National Library of Medicine
PubMed Central Back Issue Scanning Project

Image Specifications and Functional Requirements for Citation Capture

Version 3.2

May, 2007

National Library of Medicine

National Institutes of Health/Health & Human Services

8600 Rockville Pike

Bethesda MD 20894

Prepared in Cooperation with Apex Publishing

198 Van Buren Street

200 Presidents Plaza

Herndon, Virginia 20170-5338

Acknowledgements

This document was prepared by Apex Publishing with the National Library of Medicine (NLM) for use in digitizing the archives of selected journals as part of NLM's PubMed Central Back Issue Scanning Project. PubMed Central is a free digital archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health (NIH), developed and managed by NIH's National Center for Biotechnology Information (NCBI) in the National Library of Medicine. PubMed Central is available at: www.pubmedcentral.gov.

Purpose and Scope of this Document

The purpose of this publication is to document all requirements for image capture, article identification, and citation creation for the materials scanned as part of the PubMed Central Back Issue Scanning Project. It is subject to change, and may specify requirements that are only pertinent to records created for NLM's PubMed Central database.

Contacts

For questions about this document or about the PubMed Central Back Issue Scanning Project, please write to: pmc@ncbi.nlm.nih.gov

Table of Contents

1.0 Project Overview

1.1 Summary

This project involves capturing bibliographic citation data (XML) and full article text (OCR) as well as creating article level PDF files and full page images of each page (TIFF) that will be added to the customer database containing a digital archive of life sciences journal literature.

The customer database already contains bibliographic citation data for some of the articles. The customer will provide a list of articles that are already in their database and for which the XML data does not need to be created.

1.2 Definitions

1.2.1 Article-Level

The term **article-level** defines a set of pages of the same page type (e.g. article, administrative content, advertisement and table of content pages) and grouped according to the logical boundaries of the item in the source. For example, all pages of a specific article are grouped together into a single article-level item. Similarly, all pages of administrative content are grouped together into a single article-level item.

1.3 Project Deliverables

The deliverables for this project include:

1.3.1 A 600-dpi bitonal TIFF for each issue page. Issue pages that contain text from more than one article (see Item 5.2.2) are duplicated in the corresponding directories for each article.

1.3.2 A cropped color/grayscale TIFF image of each article illustration at 300-dpi, 24-bit color or 8-bit grayscale (Packbits compression). Illustration images are delivered **only** for article page types (see Item 5.2.2).

1.3.3 An article-level PDF file of one or more related pages with composite images for pages that contain a mixture of black and white, grayscale and color. Pages in the PDF appear in reading order sequence, typically the same sequence as the original source. This type of PDF file is delivered **only** for article page types (see Item 5.2.2).

1.3.3.1 For full-color articles (see Item 5.2.2.5), the article-level PDF will contain color page images.

1.3.3.2 For TOCs appearing on full-color front covers, the TOC-level PDF will contain color page images, *when requested by the customer-provided style sheets OR by ACV specifically*.

1.3.4 An article-level PDF file of one or more related pages comprised of the bitonal full-page images. Pages in the PDF appear in reading order sequence, typically the same sequence as the original source. This type of PDF is delivered only for non-article page types (see Items 5.2.2.3, 5.2.3.5, 5.2.5 and 5.2.6)

1.3.5 An article-level ASCII file of the text of each article page, created using OCR technology (unedited). If any page in the article contains text from more than one article, the OCR text file is edited to exclude any text belonging to other articles. Autozoning generally achieves column recognition. However, iIn batches that this is not consistently achieved, the OCR files should be corrected to obtain basic page column recognition.

1.3.6 An article-level XML-tagged citation for each article for which the customer does not already have citation data

1.3.7 A “place-holder” XML file for each article for which the customer already has citation data

1.3.8 An index file for each CD-ROM or DVD listing the name and MD5 checksum of each file

1.3.9 A file listing the PMID of any articles contained in the customer-exclusion list, but not present in the deliverable (see Item 1.7)

1.3.10 An inventory file mapping the contractor-assigned page image filenames (see Item 5.6.2) to source page number.

The inventory file will be issue-specific and contain a header with issue-level information. The format of the header is as follows:

```
<path="ppp"><journal="jjj"><vol="vvv"><issue="iii"><date="ddd">
```

where

ppp:	path to the top level directory for the journal issue
jjj:	journal code (see Item 5.6.1.1)
vvv:	volume number as per source
iii:	issue number as per source
ddd:	the issue date as per source

1.3.11 An inventory file mapping figure sequence to illustration images. This file will map illustration image file to figure label and sequence. It will be formatted as a tab-delimited ASCII file with the following fields:

1.3.11.1 File Header

The inventory file will be issue-specific and contain a header with issue-level information. The format of the header will be as per Item 1.3.10.

1.3.11.2 Image Filename

Contains the filename of the illustration image.

1.3.11.3 Figure Label

Containing the figure label as per source, e.g. “Fig. 5”

Figures may appear without any numbering in the figure labels, these figure labels will have to be generated according to the following guidelines:

p*nnn-a*

where:

p:	literal page number prefix
nnn:	page number as per source
-a:	optional alphabetic image sequence identifier

1.3.11.4 Figure Sequence

Containing the sequential identifier of the illustration. For example, “Figures 5, 6 and 7” on the same page would be sequenced as “1, 2 and 3”

1.4 Source Material

The source will be supplied primarily as bound paper journals to be disbound. The expected volume is 564,800 articles (approximately 5.7 million pages).

1.5 Logistics

1.5.1 There is no set source receipt schedule as the customer does not receive titles on a regular basis. Therefore, the source documents will be picked up from the customer’s site on pre-designated dates according to source availability.

1.5.2 The batch sizes will vary but should be approximately 3000 pages. Batches are sized such that each batch will fit on a single 4.7 GB single-sided DVD-R.

1.5.3 The XML data produced must meet the following quality standards:

- 100% Structural Integrity
- 99.995% Keying Accuracy

1.6 Project Workflow: Overview

1.6.1 Customer prepares source documents for shipping

1.6.2 The shipment will be picked up from the customer site and brought to the contractor.

1.6.3 The contractor prepares shipment to send to the Facility by confirming contents against customer-supplied packing list.

1.6.4 Upon receipt, the Facility will review each page of the source material and report any source discrepancies.

1.6.5 Facility will provide full inventory details to the contractor. The contractor will then divide source material into batches and provide inventory and schedule information to customer within five (5) business days after receipt of data at Facility.

1.6.6 Facility converts data and returns to the contractor

1.6.7 The contractor batches data, with associated issue-level ID numbers from customer-supplied packing list, to Customer to load into Database.

The issue-level ID numbers are conveyed to the customer via the shipment notification, the issue-level ID numbers are not conveyed through the deliverable inventory file (see Item 1.3.10).

1.7 Article Selection

1.7.1 The customer has article metadata for several journal articles, primarily from the more recently published issues. Therefore, only select articles require XML metadata conversion.

This list will be reconciled against the article inventory created during production. Articles excluded from conversion are identified as articles satisfying one of the following parameters:

1.7.1.1 If `<Item Name="HasAbstract" Type="Integer">` is '0' and the corresponding articles does NOT have an abstract in the source

1.7.1.2 If `<Item Name="HasAbstract" Type="Integer">` is other than '0'

1.7.2 Any article either not satisfying the above parameter or not listed in the customer-supplied file will be converted.

NOTE: Articles are excluded ONLY on the basis of the parameters described above. Any differences in the number of authors presented in either the article or the exclusion list does not have any bearing on the exclusion of articles.

1.7.3 For excluded articles, an XML file will be generated containing only the customer-assigned `<Id>` number within an Article Identifier element (see Item 7.4.2)

1.7.4 These guidelines are used for resolving discrepancies between article boundaries as defined in the source or Style Sheet requirements and as defined by the exclusion list:

1.7.4.1 General rule:

Follow the TOC for article boundaries/definition and the rules below. The specific journal Style Guide if it lists an exception to the TOC rule will always override the rules below.

1.7.4.2 **Case 1:**
TOC requires multiple articles AND exclusion List contains single article encompassing all pages of multiple articles.

Create the individual articles per the TOC or style guide. ***Match the first individual article to the exclusion list item encompassing all pages.***

Special Case:

In some cases, relying on pages cited in the exclusion item alone is not enough to determine if the exclusion item contains one or multiple articles. This is especially common in cases of letters followed by replies.

These cases can only be resolved by referring to the letter title and the authors if available. When letters and following reply letters are grouped together as a single item in the exclusion list, match the exclusion item with the first letter. All following reply letters are not linked to the exclusion list item and therefore are not excluded.

1.7.4.3 **Case 2:**
TOC requires one article AND exclusion list contains multiple entries for individual articles within the section.

Split the section into individual articles. Match the article as cited in the entry from the exclusion list. This will result in multiple articles being created for each item in the section, where some or all of the articles may be matched to the exclusion list.

Example:

TOC:	100-150
Exclusion List:	article 1: 101-120 article 2: 140-150

XML: article 1 : 100 (individual article XML)
article 2: 101-120 (<aid> only; matched to exclusion list)
article 3: 121-139 (individual article XML)
article 4 : 140-150 (<aid> only; matched to exclusion list)

In this example, a single article is listed in the TOC as 100-150. However, the exclusion list contains entries for two of the cited articles (101-120 and 140-150). Therefore, these two articles need to be matched accordingly. Additionally, articles contained in the section but NOT cited in the exclusion list should still be separated out as individual articles (100 and 121-139).

1.8 Source Discrepancies

1.8.1 During the inventory checking process for each batch, the Facility will identify any source discrepancies caused by missing, damaged and illegible pages.

1.8.2 The contractor will inform the Customer of all source discrepancies within each batch.

1.8.3 The Customer will provide replacement images for each page identified as a source discrepancy or will resolve the discrepancy in some other appropriate way.

1.8.4 The replacement image will be used as the source page in the conversion process.

1.9 Customer Delivery Requirements

All files are to be delivered on DVD-R. The source documents do not need to be returned to the customer and should be held in safe custody for one year after batch completion and acceptance. The customer will notify when the source and media can be disposed of after batch acceptance.

Files are to be in a consistent naming structure as determined by the customer. A separate directory will be delivered for each article that contains all the associated files for that article.

For each batch of data, the following deliverables will be sent:

- Deliverables listed in Item 1.3
- Conversion Log identifying any exceptional constructs that were encountered with an explanation of their treatment in the delivered files

2.0 Project Workflow

2.1 Shipment Preparation by Customer

2.1.1 Donors send Source Documents to the Customer

2.1.2 Customer will review the Source Documents for completeness and usability for scanning. The Source Documents will then be collated, boxed and labeled into a single shipment.

2.1.3 Each shipment will include an inventory list of the contents which will contain the following:

- Journal Name
- Volume/Part/Year Information
- Total Number of Volumes/Issues
- Estimated Total Number of Pages for each Journal

2.2 Shipment Preparation by the contractor

2.2.1 The contractor will arrange for a courier to retrieve the shipment from the customer.

2.2.2 Verify the shipment contents against the customer-supplied packing list and ensure that every issue is present.

2.2.3 Ship source material to production facilities.

2.3 Shipment Inventory by Facility

2.3.1 Verify the shipment contents against the packing list and ensure that every issue is present.

2.3.2 Perform a full inventory of all source material in the shipment, including actual page counts.

2.3.3 Identify any missing pages as source discrepancies

2.3.4 Provide inventory details and source discrepancies to the contractor .

2.4 Batch Inventory and Schedule

2.4.1 The contractor will use the full inventory information provided by the facility to divide the content of the shipment into production batches as per Item 1.5.2.

2.4.2 Provide batch inventory and production schedule to customer.

2.4.2.1 The batch inventory will detail the contents of each batch including:

- Contractor assigned batch number
- Journal title of batch content
- Volume/issue numbers of batch content
- Issue-level ID number supplied by Customer in source packing list

2.4.2.2 The production schedule will detail the schedule for the following:

- Delivery dates for each batch (referenced by the contractor assigned batch number)
- Pending unresolved source discrepancies and/or technical queries
- Adjusted delivery dates, if needed

2.5 Source Discrepancies

The contractor will report any missing pages noted as source discrepancies to the customer.

Discrepancies should be reported to the Customer on a weekly basis. Resolution of the problems should be returned to the contractor within the following week unless entire issue replacements are needed. Issue replacements will take longer.

2.5.1 Discrepancies should only be reported for damaged pages of articles and administrative material, missing covers and TOCs not noted on the packing list, and missing pages determined in the full inventory conducted by the contractor. No missing advertising pages should be reported as a discrepancy. Replacement pages with spots and discoloration should only be request when they obscure text.

2.5.2 Replacement pages can be provided in one of two formats:

2.5.2.1 Hardcopy

- Replacement pages provided as hardcopy will yield the best results in the final converted product as the hardcopy replacement page will undergo the same exact conversion process as the remainder of the issue.
- Schedule modifications may be necessary to account for the time involved in shipping hardcopy replacements to the production facility.

2.5.2.2 Image

- Image replacements are required to be 300dpi, 24-bit, color TIFF full-page images. This is necessary in order to produce results that best approximate the standard conversion process.
- The method in which replacement images will be provided remain to be determined. However, most options (e.g. email and FTP) can be easily accommodated.

2.6 In cases of unresolved source discrepancies caused by the lack of replacement pages, the source should be used “as is”, unless otherwise directed.

2.7 Production Inventory

All production inventory files are created using a key/verify data entry application. This ensures a higher accuracy than single-pass keying. The inventory files are used by software throughout the production process.

2.7.1 Create Batch Inventory File

The Batch Inventory File will detail the contents of a single production batch. The information will be relative to each issue contained within the batch and include:

- Contractor-assigned Document Number
- Journal Title
- ISSN
- Volume Number
- Issue Number
- Publication Date

2.7.2 Create Page and Article Inventory Files

The Page and Article Inventory Files will provide an inventory of each page and article within an issue. These files include the following information:

- Contractor-assigned Document Number
- Source Page Number
- Page Type
- Number of illustrations
- Type of each illustration
- Contractor-assigned Article ID of each article (or part of) appearing on page
- Presence of abstract text
- PubMed ID if metadata is already available for the article
- Whether OCR text is required

2.8 Production Process

2.8.1 Converted data is produced according to the specifications and procedures outlined in this document.

2.8.2 Batches in production are systematically checked by the Quality Assurance Group during “Process QA”

2.8.3 Batches are submitted to both the contractor and the Quality Assurance Group for “Product QA” upon completion.

2.9 Deliver Batch to Customer

2.9.1 Batches are delivered to the customer only when accepted by the Quality Assurance Group.

2.9.2 Special measures taken during production to handle non-standard situations, e.g. missing pages, will be detailed in the Conversion Report submitted with each batch delivery.

Data is usually delivered to the customer on a single DVD. However, if two DVDs are required, then label the DVDs as:

301-bbbb 1 of 2"
301-bbbb 2 of 2

The first DVD should contain the .lst file, all the .map files, and the data for some of issues in this batch. The second DVD would contain the remaining issues inside of a journal directory, but would not contain duplicates of any of the auxiliary files.

3.0 Quality Assurance

3.1 Overview

Quality assurance is attained in all the contractor's projects through two different methods. Each method is independent of each other. The methods are employed by a self-contained Quality Assurance Group that is answerable only to the contractor and not the individual production facility. The independence of the Quality Assurance Group prevents any cross-contamination and conflicts of interest when measuring the quality of the converted data.

3.2 Methodology

3.2.1 Process QA

This method is employed by a separate department within the Quality Assurance Group. Process QA involves taking 10% biased samples periodically from each step in the production process. The samples are biased towards areas that may be problematic, such as new operators, new production steps and especially complex source materials.

Any quality problems identified during Process QA are corrected during production. Quality problems that require resolution or clarification from the customer will be submitted as queries. Problems will be corrected based on the response received.

3.2.2 Product QA

This method is employed by a separate department within the Quality Assurance Group. This method is similar to the final product quality checking traditionally used in data conversion processes. Product QA involves taking a 7% random sample of the final converted data. Checks are made to ensure that the deliverable data conforms to the specifications and acceptance criteria specific to the project. Only when a batch has successfully passed Product QA can it be delivered to the customer.

