1 Introduction

Information technology and resources have become integral and indispensable to the pedagogic mission of the University of California. Members of the UC community routinely produce and utilize a wide variety of digital assets in the course of teaching, learning, and research. These assets represent the intellectual capital of the University; they have inherent enduring value and need to be managed carefully to ensure that they will remain available for use by future scholars. Within the UC system the UC Curation Center (UC3) of the California Digital Library (CDL) has a broad mandate to ensure the long-term usability of the digital assets of the University. UC3 views its mission in terms of digital curation, the set of policies and practices aimed at maintaining and adding value to authentic digital assets for use by scholars now and into the indefinite future [Abbott].

In order to meet these obligations UC3 is developing Merritt, an emergent approach to digital curation infrastructure [Merritt]. Merritt devolves infrastructure function into a growing set of granular, orthogonal, but interoperable micro-services embodying curation values and strategies. Since each of the services is small and self-contained, they are collectively easier to develop, deploy, maintain and enhance [Denning]; equally as important, since the level of investment in and commitment to any given service is small, they are more easily replaced when they have outlived their usefulness. Yet at the same time, complex curation functionality can emerge from the strategic combination of individual, atomistic services [Fisher].

The Merritt Ingest service manages the acquisition of new digital content supplied by content producers into a Merritt curation environment.

NOTE The key words “MUST”, “MUST NOT”, “REQUIRED”, “SHALL”, “SHALL NOT”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY”, and “OPTIONAL” are to be interpreted as described in RFC 2119 [RFC2119].

2 Requirements

The Merritt Ingest service MUST meet the following functional and non-functional requirements:

- The service SHALL impose minimal requirements regarding the acceptable structure of submitted digital objects.
  - The simplest acceptable submission SHALL be a single file.
  - The minimal administrative metadata that MUST accompany a submitted object are the identifiers for the submitting user agent and the Storage service and storage node in which the object will be deposited. The Storage services and nodes MAY be constrained to those authorized for the submitting user agent.
The minimal descriptive metadata that MUST accompany a submitted object is its primary ARK identifier. This requirement SHALL be fulfilled by specifying either a pre-existing identifier or the noid minter to be used in minting a new identifier.

- It SHALL be possible for multiple digital objects to be submitted in a single operation.
- Submission SHALL be possible without requiring a client-side web server.
- The service SHALL provide interfaces for both synchronous and asynchronous processing. The synchronous interface MAY be restricted to administrative use only.

3 Ingest

The Merritt Ingest service is based on the following conceptual entities, each defined in terms of its specific state properties.

- Service.
- Queue.
- Batch.
- Job.
- Profile.
- Handler.

The Ingest service supports methods that can be used to manipulate and access these entities and their state in useful ways.

All Merritt ingests are relative to one of more digital objects [Object], either pre-existing objects or objects newly established as a side-effect of the ingest operation. An ingest operation always results in a new version of the relevant object.

3.1 Service

The initial conceptual entity is the Ingest service itself, which provides a mechanism to acquire new digital content in a Merritt curation environment. The global state properties of the service MUST include:

- Service name. [ing:name]
- Service identifier, assigned to be globally unique among all UC3-controlled instantiations. [ing:identifier]
- Service description. [ing:description]
- Service implementation version. [ing:serviceVersion]
- Actionable reference to queue state. [ing:queueState]
- Actionable references to Storage service instantiations known to the service.
  - Actionable reference to Storage service state. [ing:storageState]
- Number of jobs currently in the queue. \([\text{ing:}\text{numJobs}]\)
- Total number of jobs processed. \([\text{ing:}\text{numTotalJobs}]\)
- Last submission date/timestamp. \([\text{ing:}\text{lastSubmission}]\)
- Last consumption date/timestamp. \([\text{ing:}\text{lastConsumption}]\)
- Creation date/timestamp. \([\text{ing:}\text{created}]\)
- Modification date/timestamp. \([\text{ing:}\text{lastModified}]\)
- Service specification and version. \([\text{ing:}\text{serviceScheme}]\)
- Base URI for the service method invocations. \([\text{ing:}\text{baseURI}]\)
- Support URI for the service. \([\text{ing:}\text{supportURI}]\)
- Administrative URI for the service. \([\text{ing:}\text{adminURI}]\)

Additional global service properties MAY be defined and managed by the service.

The Ingest service can be configured with knowledge of an arbitrary number of Merritt Storage services, which manage the secure, persistent storage of digital objects [Storage]. Multiple Storage services can be defined to encapsulate different underlying storage technologies, policy regimes, or administrative scope.

Jobs submitted for (asynchronous) processing are managed in a queue.

### 3.2 Queue

The queue manages jobs submitted to the service but awaiting processing. Conceptually the queue interacts with two processes: a submitter that adds jobs to the queue; and a consumer that retrieves jobs from the queue. The queue state properties MUST minimally include:

- Actionable reference to the Ingest service state. \([\text{que:}\text{ingestState}]\)
- Consumer polling interval (in seconds). \([\text{que:}\text{pollingInterval}]\)
- Number of consumer threads. \([\text{que:}\text{numThreads}]\)
- Number of jobs currently in the queue. \([\text{que:}\text{numJobs}]\)
- Total number of jobs processed. \([\text{que:}\text{numTotalJobs}]\)
- Creation date/timestamp \([\text{que:}\text{created}]\)
- Last submission date/timestamp. \([\text{que:}\text{lastSubmission}]\)
- Last consumption date/timestamp. \([\text{que:}\text{lastConsumption}]\)
- If the consumer is in a paused state, the date/timestamp at which it was paused. \([\text{que:}\text{paused}]\)
- Consumer status: running or paused. \([\text{que:}\text{status}]\)

Additional queue state properties MAY be defined and managed by the service.

