JATS, BITS, and STS

Journal Article Tag Suite
Book Interchange Tag Suite
Standards Tag Suite

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

Version 1.0 (November 2017)

JATS, BITS, and STS

JATS, BITS, and STS

Abstract

Tutorial Description

This course is targeted to people who need to make high-level decisions about JATS, BITS, and STS. If you are deciding whether, when, and how to adopt or convert to JATS, BITS, or STS, or if you want to know how they are organized and how they relate to each other, this is the class for you.

The goal of this course is to give people enough working knowledge of JATS, BITS, and STS so they can make informed business decisions and participate fully in decisions about subsetting and customizing.

Starting with a description of the original goals and current uses of JATS, BITS, and STS, we will discuss ways in which they are similar and ways in which they differ, both technically and organizationally. The key design principles are the same for all of these “cousin” tag sets: for example, they are enabling not enforcing. We will discuss the implications of these design principles for production and interchange. These tag sets are also based on the same structural principles — for example, separation of metadata from display content, nested recursive sections, and the ability to do very rich encoding of citations. Because the tag sets are all quite loose, many users find it convenient to subset the model they adopt. We will discuss the reasons for subsetting (and supersetting) the public models and methods of doing so. Finally, we will show the variety of documentation and resources available to users of JATS, BITS, and STS.

The JATS, BITS, and STS tutorial is a lecture-style course; laptops are not required.

slide 2

Who We Are

- Who am I?
  - Mulberry Technologies, Inc.
  - Tommie Usdin
- Who are You?
  - Name
  - Affiliation
  - Publishing background
  - XML background
  - JATS tag set experience

slide 3

Basic Principles

- Questions are always welcome
- Examples usually available on request
  (If you don’t see, ask)
- Based on your interests I will adjust emphasis; if I slide past something you care about, speak up

slide 4

Origins of JATS, BITS, and STS

The "begats" story

- SGML was developed and a model for books and articles was needed
- Many models were developed to fill this void,
  and all had virtues and followers
- A study was conducted to select the best, but none was anointed
- the ***NLM DTD*** was developed and widely adopted
- the NLM DTD begat ***JATS***
- JATS begat ***BITS***
- JATS begat ***ISO STS***
- ISO STS begat ***NISO STS***

slide 5

History of Desire for Shared Models

The publishing community has long wanted a “universal” model for articles, books, etc.

- AAP model developed in 1980s, then standardized as ISO 12087
- DocBook, developed for computer documentation is widely used
- DITA, developed for repurposed modular content such as user manuals is widely used

If you go back even further, SGML developed out of effort to create shared tags for typesetting books and articles

slide 6

Origins of JATS (Journal Article Tag Suite)

slide 7

Huge Collections of Journal Articles

- PubMed Central
  - US Congress declared research paid for with US $$ should be available to all
  - PubMed Central (PMC) was developed
  - NLM adopted an industry DTD -- didn't quite meet needs
  - Decided to create a model
- Others Were Contemplating Archives of Journal Articles
  - E-Journal Archival DTD Feasibility Study
  - Inera for the Harvard University E-Journal Archiving Project
  - Conclusion: one model (DTD) for all journal articles possible, but did not exist
- NLM funded development of a DTD to meet both needs:
  - Production of PMC
  - Archiving of electronic journal articles in all fields

slide 8

NLM DTD Widely Adopted

- Used to submit articles to PMC
  - converted to XML in NLM tag set after publication
  - some adopted for XML-based workflows
- Other archives and libraries adopted
- Publisher services, conversion vendors, database spinners became familiar with NLM DTD, then began to prefer, then require it

slide 9

NLM DTD Became JATS

- Many would-be users needed/wanted International Standard
- NLM gave DTD to NISO to become a Standard
- Renamed to ***JATS***, improved through committee work, now a Standard
- Very widely adopted for internal use
- Practically universal for interchange of journal articles