3.3 Product QA Acceptance Criteria

Item	Error Category	Acceptance Criteria (per sample)
1	Character accuracy	99.995%
2	Tag content accuracy • Incorrect application of XML element • Includes incorrect article type assigned as defined in CSDD Section 7.1.1.1	99.95%
3	Structural Integrity of Tags • XML not valid per DTD	100%
4	Image Quality – Page or Illustration Image • Image skewed on visual inspection • Image not consistent with source (e.g., inconsistent boundaries, pixilation, distortion, lack of sharpness and detail, etc)	99%
5A	Inventory error • Missing or duplicate file (XML, OCR, image, etc) • Citation not created for non-excluded item • Images not sequenced in “reading order” • PDF exceeds 60-page limit • OCR text does not correspond to article boundary • Individual figure delivered as multiple images	100%
5B	Inventory error • Article boundaries do not conform to CSDD, Exclusion List or Style Guide, whichever is applicable • Incorrect article type assigned as per CSDD Section 5.2 – i.e. page types	99.5%

For rejected batches, all errors that failed the batch will be corrected and the batch will be reviewed for additional instances of those error types.

The customer may request that a batch be reviewed and corrected if **both** for the following conditions are present:

a. The error is **clearly** in conflict with the CSDD, **and**

b. The error is a systematic problem that occurs throughout the batch.

Even though the customer did not identify enough errors to fail such batches, the batch will still be considered rejected and will require correction and re-delivery.

3.4 Product QA Methodology

Product QA sampling is based on the primary deliverable for the project: the article PDF file. The article PDF file is the primary deliverable as it is most representative of all deliverables for this project (XML data, full-page images, illustration images and OCR text).

A randomly generated sample of 7% of the total number of article PDF files in a batch will undergo Product QA. All acceptance criteria are determined based on the contents of this sample. Results that are below the stated accuracy requirement for any acceptance criteria will cause the batch to be rejected.

3.4.1 XML File Acceptance Determination

3.4.1.1 Character accuracy

Text contained in the article XML files in the sample is proofed against the original source material. The total number of incorrect or missing characters is measured against the total byte count in the sample, the result is multiplied against "100" to obtain the accuracy rate.:

$$\frac{\texttt{(Total Byte Count) – (No. of Incorrect Characters)}}{\texttt{Total Byte Count}}$$

3.4.1.2 Tag content accuracy

Elements in the article XML files in the sample are checked for correct content. An example of an error would be a contributor surname (<snm>) is incorrectly "tagged" as contributor given name (<gnms>). The total number of elements with incorrect or missing content is measured against the total number of elements in the sample, the result is multiplied against "100" to obtain the accuracy rate.

$$\frac{\texttt{(Total No. of Elements) – (No. of Incorrect Elements)}}{\texttt{Total No. of Elements}}$$

3.4.1.3 Structural accuracy

All XML files in the sample are parsed and validated against the then-current project DTD.

3.4.1.4 Inventory error

An article XML file is required to be generated for each article in the sample:

- Articles identified as excluded still require the minimally tagged XML file as per project specifications.
- Articles not identified as excluded must have a fully tagged XML file.

3.4.2 Full-page Image Acceptance Determination

3.4.2.1 Unsatisfactory Quality

Full-page images in the sample are reviewed for the following:

- Page image is skewed upon visual inspection
- Page image boundaries not consistent with source
- Image contains visible pixilation, distortion, color distortion or lack of sharpness and detail

The total number of poor quality page images is measured against the total number of page images in the sample, the result is multiplied against “100” to obtain the accuracy rate.:

$$\frac{(\text{Total No. of Page Images}) - (\text{No. of Poor Quality Images})}{\text{Total No. of Page Images}}$$

3.4.2.2 Inventory error

A page image is required to be generated for each page in the sample:

- Page images must appear in same sequence as source
- Page images cannot be duplicate or missing

3.4.3 Illustration Image Acceptance Determination

3.4.3.1 Unsatisfactory Quality

Illustration images in the sample are reviewed for the following:

- Illustration image boundaries not consistent with source
- Image contains visible pixilation, distortion, color distortion or lack of sharpness and detail

The total number of poor quality illustration images is measured against the total number of illustration images in the sample, the result is multiplied against “100” to obtain the accuracy rate.

$$\frac{(\text{Total No. of Illus. Images}) - (\text{No. of Poor Quality Images})}{\text{Total No. of Illus. Images}}$$

3.4.3.2 Inventory error

An illustration image is required to be generated for each illustration in the sample.

- Multipart illustration images must be created as per specifications

3.4.4 PDF file Acceptance Determination

3.4.4.1 Inventory error

An article PDF is required to be generated for each article in the sample.

- Content of article PDF must correspond to article boundary as per source
- PDF pages must appear in correct sequence as per specifications (see Item 6.10.3)
- PDF page contains OCR text
- PDF must not exceed size restriction as per specifications

3.4.5 OCR file Acceptance Determination

3.4.5.1 Unsatisfactory quality

OCR files in the sample must have been generated using auto-zoning.

3.4.5.2 Inventory Error

An OCR file is required to be generated for each article in the sample. The OCR text must correspond to article boundary

3.4.6 General Acceptance Determination

All files must be assigned the correct page type (see Item 5.1)

3.5 Process Controls

Aside from the Quality Assurance Group, there are several controls built-in to the production process. These controls allow for quality and inventory checks to be made during the production process. Most of these controls are software based and are therefore more reliable.

3.5.1 Inventory

The primary inventory system is page-based; this allows the inventory to be checked on the most logical unit for conversion. Each page is inventoried beyond a simple page count, as content-based information is provided for each and every page to be converted (see Item 2.6). The information contained in the page-level inventory file is used by software to perform "real-time" and "post-production" checks throughout the project.

3.5.2 Image Files

3.5.2.1 Scanning

Scanning is performed on high-quality, industry standard scanners. Full-page images created through scanning process are checked 100%. This check is both manual and software-assisted. The software confirms that an image has been scanned for each page listed in the page inventory file. A data analyst performs a review of all scanned images and ensures that pages were scanned in the same sequence as presented in the original source.

3.5.2.2 De-skew

Page images are checked 100% to ensure that there is no visible skew in the images. Software tools are used during this process to assist in correcting skewed images.

3.5.2.3 Illustration Images

Illustrations are identified for imaging within each full page image. Software performs real-time checks to ensure that the number of images and the type of each image matches the illustration information provided in the page inventory file. Inconsistencies will cause the software to prevent further processing of the images until the conflict is resolved.

3.5.3 OCR Files

The OCR files are created through a software process after the image files have been created.

The page and article inventory files are used by the software to ensure that an article-level OCR file is generated for each page of the issue. Article boundaries in the page inventory file will identify pages that contain text from more than one article so that text belonging to other articles can be removed.

3.5.4 PDF Files

The PDF files are created through a software process. OCR text, full page and illustration images are used to generate the PDF with hidden text.

The page and article inventory files are used by the software to create PDF files according to the specifications relative to the appropriate page type. For example, composite PDF files will be created for pages typed as “article” whereas bitonal PDF files will be created for pages typed as “advertisement”.

Additionally, article boundaries in the page inventory file will ensure that PDF files are created containing only those pages comprising an individual article.

3.5.5 XML Files

3.5.5.1 Article Exclusion

Articles to be excluded from XML metadata creation (see Item 1.7) are identified by using the customer-supplied exclusion list and the page inventory file. These articles will not be processed through keying.

3.5.5.2 Article Keying

- Article pages that contain text to be keyed are identified and printed via software.
- The content on the page is marked for keying by a data analyst specifically trained with the project specifications.
- The text undergoes at two-pass keying process to ensure 99.995% character accuracy.

3.5.5.3 Article XML Creation

- XML elements are generated by software to ensure consistent appearance and structural accuracy.
- Article boundary information in the page inventory files are checked by software to ensure that an XML file was created for each article.
- XML files are automatically parsed against the project DTD to ensure validity.

3.5.6 Delivery of Data

Converted data is prepared for delivery by software. This ensures that file and directory names are systematically generated according to specifications. This also ensures that all files are accounted for as per inventory files.

4.0 Source Description

4.1 General

The source material for this project consists of issues of various medical journals. The entire run of each journal will be converted in this project. Therefore, the material can span from as far in the past as the mid-1800s.

4.2 Source Type

The source material appears in one of two different formats:

4.2.1 **Individual issues** are expected to be the most common format of source material. Each issue is a single bound copy as typical of most periodicals.

4.2.2 **Bound copies** of multiple issues are created when the publisher decides to group several issues together into a single copy.

4.3 Source Quality

4.3.1 Variation

Variation in source quality is expected to be very high. All of these variations must be accounted for throughout the entire production process.

- Older source material can be very brittle and is yellowish in appearance.
- More recent source material is usually of better quality, but will also be very glossy in appearance.
- Bound copies of multiple issues may be printed on extremely thin paper to reduce the overall weight of the copy.

Special care must be taken to avoid damaging the source material, particularly during the imaging process.

4.4 General Content

Regardless of the binding format, most issues will be relatively consistent in content. In addition to articles (which are the majority of an issue), an issue will often contain the following:

- Covers
- Table of Contents
- Articles
- Administrative Material (author index, article index, keyword index, editorial mastheads)
- Advertisements (including advertising index)

4.5 Issue Covers

The covers of most of the issues (or the first page of an issue in bound copies) contain issue identification information:

The date information is typically located on the issue cover (see Item 4.5.4). However, if a cover is not available, use the PubMed records from the Article Exclusion List (see Item 1.7 for the same issue to obtain the following information

4.5.1 Journal title

The journal title is very prominently featured on the issue cover.

4.5.2 Volume number

The volume number can be identified with a keyword such as "Volume" or "Vol." The number may appear in either Roman or Arabic numerals. While, other issue identifiers may be optional, the volume number is expected to always be present.

4.5.3 Issue number

The issue number can be identified with a keyword such as "Issue" or "Iss." The number may appear in either Roman or Arabic numerals. The issue number is not always present in the source.

4.5.4 Issue date

The issue date can appear in various formats, containing any of the following information:

- Day information (often in numeric form only)
- Month information (often spelled-out or abbreviated)
- Year information (often in 4-digit numeric form)
- Season information (often present instead of month information)

The sequence in which this date information appears is varied. However, it is expected that the most common format would be "Month Day, Year", i.e. "November 7, 2000".

4.6 Table of Content Pages

The table of content pages lists the content of an individual issue

4.6.1 Issue Information

These pages can contain issue identification information similar to that often found in the issue covers.

4.6.2 Article Information

A single article entry can contain article title, article authors, first article page and/or a brief description. Additionally, articles can be grouped into subject categories; each category would have a distinct heading or title.

4.6.3 Multiple Table of Contents Pages

Source may be received as either individual issues or bound volumes (containing multiple issues from the same volume). Within these two types of source there may be multiple table of contents pages. These types of table of contents pages are defined as:

- Cumulative TOC: Relevant to the contents of an entire volume
- Individual TOC: Relevant to the contents of a single issue

If there are individual TOCs within a bound volume, then the individual TOC should be made part of the corresponding issue. The individual TOC would be placed in the TOC category.

If there is a cumulative TOC within a bound volume, then the cumulative TOC should be made part of the issue in which it was published or bound. Do NOT capture the cumulative TOC for each issue within the bound volume. The cumulative TOC should be captured as the issue TOC if there is no individual issue TOC. If there are multiple types of TOC's present in an issue, treat the individual TOC as the issue TOC and place the cumulative TOC with the administrative material.

4.7 Front Matter Pages

4.7.1 Issue Information

These pages can contain issue identification information similar to that often found in the issue covers.

4.7.2 Publisher information

The copyright often contains publisher information, including publisher name and location.

4.8 Article Pages

The first page of an article contains the majority of the bibliographic metadata, including:

- Article title
- Article authors
- Article footnotes
- Article copyright
- Article abstract

4.9 Errata/Correction/Retraction Articles

These articles are short paragraphs indicating corrections made for previously published articles. They must contain volume and page references to the previously published article. Each errata item is treated as a separate article.

5.0 Target Database: General Specifications

5.1 Specifications Hierarchy

The CSDD provides the general specifications to be followed for this project, however, the customer also provides Style Sheet (used to provide title-specific requirements) and Exclusion List (used for determining article exclusion) as supplementary material. When there is a conflict in the requirements between these three types of documents, then the following hierarchy should be used to resolve the conflict:

- Style Sheet
- Exclusion List
- CSDD

5.2 Issue Pages and Page Types

All issue pages are scanned as full page images (see Item 1.3.1), including outside/inside covers, tables of content, articles and advertisements. Each page is categorized as containing **any** of the page types listed below.

Page types determine how a page is converted and delivered. Pages containing multiple page types are duplicated for each of the corresponding page types.. For example, if a page contains Table of Contents, Advertising and Article text, the page is identified as having “article” page type, “advertising” page type **and** “table of contents” page type. This page will appear in the deliverable for the article, the table of contents **and** advertising.

There is no exception to this rule, all page are duplicated for each page type.

5.2.1 Outside front covers with color/grayscale

When the outside front cover of an issue contains color or grayscale material, that page is scanned in color or grayscale as appropriate and is classified as a “cover page” <u>in addition to</u> other types of content which may be present on the page. Front cover pages that do not contain color or grayscale information are not covered by this category. For purposes of PDF file creation, this type of page is grouped into the same article-level item as pages of type “Issue Covers”, discussed below.

5.2.2 Articles

Articles comprise the majority of an issue. The term is used to categorize all full-length material, as well as shorter articles such as meeting reports, commentaries, reviews, etc. With few exceptions, most entries listed in the issue table of contents are considered to be articles.

Article pages are defined as pages that contain ANY article text.

5.2.2.1 Grouped Articles

- Some articles are actually groups of separate items. In these instances, each individual item is considered an individual article.

  Grouped articles can often be identified by the presence of discrete sections of text with specific set of contributor(s). For example, a “Book Reviews” section may contain multiple reviews, each written by a different contributor. In these cases, each review would be a separate article.

    - Each individual letter in a “Letters to the Editor” section would be treated as a single article.

      Additionally, any brief editorial comment or author comment following a letter (see Exhibit 18) is treated as part of the preceding letter and not as a separate article.

    - Each correction item in an “Errata” section would be treated as a single article.
    - Each individual book review in a “Book Reviews” section would be treated as a single article.
    - Each obituary is a detailed write up on an individual’s accomplishments and has a contributor so would be a separate article.
- In some cases, the customer supplied style sheet may require that grouped articles should be treated as a single article, for example, a book review section containing multiple books reviews could be treated as a single article on an exception bases for a particular title.

5.2.2.2 Poster Abstracts

Some journal issues will contain poster abstracts of a conference proceeding. These pages contain images reproduced from a poster at the conference. These are brief abstracts summarizing presentations made during the conference proceedings. Several poster abstracts may appear grouped together on one or more pages.

The grouped abstracts are clearly not typeset text as seen for regular article text. Instead, the poster abstracts are reproductions of original documents which are reduced in size in order to fit multiple poster abstracts on a single page.

Poster abstracts are **not** treated as grouped articles. They are **not** split into individual articles. A group of poster abstracts is usually listed as a single entry in the issue table of contents. This entry is used to determine article boundaries and titles.

5.2.2.3 Proceedings Articles

Some journal issues will contain research articles presented at a conference proceeding. This is a series of articles that cover a single conference proceeding. This series include proceedings abstracts, these differ from poster abstracts, as the text is the same as the typeset text for a regular article (i.e. the text is NOT a reproduced image from a poster).

Proceedings articles **are** generally, but not always treated as part of a single article. The entire section detailing the proceedings, including the abstracts, schedules, timetable, President’s Address, etc. are all grouped together as a single article.

Sometimes a PMID occurs in the Exclusion List which applies to the entire proceedings section of an issue, or to a subset of the pages included in the proceedings section. Use the PMID in the <artid> tag in the following cases:

- For the single article where no division of the proceedings
- For the first article where the proceedings is divided into sections

Proceedings can appear as either grouped into a single article or split into sections. The customer-provided style sheet will often provide details on the treatment of proceedings articles.

In general, if the proceedings articles are listed in the TOC, then treat the entire set of proceedings articles as a single article. However, if the proceedings articles comprise the entirety of an issue, then divided the proceedings article into appropriate sections (e.g. by date, subject divisions, or association sections)

5.2.2.4 Follow-up Articles

These are articles that are related to the preceding article. These articles are often titled as "Replies" or "Discussions"". These follow-up articles are treated as individual articles, although they will be linked to the preceding article via the `<relart>` element (see Item 7.10.1)

5.2.2.5 Full-color Articles

For some articles, the customer will indicate that article pages are to be converted as full-color page images. This is generally used when the source page contains full-color backgrounds and/or full-color images mixed with text.

