Workflow of Operating Your Repository with SEAD

To connect your repository to SEAD, you need to be a registered partner repository. The following steps are needed to register and accept/publish collections from SEAD:

1. Obtain access to SEAD system
2. Register the repository profile
3. Develop a client to pull and publish the collections

1. Obtain access to SEAD system

To obtain access to SEAD system, please contact SEAD and provide the IP addresses/subnets of the repository servers that will be accessing SEAD services. The SEAD team will add those addresses to its registry of partner repositories and grant access to SEAD services for those IP addresses.

Once registered, the following endpoint will be accessible by the repository: https://seadva-test.d2i.indiana.edu/sead-c3pr/.

2. Register the repository profile

Repositories need to create a profile in the JSON-LD format (http://json-ld.org/) that will be used by the SEAD Matchmaker in its pairing between datasets and repositories. The profile should include the following information:

- Repository name ID ("orgidentifier") - a required field that will serve as the repository ID in SEAD.
- Data types (“Data Mimetypes”) - types of files that the repository will accept.
- Collection depth (“Max Collection Depth”) - maximum collection depth that the repository can accept and deposit.
- Maximum dataset size (“Max Dataset Size”) – the size (in bytes) of individual files in the collection that the repository can accept.
- “Total Size” – Total acceptable size of the collection in bytes
- “Metadata Terms“- Minimum metadata fields that a collection should contain
- “Affiliations” - Organizations that a collection should be affiliated with
- “Rights Holder IDs Required” - Whether the collections should have a valid global identifier (ORCID, Clowder, or Google ID) for the "Rights Holder" metadata.

Please refer to http://seadva-test.d2i.indiana.edu/sead-c3pr/api/researchobjects/matchingrepositories/rules for more information about these metadata types.

Following is a sample JSON-LD profile of a repository:

```
{"@context": ["http://re3data.org/",
  {
    "Max Dataset Size": "http://sead-data.net/terms/maxdatasetsize",
    "Rights Holder IDs Required": "http://sead-data.net/terms/RightsHolderIdsRequired",
    "Total Size": "tag:tupeloproject.org,2006:/2.0/files/length",
    "Max Collection Depth": "http://sead-data.net/terms/maxcollectiondepth",
    "motto": "http://bobs.asseenon.tv/terms/motto",
    "Affiliations": "http://sead-data.net/terms/affiliations",
    "Data Mimetypes": "http://purl.org/dc/elements/1.1/format",
    "Metadata Terms": "http://sead-data.net/terms/terms"
  }
],
  "Max Dataset Size": "1000",
  "@type": "repository",
  "Total Size": "10000000",
  "orgidentifier": "bob",
  "repositoryURL": "http://http://www.nationaldataservice.org/projects/labs.html",
  "Rights Holder IDs Required": true,
  "Max Collection Depth": "10",
  "motto": "Our profile is up to date, so we have to be good",
  "repositoryName": "SEAD NDS Labs Publisher (Proof-of-Concept)",
  "Affiliations": [
    "SEAD",
    "NDS Members"
  ],
  "Data Mimetypes": ["text/csv"],
  "Metadata Terms": [
    "http://purl.org/dc/terms/creator",
    "http://purl.org/dc/terms/abstract",
    "http://sead-data.net/vocab/test/doesntexist"
  ]
}
```

Once the repository profile is created, it should be registered with SEAD by sending a POST request to the following endpoint with the JSON-LD profile as the POST body.

https://seadva-test.d2i.indiana.edu/sead-c3pr/api/repositories

Once registered, the newly registered repository will be listed in the above endpoint. You will be able to receive the full profile using the following endpoint.

https://seadva-test.d2i.indiana.edu/sead-c3pr/api/repositories/<orgidentifier_of_your_repository>

3. Develop a client to pull and publish the collections

When SEAD receives publication requests from its users for a repository, it keeps a list of those requests. To access the collections and publish them, partner repositories need to implement an agent that handles the metadata and data retrieval and status updates. You can refer to https://github.com/Data-to-Insight-Center/sead2/tree/master/sead-nds-repository for a sample agent code.