“JATS is no longer one of the cool kids, it’s just what you do if you have journals.” - Jeff Beck(PMC)

slide 10

Origins of BITS (Book Interchange Tag Suite)

slide 11

Demand for JATS-compatible Book Model

- JATS users, journal publishers, also publish books
  - want to use familiar (JATS) tools for books
  - want to mix books and articles in databases and presentation systems
  - often use articles as book content (e.g., chapter or section in chapter)
- Old NLM Bookshelf model not JATS-like
- Old NLM Bookshelf model rarely used outside NLM

slide 12

BITS Developed to Meet That Need

- Book-specific metadata
- Based on JATS Green (archiving)
- More flexible than JATS because more variety in books than articles
- Tools to accommodate large documents (XInclude)
- Structures for Table of Contents and Index
- Supports cut & paste from JATS
  a JATS <article> can become a BITS <book-part> with a few tweaks
  - some metadata changes needed
  - one structural change to appendices may be needed
- Management of large books (in multiple files)
- Collections of books, e.g., series

slide 13

Origins of STS (Standards Tag Suite)

slide 14

ISO Improving Internal Production

- ISO needed to reduce cost & time to produce/publish standards
- Older ISO processes
  - word- processor based
  - slow, error filled
  - publication not for months/years after standard completed
  - electronic versions (other than PDF) expensive, error prone

slide 15

ISO STS Developed for ISO Internal Use

- Studied many XML-based options
- Did TEI-based prototype
- Considered DocBook, DITA, JATS
- Selected JATS as base
  - replaced journal metadata with standards-descriptions & local tracking
  - added standards-specific structures
    (e.g., Notes, Examples, TBX term and definition model)
  - did not remove *anything* from JATS

slide 16

Original ISO STS

- Considered to be internal tool
  - no consensus-based process to develop
  - fully modeled structures ISO uses
  - Provided basic metadata for other users
    - <reg-meta> Regional-body Metadata
    - <nat-meta> National-body Metadata
- Made public, but not as a standard

slide 17

NISO STS is Based on ISO STS

- Participants included people from many types of standards organizations
- Added structures used by variety of standards organizations (Normative Notes and Examples, Adoptions)
- Added book-like structures (Table of Contents, Index) from BITS
- Made metadata richer, more flexible, optional

slide 18

The JATS Family Timeline **(optional)**

2003	NLM DTD made public
2008	Last NLM DTD 3.0
2012	JATS 1.0 released
2012	BITS released
2012	ISO STS 1.0 (current is ISO STS 1.1)
2016	BITS 2.0 (current)
2017	JATS 1.2d1 (current Committee Draft)
2017	NISO STS 1.0 (current)

slide 19

Differences Between JATS, BITS, and STS Models

- Fewer differences than commonalities
- Administrative/ownership
- Support
- Models

slide 20

Forces for Alignment and Divergence

- Tag Sets that are in use must change
  - better meet original requirements
  - meet changing needs
  - accommodate changing environment
- Users of multiple tag sets want them to evolve in synchrony
- Users of each tag set want it to evolve meet their specific needs

slide 21

Overlapping User Groups

- Many people/organizations use 2 or more of JATS family
- Overlapping members of JATS, BITS, and STS committees
- Same maintainers and documenters

slide 22

Documented Suggestions on Creating JATS-compatible Tag Sets

- Public tag sets should be designed to outlive individual involvement
- Families of tag sets should be designed to grow
- JATS Compatibility Model Description
  http://www.niso.org/apps/group_public/download.php/16764/JATS-Compatibility-Model-v0-7.pdf

slide 23

Administration Ownership/Support

slide 24

JATS and NISO STS are ANSI/NISO Standards

- Published as Standards by NISO
- Maintained by NISO committees using ANSI/NISO procedures
  - consensus based
  - public
  - standard and revisions are approved by NISO and ANSI

There are “Standards Documents” for each

- *Do not* use the official Standards document for tagging/working with the XML documents.
- Use the Tag Libraries and other non-normative materials, discussed later.

