

March 2008

PubMed Central File Submission Specifications

Follow this and additional works at: <http://digitalcommons.unl.edu/nihpmcinfo>

 Part of the [Bioinformatics Commons](#)

"PubMed Central File Submission Specifications" (2008). *Depositors' Information for NIH PubMed Central* . 2.
<http://digitalcommons.unl.edu/nihpmcinfo/2>

This Article is brought to you for free and open access by the Research Compliance Services at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Depositors' Information for NIH PubMed Central by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

PMC File Submission Specifications

PubMed Central (PMC) exists both to provide online access to journal content and to build and maintain a high-quality, durable archive of this digital content. The latter objective drives many of the requirements here. Several years of experience working with material deposited in PMC confirms that what is adequate for displaying an article online today may not be enough for archival purposes. Please keep this in mind as you make your way through the details.

Quick Links:

[Image File Requirements](#)

[Coding Requirements for XML / SGML Files](#)

[Article Data File Naming Conventions](#)

[Articles Published Online First](#)

[File Organization, Packaging, and Delivery](#)

[Submitting Revised or Corrected Files](#)

Files to be deposited for each issue:

1. A separate XML or SGML data file for the full text of each article.
2. The original high-resolution digital image files for all figures in each article.
3. A PDF file for each article. [Note: a separate PDF which directly corresponds to the individual XML or SGML data file should be provided for each book review and/or letter.]
4. Supplementary data files (e.g., spreadsheets or video files) available with the article.

Image Quality Requirements

Generally, Authors submit raw image data files to a publishing house in various formats (ppt, pdf, tif, jpg, xml, etc.). The files are then normalized to produce print or electronic output. PMC requires the normalized output, which is high-resolution, and of sufficient width and quality to be considered archival. Images generated at low resolution for display purposes are not acceptable.

All images MUST be at or above [intended display size](#), with the following image resolutions: Line Art 800 dpi, Combination (Line Art + Halftone) 600 dpi, Halftone 300 dpi. See the [Image quality specifications chart](#) for details. Image files also must be cropped as close to the actual image as possible.

What we do not want are 72 dpi web-quality graphics in which colors are not realistic, text is illegible, or images are pixilated. These undesirable qualities are usually caused by applied compression from a jpg or gif format. Although tif and eps files are the most desirable formats for archiving, it is important to stress that the real objective is to obtain the highest quality images available, regardless of format.

PDF Quality Requirements

PDF files should not be downsampled for submission to PMC. If print quality PDFs are available, please submit them. If the journal is not printed, the resolution of the images in the PDF should be no less than: Line Art 800 dpi, Halftones 300 dpi, Color 600 dpi. All fonts used in the file need to be fully embedded. Compression for images should be lossless (zip) or highest-quality JPEG. Illustrations should be encoded as vector data with no erroneous conversion to bitmaps.

Image File Requirements

File Format Requirements

Font Requirements

The files in the PMC archive must be portable. Copyright and ownership issues aside, other archives must be able to read them, as necessary, without special software or tools. Therefore EPS or TIFF images need to be formatted in one of the following ways:

- Fully Embed the non-standard fonts in the image file, Or
- Convert the text to curves and [rasterize](#)

Figure Graphics

- [Image quality specifications chart](#)

Tables and Equations as Graphics

If equations cannot be encoded in MathML, submit them in TIFF or EPS format as discrete files (i.e., a file containing only the data for one equation). Only when tables cannot be encoded as XML/SGML can they be submitted as graphics. If this method is used, it is critical that the font size in all equations and tables is consistent and [legible](#) throughout all submissions.

- [Suggested Equation Extraction Method](#)

- [PMC Table Specifications](#)
- [PMC Equation Specifications](#)
-

Coding Requirements for XML / SGML Files

For details on data elements and XML structure, please see the Journal Publishing DTD at <http://dtd.nlm.nih.gov/publishing>. For detailed information on using the Journal Publishing DTD for submissions to PMC, please read the [PMC Tagging Guidelines](#).

Required Data Elements

Certain data elements must be present and used consistently in each XML or SGML file deposited in PMC, even if the corresponding DTD does not require them. These elements contribute to making the XML/SGML files self-documenting and more portable for archival purposes:

- Journal ISSN
- Journal ID or Journal title abbreviation
- Journal publisher
- Copyright statement, where applicable
- Volume, issue (if applicable), and article sequence number or pagination
- Issue publication date
- Article electronic publication date
- For more information on any of these elements, please see the [PMC Tagging Guidelines](#) and the [Journal Publishing DTD](#)

Including a Printed Page Image as Illustration

When an article includes an image of a printed page (e.g., a chart or a form), define that page image as an inline image, not as a standard figure in the XML/SGML.

