SEAD BagIt Service and BagIt package format for ingest to Virtual Archive

V1.0 06/28/2013

Author: Kavitha Chandrasekar

1. Introduction

SEAD offers several software tools and services that are designed to lower the barrier to data management for researchers in the long tail of science, where long-tail in this case refers particularly to hydrology, environmental science, social-ecological systems, social science, and earth science. The three primary tools are a user profile service that links people, data and publications; an active curation environment that combines research tools and smart data management (SEAD Active Curation Repository, ACR), and a federation service that serves as a single point of ingest/submission and access to long tail data for multiple academic institutional repositories, SEAD Virtual Archive (SEAD VA).

This document describes the protocol by which SEAD VA accepts research objects that are submitted to it through an interface that accepts a BagIT package.

In general, SEAD VA has a separate BagIT service that can be adapted to work with any metadata storing service. In SEAD's case, the BagIT service interacts with SEAD ACR to extract relevant metadata that SEAD ACR has about a research object, and does so in the process of constructing the BagIT payload. The BagIT object is expected to include a metadata file for scientific metadata, currently FGDC is supported. The research object once arrives at SEAD VA is queued for manual data curation. SEAD VA relies on the active curation tool, SEAD ACR, to automatically derive the temporal, geolocation, and non-scientific metadata for the research object and make this available through its SPARQL interface. This automatic metadata collection reduces the manual data curation required at time of ingest.

In SEAD v1.0 release, the publish event is triggered by a scientist within the ACR environment, and the ingest event into SEAD VA is initiated by the data curation specialist working through the SEAD VA interface.

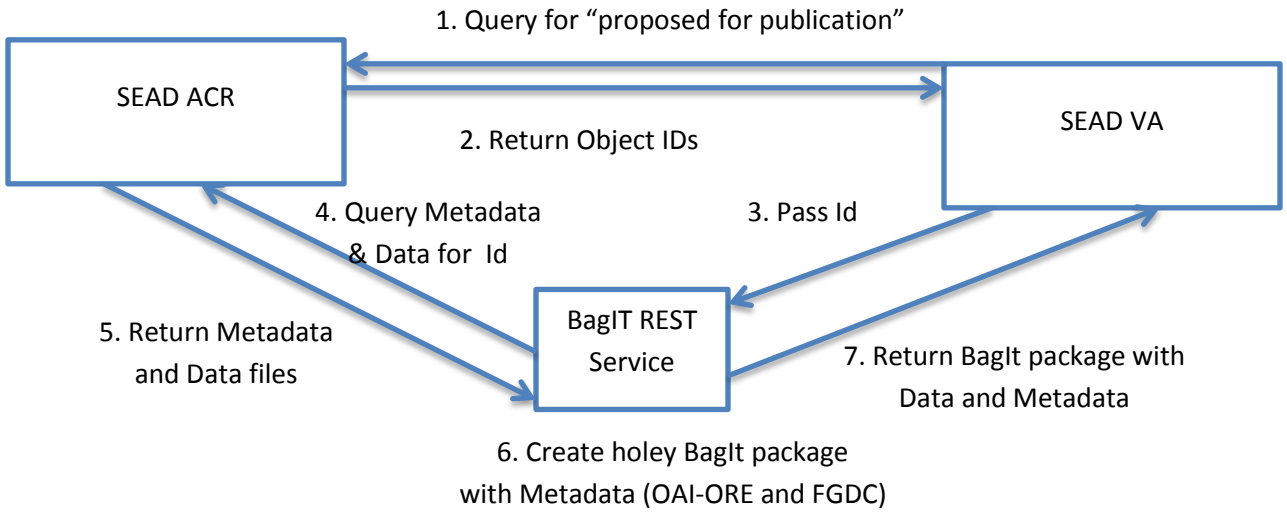
Post v1.0, research object ingest into SEAD VA will be available directly through a REST interface, allowing extensibility of the system. The Bag is being extended to include richer metadata in the ORE to describe the research object.

See source repository for documentation on BagIT interface https://github.com/Data2Insight/sead-virtual-archive.

2. SEAD BagIt Service

The *SEAD BagIt Service* supports a standard BagIt package as the means by which research objects are transferred to SEAD VA. SEAD makes use of **"holey" Bags** [http://tools.ietf.org/html/draft-kunze-bagit-04#section-4.3] as the standard package type. This was chosen 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.

The protocol by which SEAD VA accepts packages is outlined in Figure 1.

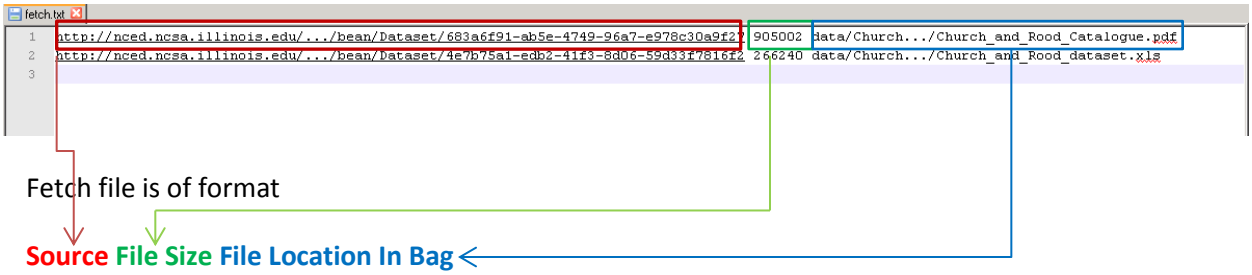


SEAD VA queries the SEAD Active Curation Repository for objects that have been selected for publication. SEAD ACR responds with a list of research object IDs. These IDs are passed to the SEAD BagIt service which queries SEAD ACR through its SPARQL endpoint for metadata about the research object. From this information a holey BagIt package is constructed that is then passed to SEAD VA. SEAD, post release v1.0, is extending the BagIt REST service to push bags directly to SEAD VA.

3. SEAD BagIt Service Format

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/] 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:



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

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

We are including a simple sample with this document for the Bag that is generated [sample/sampleBag.zip].

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:

SEAD 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 file is automatically generated by the SEAD BagIT client using metadata retrieved from ACR.

The BagIt object, once it arrives at SEAD VA, is converted to a SEAD VA SIP as part of the preservation workflow. A holey Bag is an accepted import package in VA.

4. Source Code and Installation:

The source code and installation instructions for the BagIT service can be found at https://github.com/Data2Insight/sead-virtual-archive/tree/master/SEAD-VA-extensions/services/bagItRestService.

API Description:

REST API	Function	Description
bagIt /sip	getSIP(InputStream bag)	Takes a holey bag and returns SEAD VA SIP which can be ingested into SEAD VA.
bagIt/ORE	generateORE(InputStream bag)	Generates ORE for a given holey bag.

bagIt/bag/{collectionId}	getACRBag(String collectionId)	Creates a holey bag for the given collection and returns a zipped holey bag with FGDC and ORE.

5. Future (post v.1.0)

The form of the BagIt bag is very simple and does not fully support the research object that SEAD is pioneering. SEAD and Data Conservancy are currently working on a richer data model for BagIt communication. SEAD VA is additionally looking to enable BagIt as an export package in as well.