5.2.2.6 Obituaries and Death Notices

A distinction must be made between obituaries and death notices. Obituaries are captured as individual articles, whereas death notices are captured as part of administrative material.

An obituary is typically a lengthy article detailing an individual's accomplishments. An obituary article would typically have a contributor.

However, death notices are lists of individuals who have died. The list usually consists of names, birth and death dates and place of residency. Death notices would NOT typically have contributors.

5.2.3 Table of Contents

These pages contain the table of contents page(s) of the journal issue. The tables of contents can be issue-level, volume-level, subject specific, or cumulative for several volumes.

5.2.3.1 If an issue has both a cumulative and individual TOC (see Item 4.6.3), then the individual TOC is treated as "Table of Contents" and the cumulative TOC is treated as Administrative Content

5.2.3.2 If an issue has only a cumulative TOC, then the cumulative TOC is treated as "Table of Contents".

5.2.3.3 If an issue has only an individual TOC, then the individual TOC is treated as "Table of Contents".

5.2.3.4 Advertising indices and TOCs of forthcoming issues are not treated as "Table of Contents". Advertising indices are treated as "Advertisement" (see Item 5.2.5) and TOCs of forthcoming issues are treated as "Administrative" (see Item 5.2.4)

5.2.3.5 TOC on color front covers

For some **titles**, the customer will indicate that TOCs on color fron cover pages are to be converted as full-color page images. This is generally used when the source page contains full-color backgrounds that are not sufficiently contrasted with text, thereby resulting in poor quality bitonal images.

5.2.4 Administrative Content

These pages contain various types of content, including

- Journal Masthead
- Editorial Board
- Subject or Author Indexes
  (not including Advertising Indexes)
- Instructions to Authors
- Administrative Announcements for Society Members
- Meeting Notices
  (different from meeting reports)
- Death Notices
  (lists of individuals who have died; usually consisting of names, birth and death dates, and place of residency)
- Calls for Papers
- Course Announcements
- List of Books
  (e.g. books received, selected reading, selected titles etc.)
- Book Notices (see Exhibit 17)
  (usually for review or consideration).
- Annual Report
  (If an annual report

5.2.4.1 Administrative Content and the TOC

Some administrative content, such as announcements of upcoming conferences, new books, etc., can be hard to distinguish from advertisements.

All announcement material (including book announcements, meeting announcements, course announcements and drug product announcements) should be placed with advertisements unless listed on the TOC in which case it can be placed with administrative material.

Book notices/new book lists that appear on the TOC will go with administrative material. If not on the TOC, it should be placed with advertising.

5.2.4.2 Indices

Indices are usually converted as Administrative Content. However, there is an article type for “index” (see Item 7.1.1.1). An index is treated as an article only when the issue is a supplement that comprises only indexes, i.e., the issue has no "regular" article content.

A cumulative index for a volume should be placed in the admin section of the issue with which it is bound (if the cumulative appears with the first issue of the volume, it should be placed in the admin section of the first issue; likewise, if the cumulative index appears with the last issue of the volume, it should be placed in the admin section of the last issue).

If there is a cumulative index covering multiple volumes, this should be treated as its own issue. If there are cumulative indexes containing multiple volumes associated with one issue, then this would be a mistake on the packing list and should be raised as a source discrepancy.

5.2.4.3 Subscriptions

All subscription material should be treated as Advertisement.

5.2.5 Advertisement

These pages are all the page advertisements in the issue, including partial and full-page ads. “Advertiser Index” or “Index to Advertisements” should also be captured as Advertisement (see Exhibit 16).

5.2.6 Issue Covers

These pages are the outside and inside covers of the journal issue.

5.2.6.1 Blank Covers

Blank covers are not deliverable.

5.2.6.2 Outside Front Cover

Outside front covers containing material of other page types, such as table of contents, are duplicated as bitonal pages in the appropriate file.

5.3 Source Page Numbers

The source page number for each page must be recorded in the document inventory. In some cases, the number may not be explicitly present on the source page. Use the following guidelines to resolve such cases.

5.3.1 If the source page number is inferable, then the inferred page number must be keyed within square brackets "[ ]". Inference of page numbers can be based on either the surrounding page numbers or on the page number provided in the table of contents entry. The table of contents entry is useful for inferring page numbers, when the missing source page number is the first page of the article.

Additionally, if the missing source page number occurs on the first article page of an issue and the last numbered page of the previous issue is known, then it is acceptable to infer the missing source page number using the sequence of the last known page number of the previous issue. For example, if the first page of the first article in an issue is unnumbered, but the last numbered page in the previous issue is 157, then you may infer the missing page number as [158] (see Item 5.3.2 for additional clarification).

When inferring source page numbers it is important to preserve the sequence of numbering within a single article-level item (see Item 1.2.1) or series of pages of the same the page type (see Item 5.1). In some cases, it will not be possible to infer the page number. These cases are discussed in later paragraphs.

For example, source page numbers appear as follows:

1st Administrative Content Page:	1
2nd Administrative Content Page:	2
3rd Administrative Content Page:	3
4th Administrative Content Page:	[unnumbered]

1st Article Page:	[unnumbered]
2nd Article Page	[unnumbered]
3rd Article Page	158
4th Article Page:	[unnumbered]
5th Article Page:	160

In this example, there are four unnumbered pages. The unnumbered administrative content page would be numbered [4] since it follows in the sequence of administrative content page numbers. The next two pages are article pages that precede a page whose number follows in a sequence from a previous issue. Their pages cannot be inferred since neither [5] and [6] nor [156] and [157] would be correct. (The former would imply that these are part of the administrative section, and the latter may conflict with pages from the previous issue.) The fourth article page would be numbered [159].

5.3.2 If the source page number CANNOT be inferred but is *preceded* by a numbered page within the same page type, then the last known source page number is keyed within square brackets and a sequential number suffix:

[*ssss*]-*a*

where:

ssss: is the last source page number (inferable or explicit)
a: is the sequential number of the page following

EXAMPLE: There are two pages without a source page number between Pages 234 and 235. These pages would be numbered, respectively as "234-1" and "234-2".

EXAMPLE: There is one page without a source page number between Pages [65] and 66 (where Page 65 is an inferred page). The unnumbered page would be keyed as "[65]-1".

5.3.3 If the source page number CANNOT be inferred and the unknown page number precedes the first page in a sequence of pages whose numbering may have been continued from a previous issue, use the page number of the first numbered page in this page type in square brackets followed by b*n* where “b” is the literal “b” character and *n* is a sequential number. In the following example:

1st Administrative Content Page:	1
2nd Administrative Content Page:	2
3rd Administrative Content Page:	3
1st Article Page:	[unnumbered]
2nd Article Page	[unnumbered]
3rd Article Page	158

The two article pages preceding 158 have unnumbered pages that cannot be inferred because the number 158 is continued from a previous issue. The two unnumbered pages should be respectively numbered [158]-b1 and [158]-b2.

Note that in the case, we are assuming the last numbered page of the previous issue is NOT known, therefore these pages cannot be numbered [156] and [157] (if they were known, then it is acceptable to infer the page number based on this knowledge as per Item 5.3.1) since these numbers may conflict with a previous issue, and it would not be appropriate to number them [4] and [5] because this groups them logically with the administrative content pages rather than the article pages.

5.3.4 If the source page number CANNOT be inferred AND there is no last known valid source page number, then key as follows:

nil*a*

where:

nil: is an explicit prefix
a: is the sequential number for each “nil” page within the issue

5.3.5 As the source page number is used to identify each page image to be scanned, there is a requirement for the source page number to be unique.

The uniqueness of a source page number must be maintained, regardless of its appearance in the source material. Therefore, if duplicate source page numbers are present, the duplicate source page number is keyed with a sequential numeric -suffix appended.

EXAMPLE: There are two pages numbered in the source as “45”. The second occurrence of the page number is given a numeric suffix and keyed as “45-1”.

EXAMPLE There are a total of three page numbered in the source as “27”. Each occurrence after the first page “27”, is keyed with a numeric suffix. Therefore the second page “27” is keyed as “27-1” and the third keyed as “27-2”.

NOTE: If duplicate source page numbers are present, then an analysis of the page type is required. Often, differing page types may contain duplicate page numbers. Pages with article text should retain the original source page numbers. Pages identified as non-article should be appended with the sequential numeric suffix as described above.

5.3.6 For some source pages, a single physical page will contain multiple images of reduced-size pages or “mini-pages (see Exhibits 14a and 14b). In such cases, the single physical page will have a source page number consisting of a page range of the source page numbers on the “mini-page”.

For example, in Exhibit 14a, the “mini-pages” have source page numbers of “519, 520, 521 and 522”. Therefore, the source page number of the single physical page is “519-522”. Similarly in Exhibit 14b, the “mini-pages” have source page numbers of “523, 524, 525 and 526”. Therefore, the source page number of the single physical page is “524-526”.

5.4 OCR Specifications

5.4.1 OCR Engine

ASCII text files will be created for each article-level item using Prime OCR, which is a premier OCR package used by most academic institutions creating archival-quality databases.

5.4.2 Editing

No editing will be performed on the OCR text as the good condition of the source material will generate sufficient character accuracy.

5.5 XML Specifications

5.5.1 Document Type Definition (DTD)

Article XML files are created in accordance with the Archive Header DTD (see Section 8.0).

5.5.2 Font Changes and Super/Subscript

Font changes and super/subscript characters are captured using the emphasis elements described in Item 7.16. Font changes that are applied stylistically to an entire block of text (e.g. a title completely in bold) are NOT captured. Only font changes for strings within a block of text (e.g. a single italicized word in non-italicized title) are captured using the appropriate emphasis element.

5.5.3 Extended Characters

Extended characters (that is, characters not present on a US keyboard and that fall outside of the 7-bit printable ASCII character range) are replaced by their Unicode values expressed in numeric entity form (i.e., `&#xHHH;` where `HHH` is the hexadecimal Unicode value).

5.5.4 Unknown Characters

Special non-keyboard characters that do not have a valid extended character value are replaced with the "`[unk]`" placeholder string.

5.5.5 Handwritten Text

Handwritten text is ignored in the XML metadata; no special treatment is needed.

5.5.6 Illegible Text

Illegible text appearing in the XML metadata is replaced with the "`[ill]`" placeholder string.

5.5.7 Complex Data Structures

There are no provisions in the project DTD to capture tables, lists, complex equations and images in the abstract text. Therefore, these data structures appearing in the XML metadata are replaced with a placeholder string.

The placeholder string is formatted as follows:

```
[Object: see text]
```

where *`Object`* is one of the following values:

```
Table
Formula
List
Image
```

5.6 Filename Convention

5.6.1 Standard Conventions

5.6.1.1 Journal Code

The journal identifier is code assigned to each journal title. This code is used in the final deliverable directory structure, filename convention and article metadata.

5.6.1.2 Issue Identifier

The issue identifier is a 5-digit number for each issue. This number is unique only within a journal title. Issue identifiers are NOT assigned sequentially.

5.6.1.3 Sequential Page Number

The sequential page number is a 4-digit number assigned to each page image within an issue. The sequence includes all physical pages in the issue, including covers. The sequence does NOT include blank pages as such pages are excluded from conversion. The sequence within each issue will start at "0001".

Page images are always assigned sequential page numbers according to the source sequence. However, in some cases, the page images will be re-sequenced in the PDF file as per specifications (see Item 6.10.3)

5.6.1.4 Sequential Illustration Identifier

The sequential illustration identifier is an alphabetic character assigned to each illustration appearing on the page. Identifiers are assigned to each illustration by strict sorting of the coordinates of the illustrations with respect to the bitonal full page image.

5.6.1.5 Article Identifier

The article identifier is a 4-digit sequential identifier assigned to each article within the issue. This number is unique only within a specific issue.

5.6.1.6 Article Sequence Identifier

The article sequence identifier is an optional alphabetic character assigned to an article if the article shares the same first page as another article. Identifiers are assigned to the illustration in "top-to-bottom, left-to-right" order.

This identifier is assigned to pages in the following cases:

- Page contains more than one article (see Item 5.2.2)
- Page contains an article and table of contents (see Item 5.2.3) or administrative material (see Item 5.2.3.5).

5.6.2 Full Page Images

The naming convention for full-page TIFF images (see Item 1.3.1) is as follows:

*jidiiiii-pppp***.tif**

where:

jid:	Journal code (see Item 5.6.1.1)
iiiii:	Issue identifier (see Item 5.6.1.2)
pppp:	Sequential page number (see Item 5.6.1.3)
.tif:	explicit file extension

NOTE: For color front covers (see Item 5.2.1), a suffix identifier, "-cov" is always appended. This is used to distinguish the cover page image from the table of contents page image, however, it is applied to all color front cover images for consistency.

5.6.3 Illustration Images

The naming convention for illustration images (see Item 1.3.2) is as follows:

`jidiiiii-pppp-a.tif`

where:

jid:	Journal code (see Item 5.6.1.1)
iiiii:	Issue identifier (see Item 5.6.1.2)
pppp:	Sequential page number (see Item 5.6.1.3)
a:	Sequential illustration identifier (see Item 5.6.1.4)
.tif:	explicit file extension

5.6.4 PDF Files

The naming convention for PDF files (see Items 1.3.3 and 1.3.4) is as follows:

`jidiiiii-ppppn.pdf`

where:

jid:	Journal code (see Item 5.6.1.1)
iiiii:	Issue identifier (see Item 5.6.1.2)
pppp:	Sequential page number of first article page (see Item 5.6.1.3)
n:	Optional article sequence identifier (see Item 5.6.1.6)
.pdf:	explicit file extension

5.6.5 XML Files

The naming convention for XML and XML “place-holder” files (see Items 1.3.6 and 1.3.7) is as follows:

`jidiiiii-ppppn.xml`

where:

jid:	Journal code (see Item 5.6.1.1)
iiiii:	Issue identifier (see Item 5.6.1.2)
pppp:	Sequential page number of first article page (see Item 5.6.1.3)
n:	Optional article sequence identifier (see Item 5.6.1.6)
.xml:	explicit file extension

5.6.6 OCR Files

The naming convention for article-level text files (see Item 1.3.5) is as follows:

`jidiiiii-ppppn.txt`

where:

jid:	Journal code (see Item 5.6.1.1)
iiiii:	Issue identifier (see Item 5.6.1.2)
pppp:	Sequential page number of first article page (see Item 5.6.1.3)
n:	Optional article sequence identifier (see Item 5.6.1.6)
.txt:	explicit file extension

5.6.7 Full-Color Article Files

All deliverable files associated with full-color articles (see Item 5.2.2.5) are to be named with a “-color” suffix appended to the filename.

<u>Full Page Images</u>

`jidiiiii-pppp-color.tif`

Illustration Images

There are NO illustration images for the types of pages

PDF Files

`jidiiiii-ppppn-color.pdf`

XML Files

`jidiiiii-ppppn-color.xml`

OCR Files

`jidiiiii-ppppn-color.txt`

5.7 Directory Structure

File/Directory	Description
`jid\`	Journal-level directory
`jid\iiiii`	Issue level directory
`jid\iiiii\adv\`	Directory of advertisement files
`jid\iiiii\adm\`	Directory of administrative content files
`jid\iiiii\art\`	Directory of article files
`jid\iiiii\cov\`	Directory of cover files
`jid\iiiii\toc\`	Directory of TOC files

5.7.1 Articles Directory

This directory contains all files related to an individual article. The files for each article are contained within a subdirectory named as the corresponding article identifier.