slide 25

JATS and STS Support

- Are “continuous maintenance” standards
- Users request changes/additions on NISO website forms
- Standing Committees at NISO debate and decide
- New versions of standards must be voted
  - committee draft releases
  - non-normative documentation can be revised at any time

slide 26

BITS is an NLM Project

- Sponsored and administered by NLM (National Library of Medicine)
- Committee is advisory to NLM
- No equivalent of the “Standards Documents”
- Documentation is equivalent to JATS & STS

slide 27

BITS Support

- NLM-sponsored Working Group
- List-serve collects requests

slide 28

Different Numbers of Models

slide 29

JATS has 3 Models

- Archiving (Green)
- Publishing (Blue)
- Authoring (Pumpkin)

slide 30

Journal Archiving and Interchange

'Archiving' or Green

- Most flexible of the JATS models
- Optimized as conversion target for existing documents
- Designed to retain *any information* in source documents
- Used by libraries and archives that must accept variety of materials from variety of sources

slide 31

Journal Publishing

'Publishing' or Blue

- Designed for full published articles
- Designed to enable editing and interchange of complete articles
- Imposes rules to reduce variation in files and make XML tractable/processable
- More restrictive than Archiving model

slide 32

Article Authoring

'Authoring' or Pumpkin

- Designed for creating new content
- Allows as few tagging options as possible
- Models full content of article body
- Metadata for author information but not publisher metadata or publication information

slide 33

BITS has 1 Model

- Based on JATS Archiving (Green)
- 2 top-level structures
  - book
  - book-part-wrapper
    to hold chapters, parts, sections, etc.

slide 34

STS

- 2 top-level structures
  - <standard>
  - <adoption>
- 2 Models
  - Interchange
    HTML-based tables
  - Extended
    OASIS/CALS & HTML-based tables

All Models available in several forms

Machine-readable versions available in many forms

- Grammars
  - DTDs - Document Type Definition
    the document grammar language adopted from SGML
  - XSD - W3C XML Schema
    a document grammar language developed by the W3C, popular for more structured (non-prose) documents
  - RNG - RelaxNG
    powerful but not widely implemented document grammar

  use one or all; all represent the same model
- Table Models
  - HTML-based and easily converted to HTML
  - OASIS/CALS model & HTML-based
- MathML versions
  - MathML2
  - MathML3

  Use only one MathML model; do not mix.

slide 36

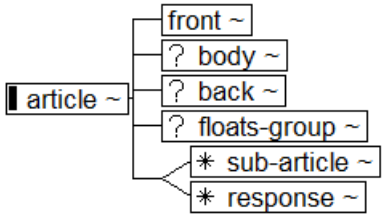
Similarities & Differences Between Suites

- Differences
  - top level structure(s)
  - metadata
  - some document-type specific body structures
- Similarities
  - text/prose models
  - separation of display content from metadata
  - reference model

