BagIt package communication from SEAD Active Curation Repository to Virtual Archive

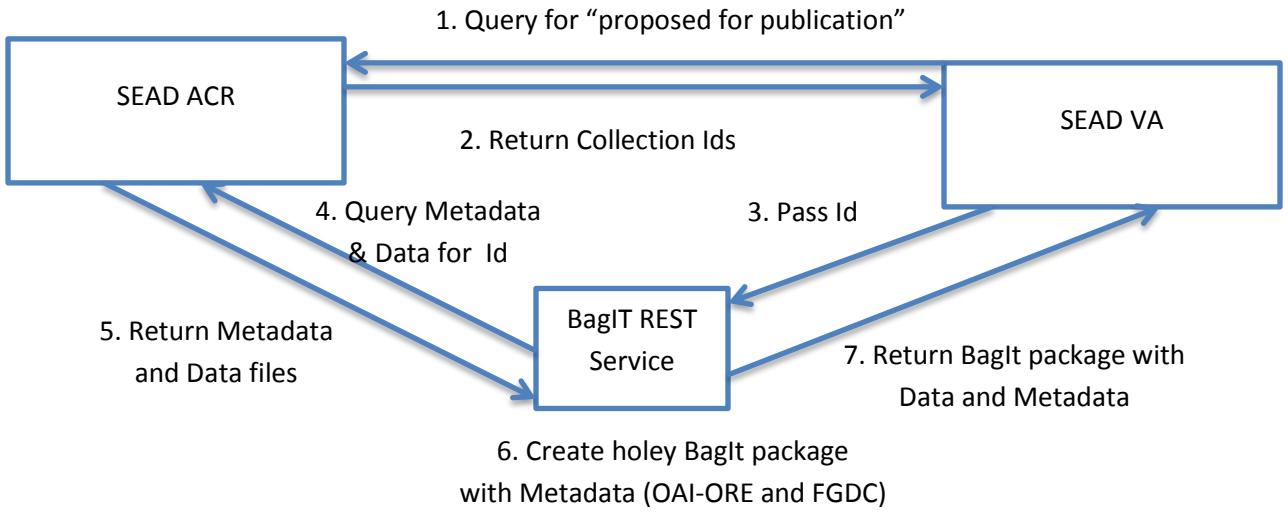
Produced by Kavitha Chandrasekar on 06/28/2013

This document is only based upon what has been implemented as an initial design to transfer contents from SEAD Active Curation Repository (ACR) to SEAD Virtual Archive (VA).

"BagIt Rest Service" has been implemented by SEAD VA as a light-weight service which sits in front of VA and queries ACR for metadata and returns a BagIt package.

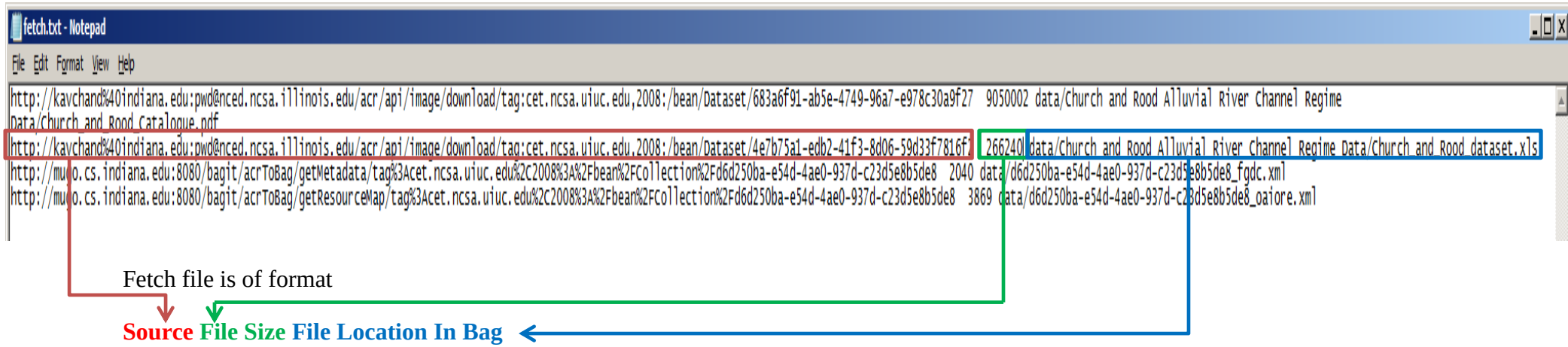
This service allows standard BagIt package to be the transfer mechanism between ACR and VA. As a first step, we make use of **"holey" Bags** [http://tools.ietf.org/html/draft-kunze-bagit-04#section-4.3] in BagIt specification to be the standard transfer mechanism. This was chosen as a first step because of the flexibility it offers for transferring of data by requiring only metadata i.e. actual source of files in a bag and not the actual data files itself. This seemed like a better method to scale when data package to be transferred is of the order of 100s of GB needs to be transferred.

The BagIt transfer between ACR and VA is described here:



The holey Bag consists of a fetch and bag-info.txt file. BagIt transfer tools from Library of Congress [http://sourceforge.net/projects/loc-xferutils/files/loc-xferutils/] can use fetch.txt and manifest-md5.txt files to generate/download the entire Bag with data. So we make sure to include these files in the bag.

A **fetch file** (fetch.txt) for Church Alluvial data Collection from ACR looks like this:



Fetch file is of format

Source File Size File Location In Bag

A **manifest file** (manifest-md5.txt) is of format:

Column1-Checksum Column2-File-location-in-Bag

We are not including a sample in this document as it is a simple file format.

The **OAI-ORE file** has the following details defined:

The name spaces used are:

dc="http://purl.org/dc/elements/1.1/"

j.0="http://purl.org/spar/cito/" (This was based on DataONE's usage of OAI-ORE)

ore="http://www.openarchives.org/ore/terms/"

1) Aggregation relationships from OAI-ORE (**ore:aggregates and ore:isAggregatedBy)**. We also allow for hierarchical relationship for aggregations in OAI-ORE, by using linking to external child OAI-ORE files.
2) We allow for external metadata files to be added as an internal/external reference (**j.0:isDocumentedBy)**
3) Identifier for files/collections (**dcterms:identifier)**
4) Title for files/collections (**dcterms:title**)
5) File Format from ACR (**dcterms:format**)

FGDC Metadata File:

We include an additional FGDC file so that description metadata about the data collection can be specified more completely using a scientific metadata format for spatial, temporal metadata, keywords with thesaurus etc. This FGDC is automatically generated based on metadata from ACR.

This BagIt is converted to SEAD-VA SIP before being going through preservation workflow and being archived in long term. A holey Bag is an accepted import package in VA and we will possibly be looking to enable BagIt as an export package in as well.