At the conclusion of ingest processing of a given job the consumer’s behavior is:

- If the polling mode is immediate, poll the queue to see if another pending job is available; if
the mode is wait, wait for the polling interval to expire before polling the queue.

- If a job is available, i.e. in pending status, at the time of polling, the job is consumed; if not, wait for the polling interval to expire before polling the queue.

Individual jobs are managed explicitly in the queue; batches are managed indirectly, with each queued job knowing the batch of which it is a member. A queue can manage an arbitrary number of jobs.

NOTE The queue entity is not strictly necessary; since there is a one-to-one relationship between the service and the queue, all of the queue properties could have been defined as ingest service properties. However, it is useful to have a queue entity so that its state, which may potentially incorporate significant numbers of job references, can be interrogated independently of the global service state, which does not then need to incorporate job references.

### 3.3 Batch

A batch encapsulates an arbitrary number of jobs that are submitted to the Ingest service in a single operation. Batches are defined for administrative convenience, since they permit a single notification to be sent regardless of the number of component jobs. In particular:

- A batch does not constitute an indivisible unit of processing; each job in a batch is processed, and succeeds or fails, independent of all other jobs and the batch itself.

- The failure of a given batch job does not cause the rollback of previously processed jobs, and does not terminate the processing of subsequent jobs.

The batch state properties MUST minimally include:

- Batch identifier, assigned to be locally-unique with the service. [bat:identifier]
- Number of jobs in the batch. [bat:numJobs]
- Number of pending jobs. [bat:numPendingJobs]
- Number of consumed jobs. [bat:numConsumedJobs]
- Number of completed jobs. [bat:numCompletedJobs]
- Actionable reference to the queue state. [bat:queueState]
- Actionable references to job states. [bat:jobStates]
  - Actionable reference to job state. [bat:jobState]
- Submission date/timestamp. [bat:submitted]
- Completion date/timestamp. [bat:completed]
- Status: pending, consumed, or completed. [bat:status]

Additional batch state properties MAY be defined and managed by the service.

A batch is in the pending state if all of its jobs are in a pending state; in the consumed state if any of its jobs are in the consumed, completed, or failed state; and in the completed state if all of its jobs are in the completed or failed state.
A batch encompasses an arbitrary number of jobs. Batches are managed by the queue indirectly, with each queued job knowing the batch of which it is a member.

3.4 Job

A job encapsulates the ingest processing of a single digital object. The successful processing of a job results in a new version of the object managed in an instance of the Merritt Storage service [Storage]. The job state properties MUST minimally include:

- Job identifier, assigned to be locally unique within the batch and queue. [job:identifier]
- Actionable reference to the queue state. [job:queueState]
- Actionable reference to the parent batch state. [job:batchState]
- Actionable reference to the profile state. [job:profileState]
- Batch identifier. [job:batch]
- Submitting user agent. [job:submitter]
- Primary identifier, if supplied as part of the submission or newly minted as part of ingest processing. [job:primaryIdentifier]
- Object creator, if supplied as part of the submission. [job:creator]
- Object title, if supplied as part of the submission. [job:title]
- Object date, if supplied as part of the submission. [job:date]
- Object local identifier(s), if supplied as part of the submission. [job:localIdentifier]
- Object expository note, if supplied as part of the submission. [job:note]
- Submission date/timestamp. [job:submitted]
- Consumption date/timestamp. [job:consumed]
- Completion date/timestamp. [job:completed]
- Optional MetaCat registration status: success or failure. [job:metacatStatus]
- Job status: pending, consumed, completed, failed, or deleted. [job:status]

The status transition diagram for jobs is:

![Status Transition Diagram](image)

Figure 1 – Job status transitions

Additional job state properties MAY be defined and managed by the service.

In addition to these state properties, a job also has a payload: the data and metadata files received from the submitting user agent that collectively define a new object version. A subset of state properties – batch identifier; job identifier; user agent; primary identifier, creator, title, date, local identifier, and note, if supplied; submission, consumption, and completion date/timestamps; and an actionable
reference to the payload – constitute a job’s *sidecar* information, which MUST be maintained throughout the job’s processing.

The specific processing performed by a job is determined by the object’s *profile*.

### 3.5 Profile

A *profile* declares the administrative categories, content type, and processing disposition associated with an object. The profile state properties MUST minimally include:

- Profile identifier, assigned to be locally unique within the Ingest service.  
  
  [pro:identifier]
- Optional profile description.  
  
  [pro:description]
- Object type: *MRT-curatorial* or *MRT-system*..  
  
  [pro:type]
- Object role: *MRT-content* or *MRT-class*..  
  
  [pro:role]
- Optional aggregate type: *MRT-collection*, *MRT-owner*, or *MRT-service-level-agreement*..  
  
  [pro:aggregate]
- Owner identifier.  
  
  [pro:owner]
- Collection identifier(s).  
  
  [pro:collection]
- Queue priority level.  
  
  [pro:priority]
- Target Storage service identifier.  
  
  [pro:storageService]
- Target storage node identifier.  
  
  [pro:storageNode]
- Target Fixity service identifier.  
  
  [pro:fixityService]
- Identifier scheme: *ARK* or *DOI*.  
  
  [pro:identifierScheme]
- Identifier namespace.  
  
  [pro:identifierNamespace]
- Identifier minter URI.  
  
  [pro:identifierMinter]
- Identifier context.  
  
  [pro:identifierContext]
- Object content model identifier.  
  
  [pro:contentModel]
- Notification email address(es).  
  
  [pro:notification]
- Notification format: *ANVL*, *CSV*, *JSON*, *RDF/Turtle*, *XHTML*, or *XML*.  
  
  [pro:notificationFormat]
- Notification callback URL (optional).  
  