The following operations need to be implemented in the agent:

Retrieve Publication Requests from SEAD

The requests queued for your repository can be retrieved from the following endpoint: https://seadva-test.d2i.indiana.edu/sead-c3pr/api/repositories/<orgidentifier>/researchobjects

If you want only the new requests (the requests that have not yet been started to process by the repository), query the following endpoint: https://seadva-test.d2i.indiana.edu/sead-c3pr/api/repositories/<orgidentifier>/researchobjects/new.

These endpoints return a list of requests that are queued for your repository. The collection identifier can be found in 'Aggregation.Identifier' element in each JSON object that represents a collection. You can retrieve the full request document using https://seadva-test.d2i.indiana.edu/sead-c3pr/api/researchobjects/<request_id>

Accept the collection and publish to a repository

To acknowledge receipt of a request, you're first **required** to send a status message to SEAD to indicate that you have started processing the particular request. For this you should send a POST request to the following endpoint: https://seadva-test.d2i.indiana.edu/sead-c3pr/api/researchobjects/<collection_id>/status

with the post body:

```
{
        "reporter": "<your_org_identifier>",
        "stage": "Pending",
        "message": "The repository is now processing this request"

}
```

Once you send this status message to SEAD with the "reporter":<your_orgidentifier> element, the /repositories/<orgidentifier>/researchobjects/new endpoint will no longer list that collection, but it will still be available in the /repositories/<orgidentifier>/researchobjects endpoint.

Process the Request

The request document contains summary information about the data to be published (e.g. title, abstract, creators, statistics about the # files, total size, included mimetypes) and information about the person making the request (e.g. ID, affiliations) and their preferences for how the request should be processed (e.g. is it a 'test'). Together, this information may be sufficient to allow you to decide to accept it and begin processing it or to reject it. To reject it at this stage or any later stage, you POST status with the "stage": "failure". SEAD will assume that no further processing will be done by the repository and will report the failure to the user in their Project Space.

If you decide to process the request, your next step would be to retrieve the OAI OREMap (a single JSON-LD file that has the complete set of metadata SEAD knows about the data to be published). The location of the OREMap is given in the request as the Aggregation.@id element. As currently implemented, this is always a URL of the form: https://seadva-test.d2i.indiana.edu/sead-c3pr/api/researchobjects/<collection_id>/oremap , but your implementation should not rely on this and should retrieve the link from the request document.

The OREMap describes the structure of the data to be published using a combination of ORE and Dublin Core terms: In ORE terms, the data is represented by an Aggregation that includes a flat list of 'AggregatedResource's listed by ID in the "aggregates" field. SEAD then supplements this with DC Terms "Has Part" relationships that describe the intended hierarchical structure of the data. (This hybrid approach allows a single OREMap to describe the whole publication, while also providing a folder/file - style hierarchy that can be used to structure the archival copy.). A basic strategy for parsing this structure would be to read the metadata for the "Aggregation", find the list of "Has Part" entries, and begin recursively retrieving and parsing the AggregatedResources from the flat "aggregates" list. Each AggregatedResources object will have some metadata including a type and, for collections(v1.5)/folders(v2.0) a further "Has Part" list of its children. Datasets(1.5)/Files(2.) have no subparts and represent an item with an associated content stream (e.g. file contents). This metadata can be used to make decisions about how to structure the data publication in your repository.

The OREMap does not include the file content itself. When your repository is ready, the content for individual objects can be retrieved using the "Aggregate dResource.similarTo" URLs.

SEAD requires that repositories preserve all metadata in the OREmap, but does not define how this is done. Some repositories may be able to parse all metadata in the OREmap and map it to their internal structure/vocabularies. If this is not possible, preserving the OREmap as a file within the data publication is also acceptable. SEAD makes no assumptions about what metadata you will index or display, whether you will infer additional metadata, whether you will reject files (e.g. if one contains a virus), convert formats, or make other changes. SEAD does expect that any changes made will be guided by good curatorial and preservation practices and that the submitters are informed of and agree to any such changes. (SEAD does not define what constitutes 'good' practice and does not consider itself a party in the agreement between submitters and the repository as to what is acceptable. See SEAD's Terms of Services for details).

