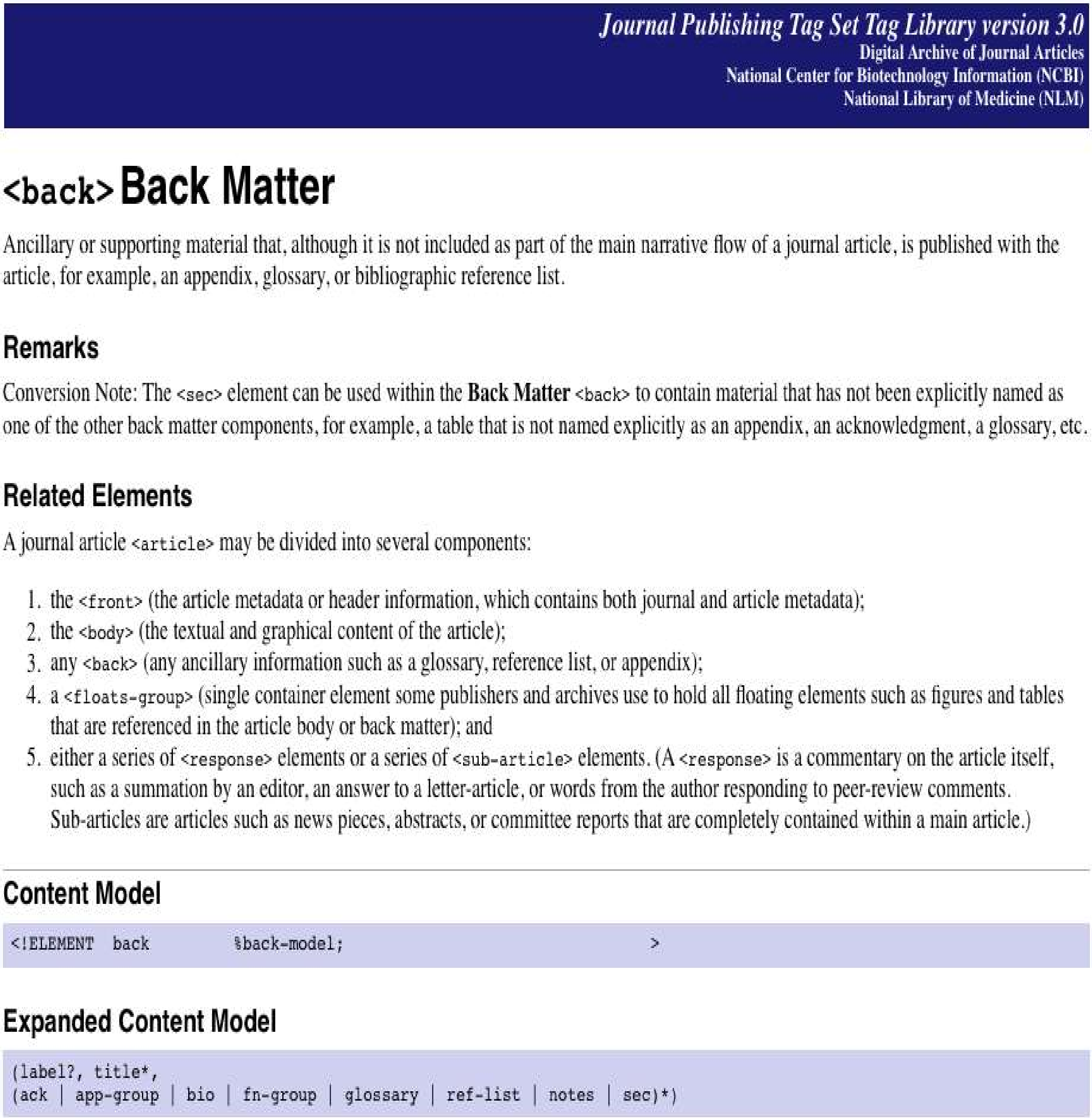
- Splash page for all 4 public Tag Sets
  `http://dtd.nlm.nih.gov/`
- Tag Library User Documentation (one per Tag Set)
  for example, `http://dtd.nlm.nih.gov/publishing/`
- FAQs
- Usable schemas
  - DTD (primary version)
  - XSD Schema
  - RELAX NG Schema
  - Tag set from which to build other tag sets
- Tagged samples
- Email feedback form
- Tools for transforming / using the XML

Tag Library: The Tag Set Documentation

(help for the user)