  [pro:callbackURL]
- Actionable references to handler states.  
  
  [pro:handlers]
  - Actionable reference to handler state.  
    
    [pro:handler]
- Creation date/timestamp.  
  
  [pro:created]
- Modification date/timestamp.  
  
  [pro:lastModified]

Additional profile state properties MAY be defined.

The aggregate type property MUST be defined only for profiles used to submit class objects, that is, those for whom the object role is *MRT-class*.

The namespace for the ARK scheme is an ARK *shoulder* of the form: “ark:/naan/shoulder”
The namespace for the DOI scheme is a DOI prefix of the form: “doi:prefix/” [DOI]. In normal Merritt operations, the identifier scheme is “ARK”; the ARK identifier namespace is “ark:/13030/m5” for content objects and “ark:/13030/j2” for class objects (see [Object] for the distinction between content and class objects); and the identifier minter URI is “http://n2t.net/ezid”. The identifier context is a value that is passed to the minter to be associated with the newly minted identifier.

The primary identifier for a Merritt object MUST be an ARK. If a submission profile defines the identifier scheme as DOI, the minter MUST mint and return both a DOI and an ARK identifier.

A profile specifies an arbitrary sequence of handlers that will be applied sequentially against all submitted objects.

### 3.6 Handler

A handler encapsulates a specific subset of ingest processing. The handler state properties MUST minimally include:

- Handler name. [han:name]
- Handler identifier, assigned to be locally unique within the Ingest service. [han:identifier]
- Handler implementation version. [han:version]
- Handler description. [han:description]
- Creation date/timestamp. [han:created]
- Modification date/timestamp. [han:lastModified]

Additional handler state properties MAY be defined.

### 4 Service Interface

All Merritt services are defined in terms of abstract interfaces that can be implemented in various interactive modalities, including a procedural API with various language bindings, a command line API supported in various operating system command shells, and a RESTful API [Fielding].

State information about the various entities managed by the service MAY be requested in the following formats:

<table>
<thead>
<tr>
<th>Format</th>
<th>Extension</th>
<th>MIME type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANVL</td>
<td>.txt</td>
<td>text/anvl</td>
</tr>
<tr>
<td>CSV</td>
<td>.csv</td>
<td>text/csv</td>
</tr>
<tr>
<td>HTML</td>
<td>.html</td>
<td>application/xhtml+xml</td>
</tr>
<tr>
<td>JSON</td>
<td>.json</td>
<td>application/json</td>
</tr>
<tr>
<td>RDF/Turtle</td>
<td>.ttl</td>
<td>text/turtle</td>
</tr>
<tr>
<td>XML</td>
<td>.xml</td>
<td>application/xml</td>
</tr>
</tbody>
</table>
Table 1 – State formats

**NOTE** Until such time as a formal MIME types for the ANVL [ANVL] and Turtle [Turtle] formats are established at the IANA registry, the experimental MIME types “text/x-anvl” and “text/x-turtle” SHOULD be used, respectively.

<table>
<thead>
<tr>
<th>RESTful options</th>
<th>Command line options</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>D[=level]</td>
<td>--context context</td>
<td>Object local identifier context</td>
</tr>
<tr>
<td>h</td>
<td>--debug [level]</td>
<td>Debug level.</td>
</tr>
<tr>
<td>-I identifier</td>
<td>--identifier identifier</td>
<td>Object local identifier</td>
</tr>
<tr>
<td>M=mode</td>
<td>--mode mode</td>
<td>Queue consumer mode.</td>
</tr>
<tr>
<td>-o file</td>
<td>--output file</td>
<td>Output to file (rather than standard output).</td>
</tr>
<tr>
<td>S=status</td>
<td>--status status</td>
<td>Queue consumer status.</td>
</tr>
<tr>
<td>Accept: form</td>
<td>--response-form form</td>
<td>Response format.</td>
</tr>
<tr>
<td>Content-type:</td>
<td>--request-form form</td>
<td>Request format.</td>
</tr>
<tr>
<td>v</td>
<td>--verbose</td>
<td>Verbose response.</td>
</tr>
<tr>
<td>V</td>
<td>--version</td>
<td>Version information.</td>
</tr>
</tbody>
</table>

Table 2 – Common API options

5 Service Methods

The Ingest service SHOULD support the following methods. Each method is first defined abstractly and then in terms of RESTful and command shell APIs.

**NOTE** The RESTful API is defined in terms of HTTP request and response messages. The notations “UA” and “OS” are used to distinguish the User Agent request from the Origin Server response. Names in *italics* indicate arbitrary, rather than fixed values. Brackets “[“ and “]” enclose optional elements, parentheses “(“ and “)” enclose groups of elements, a vertical bar “|” separates alternatives; and an ellipsis “...” indicates the arbitrary repetition of the previous element.
## 5.1 Help

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Enum</th>
<th>Optional</th>
<th>[idempotent, safe]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Enum</td>
<td>Optional</td>
<td>Specific method about which help is requested.</td>
</tr>
<tr>
<td>ResponseForm</td>
<td>Enum</td>
<td>Optional</td>
<td>Response form. The supported forms SHOULD include ANVL (default for command line interfaces), JSON, RDF/ Turtle, XHTML (default for web interfaces), and XML.</td>
</tr>
<tr>
<td>RETURN</td>
<td>ResponseForm</td>
<td>Mandatory</td>
<td>Help information about the specific method or the service as a whole.</td>
</tr>
</tbody>
</table>

### SIDE EFFECTS

<table>
<thead>
<tr>
<th>ERRORS</th>
<th>Status Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td></td>
<td>Badly-formed request.</td>
</tr>
<tr>
<td>401</td>
<td></td>
<td>Unauthorized user agent.</td>
</tr>
<tr>
<td>415</td>
<td></td>
<td>Unsupported response form.</td>
</tr>
<tr>
<td>503</td>
<td></td>
<td>Service unavailable.</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>Service error.</td>
</tr>
</tbody>
</table>