Coding Open Access Articles

In the PubMed Central context, an Open Access (OA) article is one that is made available with a [Creative Commons](#) or similar license. If you submit any such OA articles to PMC, you must include a brief license statement (i.e., the applicable terms of use) in the XML and PDF of each article. These articles will be included in [PMC's Open Access Subset](#). A user may download the source files for any of these articles and reuse them according to the license statement in each article.

The [Licensing Information](#) section of the PMC Tagging Guidelines illustrates how to code the license using the NLM Journal Publishing DTD. Note that the terms of your license may be different from those in the example. If your DTD does not include a license-specific element,

this information needs to be captured in the copyright statement or as a footnote in the front matter.

Article Data File Naming Conventions

- **A key requirement is that the names of the image files and supplemental data files MUST match the names that are called out in the XML/SGML.**
- All file names must be unique within a volume.
- XML and PDF base file names must match exactly.
- See information about [File Names for OLF Articles](#).
- “Zero-fill” or “pad” volume and page numbers so that the same number of digits exists in every file.
- **Note:** The naming convention is different for container files (.zip, .tar, .gz, .tgz) than actual article data files (.xml, .pdf, .tif, etc.).
See [File Organization, Packaging, and Delivery](#)

Naming XML and PDF Files:

jour-vol-pg.ext is the PMC Naming Convention

jour is an alphanumeric identifier such as the journal abbreviation. The ISSN may be used, in addition to the alphanumeric identifier.

vol identifies a specific journal volume.

pg is the first page of the article.

- For articles that start on the same page, to make the file names unique, add a sequence letter (-a, -b, -c) to the end of the page number. Use sequence letter -a for the first article that starts on that page.
- For electronic journals that assign each article a unique article sequence number (e-ID) in place of pagination, use this e-ID as the "first page".

ext is the lowercase file type extension (.xml or .pdf).

Naming Graphics and Supplementary Files:

jour-vol-pg-tyt.ext is the PMC naming convention for GRAPHICS and SUPPLEMENTARY DATA files

jour, vol, pg, and ext are the same as above.

typ is NOT to be used with full-text article XML or PDFs, is optional, and indicates one of the following:

- g figure graphic + alphanumeric identifier
- i inline graphic + alphanumeric identifier
- s supplementary data file + alphanumeric identifier
- e equation + alphanumeric identifier

**For Example: You want to send an article from the journal
“Biological Testing”, volume 01, page 100:**

XML	biotes-01-100.xml
PDF	biotes-01-100.pdf
Graphic figures (2)	biotes-01-100-g001.tif, biotes-01-100-g002.tif
Inline graphics (2)	biotes-01-100-i001.tif, biotes-01-100-i002.tif
Supplementary files (2)	biotes-01-100-s001.tif, biotes-01-100-s002.tif
Equation graphics (2)	biotes-01-100-e001.tif, biotes-01-100-e002.tif

Articles Published Online First

These requirements apply when a journal releases selected articles online before publication release of the full issue containing those articles.

Terms Used Here

- “OLF” refers to the version of the article that is released online first.
- “Published version” refers to the article, as it is included in the full published issue.

Unique Element Linking OLF Version to Published Version

Include a valid DOI (Digital Object Identifier) in the coding of the XML/SGML for all articles. This is a means of making an unambiguous association between the OLF and published versions of an article. The DOI, which uniquely identifies an article, must be the same for the OLF and published versions of the article.

File Names for OLF Articles

jour-vol-id.ext is the PMC OLF naming convention

If pagination is not known when a journal publishes articles online, the **pg** in the file name format **jour-vol-pg.ext** must be substituted with some other unique identifier **id**.

For example:

- If the XML file for the published version will be named: **biotes-10-108.xml**
- The corresponding OLF file must be named either:
biotes-10-108.xml or **biotes-10-108.xml.olf**

Note: The **.olf** suffix is not required, but it is the only acceptable option if a journal wants to differentiate the name of the OLF version from the published version of the same article. The **.olf** extension replaces the **.xml** extension of the XML files only. All PDF, image, and supplementary files must have their proper extensions (**.pdf**, **.tif**, **.eps**, **.txt**, etc.). Other variations on the file name besides the **.olf** extension are not acceptable.

File Organization, Packaging, and Delivery

Terms Used Here

- “Container files” refers to a file format that houses the collection of article XML, PDFs, image files, and supplementary data files.