In SEAD's reference repository, data is laid out in a hierarchical structure mirroring the DC Terms Has Part relationships within an overall BagIT structure. Some of the metadata is parsed to produce BagIt manifest files, and the OREmap itself is added as an additional file in the manifest folder. SEAD follows the DataOne convention to map the @ids used in the OREMap to the local folder/file paths within the Bag. A zipped copy of this BagIT structure is the archival copy that is then stored.

As part of this process, your repository should create a persistent identifier (PID) for the data publication. SEAD does not require a specific PID system, but generally uses DOIs. Handles, or ARKs could also be used.

SEAD has the technical capability to generate DOIs on behalf of a repository, but has not committed to making this available. If your repository would like to take advantage of this, please contact us. The DOI generation service works as follows: To create a DOI (Digital Object Identifier) for the collection, send a POST request to the following endpoint: https://seadva-test.d2i.indiana.edu/sead-c3pr/api/doi

with the post body as follows:

```
{
        "target":"http://landing_page_of_the_collection",
        "metadata":
        {
                "title":"<title_of_the_collection>",
                "creator":"<creator>",
                "pubDate":"<publication date>"
        },
        "permanent":"false"
}
```

"target" is a required field and it should be the landing page of the collection in your repository. Set "permanent" to 'true' only if you need to create a permanent DOI. Otherwise it will create a temporary DOI which will expire in two weeks.

Communicating with the Submitters

As you process the request, you can communicate with the submitters as required by your process. As part of the request, you will get information about the submitter and the creators of the data being published. This information may be as minimal as a name string, but may also include an email address and/or identifier. You can use this information to contact users out-of-band (e,g, by sending an email, looking up a phone number using their ID, etc.). SEAD also enables you to send status messages that the submitter (and others with access in their Project Space) will be able to see in the Staging Area as a status update. In order to communicate other statuses to the submitter, you can update the status using the above mentioned endpoint and the same JSON format. You should always send your organization ID as the value for the "reporter" field, but you may send any value for "Stage" and message ("pending", "failure", and "success" are currently the only reserved terms for the "stage" - status messages with these stage values affect processing in SEAD and must be used as described in this document). These status messages could be informational ("stage":"In Review", "message":" Your submission is in review. This process usually takes 1-2 weeks.") or may be a call for action ("stage","Approval Required", "message":"Your publication has been accepted. Please visit "URL" to accept our terms and conditions and finalize your publication.").

Complete publishing the collection

Once you have successfully published the collection in the repository, you are **required** to send a final status message to the above mentioned endpoint, and with the POST body as follows:

```
{
        "reporter": "your_org_identifier",
        "stage": "Success",
        "message": "<PID of the collection>"
}
```

The value for “stage” should be “Success” and the “message” should contain the persistent identifier (PID) of the collection. For any PID system you use (DOI, ARK, Handle, ..), the URL form for the identifier, e.g. http://dx.doi.org/<id> rather than doi:<id>) is preferred as the value sent in the "message", as the returned string will then be displayed as a live link in the originating Project Space.

Once this message is received, SEAD will handle registering the metadata for the new publication with external catalog(s) (currently DataOne) and returning the PID of the publication for display in the Project Space.

PID Landing Page

Once you have published the data, SEAD assumes that your repository and the PID you generated for the data become the primary means for accessing the data. At present, SEAD retains a record of the publication event including the summary information in the publication request, and may also retain a 'live' copy of the data within the submitters' Project Space. While this information is retained, SEAD may generate web pages that list these publications (e. g. as a list of datasets published through SEAD, or as lists of datasets published by specific projects) and may associate the 'live' dataset with any published versions. All of these displays will use the URL form of the PID you supply to refer viewers to the authoritative data copy you are preserving. Conversely, you are welcome to use the information provided in the request to link back to SEAD (in general, to the specific publication event, or to the live dataset, etc.) as you find useful. This could allow visitors to your repository to learn, for example, that further versions of the data have been published, that the live copies have further annotations and/or revisions since publication, that the same project has published related datasets to other repositories, etc. If you are interested, it may be possible for you to use SEAD's APIs to retrieve some of this information and incorporate it into the landing page you maintain.