slide 37

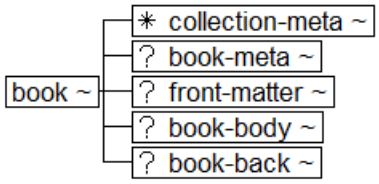
JATS Top-level Structure

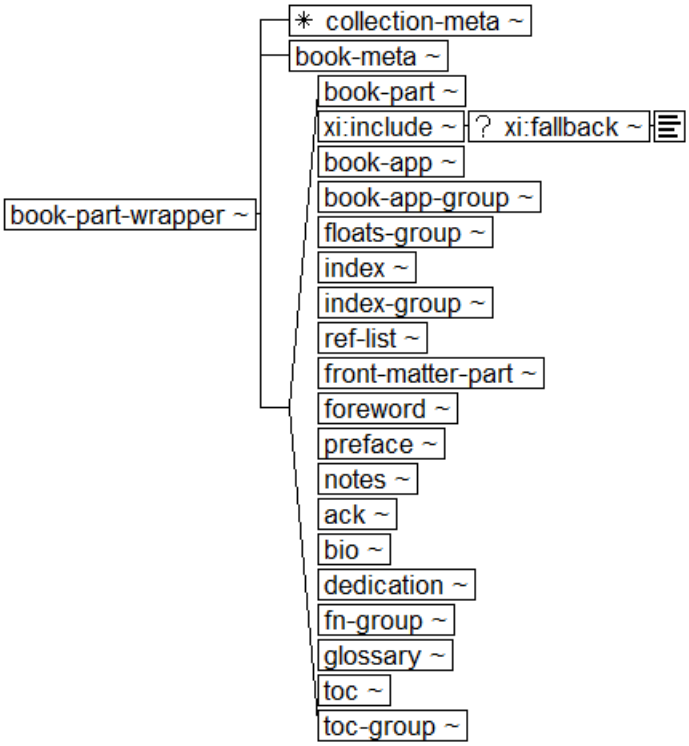
(<article>)



Two BITS Top-level Structures

(<book> and <book-part-wrapper>)

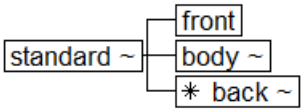


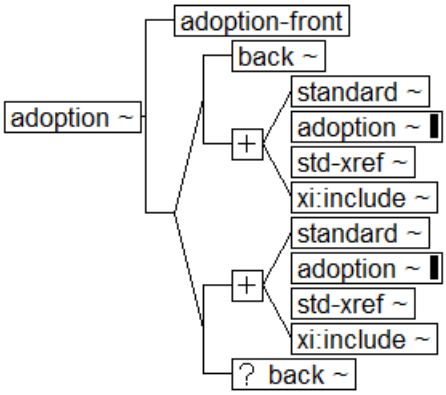


slide 39

Two STS Top-level Structures

(<standard> and <adoption>)





slide 40

JATS Metadata

- Journal metadata
  - Identification
  - Title
  - Publisher
- Article metadata
  - Identification
  - Title
  - Contributors (authors, etc.)
  - Permissions
  - Abstract(s), subject terms, keywords
  - Funding

BITS Metadata

- Collection metadata
  - Identification
  - Title of collection
  - Publisher
  - Abstract(s), subject terms, keywords
  - Role of book in collection
- Book and Book-part metadata
  - Identification
  - Title
  - Contributors (authors, etc.)
  - Permissions
  - Abstract(s), subject terms, keywords
  - Funding

slide 42

STS Metadata

- Content of the standard
  - abstract(s)
  - keywords
  - subject terms
- Identification of the Standard
  - title and parts of title
  - designators
  - wi codes
- Location in Standards life-cycle
- Standards Organization(s)
- ISO- Regional- and National-specific metadata

slide 43

Specific Document-type Structures

slide 44

Book (BITS) Specific Body Structures

- Narrative front matter (Dedication, Foreword, Preface, ...)
- Structural Table of Contents
- Recursive named book parts (volumes, books, chapters, parts, ...)
- Index
  - Structural
  - Embedded
- Questions & Answers

slide 45

Standards (STS) Specific Body Structures

- Editing instructions
- Ability to flag all content as normative or non-normative
- Terms & Definitions (2 models)
- Notes
  - Normative
  - Non-normative
- Examples
  - Normative
  - Non-normative
- Structural Table of Contents
- Index
  - Structural
  - Embedded

slide 46

Commonalities Among all JATS, BITS, and STS models

(the similarities are greater than the differences)

- Key Design Principles
- Document Structure
- Designed to be Customizable

slide 47

Key Design Principles

(designed as *interchange* tag sets)

