Guidelines and Conventions for Digital Library Collections

- **Last Updated:** 2004-05-26
- **Editor:**Peter C. Gorman
- **Version:**2.0

Directory path conventions

The basic directory structure for all types of digital library collections is the same. All files for the collection's Web site (Home and About pages, copyright page, etc.), including local graphic items such as icons or banners, must be contained in the `htdocs/` directory or in directories under it (other than `htdocs/data/`) that project staff create. No HTML documents or Web site graphics should be placed under `htdocs/data/`.

Digital Library collections will use the directories `htdocs/data/sgml/` for SGML markup and entity definition files, `htdocs/data/images/` for all images such as figures or page images, and `htdocs/data/delimited/` or `htdocs/data/tagged/` for metadata files. If further subdivision is required (e.g., if the number of files in a directory is expected to exceed 500), the subdirectory names can be anything useful for project staff, but they must appear at the level indicated by , , etc. in the diagram. Subdivision directories can be nested to any level, according to the needs of the collection.

The directory names for various sizes of image files (`Thumb/`, `Reference/`, etc.) are controlled. If a collection requires a relative image size other than those currently defined, the architecture will be expanded to include the new size (with a new, controlled directory name) for all collections.

For Electronic Facsimile collections, the subpaths under `htdocs/data/images/` and under `htdocs/data/text/` must be identical. In other words, the data element `Page-Location` must apply to both the image and to the OCR Text file.

File naming conventions

File and directory names must use only characters allowed by the UNIX operating system. Allowable characters are: `a-z A-Z 0-9 . _ -`
(Note: The characters '-' and '.' may not be used as the first character in a name.)

SGML files

The names of files containing TEI markup must follow the pattern *`[Name].[ext]`*, where *`[Name]`* is not controlled (other than the general character constraints already mentioned), and *`[ext]`* is `sgm` or `sgml`. Entity definition files, however, must be named according to

the pattern *[Issue-ID]*.ent, where *[Issue-ID]* is the value of the <idno type="Issue-ID">...</idno> in the document's <teiHeader> . It is strongly suggested that the name of the primary SGML file also include the *Issue-ID*. It follows that if two or more TEI files use the same Issue-ID value, they must share the same entity definition file.

Image files

The names of image files must follow the following pattern: *[Name][Relative-Size].[ext]*, where *[Name]* is not controlled (other than the general character constraints already mentioned), *[Relative-Size]* is one character drawn from the list of relative image sizes, and *[ext]* is a three-character filename extension corresponding to the image type.

Page images for Electronic Facsimile collections should **not** include the *[Relative–Size]* character.

If the image is to be accessed from a multimedia database, [Name] must be composed of values representing the data elements *Object-ID* and *Num–Components*.

Metadata files

Metadata files for bibliographic and multimedia databases should be named [Collection-ID].([Subcollection-ID].)Resource.txt. If media objects are exported in a separate table, it should be named [Collection-ID].Media.txt.

When a collection consists of several subcollections, the subcollections' Collection-ID values should be used.

For Electronic Facsimile collections, metadata files should be named according to the objects being described:

```
Collection.txt
Subcollection.txt
Aggregate.txt
Issue.txt
Item.txt
Page.txt
```

For other types of collections, names for metadata files are not controlled (other than the general character constraints already mentioned).

OCR Text files

OCR text files created for Electronic Facsimile collections are subject to the following constraints:

1. The names of the text files must match exactly the names of the corresponding image files, except for the filename extension. In other words, the data element `Page-Filename` must apply to both the image and to the OCR Text file.