Each article subdirectory will contain the following:

- Full-page images (see Item 5.6.2)
- Illustration images (see Item 5.6.3)
- Composite PDF file (see Item 5.6.4)
- XML file (see Item 5.6.5)
- OCR files (see Item 5.6.6)
- Full-color article files (see Item 5.6.7)*
  *only if required

5.7.2 Table of Contents Directory

This directory contains the tables of content-related files:

- Full-page images (see Item 5.6.2)
- Bitonal PDF file (see Item 5.6.4)
- OCR files (see Item 5.6.6)

5.7.3 Administrative Content Directory

This directory contains the administrative content:

- Full-page images (see Item 5.6.2)
- Bitonal PDF file (see Item 5.6.4)
- OCR files (see Item 5.6.6)

5.7.4 Advertisement Directory

This directory contains advertising-related content:

- Full-page images (see Item 5.6.2)
- Bitonal PDF file (see Item 5.6.4)

5.7.5 Issue Covers Directory

This directory contains issue covers

- Full-page images (see Item 5.6.2)
- Bitonal or Full-Color PDF file (see Item 5.6.4)
- OCR files (see Item 5.6.6)

5.7.6 Auxiliary Files

The following files will be contained within the journal directory:

- Batch-Level Checksum File (**`301-`**`bbbb`**`.lst)`**: containing the calculated checksums for each deliverable file (see Item 1.3.8).
- Batch-Level Exclusion File (**`301-`**`bbbb`**`.exc)`**: containing the PMID of any articles that were present in the customer-supplied exclusion list but not found in the deliverable data (see Item 1.3.9)
- Issue-Level Mapping File (*`jidiiiii`***`.map)`**: containing the issue-specific mapping the Contractor-assigned page image filenames to source page number (see Item 1.3.10)
- Figure Sequence Mapping File (*`jidiiiii`***`.img)`**: containing the issue-specific mapping the illustration image filenames to figure labels and sequence (see Item 1.3.11)

6.0 Target Database: Image Specifications

6.1 Full-page Images

Resolution	600 dpi (dots per inch)
Bit-depth	1-bit Black and White
File Format	TIFF (CCITT Group 4) – monostrip
Scan Area	The full printed page including any headers and footers
Image size	As per source
Orientation	Consistent with the printed source i.e. no rotation

6.2 Grayscale Illustrations

Resolution	300 dpi
Bit-depth	8-bit Grayscale
File Format	TIFF (Packbits compression) – monostrip
Scan Area	The full illustration
Orientation	Orientation of the original printing i.e. no rotation

6.3 Color Illustrations

Resolution	300 dpi
Bit-depth	24-bit Color
File Format	TIFF (Packbits compression) – monostrip
Scan Area	The full illustration
Orientation	Orientation of the original printing i.e. no rotation

6.4 Image Coordinates

The offset coordinates relative to the bitonal image are embedded in the Image Description field of the TIFF header of each illustration TIFF file. This unambiguously identifies each cropped illustration by its position on the page.

6.5 Image Quality

The image should reflect the source page. In particular, 6.5.1 The Colors in the image should match the original in hue, balance and intensity.

6.5.2 The image should match the level of focus in the printed original and should not be marred by blurriness.

6.6 Identifying Illustrations

6.6.1 General Rule

6.6.1.1 A grayscale or a color TIFF should be made for substantive illustrations present within articles ONLY.

6.6.1.2 Non-substantive illustrations include decorative headers/footers, border or text separators; as well as other stylistic artifacts such as graphic bullets and logos.

6.6.1.3 Elements such as figure labels and pointers are included in the illustration image. These elements are converted as per the parent illustration format. For example, bitonal labels in a color illustration are converted as color and NOT as bitonal.

6.6.2 Grayscale Illustration

A grayscale illustration TIFF should be made for:

6.6.2.1 Black and white photographs.

6.6.2.2 Grayscale illustrations, defined as any illustration in which there is the appearance of gray, whether this is achieved via halftone dots, cross-hatching with fine lines, or some other method.

6.6.3 Color Illustration

A color illustration TIFF should be made for:

6.6.3.1 Color photographs

6.6.3.2 Color illustrations

6.6.4 Converted Bitonal Illustration

Converted bitonal illustrations are **not** defined as “black & white” images, e.g. line art. Instead, this definition is used to identify non-article type grayscale or color illustrations occurring in both non-article and article pages, e.g. color advertisement illustration on the same page as article text. Such illustrations are converted to bitonal using a simple threshold rather than the OCR-optimized threshold value used for the rest of the page.

6.6.5 Shaded Non-Illustration

A shaded non-illustration is defined as a non-graphic element that contains shading. For example, a table with gray-shaded cells or a gray-shaded text box. These elements are converted as grayscale illustrations within the composite PDF but are **NOT** delivered as separate grayscale illustration TIFF images (see Item 1.3.2).

Therefore, a page containing a shaded non-illustration would appear as a grayscale illustration in the composite PDF page image, but would be retained as a bitonal illustration in the bitonal TIFF page image (see Item 1.3.1). However, the shaded non-illustration will be converted as a “converted bitonal illustration” (see Item 6.6.4).

When an element is identified as a shaded non-illustration the entire element is treated as a shaded non-illustration. For example, for a table with several gray-shaded cells, the entire table is converted as a grayscale illustration in the composite PDF page image, NOT only the shaded cells in the table.

6.6.6 Grayscale Conversion of Bitonal Illustrations

There are two types of bitonal illustrations that require treatment as grayscale images. In both of these types, the illustrations are clearly bitonal with NO element of gray. However, the resultant bitonal conversion of these types is not desirable as the black portions of the illustrations obscure the white portions.

For these bitonal illustration types, they are to be treated in the same fashion as “shaded non-illustration” images (see Item 6.6.5)

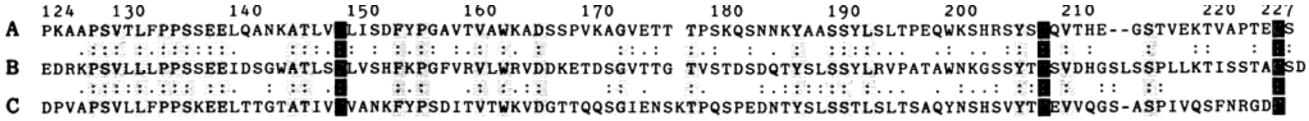
It is worth noting the impact of this treatment:

- Increased amount of illustration zoning during production
- Strong potential for subjective cases
- Inconsistent conversion of similar images (bar graph images without "tight cross-hatching" would be bitonal in the PDF, bar graph images with "tight cross-hatching" would be grayscale in the PDF)

The two types of illustrations requiring this treatment are:

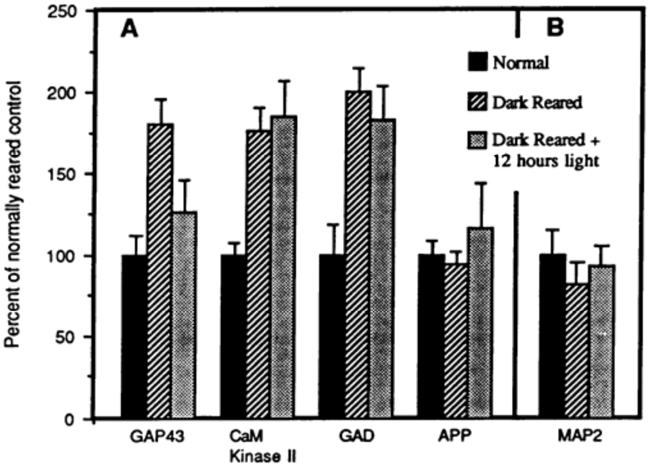
6.6.6.1 Illustrations with small instances of white text on black background

These are commonly found in gene sequence illustrations. Usually, gene sequence characters are shaded in gray for emphasis. However, in some cases, the characters are printed as white text within black boxes. In these types of illustrations, the heavy black background obscures the white characters. This problem is magnified when the source page background is not pure white, but discolored or slightly gray. When converted to bitonal, these characters are almost completely rendered as black (see example below)

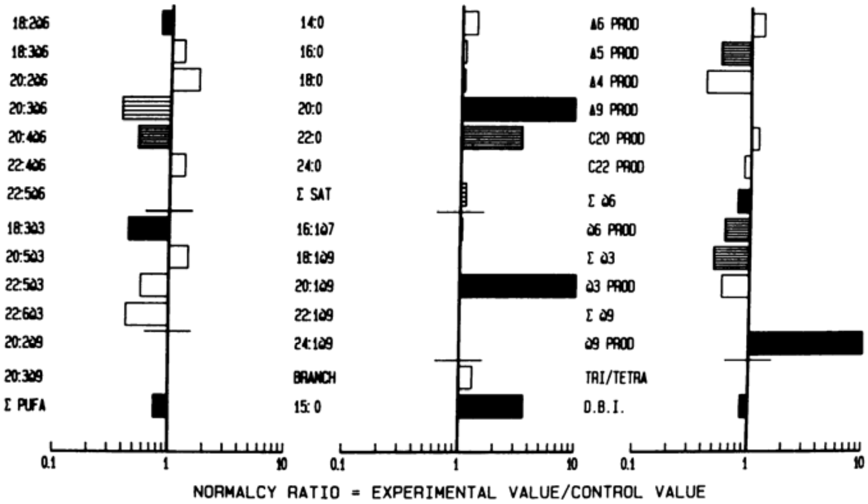

124 130 140 150 160 170 180 190 200 210 220 227
A PKAAPSVTLFPPSSEELQANKATLV LISDFYPGAVTVAWKADSSPVKAGVETT TPSKQSNNKYAASSYLSLTPEQWKSHRSYS QVTHE--GSTVEKTVAPTE S
B EDRKPSVLLLPPSSEEIDSGWATLS LVSHFKPGFVRVLWRVDDKETDSGVTTG TVSTDSDQTYSLSSYLRVPATAWNKGSSYT SVDHGSLSSPLLKTISSTA SD
C DPVAPSVLLFPPSKEELTTGTATIV VANKFYPSDITVTWKVDGTTQQSGIENSKTPQSPEDNTYSLSSTLSLTSAQYNSHSVYT EVVQGS-ASPIVQSFNRGD

6.6.6.2 Illustrations “tight” bitonal cross-hatching

Cross-hatching is frequently used in bar graphs and charts to distinguish one type of data from another. In some cases, the cross-hatching is “loose’ enough to allow clear delineation between white and black portions of the illustration. As in the example below, the middle bar within each set of bars has sufficient space between white and black portions of the illustration.



However, when the cross-hatching is “tight”, the delineation between white and black portions of the illustration is not clear enough to be rendered accurately in the bitonal image. This problem is magnified when the source page background is not pure white, but discolored or slightly gray. When converted to bitonal, these characters are almost completely rendered as black (see example below):



6.7 Multipart Figures

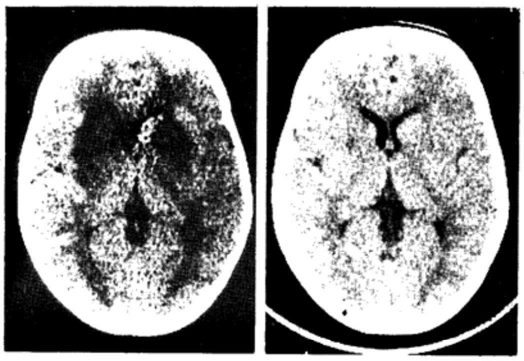
Often a figure will contain multiple parts, these multipart figures are to be captured as a single illustration image. The figure label and caption are to be used to identify whether or not a figure is a multipart figure or separate multiple figures. This structure is often reflected in the figure caption, where the single block of caption text contains delimiters for each of the figure parts.

6.7.1 Figure Label

The figure label is the primary source of figure identification. A single figure label identifies a single figure unit. The figure caption can also be used as an additional identifying marker for figures.

Directional cues are occasionally embedded within the figure caption. In such cases, the directional cues are not to be used as separate figure label identifiers. Instead, they are more similar to figure subpart identifiers, such as “Fig 1a, 1b, etc.”

In the example below, the figures are treated as a single illustration, similar to a multipart figure.



Left: initial computed tomogram shows low density areas in region of both basal ganglia and surrounding white matter. Right: a repeat scan, 10 weeks later, shows complete resolution of these areas.

6.7.2 Figure Format

6.7.2.1 Grayscale Figures

Figures that consist of grayscale figures **only** are converted as grayscale illustration images.

6.7.2.2 Color Figures

Figures that consist of color figures **only** are converted as color illustration images.

6.7.2.3 Mixed Format Figures

Figures that consist of a mixture of multiple different formats are converted according to the following guidelines:

- A figure containing **any** color figures is converted as a color illustration image
- A figure containing **any** grayscale figure, but **no** color figures is converted as a grayscale illustration image
- Where bitonal images or elements exist as a part of either a color or grayscale figure, treat them in same format as the rest of the figure (either all grayscale, or all color).

6.8 Special Case Figures

6.8.1 Overlapping figures

In cases when figures are overlapping, it is not possible to create a single rectangular image containing only the isolated figures. Therefore, it may be necessary to group multiple figures together as a single figure, as in Figs 10 and 11 below:

6.8.2 Non-rectangular figures

In cases when figures are non-rectangular, it is not possible to create a single rectangular image containing only the isolated figures. Therefore, it may be necessary to group multiple figures together as a single figure, including adjacent text, as in the example below:

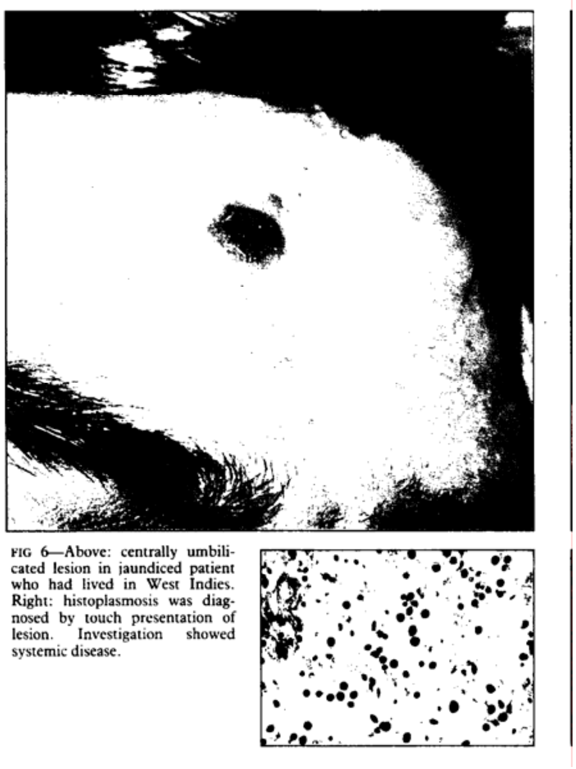


FIG 6—Above: centrally umbilicated lesion in jaundiced patient who had lived in West Indies. Right: histoplasmosis was diagnosed by touch presentation of lesion. Investigation showed systemic disease.

6.9 Pages not to be scanned

The following types of pages are not to be scanned:

6.9.1 Tissue paper or glassine leaves inserted to protect plates or illustrations, unless they contain text or are part of the pagination sequence.

6.9.2 Reader-service cards, advertising inserts, and subscription forms bound into the issue but printed on a card or other paper significantly smaller than the substantive pages of the issue, or otherwise obviously intended to be removed from the issue and used or discarded by the reader, unless the item is part of a pagination sequence or contains substantive information.

6.9.3 Blank pages, including both blank flyleaves inserted at the beginning or end of an issue or volume as well as entirely blank pages that may appear within a document. A blank page is also defined as a page that does not contain any substantive material. For example, a page containing only header, footer with no other content on the page (refer to Exhibits 15a and 15b).

6.9.4 Any non-substantive item inserted loose into the issue or volume, e.g. advertising supplements or meeting programs that are not part of a pagination sequence.

6.9.5 Any blank backside of a photocopied page, e.g. replacement pages provided by the customer or photocopied pages bound in the original source.

6.10 Article PDF Files

6.10.1 PDF with Hidden Text

All article-level PDF files except those constructed from advertisement pages will contain “hidden text” of the article. The searchable OCR text of the article will be included within the PDF file for searchability but will not be visible.

6.10.2 Article Page Types

All pages containing article text are included in a separate article-level PDF file for each article. Pages with illustrations are created as a composite of the full-page image (see Item 6.1) and individual illustration images (see Items 6.2 and 6.3). Pages appear in the article PDF file in the same sequence as they appear in the original source article, with the exception of re-sequencing for reading order. Illustrations are stored in the PDF file using JPEG compression to reduce the PDF file size.

6.10.3 Table of Contents, Administrative Content and Advertisement Page Types

All pages of table of contents, administrative content and advertisement page types are included in separate “article-level “ PDF files for each page category. Only the bitonal page images are used (see Item 6.1).

Administrative and Advertisement pages appear in the article PDF file in reading order sequence, typically the same sequence as they appear in the original source article.

6.10.4 Issue Cover pages

All Issue Cover pages are included as a single PDF file. The front cover page appears as a full-page color or grayscale image if it contains any color or grayscale material.