- Models *current* documents and process
- Descriptive not prescriptive
  - very little required, but much possible
  - current order (reading sequence) usually works
- enabled inclusion of popular vocabularies (MathML, HTML tables, CALS tables)
- DTD as well as schemas
- *Documented!*

slide 48

Models *Current* Documents and Processes

- Based on analysis of journals and journal DTDs
- Accommodates variations in habits and practice
- Extended as current practice changes
- Follows, not leads (!)
- Not designed for backfile/historical documents

slide 49

Descriptive not Prescriptive

Enabling not Enforcing

- Allows very granular markup, e.g., every semantic item in citations
- Allows very chunky markup, e.g.,
  - face markup only in citations
  - no markup inside citations

slide 50

Little Required, Much Possible

- Numeration (e.g., list item numbers) allowed in XML or not
- IDs on virtually everything allowed, not required
- Detailed metadata (e.g., history) allowed, not required
- Many ways to tag the same structure

slide 51

Current Order Usually Works

- If transforming from a local XML model
- Presentation order for body content

slide 52

Popular Vocabularies Included

- MathML
- HTML tables
- OASIS/CALS tables

slide 53

Similarities in Basic Document Structure

(same in JATS, BITS, STS)

- Separate metadata from narrative (user) content
- Nested recursive sections
- Document bodies are very similar
- All 3 use same lower-level structures (blocks and inlines)
- Ability to tag at varying levels of granularity

slide 54

Block and Inline Structures

JATS, BITS, STS use *the same* text markup

- Full text and graphics of the body, including:
  - structural items (sections, paragraphs, lists)
  - figures and tables
  - content items (such as genus-species, gene)
  - typographical highlighting (bold, small caps)
  - sidebars and text boxes
  - internal pointers to figures, tables, etc.
  - external pointers to related material such as databases
- Bibliographic references
- Appendices

slide 55

Body Structures

- Paragraph-level stuff (tables, figures, etc.), followed by
- Sections (<sec> which are recursive), followed by
- Optional signature block (<sig-block>)

slide 56

Flexible Markup for Special Semantics

(One tag set can never name it all)

- Open ended elements
  - <named-content>
  - <styled-content>
  - attributes name the semantics
- Generic metadata name/value pairs (<custom-meta-wrap>)

slide 57

Resources to Support Use

- All three tag sets are heavily documented
  - Tag Libraries
  - Sample documents
- Many independent resources
  - Discussion Lists
  - JATS 4 Reuse
  - Conference proceedings
  - STS support group coming soon

Tag Libraries (in HTML and linked)

The first place to look; often the only thing needed

- Available online:
  - JATS Green https://jats.nlm.nih.gov/archiving/tag-library/
  - JATS Blue https://jats.nlm.nih.gov/publishing/tag-library/
  - JATS Pumpkin https://jats.nlm.nih.gov/articleauthoring/tag-library/
  - BITS https://jats.nlm.nih.gov/extensions/bits/tag-library/
  - STS (from) http://www.niso-sts.org/
- Downloadable copies available:
  - JATS Green ftp://ftp.ncbi.nlm.nih.gov/pub/jats/archiving/1.1/
  - JATS Blue ftp://ftp.ncbi.nlm.nih.gov/pub/jats/publishing/1.1//
  - JATS Pumpkin ftp://ftp.ncbi.nlm.nih.gov/pub/jats/articleauthoring/1.1/
  - BITS ftp://ftp.ncbi.nlm.nih.gov/pub/jats/extensions/bits/2.0/
  - STS (from) http://www.niso-sts.org/
- Demonstration/walk-through of Tag Library

slide 59

Sample Documents

- Fragments of sample documents in tag libraries
  - 1 or more on each element
  - sometimes also on attribute entries
- Complete tagged samples linked from JATS Blue tag libraries
- Tagged samples of STS on niso-sts.org (soon)
- Live JATS documents:
  - PubMedCentral open access subset: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
  - Elementa articles available to download in JATS:
    - https://www.elementascience.org/
    - find article (search, select)
    - clock "Download" in horizontal bar
    - select "XML"
  - PLOS articles available to download in JATS:
    - https://www.plos.org/
    - find article (search, select)
    - beside "Download PDF" click down arrow, select "XML"

