Filename Requirements for Digital Objects

> **Status**
> This page is **finalized**. Any modifications should be approved by the infrastructure group.

As we create new collections, it is useful if there is consistency in the filenames assigned to digital objects. Each collection will impose its own restrictions on filenames, but following these requirements will ensure basic consistency across collections and make later processing of the files much easier.

Files must be named according to these requirements to be ingested into the Fedora repository. These requirements must be used for all new DLP collections. Older collections that do not meet the requirements will be renamed when it becomes convenient. Note that these requirements apply to "digital objects" only. Other supporting materials (documentation, html pages, etc.) may be named according to different standards. If supporting materials are to be stored in the repository, they may be collected into a tar or zip file whose name meets these requirements.

The need for standardized filenames

The needs to be fulfilled by enforcing requirements on filenames are (in order of importance):

1. Ease of identification. During the life of a digital file, it moves through various locations. These moves may be due to manual processing or automatic processing. A file may stay in a given location for an arbitrary length of time, during which human memory of its origin fades. At all points in the life of the file, humans and automatic processes must be able to easily identify which digital object the file belongs to and the position it occupies within that object. As a side effect of this, the filename will facilitate the process of locating metadata about the file.
2. Automatic processing systems must be able to make basic assumptions about the filenames they will process. Developers should not be concerned with complex processing to handle special characters. (Although as a security matter, it is good practice to verify that a filename conforms to these requirements before initializing an automatic process.)

Requirements for filenames

Absolute requirements

All files must conform to the following requirements:

- Each filename must contain an identifier that uniquely specifies a single digital object within the parent collection.
- If a digital object consists of multiple files, each filename must contain the object's identifier, along with a unique sequence number.
- Each filename must be fully specified. It cannot just be a sequence number that is dependent on location within a directory structure for context.
  Rationale: Files are often moved between locations for processing or testing. It is not always feasible to move an entire directory structure with the file, so all necessary information must be in the filename itself.
- Filenames must not include spaces.
  Rationale: There are many instances where using a space in a filename can cause programs to misbehave. Automatic processing as well as human access to the file becomes more difficult when spaces are involved.
- The first character of the filename must be an ASCII letter ('a' through 'z' or 'A' through 'Z').
  Rationale: Many programming and metadata languages place this restriction on their identifiers. Filenames should be usable as identifiers in these languages (e.g., section ID's in a METS document).
- The "base" filename may include **only** ASCII letters ('a' through 'z' and 'A' through 'Z'), ASCII digits ('0' through '9'), hyphens, underscores, and periods. No other characters are permitted.
  Rationale: Characters from other character sets can be difficult to read, depending on program support and available fonts. Many operating systems and programs are unable to correctly process non-ASCII characters. Punctuation and other ASCII characters not listed here may have special meanings, depending on the context; files using these characters may cause unexpected problems.
- The "base" filename must be followed by a single period and a suitable extension to specify the type of file. The extension should consist of three letters (e.g., jpg, txt, xml, tif), but longer extensions are permissible if they are widely used (e.g., html, tiff, djvu, aiff).
  Rationale: Whenever possible, the extension should make sense to a human. On systems where the file extension dictates automatic behaviors, the file should exhibit the expected behavior.
- A derivative file must have the same name as the master file, except the "base" filename should have an indication of the derivative's type appended (e.g., "full" or "screen" for images, an indication of the bitrate for audio files). Derivative files will typically have a different file type, and therefore a different extension, than the master file.
  Rationale: It should always be easy to identify files with master-derivative relationships.

Best practices

The following "best practices" should be followed whenever possible. If one of these practices is not followed, the change should be well

documented, with a description of the reasons for not following the practice.