6.10.5 PDF Page Count Restriction

PDF files are restricted to a page count of sixty (60) pages. This limitation is generally applied to non-research articles, such as:

- Bibliographies
- Poster abstracts
- Proceedings of meetings
- Indices

Non-research articles exceeding the page count restriction are split into multiple articles according to the following guidelines.

- Use logical breakpoints (such as by day, session or topic, subject grouping, alphabetic index entry) to split article.
- If the article still exceeds the page count restriction, then the article is to be split arbitrarily at 40-page increments.

If the article exceeding the page count restriction is a research article, then the article should NOT be split.

7.0 Target Database: XML Element Description

7.1 Root Element

7.1.1 Article `<art>`

`(fm)`

The `<art>` element is the highest document-level container element.

7.1.1.1 The “`art_type`” attribute contains one of the following valid values indicating article type (refer to Exhibit 20.zip for examples of select article types).

abstract The article itself is an abstract (of a paper or presentation), usually that has been presented or published separately.

art_commentary An article whose subject or focus is another article or articles; this article comments on the other article(s). For example, with a controversial article, the editors of the publication might invite an author of the opposing opinion to comment on the first article, and publish the two together. Another article categorized as this type is a discussion, an invited discussion related to a specific article or issue (see Exhibits 19a, 19b and 19c).

book_review Review or analysis of one or more printed or online books (Note that product reviews are a separate type.)

correction A modification, or correction of previously published material (sometimes called “errata") Similar value "addendum" merely adds to previously published material

If the related article information cannot be identified, then the correction should still be typed as a “correction” article.

editorial Opinion piece, policy statement, or general commentary, typically written by staff of the publication. Note: similar value "art_commentary" is reserved for a commentary on a specific article or articles.

introduction An introduction to the publication, a series of articles within the publication, etc., typically for a special section or issue

letter Letter to the publication, typically commenting upon a published item

meeting_report Report of a conference, symposium, or meeting

news News item

other Not any of the article types explicitly named in this list Including, but not limited to articles that fit the following:

addendum: A published item that adds additional information or clarification to another item Similar value "correction" corrects an error in previously published material

announcement: Material announced in the publication (may or may not be directly related to the pub)

calendar: A list of events

in_brief: Summary of items in the current issue

obituary: Announcement of a death or appreciation of a colleague who has recently died

product review: Description, analysis, or review of a product or service, for example a software package (note that book review is a separate type)

research_art Research article or an article of the following types:

brief_report: A short and/or rapid announcement of research results

case_report: Case study, case report, or other description of a case

review article: Review or state-of-the-art summary article

retraction Retraction of previously published material

poster Poster abstract article as per Item 5.2.2.2

index List of bibliographical information or citations arranged usually in alphabetical order of some specified datum (such as author, subject, or keyword). Also, a list of items (such as topics or names) treated in a printed work that gives for each item the page number where it may be found. An index is treated as an article only when the issue is a supplement that comprises only indexes, i.e., the issue has no "regular" article content.

reply Reply to a letter or commentary, typically by the original author commenting upon the comments

If the letter being replied to cannot be identified or if multiple letters are being replied to, then the letter is typed as “letter” and not as “reply”

cover Cover of the journal issue as per Item 5.2.6

admin Administrative material as per Item 5.2.3.5

advert Advertisements as per Item 5.2.5

toc Tables of contents as per Item 5.2.3

NOTE: Filler material is extraneous short paragraphs or images found on a page. It does not relate to the subject of the articles around it. It is used to fill white space on a page. Many filler items are historical information. If a journal contains filler, it will be noted on the style sheet and examples will be provided. Filler should be captured as part of the preceding article. The filler text will be captured in the OCR text of the article only and not indicated in the article metadata (i.e. filler title or author information).

7.1.1.2 The "`lang`" attribute contains a two-character code indicating the language of the article. The list of valid language codes are available at http://www.ncbi.nlm.nih.gov/entrez/query/static/spec.html#SubsetofLanguageCodes.

The default value of this attribute is EN (English).

7.2 Front Matter Elements

7.2.1 Front Matter `<fm>`

```
(jmeta, ameta, notes?)
```

This element is the container element for metadata relevant to the article. The metadata is divided into journal-level, article-level and footnote categories.

7.3 Journal Metadata Elements

7.3.1 Journal Metadata `<jmeta>`

```
(jid*, issn*)
```

This element is the container element for journal-level metadata.

7.3.2 Journal Identifier `<jid>`

```
(#PCDATA)
```

This element contains a journal code provided by the customer used for identification (see Item 5.6.1.1). The "`jid_type`" attribute indicates the identifier type. If journal codes of multiple types are provided, then this element is repeated. There are only two valid values for this attribute:

7.3.2.1 "`pmc`" is used for identifiers assigned by PMC (PubMed Central)

7.3.2.2 "`nlm_ta`" is used for identifiers assigned by PubMed/Medline. This attribute will not be used in this project.

7.3.3 International Standard Series Number `<issn>`

```
(#PCDATA)
```

This element contains the ISSN of the journal.

7.4 Article Metadata Elements

7.4.1 Article Metadata `<ameta>`

```
(aid*, agrouping?, titlegrp?, trans-title*, (contribgrp | aff)*,
aunotes?, pubdate*, volume?, issue?, fpage, lpage?, range?,
stringrange?, relart*, cpyrt?, abs*, kwdg*)
```

This is the container element for article-level metadata.

7.4.2 Article Identifier <aid>

(#PCDATA)

This element contains the article identifier.

The "pubid_type" attribute indicates the type of article identifier. The "doi" type is used for Digital Object Identifiers, but will not be used in this project. The "pmid" type is used for PubMed article identifiers, which will be used only when provided in the customer-supplied Article Exclusion list (see Item 1.7)

The PubMed article identifier is provided in this element in the following instances:

- An article has been excluded as per the customer-supplied Article Exclusion list
- An article that is present in the customer-supplied Article Exclusion list, but the article has an abstract in the source that has not been identified in the list
- A proceedings article (see Item 5.2.2.3) contains the PubMed article ID of an article contained within the proceedings article, typically this would be "The President's Address"

7.4.3 Copyright Statement <cpyrt>

(#PCDATA | %emphasis;)*

This element is no longer captured.

7.5 Article Grouping Data Elements

7.5.1 Article Grouping Data <agrouping>

(subj_group)

This is the container element for article grouping data. Articles are often grouped together in the issue Table of Contents under distinct categories. The categories are based on the subject matter or the type of articles grouped within. In instances where multi-level grouping occurs, this element allows for only the top-level category.

Footnotes in article group headings are not captured.

7.5.2 Subject Grouping Name <subj_group>

(subject)

This is the container element for the subject of the article grouping category.

7.5.3 Subject <subject>

(#PCDATA | %emphasis;)*

This element contains the subject of the article-grouping category.

7.6 Title Group Elements

7.6.1 Title Group <titlegrp>

(atitle, subtitle?)

This is the container element for the article title.

7.6.1.1 The “`lang`” attribute contains a two-character code indicating the language of the translated title. The list of valid language codes are available at http://www.ncbi.nlm.nih.gov/entrez/query/static/spec.html#SubsetofLanguageCodes.

The default value of this attribute is EN (English).

7.6.2 Article Title `<atitle>`

`(#PCDATA | %emphasis; | xref)*`

This element contains the article title. The text and formatting of the article title is obtained from the corresponding entry in the table of contents. TOC headings are not captured as part of article titles.

7.6.2.1 Footnotes in titles

Footnote references in the article title are linked to the corresponding `<fn>` element (see Item 7.14.5) within `<notes>` (see Item 0) via the `<xref>` element (see Item 7.14.1). Footnote references are obtained from the article title as per the first page of the article. This is necessary as footnote references will not appear in the corresponding entry in the table of contents.

7.6.2.2 Titles in split articles

For articles that exceed the PDF page count restriction (see Item 6.10.5) generate this element based on the following guidelines:

- The title is as per standard title conventions, e.g. generated from the corresponding entry in the table of contents
- The subtitle (see Item 7.6.2.3) element is used to distinguish each part of the split article.

7.6.2.3 Titles in book review articles

This element contains the explicit specific title given to a book review, if present. If no explicit title is present, then this element contains the title of the first (or only) book being reviewed.

If a translated title is provided in the book review along with the original title, this does not appear as either title or translated title. Instead, the translated title is omitted.

- If the book review has an explicit title, then use this title. For example, a review article of “Ancients and Moderns in the Medical Sciences by Roger French” may be entitled as “A look at medicine through the ages”. In such a case, the explicit title would be the title of the article.
- In some instances, an explicit title for each book review will be listed as an entry in the table of contents. In these cases, use the title as it appears in the table of contents as per Item 7.6.2
- If the book review does not have an explicit title, then the title book being reviewed is the title of the article. If there is more than one book being reviewed, the use the title of the first book listed.
- If multiple book reviews are clubbed together as a single article, due to either customer-supplied style sheet requirements or the lack of individual contributors for each book review, then use the generic “Book

Reviews" as the article title and not the title of the first book listed in the first review.

7.6.2.4 Titles in letter articles

- If the letter has an explicit title, then use this title.
- In some instances, an explicit title for each letter will be listed as an entry in the table of contents. In these cases, use the title as it appears in the table of contents as per Item 7.6.2
- In some instances letters are grouped by a subject title. Each letter in the group should be captured separately and given the subject title.
- If the letter does not have an explicit title, then use the title of the section as the title for each letter, e.g. "Correspondence", "Letters to the Editor", etc.
- If multiple letters are grouped under one group title and each letter does not have an individual title, then use the group title as the title for each letter. For example, group heading of Toxic Shock Syndrome has three letters found under it with no specific title listed for each. Each of the three letters would have the title of Toxic Shock Syndrome.

  In Exhibit 21, there is a single title "Ichthyosis and Hypnosis" followed by two separate letters (one by "A.A. Mason" and the other by "John Freeman"). In this case, each letter would have the same tilte "Ichthyosis and Hypnosis"

7.6.2.5 Titles in correction articles

- If the correction article has an explicit title, mark to key the explicit title as the article title.
- If the correction article does not have an explicit title, mark to key the generic group heading (Erratum, Correction, etc.) as the article title and as the group heading.
- If the correction title does not have either an explicit title or a generic group heading, then mark to key the title of the corrected article as the article title.

7.6.3 Article Subtitle `<subtitle>`
`(#PCDATA | %emphasis; | xref)*`

This element contains the article subtitle. The text and formatting of the article subtitle is obtained from the corresponding entry in the table of contents.

This element contains the article subtitle. Footnote references in the article subtitle are linked to the corresponding `<fn>` element (see Item 7.14.5) within `<notes>` (see Item 0) via the `<xref>` element (see Item 7.14.1).

If the corresponding entry in table of contents does NOT contain the article subtitle, but the subtitle is present in the article, then the formatting of the article subtitle should match the formatting of the article title in the table of contents entry. For example, if the article title in the table of contents entry is in Title Case, but the subtitle appears in the article as UPPERCASE, then the formatting of the subtitle should be Title Case to match the formatting of the article title in the table of contents entry.

For articles that exceed the PDF page count restriction (see Item 6.10.5), generate this element based on the following guidelines:

- Use the distinguishing identifiers for logical break points (such as by day, session or topic, subject grouping, alphabetic index entry).

  For example:

  ```
  <titlegrp>
  <atitle>The American Society for Cell Biology 34th annual
  meeting. San Francisco, California, December 10-14, 1994.
  Abstracts </atitle>
  <subtitle>Sunday, Poster Sessions, Part I</subtitle>
  </titlegrp>

  <titlegrp>
  <atitle>The American Society for Cell Biology 34th annual
  meeting. San Francisco, California, December 10-14, 1994.
  Abstracts </atitle>
  <subtitle>Monday, Minisymposia</subtitle>
  </titlegrp>

  <titlegrp>
  <atitle>Cumulative Author Index</atitle>
  <subtitle>A-M</subtitle>
  </titlegrp>

  <titlegrp>
  <atitle>Subject Index, vols. 1-10</atitle>
  <subtitle>A-M</subtitle>
  </titlegrp>
  ```

- If no identifiers are available, then use the following generic content:

  ```
  Part N
  ```

  where:

`Part`:	explicit label
N:	Uppercase Roman numeral starting with "`I`"

7.6.4 Article Translated Title `<trans-title>`
`(atitle, subtitle?)`

This element is a wrapper element for translations of article titles/subtitles.

7.6.4.1 The "`lang`" attribute contains a two-character code indicating the language of the translated title. The list of valid language codes are available at http://www.ncbi.nlm.nih.gov/entrez/query/static/spec.html#SubsetofLanguageCodes.

The default value of this attribute is EN (English).

7.7 Contributor Group Elements

7.7.1 Contributor Group `<contribgrp>`

`(contrib+, etal?)`

This is the container element for article contributor information. Contributor names, both person and group contributors, appear as Initial Caps within these elements, with the exception of acronyms, e.g. FBI, CIA, etc.

All contirbutors are captured within a single <contribgrp>, even if multiple authors are linked to different affiliations in the source.

For example:

Carol P. Herbert
Department of Family Medicine, University of British Columbia, Vancouver, BC

Elizabeth A. Lindsay
Community Health Research Unit, Faculty of Medicine, University of Ottawa, Ottawa, Ont.

would appear in the XML as :

```
<contribgrp>
<contrib contrib-type="author">
<name>
  <snm>Herbert</snm>
  <gnms>Carol P.</gnms>
  </name>
</contrib>
<contrib contrib-type="author">
<name>
  <snm>Lindsay</snm>
  <gnms>Elizabeth A.</gnms>
  </name>
</contrib>
</contribgrp>
```

7.7.2 Contributor `<contrib>`

`( (collab | (name, degrees*) ), role*, xref* )*`

This element contains the information for a single contributor to the article.

- The “`contrib_type`” attribute indicates whether the contributor is either an “`author`” or “`editor`”.
- For book reviews, this element contains the name of the reviewer and NOT the name of the book author. This element does not contain the name of the reviewer, if the name appears only as initials. Additionally, if a review author name is provided only as initials, do not capture any contributor information.
- Only main authors are captured within this element. Contributors defined as providing assistance, research or support are not captured.
- If an author name is presented consisting of initials only, do not capture as contributor. Instead, the initials are omitted.

- For multiple articles grouped together as a single article (see Item 5.2.2.1), do not capture each individual author. Capture only contributors which are applicable to entire set of multiple articles grouped together, such as section editors.

7.7.3 Collaboration <collab>

(#PCDATA | %emphasis;)*

This element contains the name of a group or non-person contributor.

7.7.4 et al <etal>

(#PCDATA)

This element contains the string "et al", if present in the source.

7.7.5 Degree(s) <degrees>

(#PCDATA | %emphasis;)*

This element is no longer captured.

7.7.6 Role <role>

(#PCDATA | %emphasis;)*

This element is no longer captured.

7.8 Person Name Elements

Person names are normalized in these elements as Initial Caps, regardless of letter-casing in original source.

7.8.1 Name of Person <name>

(snm, gnms?, suf?)

This is the container element for person name information.

7.8.2 Surname <snm>

(#PCDATA | %emphasis;)*

This element contains the surname or last name of the individual. If the parts of the person name cannot be explicitly identified, then the entire name appears within this element. If only one name is present, then that name appears within this element.

7.8.3 Given Names <gnms>

(#PCDATA | %emphasis;)*

This element contains the given names, including first and middle names, of the person.

7.8.4 Suffix <suf>

(#PCDATA | %emphasis;)*

This element contains the name suffixes of a person, e.g. Sr., Jr., III, etc.

7.9 Article Enumeration Elements

7.9.1 Publication Date <pubdate>

(%date_model;)

This element contains the date of publication of the article. When publication date is provided as a range and no other single publication date is present in the source, then the **last** date of the range is used to populate the <day>, <month> <season> and/or <year> elements. The full date range appears within the <stringdate> element.

The content model of the %date_model; entity is as follows:

```
(((day?, month?) | season)?, year?, stringdate?)
```

- <day> (see Item 7.15.1)
- <month> (see Item 7.15.2)
- <season> (see Item 7.15.3)
- <year> (see Item 7.15.4)
- <stringdate> (see Item 7.15.5)

7.9.2 Volume <volume>
(#PCDATA | %emphasis;)*

This element contains the volume number of the issue containing the article. Roman numerals in the source appear as corresponding Arabic numerals in this element.

