AIC Use Case: De-duplication

Title (Goal)	De-duplication of non-RDF (binary) resources
Primary Actor	Repository manager / depositor
Scope	Component
Level	
Author	Stefano Cossu
Story	As a staff user who ingests and/or manages contents in a Fedora repository, I want to make sure that my colleagues and I are not ingesting the same file in two different locations. While Fedora does a good job at avoiding binary content duplication under the hood, there is still a use for avoiding the creation of two separate resources referring to the same content. Users may add metadata and relationships to either one resource, making it harder to get consistent information.

Web Resources and Interactions

This extension would initiate a process after a client sends a POST or PUT request containing a binary file, and before this file is persisted as a Fedora resource.

The process calculates the checksum of the incoming file and queries a triplestore or Solr index to check if a resource with the same checksum exists.

If such resource does not exist, the extension proceeds with the ingestion process or other pre-ingest extensions.

If such resource exists, a 409 Conflict error is returned indicating the location of the duplicate image and the ingest process is aborted.

The extension must accept a header or POST parameter that allows forcing duplication. When this parameter is set, the extension may add a header to the response with the location of the duplicate file, but lets the ingest process proceed. This allows the client to issue an "Are you sure?" warning to the end user.

Preconditions

The API extension architecture should expose an event hook that allows to redirect the ingestion flow before the resource is persisted.

Deployment or Implementation notes

Search for checksums via indexes is probably the only feasible way, but it does not absolutely guarantee de-duplication due to the asynchronous nature of these. A routine scan for duplicates may be advisable.

Proposed Requirements

Calculation of checksum may leverage existing Fedora functionality.

API-X Value Proposition

This extension has similar requirements as validation extensions. Depending on how these are designed, this may become a particular case of validation. The barrier to implement de-duplication as such may be quite low.

This functionality is crucial to ensure consistency and discoverability of data in repositories where many users from many different departments may have the same files and they want to ingest them.