

Global Format Registry Use Case: Validation of Data files in a SIP

Use Case ID	NYU-1
Description	A software agent for an OAIS-compliant repository, presented with a data file with its format tentatively identified by technical metadata contained with a SIP, queries the Registry for representation information necessary to confirm the data format for the file, and uses that information to validate the file type prior to ingesting the content into the repository.
Actors	<i>Registry</i> – provides representation information regarding a particular file format <i>Repository ingest agent</i> – requests representation information from the registry for a particular file format
Assumptions	Registry must possess representation information regarding particular file format. Representation must be stored in such a way that it can be disseminated in a form suitable for machine processing. Registry must have legal authority to disseminate representation information regarding format to requesting agent. Repository Agent must possess sufficient information regarding data file within a SIP to formulate a precise request to Registry. Repository Agent must be able to use returned representation information to confirm format of data file
Preconditions	Registry and Repository agent must share application protocol for request/dissemination of representation information.
Triggers	A SIP is submitted to repository for ingest, at which point a software agent attempts to tentatively identify the format of included data files packaged within the SIP, and requests representation information from the registry to confirm its identification
Primary Scenario	Step 1 -- Repository Agent receives SIP and decomposes it to identify the data files contained within the SIP, and tentatively identify a specific data format for each file.
	Step 2 – For each tentatively identified data format, Repository Agent submits a query to Registry for Representation Information needed to confirm the data files conformance to the data format (see Use Case 2)
	Step 3 – Registry returns requested representation information.
	Step 4 –Repository Agent uses returned information to confirm data format for each file
	Step 5 – Assuming that all data files conform to their purported data formats, ingest process proceeds. If the data files do not conform, ingest halts and the SIP is set aside for human review.
Primary Result	The process minimizes human processing of the SIP and ensures that data files ingested into the system have had their data format properly identified before ingest proceeds.
Post-Conditions	
Non-functional requirements	
Notes	

Issues	Highly detailed representation information will need to be stored by the registry, and this information may vary significantly from data type to data type. The information that a software agent will require to determine whether a document is a TEI P2 document will differ a great deal from what it will need to determine whether a document is an OSIRIS data dictionary file.
--------	--