The volume information is typically located on the issue cover (see Item 4.5.4). However, if a cover is not available, use the PubMed records from the Article Exclusion List (see Item 1.7 for the same issue to obtain the volume information

7.9.3 Issue <issue>
(#PCDATA | %emphasis;)*

This element contains the issue number of the issue containing the article. Roman numerals in the source appear as corresponding Arabic numerals in this element.

The issue information is typically located on the issue cover (see Item 4.5.4). However, if a cover is not available, use the PubMed records from the Article Exclusion List (see Item 1.7 for the same issue to obtain the issue information

- Often, the issue number will be readily available in the source. Case-specific resolutions will be provided whenever the issue number is not explicit in the source.

- The word “and” appearing in paired issue numbers (e.g. 1 and 2) appears as a hyphen in this element (e.g. 1-2)

- Part and supplement designators appear in this element as “Pt.” and “Suppl.”, respectively.

- Supplement issues may have the issue information indicated as a supplement to a previous issue, i.e “Supplement to Vol. 5, Iss. 15”. However, in a minority of cases, the supplement would simply point to a publication date of the “parent” issue, i.e. “Supplement to July/Aug 1987”. In these cases, identify the “parent” issue in the source received. Typically, it would be in the same batch. Once identified, capture the issue number from the “parent” issue accordingly.

7.9.4 First Page <fpage>
(#PCDATA)

This element contains the page number of the first page of the article. The value of the optional “seq” attribute is a sequence letter used to distinguish multiple articles beginning on the same page. For example, two articles that begin on page “25” will have the following <fpage> elements:

```
<fpage seq="a">25</fpage>
```

```
<fpage seq="b">25</fpage>
```

Pages without printed page numbers are handled as per the source page guidelines.

7.9.5 Last Page `<lpage>`
`(#PCDATA)`

This element contains the page number of the last page of the article.

Pages without printed page numbers are handled as per the source page guidelines

7.9.6 Page Range `<range>`
`(subrange+)`

This optional element is used only when the article pages are in a non-contiguous sequence. This element is a wrapper element for the empty <subrange> element.

7.9.7 Page Subrange `<subrange>`
`(EMPTY)`

This empty element is repeatable and can represent multiple non-contiguous page ranges. This element has attributes to indicate the page numbers in the range.

7.9.7.1 begin

This attribute contains the page number of the first page in the range. If this attribute is present, then the “end” attribute must also be present.

7.9.7.2 end

This attribute contains the page number of the last page in the range. If this attribute is present, then the “begin” attribute must also be present.

7.9.7.3 single

This attribute contains the page number when only a single page is contained within the range. If this attribute is present, then neither “begin” nor “end” attributes can be present.

The following are examples of how this element is used for different types of pagination:

For the pagination “14-16, 19-22”:

```
<fpage>14</fpage>
<lpage>22</lpage>
<range>
<subrange begin="14" end="16"/>
<subrange begin="19" end="22"/>
</range>
```

For “146, 148”

```
<fpage>146</fpage>
<lpage>148</lpage>
<range>
<subrange single="146"/>
<subrange single="148"/>
</range>
```

For “100-101, 105, 107-120”

```
<fpage>100</fpage>
<lpage>120</lpage>
<range>
<subrange begin="100" end="101"/>
<subrange single="105"/>
<subrange begin="107" end="120"/>
</range>
```

7.10 Errata/Correction/Retraction Element

A correction is defined as a correction pertaining to a specific article. Therefore, a correction may have more than one item of correction, but if all pertains to a specific article, then it is considered as a single correction.

Each correction is marked to be keyed as a separate article. A single item can be identified by the corresponding volume and page citation in the correction. The related article should only be tagged for articles in the SAME journal.

Some articles may be followed with a “Note Added in Proof” or “Addendum”. These are corrections made to the immediately preceding article. They are not true corrections, as these are corrections made for the current article (not a previously published article). Therefore, these are NOT treated as separate errata articles, but are grouped with the preceding article as part of the article text.

7.10.1 Related Article `<relart>` `(EMPTY)`

This is an empty element containing reference information for article corrections (see Item 4.9).

7.10.1.1 The value of the “`vol`” attribute is the volume number of the corrected article. The content of this attribute must be formatted as Arabic numerals.

7.10.1.2 The value of the “`fpage`” attribute is the first page number of the corrected article.

7.10.1.3 The value of “`seq`” attribute is the alphabetic identifier indicating the sequence of the article, when more than one article appears on the page. The alphabetic identifier is similar to the article sequence identifier (see Item 5.6.1.6).

This attribute is populated **only** when the `<relart>` is used to link a “reply” letter to the original letter (see Item 7.1.1.1) **and** both letter and reply must appear within the same issue.

7.10.1.4 The value of the “`type`” attribute is an indicator of the type of correction

The <relart> element has a "type" attribute with the following allowed values. The "type" attribute indicates the nature of the related article containing the <relart> element.

- `corrected.art`:

  The corrected.art is used to identify the article that contains the correction. It is used when the erratum consists of only the correction portion of the article.

- `original.art`:

The original.art is used to identify the original article that is now republished. This is used when there are references in the Erratum to both the original article **and** the republished article.

- `republication`:

  The republication is used to identify a republished article. The article is republished as a correction to the original article. This type is used only when the original article has been corrected and republished in its entirety as indicated by the Erratum.

- `retracted.art`:

  The retracted.art is used to identify an article that has been retracted. This indicates that the article has been "removed" from consideration. This may be used in conjunction with republication, if the article is subsequently republished.

- `articleref`:

  The articleref is used when the Erratum does not fit any of the types described above.

7.11 Abstract Elements

7.11.1 Abstract <abs>

`(p | sec)*`

This element contains the article abstract. The abstract is a brief description of the subject and contents of the article.

7.11.2 Section <sec>

`(title, (p | sec)*)`

This element contains the sections of an abstract. Occasionally, an abstract will be divided into subject-based sections.

7.11.3 Title <title>

`(#PCDATA | %emphasis;)*`

This element contains the title of an individual section. Common section titles include, "Materials and Methods", "Summary", and "Conclusions".

7.12 Keyword Elements

7.12.1 Keyword Group <kwdg>

`(kwd+)`

This element contains a single set of article keywords.

7.12.2 Keyword <kwd>

`(#PCDATA | %emphasis;)*`

This element contains a single article keyword.

7.13 Paragraph Elements

7.13.1 Paragraph <p>

(#PCDATA | %emphasis; | email | xref)*

This element contains a single paragraph of text.

7.13.2 Email Address <email>

(#PCDATA)

This element contains an individual email address.

7.14 Cross Reference Elements

7.14.1 Cross-Reference <xref>

(#PCDATA | %emphasis;)*

This element is used to contain cross-reference pointers and establish a link to the referenced element.

The “ref_type” attribute indicates the type of element being cross-referenced, either “aff” or “fn” (see Items 7.14.2 and 7.14.5). The “rid” attribute contains the unique id value of the referenced element. Cross-reference pointers contained within this element are not tagged within superscript element (see Item 7.16).

7.14.2 Affiliation <aff>

(#PCDATA | %emphasis; | xref)*

This element is no longer captured.

7.14.3 Notes <notes>

(fn+)

This element contains footnotes pertaining to the entire article. Article-level footnotes include article title footnotes. The value of the “notes_type” attribute is always “footnotes”.

Author-related footnotes are NOT contained within this element. Instead, author-related footnotes are contained within the <aunotes> element (see Item 7.14.4).

7.14.4 Author Notes <aunotes>

(fn+)

This element is no longer captured.

7.14.5 Footnote <fn>

(p+)

This element contains the footnote text of either the article-related <notes> (see Item 0) element.

The value of the “id” attribute is the unique identifier for the footnote text. This “id” is used for linking to <xref> pointers (see Item 7.14.1)). The format of the “id” attribute is the prefix “fn” for footnotes or “an” for author notes followed by a sequential number unique within the article.

The value of the “symbol” attribute is the string used as the reference pointer in the text. The symbol character itself does not appear as part of the footnote text.

Only capture the footnotes that have superscripts associated with the title of an article.

7.15 Date Elements

7.15.1 Day `<day>`

`(#PCDATA)`

This element contains the two-digit day of the month.

7.15.2 Month `<month>`

`(#PCDATA)`

This element contains the two-digit month of the year.

7.15.3 Season `<season>`

`(#PCDATA)`

This element contains the name of the season.

7.15.4 Year `<year>`

`(#PCDATA)`

This element contains the 4-digit year

7.15.5 Date as a String `<stringdate>`

`(#PCDATA)`

This element contains the date “spelled-out” as a single string, as it appears in the source.

Although the stringdate element is optional in the DTD, the contractor will always supply this element with the date as it appears on the source. The other date elements will be provided as well whenever possible.

In this element, only the following characters are allowed. Any character not appearing in this list is converted as an empty space:

- Alphanumeric characters (A-Z, 0-9)
- Empty space
- Comma (,)
- Hyphen (-)
- Period (.)
- Slash (/)

For supplement issues, the date on the cover may be the date of the conference proceeding covered in the supplement. In these cases, the year of publication should be used as the date. Identify the year of publication by reviewing the year information of issues preceding or following the supplement issue.

7.16 Emphasis Elements

All emphasis (defined as bold, italic, underline and small caps) are captured ONLY in abstract, if the emphasis is not consistent, Do NOT capture emphasis within any other element, particularly within article titles.

For example, an abstract entirely in italics would not have italics captured. However, for an abstract, in which some words are italicized and the rest not-italicized, the italicized words would be captured with italic emphasis. If an abstract is entirely italicized with the exception of a few key words, then this is a case of reverse emphasis. In such a case, the majority italicized text would be treated as normal, whereas the key words that are not italicized in the source would be treated as italicized.

- Bold `<b> (#PCDATA | %emphasis;)*`
- Italic `<it> (#PCDATA | %emphasis;)*`
- Small Caps `<sc> (#PCDATA | %emphasis;)*`
- Subscript `<sub> (#PCDATA | %emphasis;)*`
- Superscript `<sup> (#PCDATA | %emphasis;)*`
- Underline `<ul> (#PCDATA | %emphasis;)*`

8.0 Appendix A: Archive Header DTD

```
<!-- ============================================================= -->
<!--  TITLE:     NLM Archive Header  DTD                           -->
<!--  VERSION:   1.1                                               -->
<!--  DATE:      September 12, 2003                                -->
<!--                                                               -->
<!-- ============================================================= -->
<!-- ============================================================= -->
<!--                   CHANGE HISTORY                              -->
<!-- ============================================================= -->
<!--
      May 21, 2003
            1. Added <xref> to <aff>

      March 17, 2003
            1. Removed <history>
            2. Removed <publisher>
            3. Added <relart> for storing related article information
               for Errata (corrections and retractions)

      September 12, 2003
            1. Added <subtitle> to <titlegrp>

      January 23, 2004
            1. Added value "poster" to @art_type
            2. Added lang attribute to <art>
            3. Added <trans-title> and <trans-abs> to <ameta>, both
               with a required lang attribute.

      January 29, 2004
            1. Added optional lang attribute to <titlegrp> and <abs>

      February 3, 2004
            1. Added type attribute to <relart> with default value
               of 'original.art'

      February 5, 2004
            1. Removed optional lang attribute from <abs>
            2. removed <trans-abs>
            3. Added <xpage> with @begin and @end

      February 12, 2004
            1. Removed <xpage>
            2. Added <range> and <subrange>
            3. Added @single, @begin, and @end to <subrange>
            4. Added <stringrange>

      March 23, 2004
            1. Added 'articleref' to @type list for <relart>
      July 9, 2004
            1. Added "reply" as a value for @art_type.
      2. Added @seq to <relart> for page seqence letters

      June 14, 2006
         1. Added xref to subject to allow for footnotes to subject.
                                                                      -->
<!-- ============================================================ -->
<!--                    PARAMETER ENTITIES                         -->
<!-- ============================================================ -->
<!--                    EMPHASIS/RENDITION ELEMENTS                -->
<!ENTITY % emphasis "b | it | sc | sub | sup | ul">
<!ENTITY % lang-reqired "lang             CDATA             #REQUIRED">
<!ENTITY % lang-optional "lang                  CDATA             #IMPLIED">
<!-- The lang attribute indicates that language of the contents
     of a given element.

        When the lang attribute is not used for article, it will be
        assumed that the language is English.

        All abstracts should be tagged in English. Only English

        Values should be from the ISO 639 standard for language codes.
        A subset of this list is available here:

http://www.ncbi.nlm.nih.gov/entrez/query/static/spec.html#SubsetofLanguageCodes

        EN  English
        ZH  Chinese
        EO  Esperanto
        FR  French
        DE  German
        JA  Japanese
        RU  Russian
        ES  Spanish


        An article in English that has an English title and a Spanish
        title does not need the @lang on <art>. The Spanish
        title should be tagged as <trans-title lang="SP">.

        An article in French that has an English title and abstract
        and a French title should be tagged as <art lang="FR">,
        <titlegrp>(French title),
        <trans-title lang="EN">, and <abs>(English abstract).     -->
<!--                    DATE ELEMENTS MODEL                        -->
<!--                    The content models for elements that describe
                        dates, such as Publication Date <pubdate>.
                                                The <stringdate> element holds
dates for
                                                which months and years are not
given, for
```

```
                                                example "first quarter", "spring",
etc.    -->
<!ENTITY % date_model "(((day?, month?) | season)?,
                          year?, stringdate?)">
<!-- ============================================================= -->
<!--                    ROOT ELEMENT                               -->
<!-- ============================================================= -->
<!--                    ARTICLE                                    -->
<!--                    For the archive headers, the article
                        consists of only one element: frontmatter. -->
<!ELEMENT art (fm)>
<!ATTLIST art
   art_type (abstract | art_commentary | book_review | correction | editorial |
introduction | letter | meeting_report | news | other | poster | research_art |
reply | retraction | cover | admin | advert | toc | index | filler) #IMPLIED
   %lang-optional;
>
<!--         art_type   What kind of article is this?
                        Note: When the article is a commentary on
                        another article, for example a correction or
                        addendum, this attribute is metadata for the
                        commentary itself, it does NOT define the
                        kind of article that is being corrected or
                        amended.
                        Authoring Note: All articles
                        should have types assigned if possible.
                          abstract   The article itself is an
                                     abstract (of a paper or
                                     presentation), usually that
                                     has been presented or published
                                     separately.
                          art_commentary
                                     An item whose subject or focus
                                     is another article or articles;
                                     this article comments on the
                                     other article(s) (For example,
                                     for a controversial article, the
                                     editors of the publication
                                     might invite an author of the
                                     opposing opinion to comment on
                                     the first article, and publish
                                     the two together.) Or a
                                     discussion (Invited discussion
                                     related to a specific article
                                     or issue).
                          book_review
                                     Review or analysis of one or more
                                     printed or online books (Note
                                     that product reviews are a
                                     separate type.)
                          correction A modification, or
                                     correction of previously
                                     published material (sometimes
```

called "errata") Similar value
"addendum" merely adds to
previously published material

editorial Opinion piece, policy statement,
or general commentary, typically
written by staff of the
publication. Note: similar value
"art_commentary" is reserved
for a commentary on a specific
article or articles.

introduction
An introduction to the
publication, a series of articles
within the publication, etc.,
typically for a special section
or issue

letter Letter to the publication,
typically commenting upon a
published item.

meeting_report
Report of a conference,
symposium, or meeting

news News item

other Not any of the article types
explicitly named in this list
Including, but not limited to
articles that fit the following:

research_art
Research article or an article
of the following types:

brief_report:
A short and/or rapid announcement
of research results

case_report:
Case study, case report, or other
description of a case

review article
Review or state-of-the-art
summary article

reply
a reply
to a letter. Sometimes
listed as
"Author's reply".

