Gap Analysis of the Document Repository

Criteria for the technology platform used to support Trust and Identity's document repository were outlined in The Document Repository Service, prior to the first implementation of that service, the Trust and Identity Document Repository Index. As an expedient, Confluence was used for that "proof of concept" implementation, and DOIs were established for the documents to establish persistent URLs, allowing the documents to be moved to a new repository at a later time. Now that the repository's public release is imminent, it's time to understand how well our original criteria were met.

In addition, it is important to bear in mind that the Document Repository methodology (including assignment of DOIs) that has been developed for Trust and Identity may be eventually expanded to be leveraged by other Internet2 divisions.

Contents

Gap Analysis

To quote from The Document Repository Service:

> 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.

This is followed by a list of criteria for the section of the technology platform, as outlined in the following table.

	Criteria from The Document Repository Service	Status	Analysis of the Trust and Identity Document Repository Index
1	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.	⊖	*(Internet2 has not formally specified a target for how long these documents should be preserved. Informally, though, it's been recognized that it should be possible to preserve these documents even if Internet2 ceases to exist.)* Internet2's Confluence service is not a good choice for long-term preservation. While a geographically distant copy is maintained for disaster recovery, there is no provision for fixity to address "bit rot," nor are point-in-time backups maintained to recover from human or software error.
2	It must be possible to search documents based on their metadata.	⊖	Confluence can perform full-text search, although there are not strong methods for limiting the scope of pages to be searched, so it is not possible to search only the repository.
3	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 possible to create multiple indexes, but the filter criteria for each index cannot be aligned with specific metadata fields.
4	It is desirable that the platform support full-text search across all documents.	⊖	The documents are stored as attachments, so they are not searchable.
5	When there are multiple versions of a document, the platform must provide clear indication of which version is current, and which have been deprecated.	⊖	Metadata fields were established to address this. [Update 3/30/2018: The metadata fields do not provide clear indication, so this was change to a negative status from positive.]
6	The platform must support stable, unchanging URLs for documents, as well as a strategy for maintaining URL validity after a platform change.	⊕	A DOI prefix, 10.26869, was acquired for the repository, enabling the creation of unchanging URLs of the form http://doi.org/10.26869/TI.5.1.
7	It is desirable that documents' URLs be readily mapped from the documents' identifiers.	⊕	The document's Repository ID (*e.g.*, TI.5.1) is easily identifiable within its URL.
8	The platform must provide very high disaster recovery capabilities.	⊕	Internet2 maintains a backup server geographically distant disaster recovery site.
9	The platform must support point in time backup and recovery, largely to recover from human or software errors.	⊖	There are no provisions for point in time backup and recovery.
10	The platform must provide high availability and good responsiveness.	⊕	Internet2's Confluence service provides high availability and good responsiveness.

12	The platform must support the following administrative functions: Upload documents.	⊕	Confluence provides upload capability for attachments. The size is limited (tens of megabytes), but that has been sufficient for nearly all of our documents to date.
13	The platform must support the following administrative functions: Remove documents.	⊕	Documents can be removed from the repository. It is not possible to recover them, however, if the removal was in error.
14	The platform must support the following administrative functions: Manage metadata for documents.	⊕	Metadata is managed through normal wiki page editing. There is no automatic validation or enforcement of specific fields, so editors must be careful.
15	The platform must support the following administrative functions: Extend the types of metadata available for documents.	⊖	This is possible but requires manual editing of all documents' metadata to ensure consistency across all documents.
16	The platform must support the following administrative functions: Manage multiple versions of a document with a stable URL for the current version.	⊖	Multiple versions of documents are handled as multiple documents, each with its own metadata. There is no provision for a stable URL that maps to the current version as new documents become "current." In the "TI.<document>.<version> Repository ID, the one with the highest <version> is current.
17	It is desirable that the platform support automation of administrative workflows.	⊖	Confluence is limited with respect to its automation capabilities. Support for WebDAV would improve the ability to automate tasks.
18	The platform must be accessible, compliant with WCAG and other applicable standards.	⊖	Atlassian is committed to accessibility for Confluence, and provides a VPAT. Most applicable VPAT criteria, however, are marked "Supports with exceptions," and there are many unresolved JIRA issues linked as exceptions.
	Criteria Discovered During Implementation		
19	The platform must be capable of registering DOIs with CrossRef, based on metadata that has been entered in the repository.	⊖	CrossRef provides a RESTful API for registering DOIs, but nothing exists for our Confluence repository to use it.

Alternatives for the Future

Partner with an Existing Repository Service

Probably the most effective way to achieve most or all of the criteria listed above would be to partner with a University or other institution that already operates a document respository. The number of documents Internet2 and its community will contribute to the repository will be very small, as compared to a university repository, and yet the degree of service and protection of the documents Internet2 desires is very similar. Such a partnership could be of great benefit to Internet2.

Acquire and Operate Repository Software

Absent a partnership, Internet2 must strike out on its own. There is some open source software, such as ePrints from the University of Southampton, which would likely meet more of our criteria. Internet2 could survey what's available and make a selection. Unfortunately, much of this software may be aligned to the practices of a particular institution, so viable alternatives for Internet2 may be few or nonexistent.

Enhance the Existing Confluence-Based Repository Platform

Little can be done about the limitations of the Confluence-based repository, particularly regarding its ability to support:

- Searches related to specific metadata fields
- Full-text search of the documents themselves
- Accessibility

There are things that could be done to improve long-term preservation of the documents and to ease the administrative burden:

- Long-term preservation
  - Implement point in time backup for the documents and their metadata, setting a retention schedule to address the risk of human and software error.
  - Compute fixity hashes of all documents and store them in the documents' metadata. Perform periodic checks of the hashes on a scheduled coordinated with the point in time backups. This will require software development, although computation and verification of hashes (*e.g.*, SHA-n) is available off the shelf.
- Administration
  - Enable WebDAV access to the repository to facilitate the implementation of software tools to support administrative functions.
  - Develop software to support the following functions:
    - Register a DOI for a specified document metadata page.
    - Create the metadata page for a new document, register its DOI, and compute its fixity hashes.
    - *Ad hoc* scripts to perform global changes to the metadata pages.

Preservation Past the Lifetime of Internet2

Preservation of documents after a possible (but hopefully improbable) demise of Internet2 requires agreements and technology that facilitate transfer of stewardship in that eventuality. Software like LOCKSS (Lots Of Copies Keep Stuff Safe) and partnerships like the Digital Preservation Network ("...an independent organization under the umbrella of the not-for-profit organization Internet 2 (*sic*)...") can help address this issue. An existing university repository service will undoubtedly have dealt with this issue, but Internet2 should join such partnerships if it decides to operate its own repository.