- While periods are permissible in "base" filenames, it is *highly recommended* that they be avoided.
  Rationale: Some programs assume that there is only a single period in a filename, and will behave strangely if multiple periods are present.
- It is preferable that all letters in a filename be lowercase. If a filename includes consecutive human-readable words, they may be denoted by CamelCase (e.g., wnp-04-RoyalSociety-ncn-t123.tif). This is expected to be relatively rare, though.
  Rationale: Lowercase letters aid human readability and make it easier to type the filename. In collections where filenames contain many human-readable words, CamelCase aids readability.
- Portions of the filename should indicate more specific detail as they are read from left to right. That is, the far left portion of the name should indicate the class of item, the next portion should be the item-specific ID, followed by a page/section number, and ending with the indication of derivative size. (Any of these portions that do not apply to the current file may be omitted.)
  Rationale: Alphabetical listings of files make more sense with this organization.
- Distinct portions of the filename should be separated by hyphens.
  Rationale: Separating the portions makes the filename both easier to read and easier to process automatically. Hyphens are slightly easier to type than underscores, and maintain consistency with our existing collections. Filenames that include dates may have the date portion follow the ISO 8601 standard. Note that it is reasonable for the "identifier" portion of the filename to retain underscores in identifiers from external sources, as in ihs-SHMU_01_13-01-05.tif. This reduces confusion when locating items provided by other institutions.
- While it is permissible for two different collections to contain files with identical names, this should be avoided.
  Rationale: It will not be possible to know of all filenames in use. Nonetheless, identical names can be confusing, and care should be taken to reduce the probability of identical names.
- Page numbers should be padded with leading zeros so that all filenames in a collection have the same number of characters for the page number portion. In most cases, this will be two or three digits.
  Rationale: This forces pages to display in the correct order when listed alphabetically, and provides more visual consistency when scanning a long list of files.
- When creating filename standards for a new collection, the standards should be based on existing collections/objects with similar characteristics.
  Rationale: Minimizing the variability in filename standards eases both automatic and manual processing.
- Whenever possible, the ditital object's "primary" identifier (the identifier appearing in the filenames) should correspond to an identifier in use for the original (physical) object. If the format of the primary identifier conflicts with the absolute filename requirements, appropriate changes should be made. If the format of the primary identifier conforms to the absolute filename requirements but violates best practices, it may be left intact.
  Rationale: It should be easy to determine the relationship between digital files and physical objects. This is easier if the identifier in the filename is as similar as possible to the identifier associated with the physical object.

Exception to the requirements

Files that are stored within a closed system (e.g., Fedora, Variations, DLXS) do not need to follow these requirements if the system's internal processing dictates another scheme. For example, Fedora must manage its own filenames to ensure proper version control. However, it is recommended that systems written by the DLP follow the requirements when possible, and closed systems should provide a method for converting internal filenames to "external" names that meet the requirements.

About NOTIS-style identifiers

The old NOTIS system generated IDs consisting of three letters and four numbers (like VAA1234). To easily connect with IUCAT, we have continued using IDs of this form for IU holdings that do not belong to a special collection. Items that use these IDs always include them as the "identifier" portion of the filename. These identifiers are unique across all DLP holdings. We would hope they are unique across the IU library system, but we have no way of knowing if other groups have adopted conflicting standards that emulate the old NOTIS system.

Sample filenames

Collection	Filename	Notes
IN Harmony	ihs-SHMU_01_13-01-05.tif	ihs stands for Indiana Historical Society, SHMU_01_13 is an identifier local to the Historical Society, 01 indicates the first (physical) copy of the item was used for digitization, and 05 indicates this file is the 5th page of the item.
IN Harmony	ihs-SHMU_01_13-01-05-full.jpg	Same as above, but this is a derivative file at the "full" size.

Variations	aeg9051c.wav	A NOTIS-style identifier. The sequence number is actually an alphabetic character, in this case a "c" indicating that the file is third in a set of files for this item.
Variations	aeg9051c-192k.mov	Same as above, but a derivative file encoded at 192kbps.
Hoagy Carmichael photos	ATM-MC2-3-11-30-p1-screen.jpg	A sample derivative file.
Cushman	P15754.tif	Files in the Cushman collection are all of one type, and they don't consist of pages, so the filename only consists of the local identifier.

Filename standards currently in use

- IN Harmony
- Sound Directions