retraction Retraction of previously

```
                                      published material           -->
<!-- ============================================================= -->
<!--                    FRONT MATTER ELEMENTS                      -->
<!-- ============================================================= -->
<!--                    FRONT MATTER                               -->
<!--                    The metadata concerning an article, such as
                        the name and issue of the journal in which it
                        appears and the name and author(s) of the
                        article.
                        In some journal DTDs this is called the
                        header information, and it includes metadata
                        concerning the journal <jmeta> and metadata
                        concerning the issue of the journal and the
                        individual article <ameta>.                -->
<!ELEMENT fm (jmeta, ameta, notes?)>
<!-- ============================================================= -->
<!--                    JOURNAL METADATA                           -->
<!-- ============================================================= -->
<!--                    JOURNAL METADATA                           -->
<!--                    Metadata that identifies the journal in which
                        the article was published                  -->
<!ELEMENT jmeta (jid*, issn*)>
<!-- ============================================================= -->
<!--                    JOURNAL METADATA ELEMENTS                  -->
<!-- ============================================================= -->
<!--                    JOURNAL IDENTIFIER                         -->
<!--                    Short code that represents the journal; used
                        as an alternative to or short form of the
                        journal title; used for identification of
                        the journal domain.                        -->
<!ELEMENT jid (#PCDATA)>
<!ATTLIST jid
   jid_type (pmc | nlm_ta) #REQUIRED
>
<!--         jid_type   Indicates whose identifier this is, for
                        example, "pubid" for a publisher's
                        identifier or "pmc",  Values include:
                          pmc    Identifier assigned by pmc, for
                                 example, the PubMed Central journal
                                 abbreviation such as "pnas", "mbc",
                                 "nar", "molcellb", which may be the
                                 same as the abbreviated journal
                                 title
                          nlm_ta Identifier assigned by the
                                 PubMed/Medline, and is typically
                                 the journal abbreviation, for
                                 example, "Mol Biol Cell", "Nucleic
                                 Acids Res", which may be the
                                 same as the abbreviated journal
                                 title.                            -->
<!--                    International Standard Series Number       -->
<!ELEMENT issn (#PCDATA)>
<!--                    NOTES                                      -->
```

```
<!--                    A container element for the article-level
                        footnotes. (Any footnotes to authors should
                        be in the <aunotes> element.)              -->
<!ELEMENT notes (fn+)>
<!--         notes_type To identify the type of note. This should be
                        "footnotes." The attribute does not have to
                        be added in the XML.                       -->
<!ATTLIST notes
   notes_type (footnotes) "footnotes"
>
<!-- ============================================================= -->
<!--                    ARTICLE METADATA                           -->
<!-- ============================================================= -->
<!--                    ARTICLE METADATA                           -->
<!--                    Metadata that identifies this article      -->
<!ELEMENT ameta (aid*, agrouping?, titlegrp?, trans-title*, (contribgrp | aff)*,
aunotes?, pubdate*, volume?, issue?, fpage, lpage?, range?, stringrange?, relart*,
cpyrt?, abs*, kwdg*)>
<!-- ============================================================= -->
<!--                    ARTICLE IDENTIFIER                         -->
<!-- ============================================================= -->
<!--                    ARTICLE IDENTIFIER                         -->
<!ELEMENT aid (#PCDATA)>
<!--         pubid_type  This is either "pmid" for a PubMed ID or
                         "doi."

                          doi    - Digital Object Identifier
                          pmid   - PUBMED ID (see
                                   www.ncbi.nlm.nih.gov/entrez/
                                   query.fcgi?db=PubMed)           -->
<!ATTLIST aid
   pubid_type (pmid | doi) #IMPLIED
>
<!-- ============================================================= -->
<!--                    ARTICLE GROUPING DATA (ARTICLE METADATA)   -->
<!-- ============================================================= -->
<!--                    ARTICLE GROUPING DATA                      -->
<!--                    Container for elements that may be used to
                        group articles into related clusters       -->
<!ELEMENT agrouping (subj_group)>
<!--                    GROUPING ARTICLES IN TITLED CATEGORIES
                        For some journals, articles are grouped into
                        categories, with the category indicated in
                        the article's display.
                        Sometimes the grouping or category refers
                        to the type of article, such as "Essay",
                        "Commentary", or "Article".  Sometimes the
                        grouping refers to subject areas, such as
                        "Physical Sciences", "Biological Sciences",
                        or "Social Sciences". Sometimes the grouping
                        refers to topics within the larger subject
                        areas, such as "Applied Math", "Biology", or
                        "Chemistry".
                        In a printed journal as well as on the PMC
                        website, articles may be grouped or arranged
                        under these headings (here are all the
                        Essays, here are all the Biology articles,
                        etc.) Some journals divide articles into
                        three layers of grouping, some into two, and
                        some into only one.

                        For the Archive headers, we will use only the
                        top level.

                        A one-level grouping will be
                         <subj_group>
                          <subject>Retraction</subject>
                         </subj_group>
                        or, alternatively
                         <subj_group>
                          <subject>Essay</subject>
                         </subj_group>                             -->
<!ELEMENT subj_group (subject)>
<!--                    SUBJECT GROUPING NAME                      -->
<!--                    The name of one of the subject groups used
                        to describe an article.  Such groups are
                        used, typically, to provide headings for
                        groups of articles in a printed or online
                        generated Table of Contents.               -->
<!ELEMENT subject (#PCDATA | %emphasis; | xref)*>
<!-- ============================================================= -->
<!--                    TITLE GROUP ELEMENTS (BIBLIOGRAPHIC)       -->
<!-- ============================================================= -->
<!--                    TITLE GROUP                                -->
<!--                    Wrapper element to hold the various article
                        titles.
                        A footnote referenced in the title should
                        appear in <notes> at the end of <fm>.      -->
<!ELEMENT titlegrp (atitle, subtitle?)>
<!ATTLIST titlegrp
   %lang-optional;
>
<!--                    TRANSLATED TITLE GROUP                     -->
<!--                    Wrapper element to hold a translation of the
                        article title.

                                                The lang attribute must appear and
should
                                                appear on <titlegrp> when <trans-
title>
                                                is present
-->
<!ELEMENT trans-title (atitle, subtitle?)>
<!ATTLIST trans-title
   %lang-reqired;
```

```
>
<!--                    ARTICLE TITLE                              -->
<!ELEMENT atitle (#PCDATA | %emphasis; | xref)*>
<!--                    ARTICLE SUBTITLE                           -->
<!ELEMENT subtitle (#PCDATA | %emphasis; | xref)*>
<!-- ============================================================= -->
<!--                    AUTHOR AND EDITOR GROUP ELEMENTS           -->
<!-- ============================================================= -->
<!--                    CONTRIBUTOR GROUP                          -->
<!--                    Wrapper element for information concerning
                        a grouping of contributors, such as primary
                        authors                                    -->
<!ELEMENT contribgrp (contrib+, etal?)>
<!--                    CONTRIBUTOR                                -->
<!--                    Wrapper element to contain the information
                        about a single contributor, for example an
                        author or editor.
                                                                   -->
<!ELEMENT contrib ((collab | (name, degrees*)), role*, xref*)*>
<!--         contrib-type
                        What was the contribution of this person,
                        for example: author or editor.             -->
<!ATTLIST contrib
   contrib-type (author | editor) "author"
>
<!ELEMENT collab (#PCDATA | %emphasis;)*>
<!--                    ET AL                                      -->
<!--                    This element should only be used when the
                        text "et al" appeared in the text. All
                        authors should be tagged.  -->
<!ELEMENT etal (#PCDATA)>
<!--                    DEGREE(S)                                  -->
<!--                    Academic degrees or professional
                        certifications                             -->
<!ELEMENT degrees (#PCDATA | %emphasis;)*>
<!--                    ROLE OR FUNCTION TITLE OF CONTRIBUTOR      -->
<!--                    A title or the role of a contributor
                        (such as an author) in this work. For example,
                        Editor-in-Chief, Contributor, Chief
                        Scientist, Photographer, Research Associate,
                        etc.
                        Remarks: Information on the role or type of
                        contribution is collected in two places,
                        in the "contrib_type" attribute on the
                        Contributor element and in the Role element.
                        For example, the Contributor attribute might
                        have a value of "editor", while the content
                        of the role element could be "Associate
                        Editor". As another example, the contributor
                        attribute might be "author" and the role
                        element might be "Principle Author".
                        The <role> element is also more likely to
                        appear on screen or in print than the
                        contributor attribute value.               -->
<!ELEMENT role (#PCDATA | %emphasis;)*>
<!--                    AFFILIATION                                -->
<!--                    Name of a institution or organization such as
                        a university or corporation.
                        Authoring and Conversion Note: In a typical
                        case, the "id" attribute will be pointed to
                        by one or more contributors.
                        Conversion Note: Any explicitly tagged numbers
                        or symbols for author linkages should be
                        discarded, as the linkage will be recreated
                        from the "id" connection.                  -->
<!ELEMENT aff (#PCDATA | %emphasis; | xref)*>
<!ATTLIST aff
   id ID #IMPLIED
>
<!--         id         Unique identifier so that the affiliated
                        institution may be referenced, for example
                        by a contributor                           -->
<!-- ============================================================= -->
<!--                    PERSON'S NAME ELEMENTS (BIBLIOGRAPHIC)     -->
<!-- ============================================================= -->
<!--                    NAME OF PERSON                             -->
<!--                    Wrapper element for personal names.
                        Authoring or Conversion Note: If the name
                        parts are unknown or untagged, names should
                        be placed within the Surname element <snm>.
                        Design Note: The tag abuse of overloading the
                        Surname tag is likely to lead to better
                        searching in a repository than merely
                        leaving the person's name untagged.        -->
<!ELEMENT name (snm, gnms?, suf?)>
<!--                    SURNAME                                    -->
<!--                    The surname of an individual.  If there is
                        only one name, for example, "Cher" or
                        "Pele", that is considered to be a surname
                        for consistency purposes.                  -->
<!ELEMENT snm (#PCDATA | %emphasis;)*>
<!--                    GIVEN (FIRST) NAMES                        -->
<!--                    Includes all given names for a person, such
                        as the first name, middle names, maiden
                        name if used as part of the married name,
                        etc.)                                      -->
<!ELEMENT gnms (#PCDATA | %emphasis;)*>
<!--                    SUFFIX                                     -->
<!--                    Text used as a suffix to a person's name, for
                        example: Sr. Jr. III, 3rd                  -->
<!ELEMENT suf (#PCDATA | %emphasis;)*>
<!--                    AUTHOR NOTE GROUP                          -->
<!--                    Footnotes to authors or notes about authors
                        (and, potentially other contributors) are
                        collected in the Author note group.
                        References to these footnotes are made
                        using the <xref> element.                 -->
<!ELEMENT aunotes (fn+)>
<!--                    PUBLICATION DATE                           -->
<!--                    Date of publication or release of the
                        material in one particular format.         -->
<!ELEMENT pubdate (%date_model;)>
<!--                    VOLUME                                     -->
<!ELEMENT volume (#PCDATA | %emphasis;)*>
<!--                    ISSUE                                      -->
<!ELEMENT issue (#PCDATA | %emphasis;)*>
<!--                    FIRST PAGE                                 -->
<!--                    The page number on which the article starts,
                        for print journals that have page numbers  -->
<!ELEMENT fpage (#PCDATA)>
<!ATTLIST fpage
   seq CDATA #IMPLIED
>
<!--         seq        Used for sequence number or letter for
                        journals (such as continuous makeup journals)
                        with more than one article starting on the
                        same page.

                        Use seq="a", seq="b", etc. to define articles
                        that begin on the same <fpage>             -->
<!--                    LAST PAGE                                  -->
<!--                    The page number on which the article ends,
                        for print journals that have page numbers  -->
<!ELEMENT lpage (#PCDATA)>


<!--                                      PAGE RANGE -->
<!--                                      Used to specify pagination if not
continuous.
                        Each <range> should contain at least 2
                        <subrange> elements. If attribute begin is used
                        then attribute end should be specified. If
                        attribute single is used, no other attributes
                        should be specified on that subrange. See
                        examples.

                        For the pagination 14-16, 19-22:
                              <fpage>14</fpage>
                              <lpage>22</lpage>
                              <range>
                                 <subrange begin="14" end="16"/>
                                 <subrange begin="19" end="22"/>
                              </range>

                        For 146, 148
                              <fpage>146</fpage>
                              <lpage>148</lpage>
                              <range>
                                 <subrange single="146"/>
                                 <subrange single="148"/>
                              </range>

                        For 100-101, 105, 107-120
                              <fpage>100</fpage>
                              <lpage>120</lpage>
                              <range>
                                 <subrange begin="100" end="101"/>
                                 <subrange single="105"/>
                                 <subrange begin="107" end="120"/>
                              </range>
                   -->

<!ELEMENT range        (subrange+) >
<!ELEMENT subrange     EMPTY       >
<!ATTLIST subrange
      begin       CDATA       #IMPLIED
      end         CDATA       #IMPLIED
      single            CDATA       #IMPLIED
      >

<!--                                      STRING RANGE -->
<!--                    PAGE RANGE AS A STRING                           -->
<!--                    This is a representation of the page range as a
                        string. Used for pages which are not continuous and
                                                contain text. <stringrange> is NOT
a substitute for
                                                the coding of pages using <fpage>,
<lpage>, or <range>.
                                                It should only be used when
additional information
                                                text) must be captured with the
page range information.

                                                For the pagination: 4-5, Author's
reply 6:
                                                <fpage>4</fpage>
                                                <lpage>6</lpage>
                                                <stringrange>4-5, Author's reply
6</stringrange>

                                                For 22-25; Discussion 30:
                                                <fpage>22</fpage>
                                                <lpage>30</lpage>
                                                <range>
                                                  <subrange begin="22 end="25"/>
                                                  <subrange single="30"/>
                                                </range>
                                                <stringrange>22-25; Discussion
30</strinrange> -->

<!ELEMENT stringrange (#PCDATA)>
```

```
<!--                    COPYRIGHT STATEMENT                        -->
<!--                    Copyright notice or statement, suitable for
                        printing or display.                       -->
<!ELEMENT cpyrt (#PCDATA | %emphasis;)*>
<!-- ============================================================= -->
<!--             RELATED ARTICLE INFORMATION                       -->
<!-- ============================================================= -->
<!--                    RELATED ARTICLE INFORMATION                -->
<!--                    This tag is used in Errata (/art/@art_type=
                        "correction" or "retraction") to store
                                                bibliographic information
pertaining to the
                                                corrected article(s).

                                                There are two required attributes:

                                                vol  -  the volume number of the
target
                                                        (corrected) article.

                                                fpage - the first page number of
the target
                                                        (corrected) article.

                                                type - the type of link or article
being
                                                       linked to.

                                                             corrected.art  - use
for corrections
                                                             retracted.art  - use
for retractions
                                                             republication  - use
to point to a new

printing of an article
                                                             original.art   - use
for the original

printing of a repub-

            lished article

-->
<!ELEMENT relart EMPTY>
<!ATTLIST relart
   vol CDATA #REQUIRED
   fpage CDATA #REQUIRED
   type (corrected.art | retracted.art | original.art | republication | articleref)
"original.art"
```

```
      pmid  CDATA #IMPLIED
        seq     CDATA   #IMPLIED
>
<!-- ============================================================= -->
<!--                    ABSTRACT                                   -->
<!-- ============================================================= -->
<!--                    ABSTRACT                                   -->
<!ELEMENT abs (p | sec)*>

<!ELEMENT sec (title, (p | sec)*)>
<!ELEMENT title (#PCDATA | %emphasis;)*>
<!-- ============================================================= -->
<!--                    KEYWORD ELEMENTS                           -->
<!-- ============================================================= -->
<!--                    KEYWORD GROUP                              -->
<!--                    Container element for one set of keywords
                        used to describe a document.               -->
<!ELEMENT kwdg (kwd+)>
<!--                    KEYWORD                                    -->
<!--                    One subject term, critical expression, key
                        phrase, abbreviation, indexing word, etc.
                        that is associated with the whole document
                        and can be used for identification and
                        indexing purposes.
                        There maybe several sets of keywords,
                        identified by language or vocabulary source
                        at the Keyword Group level <kwdg>.
                        Conversion Note: Keywords are not allowed to
                        nest.                                      -->
<!ELEMENT kwd (#PCDATA | %emphasis;)*>
<!-- ============================================================= -->
<!--                    PARAGRAPH-LEVEL ELEMENTS                   -->
<!-- ============================================================= -->
<!--                    PARAGRAPH                                  -->
<!ELEMENT p (#PCDATA | %emphasis; | email | xref)*>
<!ELEMENT email (#PCDATA)>
<!--                    X(CROSS) REFERENCE                         -->
<!--                    Used for any kind of internal article
                        referencing. The content of the reference
                        (if present) will be displayed as the link.
                        This element may be used to anything that
                        has an "id".  The "ref-type" attribute says
                        what the reference is pointing to.         -->
<!ELEMENT xref (#PCDATA | %emphasis;)*>
<!ATTLIST xref
   ref_type (fn | aff | aunote) #IMPLIED
   rid IDREFS #IMPLIED
>
<!-- ============================================================= -->
<!--                       FOOTNOTES                               -->
<!-- ============================================================= -->
<!--                    FOOTNOTE                                   -->
<!--                    Footnotes can appear in <aunotes> when they
```

```
                        contain information related to authors or
                        in <notes> when they refer to other elements
                        (title, affiliations, abstract).          -->
<!ELEMENT fn (p+)>
<!--         id         Unique identifier for the element. A cross-
                        reference will point to this ID.
             symbol     The footnote symbol entered as a keyboard
                        character (*,@,#) or a character entity
                        (&dagger;, &sect;, &para;).               -->
<!ATTLIST fn
   id ID #IMPLIED
   symbol CDATA #IMPLIED
>
<!-- ============================================================ -->
<!--                    DATE ELEMENTS                             -->
<!-- ============================================================ -->
<!--                    DATE                                      -->
<!--                    The elements <day>, <month>, and <year> should
                        ALWAYS be numeric values. The date may be
                        represented as a string in <stringdate>, but
                        the numeric values should be present whenever
                        possible.                                 -->
<!ELEMENT date (%date_model;)>
<!ATTLIST date
   date_type (accepted | received | rev_request | rev_recd) #IMPLIED
>
<!--         date_type  Attribute should only be used if the date
                        is one of the known types, otherwise omit
                        the attribute. Values are:
                          accepted    _ Date manuscript was
                                        accepted
                          received    _ Date manuscript received
                          rev_request _ Date revisions were
                                        requested or manuscript
                                        was returned
                          rev_recd    _ Date revised manuscript
                                        was received              -->
<!--                    DAY                                       -->
<!--                    The numeric value of a day of the month, used
                        in both article metadata and inside a citation,
                        in two digits as it would be stated in the "DD"
                        in an international date format YYYY-MM-DD, for
                        example "03", "25".                       -->
<!ELEMENT day (#PCDATA)>
<!--                    MONTH                                     -->
<!--                    Names one of the months of the year. Used in
                        both article metadata and inside a citation,
                        this element may contain a full month
                        "December", an abbreviation "Dec", or,
                        preferably,a numeric month "12".
                        Authoring and Conversion Note: For ease in
                        comparisons and searching, many archives
                        prefer that months be converted to numeric
                        values:
                          1 = January
                          2 = February
                          3 = March, etc.                           -->
<!ELEMENT month (#PCDATA)>
<!--                    SEASON                                      -->
<!--                    Season of publication, such as "Spring"     -->
<!ELEMENT season (#PCDATA)>
<!--                    YEAR                                        -->
<!--                    Year of publication, which should be expressed
                        as a 4-digit number: "1776" or "1924"       -->
<!ELEMENT year (#PCDATA)>
<!--                    DATE AS A STRING                            -->
<!--                    This is a representation of the date as a
                        string. Usually used for dates for which
                        months and years are not given, but may be
                        used for any date as a string(i.e. "January,
                        2001" "Fall 2001" "March 11, 2001".
                        It is better practice to tag the year
                        and month as numbers with a date such
                        as "January, 2001" or "March 11, 2001".     -->
<!ELEMENT stringdate (#PCDATA)>
<!-- ============================================================= -->
<!--                    EMPHASIS/RENDITION CLASS ELEMENTS           -->
<!-- ============================================================= -->
<!--                    BOLD                                        -->
<!--                    Used to mark text that should appear in bold
                        face type for print or display              -->
<!ELEMENT b (#PCDATA | %emphasis;)*>
<!--                    ITALIC                                      -->
<!--                    Used to mark text that should appear in
                        italic type on output.                      -->
<!ELEMENT it (#PCDATA | %emphasis;)*>
<!--                    SMALL CAPS                                  -->
<!--                    Used to mark text that should appear in a
                        font which creates smaller capital letters
                        on output.                                  -->
<!ELEMENT sc (#PCDATA | %emphasis;)*>
<!--                    SUBSCRIPT                                   -->
<!--                    A number or expression that is set lower
                        than the baseline and slightly smaller,
                        to act as an inferior or subscript          -->
<!ELEMENT sub (#PCDATA | %emphasis;)*>
<!--                    SUPERSCRIPT                                 -->
<!--                    A number or expression that is set higher
                        than the baseline and slightly smaller,
                        to act as a superior or superscript         -->
<!ELEMENT sup (#PCDATA | %emphasis;)*>
<!--                    UNDERLINE                                   -->
<!--                    Used to mark text that should appear with
                        a horizontal line beneath it for display
                        or print                                    -->
<!ELEMENT ul (#PCDATA | %emphasis;)*>
```

```
<!-- ============================================================= -->
<!--                    ISO STANDARD SPECIAL CHARACTER SETS DEFINED-->
<!-- ============================================================= -->
<!--                    ISO STANDARD ADDED LATIN 1                 -->
<!ENTITY % ISOlat1 PUBLIC
"-//W3C//ENTITIES Added Latin 1 for MathML 2.0//EN"
"xmlchars/isolat1.ent">
<!--                    ISO STANDARD ADDED LATIN 2                 -->
<!ENTITY % ISOlat2 PUBLIC
"-//W3C//ENTITIES Added Latin 2 for MathML 2.0//EN"
"xmlchars/isolat2.ent">
<!--                    ISO BOX AND LINE DRAWING                   -->
<!ENTITY % ISObox PUBLIC
"-//W3C//ENTITIES Box and Line Drawing for MathML 2.0//EN"
"xmlchars/isobox.ent">
<!--                    ISO STANDARD DIACRITICAL MARKS             -->
<!ENTITY % ISOdia PUBLIC
"-//W3C//ENTITIES Diacritical Marks for MathML 2.0//EN"
"xmlchars/isodia.ent">
<!--                    ISO STANDARD NUMERIC AND SPECIAL GRAPHIC   -->
<!ENTITY % ISOnum PUBLIC
"-//W3C//ENTITIES Numeric and Special Graphic for MathML 2.0//EN"
"xmlchars/isonum.ent">
<!--                    ISO STANDARD PUBLISHING                    -->
<!ENTITY % ISOpub PUBLIC
"-//W3C//ENTITIES Publishing for MathML 2.0//EN"
"xmlchars/isopub.ent">
<!--                    ISO STANDARD GENERAL TECHNICAL             -->
<!ENTITY % ISOtech PUBLIC
"-//W3C//ENTITIES General Technical for MathML 2.0//EN"
"xmlchars/isotech.ent">
<!--                    ISO STANDARD GREEK LETTERS                 -->
<!ENTITY % ISOgrk1 PUBLIC
"-//W3C//ENTITIES Greek Letters//EN"
"xmlchars/isogrk1.ent">
<!--                    ISO STANDARD MONOTONIKO GREEK              -->
<!ENTITY % ISOgrk2 PUBLIC
"-//W3C//ENTITIES Monotoniko Greek//EN"
"xmlchars/isogrk2.ent">
<!--                    ISO STANDARD GREEK SYMBOLS                 -->
<!ENTITY % ISOgrk3 PUBLIC
"-//W3C//ENTITIES Greek Symbols for MathML 2.0//EN"
"xmlchars/isogrk3.ent">
<!--                    ISO STANDARD ALTERNATIVE GREEK SYMBOLS     -->
<!ENTITY % ISOgrk4 PUBLIC
"-//W3C//ENTITIES Alternative Greek Symbols//EN"
"xmlchars/isogrk4.ent">
<!--                    ISO STANDARD RUSSIAN CYRILLIC              -->
<!ENTITY % ISOcyr1 PUBLIC
"-//W3C//ENTITIES Russian Cyrillic for MathML 2.0//EN"
"xmlchars/isocyr1.ent">
<!--                    ISO STANDARD NON_RUSSIAN CYRILLIC          -->
<!ENTITY % ISOcyr2 PUBLIC
```

```
"-//W3C//ENTITIES Non-Russian Cyrillic for MathML 2.0//EN"
"xmlchars/isocyr2.ent">
<!--                    ISO STANDARD MATH ALPHABETS (SCRIPT)       -->
<!ENTITY % ISOmscr PUBLIC
"-//W3C//ENTITIES Math Alphabets: Script for MathML 2.0//EN"
"xmlchars/isomscr.ent">
<!--                    ISO STANDARD ADDED MATH SYMBOLS
                           (ARROW RELATIONS)                       -->
<!ENTITY % ISOamsa PUBLIC
"-//W3C//ENTITIES Added Math Symbols: Arrow Relations for MathML 2.0//EN"
"xmlchars/isoamsa.ent">
<!--                    ISO STANDARD ADDED MATH SYMBOLS
                           (BINARY OPERATORS)                      -->
<!ENTITY % ISOamsb PUBLIC
"-//W3C//ENTITIES Added Math Symbols: Binary Operators for MathML 2.0//EN"
"xmlchars/isoamsb.ent">
<!--                    ISO STANDARD ADDED MATH SYMBOLS
                           (DELIMITERS)                            -->
<!ENTITY % ISOamsc PUBLIC
"-//W3C//ENTITIES Added Math Symbols: Delimiters for MathML 2.0//EN"
"xmlchars/isoamsc.ent">
<!--                    ISO STANDARD ADDED MATH SYMBOLS
                           (NEGATED RELATIONS)                     -->
<!ENTITY % ISOamsn PUBLIC
"-//W3C//ENTITIES Added Math Symbols: Negated Relations for MathML 2.0//EN"
"xmlchars/isoamsn.ent">
<!--                    ISO STANDARD ADDED MATH SYMBOLS (ORDINARY) -->
<!ENTITY % ISOamso PUBLIC
"-//W3C//ENTITIES Added Math Symbols: Ordinary for MathML 2.0//EN"
"xmlchars/isoamso.ent">
<!--                    ISO STANDARD ADDED MATH SYMBOLS
                           (RELATIONS)                             -->
<!ENTITY % ISOamsr PUBLIC
"-//W3C//ENTITIES Added Math Symbols: Relations for MathML 2.0//EN"
"xmlchars/isoamsr.ent">
<!--                    ISO STANDARD MATH ALPHABETS (FRAKTUR)      -->
<!ENTITY % ISOmfrk PUBLIC
"-//W3C//ENTITIES Math Alphabets: Fraktur for MathML 2.0//EN"
"xmlchars/isomfrk.ent">
<!--                    ISO STANDARD MATH ALPHABETS (OPEN FACE)    -->
<!ENTITY % ISOmopf PUBLIC
"-//W3C//ENTITIES Math Alphabets: Open Face for MathML 2.0//EN"
"xmlchars/isomopf.ent">
<!-- ============================================================= -->
<!--                    ISO SPECIAL CHARACTER SETS INVOKED         -->
<!-- ============================================================= -->
%ISOlat1;%ISOlat2;%ISObox;%ISOdia;%ISOnum;%ISOpub;%ISOtech;%ISOgrk1;%ISOgrk2;%ISOgr
k3;%ISOgrk4;%ISOcyr1;%ISOcyr2;%ISOamsa;%ISOamsb;%ISOamsc;%ISOamsn;%ISOamso;%ISOamsr
;%ISOmscr;%ISOmfrk;%ISOmopf;
```