- Each Tag Set has its own Tag Library
- HTML available online
  - Element pages (one per element)
  - Attributes pages (one per attribute)
  - Context table (where can an element be used)
  - Hierarchy diagrams (tree structures for elements)
  - Essays on common tagging practice
  - Report on what is new or changed in Version 3.0
  - Index (with synonyms and use for terms)
  - Implementors notes

Sample Tag Library Page (screen shot page 1)



Journal Publishing Tag Set Tag Library version 3.0
Digital Archive of Journal Articles
National Center for Biotechnology Information (NCBI)
National Library of Medicine (NLM)

`<back>` Back Matter

Ancillary or supporting material that, although it is not included as part of the main narrative flow of a journal article, is published with the article, for example, an appendix, glossary, or bibliographic reference list.

Remarks

Conversion Note: The `<sec>` element can be used within the **Back Matter** `<back>` to contain material that has not been explicitly named as one of the other back matter components, for example, a table that is not named explicitly as an appendix, an acknowledgment, a glossary, etc.

Related Elements

A journal article `<article>` may be divided into several components:

1. the `<front>` (the article metadata or header information, which contains both journal and article metadata);
2. the `<body>` (the textual and graphical content of the article);
3. any `<back>` (any ancillary information such as a glossary, reference list, or appendix);
4. a `<floats-group>` (single container element some publishers and archives use to hold all floating elements such as figures and tables that are referenced in the article body or back matter); and
5. either a series of `<response>` elements or a series of `<sub-article>` elements. (A `<response>` is a commentary on the article itself, such as a summation by an editor, an answer to a letter-article, or words from the author responding to peer-review comments. Sub-articles are articles such as news pieces, abstracts, or committee reports that are completely contained within a main article.)

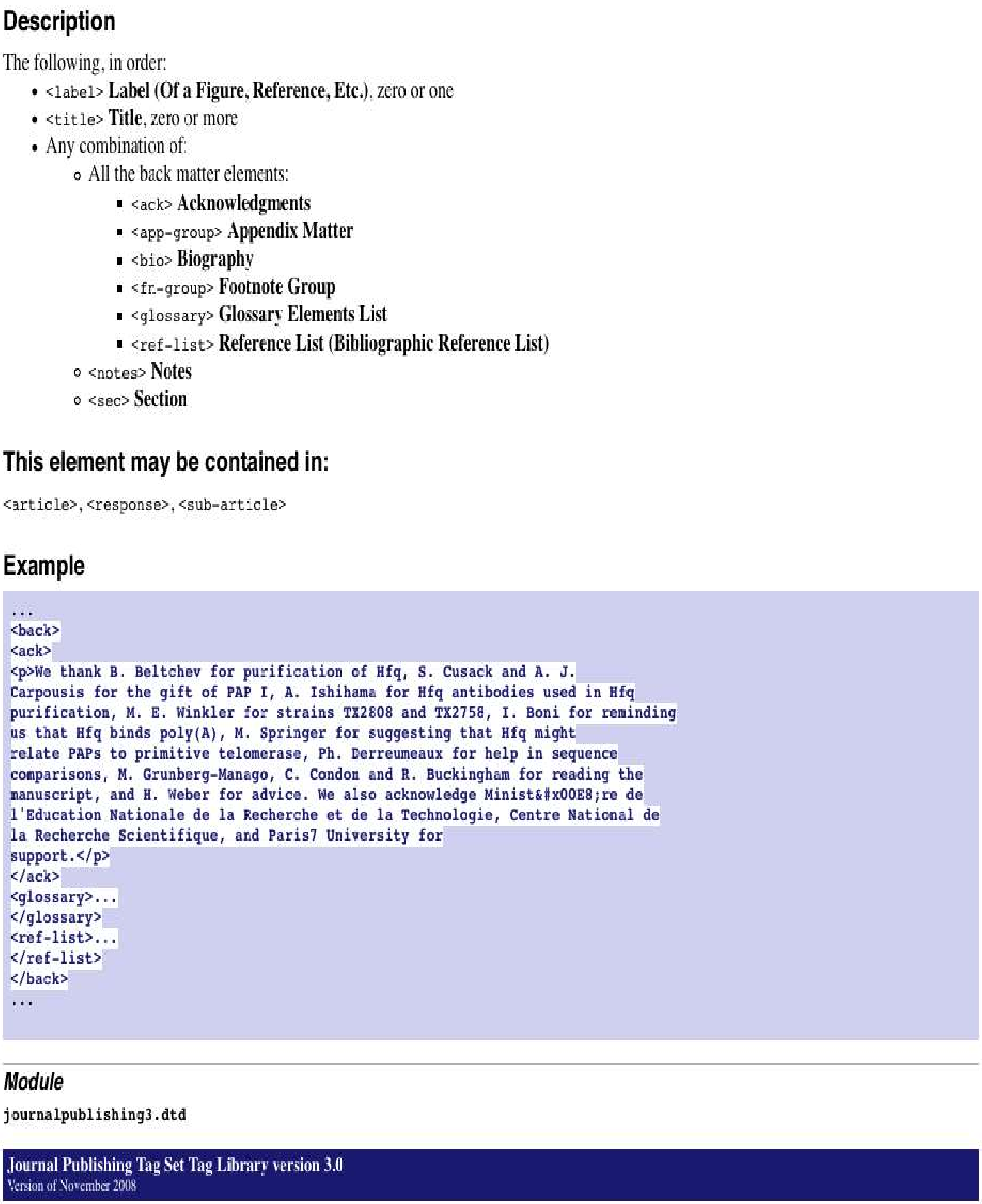
Content Model

```
<!ELEMENT  back         %back-model;                          >
```