- **RESTful API**
  
  UA: GET /help HTTP/1.x  
  UA: Host: ingest.cdlib.org  
  UA: Accept: response/form  
  UA:  
  
  OS: HTTP/1.x 200 OK  
  OS: Content-type: response/form  
  OS:  
  OS: help  

- **Command line API**

  % ingest -h [-t form] [-o file]  
  % ingest help [-t form] [-o file]
5.2 Get Service State

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Get-service-state</th>
<th>[idempotent, safe]</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>No argument.</td>
<td></td>
</tr>
<tr>
<td>ResponseForm</td>
<td>Enum</td>
<td>Optional</td>
</tr>
<tr>
<td>RETURN</td>
<td>Response form</td>
<td>Mandatory</td>
</tr>
</tbody>
</table>

SIDE EFFECTS

<table>
<thead>
<tr>
<th>ERRORS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>Badly-formed request.</td>
</tr>
<tr>
<td>401</td>
<td>Unauthorized user agent.</td>
</tr>
<tr>
<td>415</td>
<td>Unsupported response form.</td>
</tr>
<tr>
<td>503</td>
<td>Service unavailable.</td>
</tr>
<tr>
<td>500</td>
<td>Service error.</td>
</tr>
</tbody>
</table>

- **RESTful API**
  
  UA: GET /state HTTP/1.x
  UA: Host: ingest.cdlib.org
  UA: Accept: response/form
  UA:

  OS: HTTP/1.x 200 OK
  OS: Content-type: response/form
  OS:
  OS: state

- **Command line API**

  % ingest getServiceState [-t form] [-o file]
5.3 Get Queue State

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Get-queue-state</th>
<th>[idempotent, safe]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResponseForm</td>
<td>Enum</td>
<td>Optional</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Response form. The supported forms SHOULD include ANVL (default for command line interfaces), JSON, RDF/Turtle, XHTML (default for web interfaces), and XML.</td>
</tr>
<tr>
<td>RETURN</td>
<td>Response form</td>
<td>Mandatory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consumer state.</td>
</tr>
<tr>
<td>SIDE EFFECTS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ERRORS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>Badly-formed request.</td>
</tr>
<tr>
<td>401</td>
<td>Unauthorized user agent.</td>
</tr>
<tr>
<td>415</td>
<td>Unsupported response form.</td>
</tr>
<tr>
<td>503</td>
<td>Service unavailable.</td>
</tr>
<tr>
<td>500</td>
<td>Service error.</td>
</tr>
</tbody>
</table>

- **RESTful API**

  UA: GET /state/queue HTTP/1.x
  UA: Host: ingest.cdlib.org
  UA: Accept: response/form
  UA:

  OS: HTTP/1.x 200 OK
  OS: Content-type: response/form
  OS:
  OS: state

- **Command line API**

  `% ingest getQueueState [-t form] [-o file]`
## 5.4 Get-Batch State

<table>
<thead>
<tr>
<th>METHOD Get-batch-state</th>
<th>[idempotent, safe]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Batch</strong></td>
<td>Identifier</td>
</tr>
<tr>
<td><strong>ResponseForm</strong></td>
<td>Enum</td>
</tr>
<tr>
<td><strong>RETURN</strong></td>
<td>Response form</td>
</tr>
<tr>
<td><strong>SIDE EFFECTS</strong></td>
<td>—</td>
</tr>
</tbody>
</table>

### ERRORS

- 400 Badly-formed request.
- 401 Unauthorized user agent.
- 404 Consumer not found.
- 404 Batch not found.
- 415 Unsupported response form.
- 503 Service unavailable.
- 500 Service error.

### RESTful API

UA: GET /state/queue/batch HTTP/1.x
UA: Host: ingest.cdlib.org
UA: Accept: response/form
UA:

OS: HTTP/1.x 200 OK
OS: Content-type: response/form
OS:
OS: state

### Command line API

% ingest getBatchState batch [-t form] [-o file]
5.5 Get Job State

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Get-job-state</th>
<th>[idempotent, safe]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch Identifier</td>
<td>Mandatory</td>
<td>Batch identifier.</td>
</tr>
<tr>
<td>Job Identifier</td>
<td>Mandatory</td>
<td>Job identifier.</td>
</tr>
<tr>
<td>ResponseForm</td>
<td>Enum</td>
<td>Optional</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Response form. The supported forms SHOULD include ANVL (default for command line interfaces), JSON, RDF/Turtle, XHTML (default for web interfaces), and XML.</td>
</tr>
</tbody>
</table>

**RETURN**

| Response form | Mandatory | Job state. |

**SIDE EFFECTS**

- **ERRORS**

<table>
<thead>
<tr>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>Badly-formed request.</td>
</tr>
<tr>
<td>401</td>
<td>Unauthorized user agent.</td>
</tr>
<tr>
<td>404</td>
<td>Consumer not found.</td>
</tr>
<tr>
<td>404</td>
<td>Batch not found.</td>
</tr>
<tr>
<td>404</td>
<td>Job not found.</td>
</tr>
<tr>
<td>415</td>
<td>Unsupported response form.</td>
</tr>
<tr>
<td>503</td>
<td>Service unavailable.</td>
</tr>
<tr>
<td>500</td>
<td>Service error.</td>
</tr>
</tbody>
</table>

- **RESTful API**

  UA: GET /state/queue/batch/job HTTP/1.x
  UA: Host: ingest.cdlib.org
  UA: Accept: response/form
  UA:

  OS: HTTP/1.x 200 OK
  OS: Content-type: response/form
  OS:
  OS: state

- **Command line API**

  % ingest getBatchState batch job [-t form] [-o file]
### 5.6 Submit

<table>
<thead>
<tr>
<th>METHOD Submit</th>
<th>[non-idempotent, unsafe]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submitter</td>
<td>Agent</td>
</tr>
<tr>
<td>Filename</td>
<td>List of String</td>
</tr>
<tr>
<td>File</td>
<td>List of Octet-stream</td>
</tr>
<tr>
<td></td>
<td>List of Tar</td>
</tr>
<tr>
<td>Type</td>
<td>Enum</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile</td>
<td>Identifier</td>
</tr>
<tr>
<td>PrimaryIdentifier</td>
<td>Identifier</td>
</tr>
<tr>
<td>DigestType</td>
<td>Enum</td>
</tr>
<tr>
<td>DigestValue</td>
<td>String</td>
</tr>
<tr>
<td>Creator</td>
<td>String</td>
</tr>
<tr>
<td>Title</td>
<td>String</td>
</tr>
<tr>
<td>Date</td>
<td>String</td>
</tr>
<tr>
<td>LocalIdentifier</td>
<td>String</td>
</tr>
<tr>
<td>Note</td>
<td>String</td>
</tr>
<tr>
<td>ResponseForm</td>
<td>Enum</td>
</tr>
</tbody>
</table>
include ANVL (default for command line interfaces), CSV (implied for multi-job batches), XHTML (default for web interfaces), JSON, RDF/Turtle, and XML.

**RETURN**

<table>
<thead>
<tr>
<th>Response form</th>
<th>Mandatory</th>
<th>Submission notification.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RETURN</strong></td>
<td><strong>Response form</strong></td>
<td><strong>Mandatory</strong></td>
</tr>
<tr>
<td>SIDE EFFECTS</td>
<td>A single unique batch identifier is minted and associated with all jobs defined in the submission package, each with its own unique object identifier (either supplied or minted) and <em>pending</em> status. All jobs (including their payloads and sidecar information) are placed into the queue for asynchronous processing with the queue priority defined in the profile. In addition to the synchronous submission notification response (documenting all batch jobs), an equivalent asynchronous email message is sent to all email addresses specified in the profile (as described in Section § 5.6.5, below).</td>
<td></td>
</tr>
</tbody>
</table>

**ERRORS**

<table>
<thead>
<tr>
<th>Error Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>Badly-formed request.</td>
</tr>
<tr>
<td>401</td>
<td>User agent not authorized.</td>
</tr>
<tr>
<td>415</td>
<td>Unsupported package type.</td>
</tr>
<tr>
<td>413</td>
<td>Submission too large.</td>
</tr>
<tr>
<td>400</td>
<td>Empty submission.</td>
</tr>
<tr>
<td>400</td>
<td>Package digest verification failed.</td>
</tr>
<tr>
<td>404</td>
<td>Profile not found.</td>
</tr>
<tr>
<td>404</td>
<td>Minter not found</td>
</tr>
<tr>
<td>404</td>
<td>Storage service not found.</td>
</tr>
<tr>
<td>404</td>
<td>Storage node not found.</td>
</tr>
<tr>
<td>415</td>
<td>Unsupported response form.</td>
</tr>
<tr>
<td>503</td>
<td>Service unavailable.</td>
</tr>
<tr>
<td>500</td>
<td>Service error.</td>
</tr>
</tbody>
</table>

* Only meaningful for single file submission.

** Multiple local identifiers can be supplied as a semicolon-separated list. Any semicolon embedded in an identifier MUST be represented in the standard ERC escape notation “%sc”.

- **RESTful API**

  **NOTE** RESTful requests are formatted assuming an underlying HTML form [HTML] and using the “multipart-form-data” content type [Multipart].

  UA: POST /submit HTTP/1.x
  UA: Host: ingest.cdlib.org
  UA: Accept: response/form
  UA: Content-type: multipart/form-data; boundary=boundary
  UA:
  UA: --boundary
  UA: Content-disposition: form-data; name="submitter"
  UA:
  UA: user
  UA: --boundary
  UA: ( Content-disposition: form-data; name="file"; filename="filename" )
  UA: Content-type: application/octet-stream | application/x-gzip |
UA: application/x-tar | application/zip | text/checkm
UA: file }
UA: ( Content-disposition: form-data; name="file"
UA: Content-type: multipart/mixed; boundary=boundary_2
UA:
UA: -- boundary_2
UA: Content-disposition: file; filename="filename"
UA: Content-type: application/octet-stream | application/x-gzip |
UA: application/x-tar | application/zip | text/checkm
UA:
UA: file
UA: -- boundary_2
UA: ...
UA: --boundary_2 )
UA: --boundary
UA: Content-disposition: form-data; name="type"
UA:
UA: type
UA: --boundary
UA: Content-disposition: form-data; name="profile"
UA:
UA: profile
UA: --boundary
UA: [ Content-disposition: form-data; name="primary-identifier"
UA:
UA: identifier
UA: --boundary ]
UA: [ Content-disposition: form-data: name="digest-type"
UA:
UA: digest-type
UA: --boundary
UA: Content-disposition: form-data; name="digest-value"
UA:
UA: digest-value
UA: --boundary ]
UA: [ Content-disposition: form-data; name="creator"
UA:
UA: creator
UA: --boundary ]
UA: [ Content-disposition: form-data; name="title"
UA:
UA: title
UA: --boundary ]
UA: [ Content-disposition: form-data; name="date"
UA:
UA: date
UA: --boundary ]
UA: [ Content-disposition: form-data: name="local-identifier"
UA:
UA: identifier[; identifier[; ...]]
UA: --boundary ]
UA: [ Content-disposition: form-data: name="note"
UA:
UA: note
UA: --boundary ]
In general, the object primary identifier is left unspecified to indicate that a new object is being submitted (a unique primary identifier will be minted automatically during Ingest processing); the primary identifier is specified to indicate that a new version of a pre-existing object is being submitted. An arbitrary number of additional local identifiers, meaningful in some external curatorial context, can be specified in either case.