slide 60

Discussion Lists

- JATS-List https://www.mulberrytech.com/JATS/JATS-List/
- NISO-STS-List https://www.mulberrytech.com/STS/NISO-STS-List.html

slide 61

PubMed Central Guidelines and Tools

- Describes requirements *in addition to* those expressed in the DTDs or schemas
- PMC Guidelines
  - graphics requirements
  - file naming and packaging
  - coding requirements, e.g., required content
  - http://www.pubmedcentral.nih.gov/about/PMC_Filespec.html
- PMC Style Checker
  - Automatically checks much of PMC Guidelines
  - http://www.pubmedcentral.nih.gov/utils/style_checker/stylechecker.cgi

slide 62

JATS4R (JATS for Reuse)

- Independent group of publishers working to
  - develop recommendations for tagging content in JATS XML
  - share best practice examples
  - create and share validation tools that check XML against JATS4R recommendations

Resources include:

- JATS4R online validator tool (http://jats4r.org/validator/) running Schematron to flag discrepancies
- JATS4R stored XML examples (https://github.com/JATS4R/JATS4R-Participant-Hub/tree/master/examples)

slide 63

Conference Proceedings

- JATS-Con Proceedings (Journal Article Tag Suite Conference)
  https://www.ncbi.nlm.nih.gov/books/NBK65129/
- Balisage conference proceedings
  https://www.balisage.net/Proceedings/index.html

slide 64

Decisions You Need to Make

- Choosing to use JATS/BITS/STS (and which one or ones) is one level of decision
- Even when you know you want JATS-based,
  other high-level decisions need to be made
  - subsetting & supersetting
  - enhancing the XML (or not)
  - how to tag citations
  - adopting coding guidelines
  - layers of validation
  - DTD, XSD, RNG

slide 65

Decisions: Subsetting & Supersetting

- Models designed to be customized
- How-to in Tag Libraries
- More guidance at NISO:
  JATS Compatibility Meta-Model Description
  http://www.niso.org/apps/group_public/download.php/16764/JATS-Compatibility-Model-v0-7.pdf

slide 66

Subset

- All document valid to subset valid to public model
- No need to share subset (or fact of subset) when sharing documents
- Public tools will still work

slide 67

Subset To

- Remove variation you don’t need
- Tighten loose models (require some elements!)
- Remove element & attributes you won’t use
- Decide on one way to do something, and make the others illegal
- Provide specific lists of attribute values instead of allowing anything

slide 68

Value of Subsetting

- Reduce complexity of applications
- Increase ease of use for editors
- Faster implementation of display and management tools
- Increase coherence and usability of document set
- Prohibit “creative” tagging in your database

slide 69

Supersetting

- Add metadata needed for internal purposes
- Add named structures key to your business
- Take advantage of public tools, customize instead of growing own

(Getting valid JATS back means a transform or removal of proprietary structures)

slide 70

Decisions: Enhancing the XML

- Enhancements: information in the XML that the user (typically) does not see
- The converse of Generated Text: information (e.g., labels) not in the XML but seen by the users
- Common enhancements:
  - accessibility information
  - alternative versions of graphics for print and screen use
  - Digital Object Identifiers (DOIs)
  - contributor and organization identifiers (e.g., ORCHID, ISNI, Ringgold)
  - semantic enrichment (tagging particular concepts in text, RDFa attributes, etc.)