Expanded Content Model

```
(label?, title*,
(ack | app-group | bio | fn-group | glossary | ref-list | notes | sec)*)
```

Sample Tag Library Page (screen shot page 2)



Description

The following, in order:

- `<label>` **Label (Of a Figure, Reference, Etc.)**, zero or one
- `<title>` **Title**, zero or more
- Any combination of:
  - All the back matter elements:
    - `<ack>` **Acknowledgments**
    - `<app-group>` **Appendix Matter**
    - `<bio>` **Biography**
    - `<fn-group>` **Footnote Group**
    - `<glossary>` **Glossary Elements List**
    - `<ref-list>` **Reference List (Bibliographic Reference List)**
  - `<notes>` **Notes**
  - `<sec>` **Section**

This element may be contained in:

`<article>`, `<response>`, `<sub-article>`

Example

```
...
<back>
<ack>
<p>We thank B. Beltchev for purification of Hfq, S. Cusack and A. J.
Carpousis for the gift of PAP I, A. Ishihama for Hfq antibodies used in Hfq
purification, M. E. Winkler for strains TX2808 and TX2758, I. Boni for reminding
us that Hfq binds poly(A), M. Springer for suggesting that Hfq might
relate PAPs to primitive telomerase, Ph. Derreumeaux for help in sequence
comparisons, M. Grunberg-Manago, C. Condon and R. Buckingham for reading the
manuscript, and H. Weber for advice. We also acknowledge Minist&#x00E8;re de
l'Education Nationale de la Recherche et de la Technologie, Centre National de
la Recherche Scientifique, and Paris7 University for
support.</p>
</ack>
<glossary>...
</glossary>
<ref-list>...
</ref-list>
</back>
...
```

Module

`journalpublishing3.dtd`

Journal Publishing Tag Set Tag Library version 3.0
Version of November 2008

NLM Provided Support Tools

`http://dtd.nlm.nih.gov/tools/`

- Conversion
  - from any version of the Archiving Tag Set
  - into version 3.0
- Publishing Previews (previous versions)
  - Preview XSLT (makes HTML using XSLT)
  - Preview XSL-FO (makes PDF via XSL-FO using XSLT)
- Publishing Previews for 3.0 (online soon)
  (drafts on your giveaway flashdrive *now*)
  - Preview HTML
  - Preview PDF

Other NLM-provided Tools

- PMC XML Validator
  - Validates any XML to a DTD
  - special “hooks” for NLM Journal DTDs
  - http://www.pubmedcentral.nih.gov/utils/validate/xmlcheck.cgi
- Equation preview tool
  - Make a GIF image from
    - TeX math
    - MathML tagged math
  - http://www.pubmedcentral.nih.gov/utils/mathtool/mathtool.cgi

Where to Look for Information

- The home page for the Tagset (the Suite) and the Archiving DTD: `http://dtd.nlm.nih.gov`
- 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`
- Archival DTD Documentation`http://dtd.nlm.nih.gov/tag-library/1.1/index.html`
- Publishing DTD Documentation`http://dtd.nlm.nih.gov/publishing/tag-library/1.1/index.html`
- PubMedCentral: `http://www.pubmedcentral.gov/`
- 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`