File Transfer Guidelines

Overview:

File-Transfer means preservation of items provided by the content owner. In order to support this method of preservation, the following general conditions must be met:

1. Ability to deposit content to the CLOCKSS archive either by staging to an account we provide or by providing access to a deposit location we can access (FTP, AWS bucket, etc).
2. Uniquely-named deliveries, including full text and supplemental materials as well as metadata identifying the contents.
3. Full-text content in a standard readable format (ex PDF) as part of each delivery.
4. Clear, consistent file naming convention within each delivery that supports programmatic identification of the delivered content from the metadata files.
5. Ability to suppress pre-publication content and to defer delivery until publication units are complete.
6. Notification to the delivery alias when a delivery has occurred and a listing of files delivered with updates clearly identified.
7. Delivery of clearly identified “duplicative content” only for corrections and updates.
8. Delivery in non-proprietary formats.

Details:

Required Metadata:

Metadata must be in a machine-readable text format, such as XML in a standard schema, or RIS.

Some examples of standard XML schemas for scholarly publication materials are JATS, ONIX, PubMed, and Crossref.

The metadata for scholarly content should include: DOI, publication, publication date, ISSN or ISBN and as appropriate: publication title, item title, series title, volume, issue and page number.

For items other than scholarly published content, the metadata should provide enough information to uniquely identify the item - DOI, if applicable, or proprietary identifier, provider and so forth.

Delivery format, file naming and content identification:

Only final content version (not pre-publication) should be delivered.

Content files should come in only one non-text format, such as PDFs without redundant alternate versions such as manuscript, epub, mobi.

Updates to previously delivered content would include just the modified items.

Content can be delivered unpacked and organized in to a hierarchy of directories or in non-proprietary archive formats such as zip or tar.

Deliveries in tar or zip bundles should be completely self-contained. Multiple archives should not need to be expanded to map metadata to content files.

There needs to be a way to map information in the delivered metadata file(s) to the corresponding content file. For example:

- if there is one metadata file for each content file, the two files could use the same base filename.
- If there is one metadata file providing information for multiple content files, there might be information in the metadata that specifies the content filename or the content filename could use some consistent pattern of information:
    - a filename that is defined in the XML
    - <content_filename>fred_smith_2002.pdf</content_filename>
    - a filename built up from journal_id + volume + start_page as identified in the XML, e.g., foo_17_101.pdf
    - a filename that corresponds to the doi of the article in question with underbars in place of "/", so “10.1111/foo/xxx2018.1" would be, e.g., 10.1111_foo_xxx2018.1.pdf
    - a filename that uses the publisher internal identifier for the content, e.g., xxx2019.1.pdf

In our experience, the use of portions of the article title or author name do not work as well because of variance encoding, punctuation or potential lack of uniqueness, but so long as the approach is consistent and we can extract the information we need, we can work with a naming convention that fits within the publisher's processes. The files might be co-located within the delivery or they might be in relative directory locations, so long as the layout remains consistent over time.