slide 71

Accessibility Information

- Increasing requirements for accessibility information
- JATS provides the structures:
  - <alt-text> & <long-desc> elements on graphical and tabular structures
  - @alt attribute on abbreviations, labels, cross-references, as needed
  - <alternatives> for some math, audio, video, emoticons
- See chapter in JATS Tag Libraries
- *Be aware that this may require content experts*
- *Have the resources to carry through if you start*

slide 72

Multiple Versions of Graphics

- High resolution for print
- Moderate resolution for print
- Thumbnail for hand-held devices or navigation
- *If you want readers to reuse your graphics given them appropriate versions*
- *If graphics are critical to understanding your content provide good-enough versions*
- *Provide fast-loading versions of graphics so users don't wait for your content, let them ask for big/slow versions*

slide 73

Decision: Tagging Bibliographic Citations

2 ways to tag citations in JATS:

- Element Citation <element-citation>
- Mixed Citation <mixed-citation>

(don't use NLM Citation <nlm-citation>, holdover from old NLM DTD)

slide 74

Mixed versus Element Citations

Mixed Citation

- Content, punctuation, spacing in the XML
- Tag as much, or as little, of citation as you want
- Expect display to be *as tagged*
- "Copyediting citations" is responsibility of document creation/editing

Element Citation

- All content inside XML tags (no loose text or punctuation)
- Punctuation & spacing provided by display software
- Works well for expected types of citations
- Works poorly for unexpected content

slide 75

An Example Citation

from "PubMed Central Tagging Guidelines"

Petitti DB, Crooks VC, Buckwalter JG, Chiu V. Blood pressure levels before dementia. Arch Neurol. 2005 Jan;62(1):112-116.

- Display form of a citation of a journal article (in NLM style)
- It looks obscure but readers learn to parse this syntax.
- Embedded information objects useful for machine analysis
  - Author names
    - family/last/surname
    - given/first name(s)
    - initials
  - Journal title, volume, issue, date
  - Article title, page numbers

slide 76

Sample Element Citation

Tagged as elements with no free text:

```
<element-citation publication-type="journal" publication-format="print">
<name>
<surname>Petitti</surname><given-names>DB</given-names>
</name>
<name>
<surname>Crooks</surname><given-names>VC</given-names>
</name>
<name>
<surname>Buckwalter</surname><given-names>JG</given-names>
</name>
<name> <surname>Chiu</surname><given-names>V</given-names>
</name>
<article-title>Blood pressure levels before dementia</article-title>
<source>Arch Neurol</source>
<year>2005</year>
<month>Jan</month>
<volume>62</volume>
<issue>1</issue>
<fpage>112</fpage>
<lpage>116</lpage>
</element-citation>
```

slide 77

Sample Mixed Citation

Tagged with the punctuation and spacing the editors require:

```
<mixed-citation
  publication-type="journal" publication-format="print">
<string-name><surname>Petitti</surname>
 <given-names>DB</given-names></string-name>,
<string-name><surname>Crooks</surname> <given-names>VC</given-names></string-
name>,
<string-name><surname>Buckwalter</surname>
 <given-names>JG</given-names></string-name>,
<string-name><surname>Chiu</surname> <given-names>V</given-names></string-name>.
 <article-title>Blood pressure levels before dementia</article-title>.
<source>Arch Neurol</source>.
 <year>2005</year> <month>Jan</month>;<volume>62</volume>(<issue
>1</issue>):<fpage>112</fpage>–<lpage>116</lpage>.</mixed-citation>
```

slide 78

Citations will get Punctuation and Spacing

- Always assume XML documents will be displayed to people
- If punctuation and spacing not in XML, display engine will supply it
- Using Element Citation means relying not just on the kindness of strangers but on their good will, intuition, and programming skill
- Use <mixed-citation> and <string-name>

slide 79

Citations of Standards

- To journals & books, standards are cited like books with one standard-specific structure (std-organization)
- To standards, standards are cited differently, with structures for
  - Standard ID
  - Standard Reference Designation

slide 80

Decisions: Adopting Coding Guidelines

There are many "guidelines" for how to use JATS & how to tag articles in XML

- Funder may have requirements
- JATS Tag Libraries
- PubMed Central Guidelines
- Archives/repositories
- Datacite
- Force11
- JATS4R
- Your service providers
- NISO recommended practices
  - Recommended Practices for Online Supplemental Journal Article Materials
  - Access and License Indicators

slide 81

Following Guidelines

- Distinguish between requirements and guidelines
  - funding source
  - business partner
  - advocates for improved interchange
- Do NOT assume all guidelines apply to you
- Do NOT allow others to spend your money for their convenience UNLESS it matches your values
- Balance your costs with benefits to your constituents

slide 82

Decisions: Tag Set Validation and Checking

(optional)

