URI & File Naming Plan

Scholars Portal URI and File Naming Policy

1. Policy Statement

URIs created by Scholars Portal

- Scholars Portal uses a systematic convention to generate unambiguously unique identification for digital objects within its repository. This convention will create a stable name or reference to an object that can be permanently associated with that object, regardless of future changes to organizational structure or to digital access protocols.

- This is in conformance with section 4.2.4 of Metrics for Digital Repository Audit and Certification (CCSDS, June 2009) which states that a compliant repository “shall have and use a convention that generates persistent, unique identifiers for all AIPs” and “its components.”

- This convention will ensure that “each AIP can be unambiguously found in the future” and that “each AIP can be distinguished from all other AIPs in the repository”

2. Implementation

2.1. Journal articles

2.1.1. Scholars Portal URIs are consistently constructed in the following manner:

- /<ISSN>/v<volume number>i<issue number padded to four digits>/<article hash>

- The article hash is generated by concatenating the starting page number of the article, an underscore character, the first letter of the first six words in the article title, and the first letter in the last six words in the article title. In cases where there are not enough words in the article title to construct to this specification, the first letters of each word in the title are used.

Examples:

- ***Article***: DNA-Directed Self-Assembly of Gold Nanoparticles onto Nanopatterned Surfaces: Controlled Placement of Individual Nanoparticles into Regular Arrays - ACS Nano (October 2010), 4 (10), pg. 6153-6161

- ***URI***: /19360851/v04i0010/6153_dsognooinira

- ***Article***: A possible chemical burn to the scalp following hair highlights - Burns (June 2005), 31 (4), pg. 530-531

- ***URI***: /03054179/v31i0004/530_apcbttsfhh

In the case of a collision, the URI that was generated will be appended with an underscore (_) and a sequential number beginning with one. This number will increment for each duplicate URI.

Example:

- ***Article:*** Book Review - Journal of the Franklin Institute (September 1944), 238 (3), pg. 224-224
- ***URI:*** /00160032/v238i0003/224_br

- ***Article:*** Book Review - Journal of the Franklin Institute (September 1944), 238 (3), pg. 224-224
- ***URI:*** /00160032/v238i0003/224_br_1

In the case of a replacement article, the new copy of an article will supersede the old and claim the original identifier. The old copy of the article will retain the original identifier with “_old1” appended to the end. In the case of subsequent replacements, the replaced article will be appended with "_old<X>", where <X> is the next available integer.

Example:

- ***Article:*** The Myth of the Unicorn - Diogenes (September 1982), 30 (119), pg. 1-23

- ***URI:*** /03921921/v30i0119/1_tmotu

- ***Old Article:*** The Myth of the Unicorn - Diogenes (September 1982), 30 (119), pg. 1-23

- ***URI:*** /03921921/v30i0119/1_tmotu_old1

The URI is used not only as the unique identifier for the item, but also as a path to the item's file in Marklogic on both the search index and the preservation database, it provides a link between the preservation engine and the Scholars Portal search engine. The Marklogic path takes the form of:

http://journals2.scholarsportal.info/details.xqy?uri=/03054179/v31i0004/530_apcbttsfhh.xml

2.1.2. URIs for individual files and events created by Scholars Portal

Scholars Portal URIs for individual files and events are consistently constructed in the following manner:

- The URI of the parent object and a hash generated from the current date/time are concatenated. PDF files are identified by adding “pdf_fulltext” after the parent object URI and then the current date\time hash. While XML files are identified by adding “xml_fulltext” after the parent object URI and then the current date\time hash. In this way, there should be no chance of collisions.

Examples:

- ***Parent object:*** A C. elegans LSD1 Demethylase Contributes to Germline Immortality by Reprogramming Epigenetic Memory - Cell (April 2009), 137 (2), pg. 308-320

- ***Parent object URI:*** /00928674/v137i0002/308_aceldcgibrem

- ***PDF Fulltext:*** /00928674/v137i0002/308_aceldcgibrem/pdf_fulltext/1303399300907
- ***XML Fulltext:*** /00928674/v137i0002/308_aceldcgibrem/xml_fulltext/130339930294
- ***Other files:*** /00928674/v137i0002/308_aceldcgibrem/1303399302709
- ***Events:*** /00928674/v137i0002/308_aceldcgibrem/1303399314628

2.1.3. File system path structure for content files

- Content objects are stored in the same directory structure in which they arrived, which is transferred to a filesystem specific to the collection (e.g., ejournals1), in a directory specified for the publisher and named at the time a loader script is written. Both the filesystem name and the publisher directory name are followed by a sequential identifier, starting with 1. Publisher directories are limited in space to 200GB, so when the directory reaches that size, a new one is created, and the sequential identifier is incremented. (e.g. kluwer1, kluwer2, kluwer3)

- When the current filesystem reaches 2 TB in size, it is unmounted and remounted as a read-only volume. At this time, a new filesystem is created, and the filesystem's sequential identifier is incremented. (e.g. ejournals1, ejournals2) Any publisher directories that were on the old filesystem are considered closed, and the sequential identifier will be incremented for the next directory created for that publisher.

Please see Move Event Diagram

Example paths:

- /mnt/pillar/ejournals2/wiley5/

- /mnt/pillar/ejournals3/wiley6/

- /mnt/pillar/ejournals3/ieee2/

3. References

3.1. Metrics for Digital Repository Audit and Certification (2009) CCSDS XXX.0-R-1. Red Book. Issue 1. June 2009.

4. Document History

Version	Date	Change	Author
0.1	08/23/11	Draft created	Aurianne Steinman
0.2	09/01/11	Draft formatted	Aurianne Steinman
0.3	09/20/11	Suggested edits	Aurianne Steinman

File	Modified
Microsoft Word 97 Document SP.URIFilenaming.doc	Dec 15, 2011 by Aurianne Steinman