Quality Control Specifications

Scholars Portal Quality Control Specifications

1. Policy Statement

Scholars Portal is committed to ensuring that the integrity of digital objects within the repository is maintained.

1.1. Quality Control Standards

- Upon ingest, SP requires that publishers provide their content in PDF format and in XML or SGML (in a format agreed upon between SP and the publisher) containing descriptive metadata, and full-text content is strongly encouraged.
- Every time the digital object is moved during the ingest process, a fixity check is performed. This ensures that the file has been transferred correctly, and not become corrupted in the process.
- Any errors are recorded automatically in an error log, as well as in the publisher problem directory. A notification of error is emailed immediately to the metadata librarian. The errors are troubleshot and corrected as soon as possible.
- Descriptive metadata is normalized to a SP-specific profile of the NIH Journal Archive & Interchange Tag set. Transformed metadata is validated against a DTD to ensure its compliance.

1.2. Organizational Responsibility

Please refer to the Scholars Portal Roles and Responsibilities document for delineation of responsibilities by staff member.

Please refer to the Scholars Portal Organizational Chart for the structure of the organization.

2. Implementation Examples

2.1. Procedures for data integrity testing:

2.1.1. Test during pull script (see Pull Script Detail):

- After a new dataset is saved into the Ejournals FTP in Pillar, it is retrieved and the file size is compared to that of the original copy held in the publisher FTP server.
- If the file size matches, the script makes a record of the FTPed files with the file name, size, and current date and adds the file name to the FTP downloaded log file.
- If the file size does not match, the script sets the error flag and increments the try count. Once the try count hits three and there is still an error flag, the file is deemed corrupt and an email is sent to loaders@scholarsportal.info.

2.1.2. Test during the preparation of datasets (see Prepare Datasets Detail):

- New datasets are retrieved from the Ejournals FTP in Pillar and decompressed.
- If decompression is successful, the script checks for the file name in the publisher error log and removes the file name from the publisher problem directory if it exists.
- If there is an error during decompression, the script writes the file name to the publisher error log and moves the error file to publisher problem directory. The zip file information is then emailed to JIRA.

2.1.3. Test during ejournals loader (see Manual Log File Check):

- New datasets are converted from the publisher XML/SGML into NLM XML, given a URI, and inserted into the ejournals database. Any errors are recorded in a log file.
- Either the SP programmer or metadata librarian manually checks the log file for errors during the conversion and insertion into the ejournals database of each dataset.
- If the conversion or the insertion failed, the SP programmer or metadata librarian investigate the error log file to find out where the error occurred.
- If the error is due to a publisher problem, the publisher reloads the SIPs.
- If the error is due to a loader problem, a reworking of the loader script is necessary.

2.2. Control of incoming data (see Pull Script Detail):

- Depending on the publisher, incoming data is either pulled or pushed from the publisher FTP into SPs ejournals FTP in Pillar.
- If it is pushed in by the publisher, they send new content to the location that SP gives them and notifies the SP programmers or metadata librarian.
- If it is pulled in by SP's pull script, the process is activated daily at 11am.
- The pull script loads configuration properties of the publisher and connects to the publisher FTP server using the FTP username and password.

- The script retrieves the file names from the publisher FTP server and compares them to names in the FTP downloaded log file to look for new file names.
- If a new file name is found in the publisher FTP server, the script creates a new dataset and saves it in the ejournals FTP in Pillar.
- SP ingests new files at a controlled rate.

2.3. Explanation of repository workflow (see Workflow Charts)

2.4. Metadata fields with quality control information (see Data Dictionary):

- objectCharacteristics – contains technical properties of a file or bitstream that are applicable to all or most formats. It includes the following:
- fixity – information used to verify whether an object has been altered in an undocumented or unauthorized way
- size – size in bytes of the file or bitstream stored in the repository
- format – identification of the format of a file or bitstream where format is the organization of digital information according to preset specifications
- storage – information about how and where a file is stored in the storage system

3. Related Documents

- **Please refer to** the Document Checklist for related Scholars Portal Policies.
- **Please refer to** Audit and Certification of Trustworthy Digital Repositories (TRAC) and Reference Model for an Open Archival Information System (OAIS) for related external documents.

4. Definitions

- **Please refer to** Acronyms and Glossary for Scholars Portal definitions.

5. Document History

Version	Date	Change	Author
0.1	10/07 /11	Draft created	Aurianne Steinman
0.2	10/13 /11	Steve's edits	Aurianne Steinman
0.3	10/28 /11	Edits by Steve	Steve Marks
0.4	11/02 /11	Updated doc file, minor textual corrections	Karl Nilsen

File	Modified
Microsoft Word 97 Document Quality Control Plan.doc	Nov 02, 2011 by Karl Nilsen