- Well-formed documents are XML but...
- Well-formed is not enough
- If it isn't valid; it isn't really usable
- But valid is not enough either
- Many users (especially those who collect content from many sources) have additional constraints

slide 83

Customizations Require Local Rules

(Not enforced by DTD or described in Tag Library)

- One of the `<article-id>`s must be a `DOI`
- DOI must start with your corporate prefix
- Every citation (mixed or element) of *`ref-type="journal"`* must have an author
  `<person-group>` or `<name>` or `<string-name>`
- There is a limited set of values for "`article-type`" attribute
- Every `<ref>` of *`ref-type="book"`* must have a `<publisher-name>`

slide 84

Other Rules-checking

- Homegrown software (perl, javascript, et al.)
- Schemas can add data-typing and element content constraints (regular expressions)
- Schemas can add `OR` group (bag) constraints
  - 3 optional repeatable elements
  - one can only occur once
  - other two can be as many as you like
- Schematron

slide 85

Validation with Schematron

Schematron can be thought of as ...

- Way to test XML documents
- Rules-based validation language
- Way to specify and test statements about your XML
  - elements
  - attributes
  - content
- Cool report generator

All of the above!

slide 86

Reasons to Use Schematron

- Business/operating rules that other constraint languages can't enforce
- Different requirements for different stages of the document lifecycle
- Local or temporary constraints (not in base schema)
- Unusual (but not illegal) variation
- No DTD or schema
- Ad hoc querying and discovery

slide 87

Schematron is an XML Vocabulary

- A Schematron "program" is a well-formed XML document
- Elements in the vocabulary are "commands" in the language
- A Schematron "schema" specifies
  - tests to be made on your XML
  - messages you get back if the tests succeed or fail

Schematron Provides the World's Best Error Messages: You write them!

slide 88

Validation in the Grammar (DTD)

- Naming all elements and attributes
- Some sequences as element content
- Some attribute value lists

Separation of concerns:

“Treat the DTD as a coarse whitelist of elements which may be present, and Schematron as a finer blacklist of structures which are not desired; this distinction makes demands on both developers and users to distinguish between the two categories of error.” - Mike Eden and Tom Cleghorn.

An Implementation of BITS: The Cambridge University Press Experience

slide 89

Validation with XSLT, perl, and...

- Schematron can report patterns in a collection of documents
- But so can other languages

Don’t forget, there are lots of options

- False-color proofs for human editing
- XSLT reports on how many of element X in context Y are in your data
- Make HTML or PDF to check look-and-feel
- XQuery and XSLT are both good at “bring me back all the...”
- XPath in an editor is not a bad way to learn

slide 90

Decisions: DTD, XSD, RNG

- There 3 formats for XML grammars: DTD, XSD, & RNG
- All JATS tag sets are available in DTD, XSD, & RNG format
- The information content of each is equivalent
- *Use the most convenient at the moment*
- *Switch among them if convenient*
- *If a tool you want to use prefers one format, use it. At least while using that tool.*
- *This is unimportant.* ***Do NOT spend time or energy on this.***

slide 91

Final Questions

slide 92

Colophon

- Slides and handouts created from single XML source
- Slides projected from HTML generated from XML using XSLT
- Print copy created from the same XML source (*using a draft process*)
  - XSLT transform generates XHTML
  - Antenna House Formatter makes PDF from:
    - XHTML
    - CSS3 (slightly extended)
    - Graphics sizing table