The Document Repository Service

(9/23/2016)

Repository ID: TI.2.1
Authors: Tom Barton
Emily Eisbruch
David Walker <https://orcid.org/0000-0003-2540-0644>
Ann West
Mike Zawacki
Sponsor: Internet2
Superseded documents: (none)
Proposed future review date: March 1, 2017
Subject tags: admin, service

Introduction
The Repository Librarian
The Repository Platform
Approved Preservation Formats
Document Metadata
Other Uses of the Repository Platform

Introduction

The document stewardship life-cycle described in Trust and Identity Document Stewardship requires the existence of a Document Repository Service. This Service includes both a technical platform for the repository itself and administrative/curatorial functions performed by a Librarian. It is the purpose of the Document Repository Service to:

- retain authentic copies of all documents that have been placed in it, and
- provide basic discovery services to help people find those documents.

This document describes the components of that Service.

The Repository Librarian

The repository service is more than a technology platform. It requires someone to manage the document collections stored in the Repository. This Document Repository Librarian will perform tasks associated with the Document Repository, such as:

- assign Repository IDs
- provide guidance and oversight for the assignment of subject tags
- process submissions to the repository, supplying metadata values, such as Repository ID and current status, and ensuring that other metadata values have been provided
- monitor submissions to identify and perform required updates for other documents' metadata regarding deprecation and related documents
- monitor expiration dates to inform sponsors of upcoming reviews
- manage Document Object Identifiers for documents in the repository
- manage the assignment of subject tags and review old documents periodically in light of new tags
- manage the relationship with the technical support organization for the Repository Platform

The Repository Platform

The repository technology platform must address the needs of the service. This includes:

- Facilitation of long-term preservation of documents with permanent identifiers and URLs
- Support for the Document Metadata described below
- Support for document discovery, based on metadata elements and full-text search

The platform must provide appropriate availability and survivability of its content, the metadata, and the Service's administrative functions.

Selection of the repository platform will be done after review of this document. The following are evaluation criteria:

- The platform service must be designed to preserve the repository's contents for as long as Internet2 has an interest in maintaining access to its documents.
- It must be possible to search documents based on their metadata.
- It is desirable that it be possible to create document indexes based on metadata searches. (*E.g.*, create an index containing all documents authored by the InCommon TAC, or all documents with a #SAML tag.)
- It is desirable that the platform support full-text search across all documents.
- When there are multiple versions of a document, the platform must provide clear indication of which version is current, and which have been deprecated.
- The platform must support stable, unchanging URLs for documents, as well as a strategy for maintaining URL validity after a platform change.
- It is desirable that documents' URLs be readily mapped from the documents' identifiers.
- The platform must provide very high disaster recovery capabilities.
- The platform must support point in time backup and recovery, largely to recover from librarian errors.
- The platform must provide high availability and good responsiveness.
- The platform must support the following administrative functions:
    - Upload documents
    - Remove documents
    - Manage metadata for documents
    - Extend the types of metadata available for documents
    - Manage multiple versions of a document with a stable URL for the current version
- It is desirable that the platform support automation of administrative workflows.

Potential candidates for the platform include:

- The Internet2 Spaces Wiki
- Internet2's Django CMS
- Internet2's Box service
- Professionally-operated preservation repositories, such as that run by the California Digital Library or PSU's ScholarSphere
- Open source repository software, such as ePrints from the University of Southampton
- A geographically-mirrored web service
- Commercial offerings, such as Sharepoint, Google Docs, *etc*.

It is not necessary that a single repository be used for both access and preservation. An alternative strategy could be to operate an access repository, and then contract with one or more (possibly dark) preservation repositories to retain copies of the documents in the access repository, linking the two with Document Object Identifiers (DOI).

Approved Preservation Formats

- Documents must be plain text, PDF, or HTML. PDF documents must be accompanied by a plain text or HTML representation of the document.
- Other file formats may be placed in the Document Repository, but the Service will not expend effort to preserve their information integrity if technical support for those formats is not commonly available at some time in the future.

Other preservation formats may be approved in the future in consideration of the following factors:

- Ability to support presentation and discovery
- Resistance to requiring ongoing conversions to maintain integrity of the information
- Availability of tools for accessing and manipulating the information
- Quality of documentation of the format
- Lack of encumbrance by licenses

Document Metadata

A number of metadata elements will be associated with a preserved document:

- Repository ID, “TI.” followed by a sequential number that is never reassigned to another document, another “.”, and a sequential version number starting at 1 (*e.g.*, “TI.46.3” would indicate the third version of document 46.)
- Title
- Author (potentially multiple people and/or a group). This will be structured to allow specification of author identifiers, such as ORCiDs.
- Sponsor
- Brief description of review process (perhaps a copy of the charge to the author)
- Current status (Review / Preserve / Legacy / Not Reviewed / Unsponsored)
- Publish date
- Document Object Identifier (DOI)
- Digital signature to authenticate the document (if provided by the Author)
- Deprecated (yes / no)
- Future review date
- Superseded documents (Repository IDs of documents that have been made obsolete by this one)
- The file format(s) of the document
- Related documents (*e.g.*, Repository IDs of published documents or Work History produced as part of the same effort)

- Location of development work (*e.g.*, URL for the wiki space for ongoing work by the authors)
- IP Framework, Copyright & License (default Creative Commons).
- Subject tags

Other Uses of the Repository Platform

It is anticipated that the Repository Platform will be used for documents other than those described in Trust and Identity Document Stewardship. Other stewardship frameworks may evolve, potentially including a framework for documents that are self-stewarded by members of Internet2's constituency. Using the same Platform for multiple frameworks will enable discovery of documents across frameworks, as well as affording more effective use of resources through sharing.

When this is done, however, a different series of Repository IDs should be assigned (*i.e.*, that don't start with "TI"), and the role of the Librarian within the new framework should be determined.