Note, however, that a variant submission workflow is also possible. If an object is associated with a local identifier, a new version can be submitted with only a local identifier specified; the Storage service maintains a mapping between object primary and local identifiers that is used by the Ingest service to retrieve the object primary identifier while processing the submission.

### 5.6.1 Object manifest

An object manifest is a Checkm manifest [Checkm] listing the location of all of the individual file components that collectively compose the object. The Checkm profile for object manifests (http://uc3.cdlib.org/registry/ingest/manifest/mrt-ingest-manifest) is defined as follows:

- The URL and target filename fields MUST be specified.
- The digest algorithm, digest value, and file size fields SHOULD be specified.
- The modification time field SHOULD NOT be specified, and will be ignored if provided.

  **NOTE** Unspecified Checkm fields MUST be left empty.

- The extension MIME type field SHOULD be specified, if known.
- The first two entries in the manifest MUST be structured comments specifying the Checkm conformance level and profile identifier.
- The next three entries SHOULD be structured comments specifying namespace prefixes and field definitions.
- The last entry in the manifest SHOULD be a structured comment explicitly representing the end of the file.
5.6.2 Batch manifest

A batch manifest is an extended Checkm manifest listing the locations of object manifests for all the objects that collectively compose the batch. The Checkm profile for batch manifests (http://uc3.cdlib.org/registry/ingest/manifest/mrt-batch-manifest) is defined as follows:

- The URL and target filename fields MUST be specified.
- The digest algorithm, digest value, and file size fields SHOULD be specified.
- The modification time field SHOULD NOT be specified, and will be ignored if provided.
- Two Checkm extension fields are defined for the object’s primary and local identifiers. If an object is being initially established by an ingest request, the primary identifier MUST NOT be specified; if a subsequent object version is being created, the primary identifier MUST be specified. In either case, at least one local identifier, meaningful in the object’s curatorial context, SHOULD be specified. An arbitrary number of local identifiers MAY be specified using a semicolon-separated list. Any semicolon embedded in a local identifier MUST be represented by the standard ERC escape notation “%sc”.
- Four Checkm extension fields are defined for the Dublin Kernel “who”, “what”, “when”, and “why” elements, corresponding to the object’s creator, title, significant date, and expository note. These fields SHOULD be specified only if they are known.
- The first two entries in the manifest MUST be structured comments specifying the Checkm conformance level and profile identifier.
- The next three entries SHOULD be structured comments specifying namespace prefixes and field definitions.
- The last entry in the manifest SHOULD be a structured comment explicitly representing the end of the file.
Note the use of the leading “@” to indicate the URLs of the individual object manifests, which are defined external to the batch manifest.

**NOTE** This interpretation of the Checkm @incl...
5.6.4 Single file batch manifest

A single file batch manifest is an extended Checkm manifest listing the locations of individual files that define all the objects that collectively compose the batch. The Checkm profile for batch manifests (http://uc3.cdlib.org/registry/ingest/manifest/mrt-single-file-manifest) is defined as follows:

- The URL and target filename fields MUST be specified.
- The digest algorithm, digest value, and file size fields SHOULD be specified.
- The modification time field SHOULD NOT be specified, and will be ignored if provided.
- Two Checkm extension fields are defined for the object’s primary and local identifiers. If an object is being initially established by an ingest request, the primary identifier MUST NOT be specified; if a subsequent object version is being created, the primary identifier MUST be specified. In either case, at least local identifier, meaningful in the object’s curatorial context, SHOULD be specified. An arbitrary number of local identifiers MAY be specified using a semicolon-separated list. Any semicolon embedded in a local identifier MUST be represented by the standard ERC escape notation “%sc”.
- Four Checkm extension fields are defined for the Dublin Kernel “who”, “what”, “when”, and “why” elements, corresponding to the object’s creator, title, significant date, and expository note. These fields SHOULD be specified only if they are known.
- The first two entries in the manifest MUST be structured comments specifying the Checkm conformance level and profile identifier.
- The next three entries SHOULD be structured comments specifying namespace prefixes and field definitions.
- The last entry in the manifest SHOULD be a structured comment explicitly representing the end of the file.

```
#%checkm_0.7
#%profile | http://uc3.cdlib.org/registry/ingest/manifest/mrt-
single-file-batch-manifest
[ #%prefix | mrt: | http://uc3.cdlib.org/ontology/mom# ]
[ #%prefix | nfo: | http://www.semanticdesktop.org/ontologies/2007/03/
  22/nfo# ]
[ #%fields | nfo:fileUrl | nfo:hashAlgorithm | nfo:hashValue |
  nfo:fileSize | nfo:fileLastModified | nfo:fileName |
  mrt:primaryIdentifier | mrt:localIdentifier | mrt:creator |
  mrt:title | mrt:date ]
url | [algorithm] | [value] | [size] | | filename | [primary]
[ | local[; local[; ...]]] [ | creator] [ | title]
[ | date] [ | note]
```

5.6.5 Notification

The submission notification MUST minimally include the following properties, which can be
represented in ANVL, CSV, RDF/Turtle, XHTML, or XML format, as indicated by the ResponseForm parameter. Although these properties are defined on a per-object basis, only a single submission notification is sent per batch, regardless of the number of jobs in that batch.

