Filling Out the Document Class

The ***<Document>*** class and its sub-classes describe both the logical content (the card-catalog information, if you like) as well as the physical storage format for potentially several different physical editions of the same document (HTML, PDF/A, and so on). Every <*Product_Document*> will contain exactly one <*Document*> class.

There is some apparent overlap between the information in this class and information that might be included as part of the <*Citation_Information*> class in the *Identification_Area*. There is a subtle distinction. The *Citation_Information* should be used to cite the document product - which may contain several physical formats of the same document, and has its own revision history separate from the logical content of the document. The attributes at the top level of the <*Document*> class provide similar information for the document itself - that is, for the source material used to create the PDS product. Often there will be overlap, but there may be cases where, for example, PDS personnel have done significant work editing or restoring a document from the primary source (scanning, editing OCR files, re-keying from paper, etc.), so that the document product will have an editor credit, but the document itself would have only an author credit.

For additional explanation, see the PDS4 Standards Reference, or contact your PDS node consultant.

For a video walkthrough of filling out the Document Class of a label, watch this video:

Filling Out the Document Class Video

To follow along, use these XML documents:

Training Document Class XML

Training Document Class (empty)

Following are the attributes and subclasses you'll find in the *Document* class, in label order.

Note that in the PDS4 master schema, all classes have capitalized names; attributes never do.

<revision_id>

OPTIONAL

This is for the revision number of the document itself, as opposed to the *version_id* in the *Identification_Area*, which is the version of the PDS4 product comprising the document.

<document_name>

OPTIONAL

Use this attribute *only* in cases where the actual title of the document is too long to fit in the *title* attribute of the *Identification_Area*, which is limited to 255 bytes.

<doi>

OPTIONAL

If the document you are labelling has previously been published, it may have been assigned a *Digital Object Identifier* (DOI) by that publisher. The value of the <*doi*> attribute is case sensitive, so copy the DOI exactly; do not include any "doi:" or "DOI:" prefix.

<author_list>

OPTIONAL

The list of authors, if any, for this document, with the principal author listed first. Names should be in "Surname, Initials" format, with a semi-colon separating names. (That is, the same format as described for the same field in the <*Citation_Information*> class from Identification Area.)

<editor_list>

OPTIONAL

The list of editors, if any, for this document, with the principal editor listed first. Names should be in "Surname, Initials" format, with a semi-colon separating names. (That is, the same format as described for the same field in the <*Citation_Information*> class from Identification Area.)

<acknowledgement_text>

OPTIONAL

When republishing a copyrighted work, it is polite (and sometimes legally required) to acknowledge the original source. Here's the place to do that, formally or informally.

Even for public domain documents, this area can be used to acknowledge sources, for example: "Reprinted from *[URL]* with the kind permission of *[web master]*."

<copyright>

OPTIONAL

If you are reprinting a copyrighted document and the copyright holder wants a formal copyright notification to be included, this is the place to put it. This is also the place to specifically state that a document is in the public domain, if that seems appropriate; or to attach a formal copyright notice to a copyrighted document in the rate case where one is created for PDS archiving.

The vast majority of documents in the PDS archive are in the public domain as they are government-funded works-for-hire.

<publication_date>

REQUIRED

This is the publication date of the document. If the document has been previously published, outside the PDS, this should be the date of first publication. For documents originating in the archive, this is generally the date when the document is considered to be public - that could be the date of the review, the date of posting on a web site, or the date the product was created, depending on the node or data preparer conventions.

<document_editions>

OPTIONAL

This is the number of <*Document_Edition*> subclasses included in this <*Document*>. While it is defined as being optional, logically it ***must*** exist and have a value of at least one.

<description>

OPTIONAL

This is where you provide a brief abstract for the document. The level of formality should follow the level of formality of the document itself. If you are republishing an article that has a formal abstract, you can simply past that in here. Note that the *<description>* in the *<Citation_Information>* should contain a short description suitable for returning in a browsable list of search results. This description, if offered, should provide a bit more information than that.

<Document_Edition>

REQUIRED

Include one of these classes for each physical edition of the document. Note that some editions will be provided in multiple files (an HTML file with separate graphics and image files, for example). In these cases all the files together comprise one "edition", and should be included in one *Document_Edition*, even if file types differ. (Instead, there will be one *<Document_File>* for each of the constituent files.)

<edition_name>

REQUIRED

You should provide a brief name for this physical edition that includes an indication of the sort of file format a user can expect to get, for use in web interfaces. You can assume that the document title will be displayed at the same time, so you don't have to repeat it. So, for example, if the *title* in the *Citation_Information* is "MERC User's Manual", then a reasonable *edition_name* value might be something like "PDF version", or "HTML with graphics".

<starting_point_identifier>

OPTIONAL

If your *Document_Edition* has more than one file (*<Document_File>* class) associated with it, then one of them will be the file where a user wanting to read the document should start. Include a *local_identifier* in that *Document_File* class, and then include that *local_identifier* here. For example, in the case of an HTML file with associated graphics, you might provide a *local_identifier* of "index" for the HTML file, and then repeat the "index" identifier here. This is used to provide automatic and programmatic access to documents online.

If your document is in a single file, there is no need to provide either a *starting_point_identifier* here or a *local_identifier* in your only *Document_File*.

<language>

REQUIRED

This attribute identifies the language of this particular edition of the document. At the moment it *must* have a value of **English**. If you have a document edition in a different language, notify your PDS consultant.

Note that all essential documentation submitted to the PDS ***must*** be in English.

<files>

REQUIRED

This attribute contains the number of *<Document_File>* classes (and thus files) comprising this physical edition of the document. It must be at least one.

<description>

OPTIONAL

If there are any comments you'd care to make about this particular edition of the document not covered elsewhere, this is the place.

<Document_File>

REQUIRED

Every *Document_Edition* must have at least one *Document_File*. Many will have only one. For those with multiple files, there is no requirement that all the files have the same format - so you can have one *Document_File* containing flat text, another containing a referenced JPEG image, and so on. Neither are these documents required to be in the same directory - the *Document_File* class allows you to provide path information. Repeat this class for each individual file in this edition of the document.

The *<Document_File>* class is an extension of the *<File>* class. It contains all the attributes required or optional in the *File* class, in the same order. See Filling Out the File Class for details. Here's an abbreviated list of the attributes in that class, for reference (required attributes are marked with [*]), with two added attributes described following:

- *file_name* [*]
- *local_identifier*
- *md5_checksum*
- *creation_date_time*
- *file_size*
- *records*
- *md5_checksum*
- *comment*

<directory_path_name>

OPTIONAL

If the file you are describing is in the same directory as the label, this attribute is optional. If it is not in the same directory as the label, it ***must*** be in a subdirectory of the directory containing the label. In that case, this attribute provides the relative path from the directory containing the label to the directory containing the file. This path ***must*** be in Linux-style notation, and must not start with a "/" character, a "." character, or a Windows-style device name (like "C:").

Note: *These constraints on* directory_path_name *are not currently enforced. This must be checked manually, so code and type carefully.*

<document_standard_id>

REQUIRED

This identifies the file format standard for this particular document file. It must be one of the defined standard values you can find on the Standard Values Quick Reference page.