9.0 Appendix B: Title Specific Style Sheets

In the following section, title specific rules and guidelines will be provided. These style sheets are to be used to as reference for converting the specified issues for each title. A style sheet may provide information for a specific year of publication range within each title. These style sheets are intended to proactively identify and resolve source variation within a given title.

A style sheet provides the following information:

9.1 Title Metadata

9.1.1 Journal Title (current)

The current title of the journal

9.1.2 ISSN

The current ISSN of the journal

9.1.3 Period of coverage

The specific years of publication covered by the style sheet

9.1.4 Previous titles, years and ISSN

Previous title metadata information, used to provide background reference information for a title.

9.1.5 Publication Frequency

Frequency of publication

9.1.6 Publisher Name and Address

Name and location of publisher

9.2 Format Information

This section provides detailed information about the format and structure of the issue contents. Guidelines for conversion of specific article types appearing in the issue are provided.

For some more complex titles, examples will be provided. Each example will be based on an individual issue within a specific publication year range. These examples should be reviewed against the corresponding source for reference.

10.0 Appendix C: Accepted Rework (AR) Procedure

Redelivery of Accepted batches can be requested for the following reasons:

a. There are no “Failed” error categories, but one or more categories have errors that the customer would prefer to have checked and reworked, OR

b. There are one or more items that are not errors based on the CSDD, but the customer would prefer to have them reworked.

The following is the procedure that should be followed for delivering "AR"

10.1 Extract the original batch from tape and place it in server.

10.2 Perform the necessary rework in the production files.

10.3 Re-process the batch according to 301 process and generate the delivery.

10.4 Identify the rework data to include in the AR-delivery. Also provide the following files:

10.4.1 one list file (XXX-XXXX.lst) containing md5 check sums and file names only for rework files included in the "AR" batch. This can be obtained from the 301-generated .lst file.

10.4.2 map and img files for each issue containing rework data. Do not include .map and .img files for issues that do NOT contain rework data

10.4.3 ALL files associated with rework article (PDF, TXT, XML, TIFs). Include ALL files, even if only one file was reworked. For example, even if the change was only in the XML file, the PDF, TXT and TIF files for the same article must also be delivered.

10.5 The "AR" batch is then zipped into a single ZIP file with the following nameing convention:

jjj-bbbbar-YYYY-MM-DD.zip

where:
jjj: job number (301/361)
bbbb: batch number
ar: literal "ar" batch indicator
YYYY-MM-DD: time stamp (e.g., 2006-02-16 for February 16, 2006)

11.0 Release Notes

301CSDD_Revision_3.2-NLM.doc: May 16, 2007

- Revised Item 1.3.3; specification for PDF pages to be in reading order sequence
- Revised Item 1.3.4; specification for PDF pages to be in reading order sequence
- Revised Item 1.3.5; specification for OCR file
- Revised Item 5.2; removed exception for duplication of advertising pages in deliverables
- Revised Item 5.7; clarified contents of each deliverable directory
- Revised Item 6.10.3; specification for PDF pages to be in reading order sequence
- Revised Item 7.4.3; <cpyrt> element no longer captured in this project
- Revised Item 7.6.2.4; added specifications for treatment of multiple letters under a single group title
- Removed Item ~~10.3.9.2~~; removed markup specifications copyright element
- Revised Item 7.16; guidelines for use of emphasis elements
- Removed Item ~~**Error! Reference source not found.**~~7.16; removed project deadline
- Revised Item 1.7.4.2; specifications for special case
- Revised Item 1.8.3; specifications for resolving source discrepancies
- Revised Item 1.9; specifications for source and media retention
- Revised Item 2.5.2.2; specifications for source discrepancies
- Revised Item 10.0 ; added definition of accepted re-delivery
- Revised Item 4.6.3; specifications for multiple table of contents pages
- Revised Item 5.2.2.3; specifications for proceedings articles
- Revised Item 5.2.4.3; specifications for subscriptions
- Revised Item 5.2.6.2; specifications for outside front covers
- Revised Item 5.7.5; specifications for Issue Covers Directory
- Revised Item 7.6.2; specifications for Article Title `<atitle>`
- Revised Item 7.6.2.3; specifications for titles in book review articles
- Revised Item 7.7.2; specifications for contributor
- Revised Item 7.14.5; specifications for Footnote `<fn>`
- Corrected Item 1.7.4.3; example revised with correct pagination
- Revised Item 3.3; revised product QA criteria
- Added Item 5.1; specifications hierarchy tree
- Revised Item 5.2.2.1; specifications for obituaries in grouped articles
- Revised Item 5.2.2.1; specifications for customer style sheet requiring treatment of grouped articles as a single article.
- Revised Item 5.2.4; specifications for Administrative (ADM) content; clarified treatment of Advertising Indexes
- Revised Item 5.2.4; specifications for Administrative (ADM) content; clarified treatment of Death Notices
- Added Item 5.2.4.1; specifications for Administrative content and the TOC
- Revised Item 5.2.4.2; specifications for treatment of cumulative indices
- Revised Item 6.7.1; provided additional clarifications for directional cues in figure labels
- Added Item 6.8; specifications for special case figures
- Revised Item 7.1.1.1; specifications for “correction” article type
- Revised Item 7.1.1.1; specifications for “reply” article type
- Revised Item 7.6.2.3; specifications for article titles for multiple book review articles clubbed together.
- Revised Item 7.7.2; specifications for contributors of multiple grouped articles treated as a single article.
- Revised Item 7.9.1; specifications for publication date as range, only last date is captured as publication date elements.
- Revised Item 7.15.5; specifications for character restrictions within <stringdate> element