- Batch identifier.
- Submitting user agent.
- Filename.
- Submission type: file, container, or object-manifest.
- Profile identifier.
- Primary identifier, or “(unas)” if not supplied as part of the submission.
- Object creator, or “(unas)” if not supplied as part of the submission.
- Object title, or “(unas)” if not supplied as part of the submission.
- Object date, or “(unas)” if not supplied as part of the submission.
- Object local identifier, or “(unas)” if not supplied as part of the submission.
- Object expository note, if supplied as part of the submission.
- File message digest (optional).
  - Digest type: Adler-32, CRC-32, MD2, MD5, SHA-1, SHA-384, or SHA-512.
  - Digest hexadecimal value.
- Submission date/timestamp.
- Status: pending.
- Actionable reference to the batch state.
- Support URI for the service.

To aid in the automated processing of mailed notifications, the subject line of the asynchronous email notification message MUST confirm to the Merritt template:

Subject: service [instance]: status -- message: extra; ...

where service is “Ingest”, instance is the service instance, “[dev]” or “[stg]” (or not provided, if production), status is “OK” or “Fail”; message is “Submission queued”; and extra is the batch identifier. For example:

Subject: Ingest: OK -- Submission queued: bid-batch-uuid

The body of the email message MUST be in ANVL format:

Submission ID: bid-batch-uuid
Job(s):
:Number of pending job(s): k
:Number of completed job(s): m
:Number of failed job(s): n

User agent: identity/name
Submission date: yyyy-mm-ddThh:mm:ss-hh:mm
Status: QUEUED

More detailed notification information is provided in an attached file using the notification format specified in the submission profile, which can be ANVL, CSV, JSON, RDF/Turtle, XHTML, or XML.

All submission notification fields MUST be specified in the CSV notification. Fields for which no values were supplied by the submitting user agent or specified by the Ingest service MUST be assigned the ANVL coded value “(:unas)” (unassigned). The first line of a CSV notification MUST define the field labels.

    Batch,Submitter,Filename,Type,Profile,PrimaryIdentifier,Creator,...
    batch,submitter,filename,type,profile,primaryidentifier,creator,...
    ...

5.6.6 HTML interface

The client-side HTML interface for the RESTful API COULD look something like:

![Client HTML interface](image)

**Figure 2** – Client HTML interface.
### 5.7 Submit Object

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Submit-object</th>
<th>[non-idempotent, unsafe]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Submitter</strong></td>
<td>Agent</td>
<td>Mandatory</td>
</tr>
<tr>
<td><strong>Filename</strong></td>
<td>String</td>
<td>Mandatory</td>
</tr>
<tr>
<td><strong>File</strong></td>
<td>Octet-stream</td>
<td>Mandatory</td>
</tr>
<tr>
<td></td>
<td>Tar</td>
<td>Tgz</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>Enum</td>
<td>Mandatory</td>
</tr>
<tr>
<td><strong>Profile</strong></td>
<td>Identifier</td>
<td>Mandatory</td>
</tr>
<tr>
<td><strong>Batch</strong></td>
<td>Identifier</td>
<td>Optional</td>
</tr>
<tr>
<td><strong>PrimaryIdentifier</strong></td>
<td>Identifier</td>
<td>Optional</td>
</tr>
<tr>
<td><strong>DigestType</strong></td>
<td>Enum</td>
<td>Optional</td>
</tr>
<tr>
<td><strong>DigestValue</strong></td>
<td>String</td>
<td></td>
</tr>
<tr>
<td><strong>Creator</strong></td>
<td>String</td>
<td>Optional</td>
</tr>
<tr>
<td><strong>Title</strong></td>
<td>String</td>
<td>Optional</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>String</td>
<td>Optional</td>
</tr>
</tbody>
</table>
| **LocalIdentifier** | String      | Optional                 | List of locally-meaningful alternative object identifiers. **
| **Note**     | String        | Optional                 | Expository note regarding the object’s creation. |
| **ResponseForm** | Enum      | Optional                 | Response form. The supported forms SHOULD include ANVL (default for command line interfaces), CSV, XHTML (default for web interfaces), JSON, RDF/Turtle, and XML. |

**RETURN**

- **Response form**
  - Mandatory
  - Job notification (documenting this single job).

**SIDE EFFECTS**

If no batch identifier is specified, a unique identifier is minted and associated with the job. The job object is submitted to the Storage service and node specified by the profile via the Storage service’s Add-version method for synchronous processing. If this is the last job of a batch to complete its processing, an ingest notification (documenting all batch jobs) is sent to all email addresses defined in the profile. All of these activities are performed by Ingester handlers, described in Section § 6.4.

**ERRORS**

- **400** Badly-formed request.
- **401** User agent not authorized.
- **415** Unsupported package type.
- **413** Submission too large.
- **400** Empty submission.
- **400** Package digest verification failed.
- **404** Batch not found.
** Multiple local identifiers can be supplied as a semicolon-separated list. Any semicolon embedded in an identifier MUST be represented in the standard ERC escape notation “%sc”.

- **RESTful API**

  **NOTE**  RESTful requests are formatted assuming an underlying HTML form [HTML] and using the “multipart-form-data” content type [Multipart].