Accepted formats

- PMC accepts ONLY **.zip**, **.tar**, **.gz**, and **.tgz**.
- PMC does not accept bzip2 compressed **.zip** files.
- PMC recommends using compression software that supports the Zip 2.0 standard file format.
- If using WinZip PMC recommends using the Maximum (portable) compression option for compatibility.

File Packaging and Delivery (Non-OLF)

- **Note:** The naming convention is different for container files (**.zip**, **.tar**, **.gz**, **.tgz**) than actual article data files (**.xml**, **.pdf**, **.tif**, etc.). See [Article Data File Naming Conventions](#).

jour-vol-issue.ext is the naming convention for CONTAINER files

- **jour** is an alphanumeric identifier such as a journal abbreviation.
- Naming **MUST** be consistent across all submissions.
- If articles are not organized into issues, the **issue** may be omitted from the container file naming convention.
- Include all files for an issue in one directory. **DO NOT** create separate subdirectories for each article in an issue inside the container file.
- Package all the files for an issue in the same container file, or create a separate container file for each class of files (XML/SGML, PDFs, Images, Supplemental Material) in an issue.
- **DO NOT** send files as attachments to an email message. Either FTP them or send them on CD/DVD or tape. To transfer the files via FTP, write to pmc@ncbi.nlm.nih.gov for an account on the PMC FTP site.
- Send an email notification for each submission including:
 - Journal title
 - Volume
 - Issue
 - Container filename(s) and size(s)
 - Address the email to: pmc@ncbi.nlm.nih.gov.
- If sending files on DVD, CD, or tape, mail them to:

PubMed Central
National Center for Biotechnology Information
National Library of Medicine
Building 45, Room 5AN12
45 Center Drive, MSC 6510
Bethesda, Maryland 20892-6510 USA

For example:

- A zip file containing articles for the journal “Biological Testing” Vol. 10, Issue 01 will be named: **biotes-10-01.zip**

File Packaging for OLF Articles

olf-jour-datestamp.ext is the naming convention for OLF CONTAINER files

- Always package and deposit OLF articles separately from published version articles.

- Package each submission of OLF articles in a .zip, .tar, .gz, .tgz file. The container file name must comprise an **olf-** prefix followed by the journal id, a time/date stamp, and the .zip, .tar, .gz, .tgz extension.
- Every container file that packages OLF files must have a name that starts with **olf-**.
- Corrections to OLF articles will have a .r1 suffix added before the **ext**.
- **Note:** The name of the replacement container file must be identical to the original (replaced) container file name, with the addition of the .r1 indicator.

For example:

- A zip file containing OLF articles for “Biological Testing” deposited on July 10, 2003 may be named:
olf-biotes-20030710.zip
- A zip file sent on July 15, 2003, which corrects OLF articles deposited on July 10, 2003 may be named:
olf-biotes-20030710.r1.zip

Submitting Revised or Corrected Files

jour-vol-issue.r1.ext

is the naming convention for REVISED CONTAINER files

Naming Revisions

- Add the suffix .r1, .r2, .r3, and so on, to indicate the sequence of replacement to the Container (.zip, .tar, .gz) filename.
- The name of the replacement article files (.xml, .pdf, .tif, etc.) must be identical to the original (replaced) file names.

Contents of Revision Packages

- Always package and deposit OLF article corrections separately from published version corrections.
- If submitting replacement files for multiple issues at the same time, use a separate zip package for each issue. DO NOT combine files for different issues or journals in the same zip package.
- When submitting corrections, be sure to submit ONLY the files that have been modified, along with their corresponding XML files.
- **For journals in the evaluation stage:** Please resubmit the entire set of files, which were previously submitted.

- **For journals that have passed the evaluation stage and are in production:** Please resubmit only the files that have been modified.

Notifying PMC

- Please notify PMC of any revised packages submitted to the PMC FTP site.
- When corrections involve volume, issue, doi, or ppub date, please explicitly include what has been changed in the email notification.

For example:

- You send the original files for “Biological Testing” volume 15, issue 1 in a zip file named: **biotes-15-01.zip**
- Three days later, you send a revised XML file for one article, in a zip file named: **biotes-15-01.r1.zip**
- Later that same day, you send replacement files for the same issue, in a zip file named: **biotes-15-01.r2.zip**
- The next week, you make more corrections, and submit: **biotes-15-01.r3.zip**

Downloaded 3/10/2008 from http://www.pubmedcentral.nih.gov/about/PMC_Filespec.html