  UA: POST /submit-object HTTP/1.x
  UA: Host: ingest.cdlib.org
  UA: Accept: response/form
  UA: Content-type: multipart/form-data; boundary=boundary
  UA:
  UA: --boundary
  UA: Content-disposition: form-data; name="submitter"
  UA:
  UA: user
  UA: --boundary
  UA: Content-disposition: form-data; name="file"; filename="filename"
  UA: Content-type: application/octet-stream | application/x-gzip | application/x-tar | application/zip | text/checkm
  UA:
  UA: --boundary
  UA: Content-disposition: form-data; name="type"
  UA:
  UA: type
  UA: --boundary
  UA: Content-disposition: form-data; name="profile"
  UA:
  UA: profile
  UA: --boundary
  UA: [ Content-disposition: form-data; name="batch"
  UA:
  UA: Batch
  UA: --boundary ]
  UA: [ Content-disposition: form-data; name="identifier"
  UA:
  UA: identifier
  UA: --boundary ]
  UA: [ Content-disposition: form-data; name="digest-type"
  UA:
  UA: digest-type
  UA: --boundary
  UA: Content-disposition: form-data; name="digest-value"
UA: digest-value
UA: --boundary ]
UA: [ Content-disposition: form-data; name="creator"
UA:
UA: creator
UA: --boundary ]
UA: [ Content-disposition: form-data; name="title"
UA:
UA: title
UA: --boundary ]
UA: [ Content-disposition: form-data; name="date"
UA:
UA: date
UA: --boundary ]
UA: [ Content-disposition: form-data; name="local-identifier"
UA:
UA: identifier[; identifier[; ...]]
UA: --boundary ]
UA: [ Content-disposition: form-data; name="note"
UA:
UA: note
UA: --boundary ]

OS: HTTP/1.x 201 Created
OS: Content-type: response/form
OS: Location: http://cutie.cdlib.org/state/queue/batch/job
OS:
OS: notification

- Command line API

  % ingest submitObject filename -T type [-t form] [-o file]

5.7.1 Notification

A job notification MUST minimally include the following properties, which can be represented in ANVL, CSV, RDF/Turtle, XHTML, or XML format, as indicated by the ReponseForm parameter.

- Batch identifier.  [not:batch]
- Submitting user agent.  [not:submitter]
- Filename.  [not:filename]
- Submission type: file, container, or object-manifest.  [not:type]
- Profile identifier.  [not:profile]
- Primary identifier, or “(:unas)” if not supplied as part of the submission.  [not:suppliedIdentifier]
- Primary identifier, or “(:unas)” if not retrieved during ingest processing.  [not:retrievedIdentifier]
- Primary identifier, or “(:unas)” if not assigned during ingest processing.  [not:assignedIdentifier]
- Object creator, or “(:unas)” if not supplied as part of the [not:creator]
submission.

- **Object title**, or “(:unas)” if not supplied as part of the submission. 

- **Object date**, or “(:unas)” if not supplied as part of the submission.

- **Object local identifier(s)**, or “(:unas)” if not supplied as part of the submission.

- **Object expository note**, if supplied as part of the submission.

- **File message digest (optional).**
  - Digest type: Adler-32, CRC-32, MD2, MD5, SHA-1, SHA-384, or SHA-512
  - Digest hexadecimal value.

- **Submission date/timestamp.**

- **Consumption date/timestamp.**

- **Completion date/timestamp.**

- **Actionable reference to Access service object state**, if successful; otherwise “(:null)”. 

- **Status**: pending, completed, or failed. 

- **Error message**, if “Status” = failed. 

- **Support URI for the service.**

...
5.7.2 HTML interface

The client-side interface for the RESTful API COULD look something like:

![Client HTML interface (single object)](image)

**Figure 3 – Client HTML interface (single object).**
5.8 Set Queue Behavior

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Set-queue-behavior</th>
<th>[idempotent, unsafe]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>Enum</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Mode</td>
<td>Enum</td>
<td></td>
</tr>
<tr>
<td>ResponseForm</td>
<td>Enum</td>
<td>Optional</td>
</tr>
</tbody>
</table>

### Status
Queue consumer status: restart or pause.

### Mode
Queue consumer mode: immediate or wait.

### ResponseForm
Response form. The supported forms SHOULD include ANVL (default for command line interfaces), JSON, RDF/Turtle, XHTML (default for web interfaces), and XML.

### RETURN
Response form Mandatory Queue state.

### SIDE EFFECTS
Queue consumer status is set to the specified value.

### ERRORS
- 400 Badly-formed request.
- 401 Unauthorized user agent.
- 404 Queue not found.
- 415 Unsupported response form.
- 503 Service unavailable.
- 500 Service error.

#### RESTful API

UA: PUT /state/queue?[M=immediate|wait][&][S=restart|pause] HTTP/1.1
UA: Host: ingest.cdlib.org
UA: Accept: response/form
UA:

OS: HTTP/1.1 200 OK
OS: Content-type: response/form
OS:
OS: state

#### Command line API

% ingest setQueueStatus [-M immediate|wait] [-S restart|pause] [-t form] [-o file]

### Implementation

The Ingest service is composed of five independent system components:

- **Submitter.** The Submitter accepts request for asynchronous batch or single object processing and creates a job for each object.
- **Queue.** The Queue manages job sidecar information for jobs awaiting processing.
- **Consumer.** The Consumer polls for the availability of a job at the head of the Queue.
- **Ingester.** The Ingester processes a single job by invoking the Storage service Add-version method.
File system. The mountable file system manages job payloads.

The Submitter and Ingester are implemented as servlets; the Consumer is a multi-threaded daemon, with a thread assigned to each consumed job. The Submitter handles asynchronous Ingest requests for batches and single objects from Ingest users; the Ingester handles synchronous Ingest requests for single objects, either directly from administratively-privileged users or from the Consumer.

6.1 Submitter

The processing of a batch is performed as a configurable sequence of handlers, whose responsibilities are to push job sidecar metadata onto the queue for each object specified in the batch, write job payload data into the shared file system, and send the submission notification.

6.2 Queue

The priority queue is implemented using ZooKeeper [ZooKeeper].

6.3 Consumer

At the conclusion of ingest processing of a given job the Consumer’s behavior is:

- If the polling mode is immediate, poll the queue to see if another pending job is available; if the mode is wait, wait for the polling interval to expire before polling the queue.
- If polling indicates that a job is available, consume it; otherwise, wait for the polling interval to expire before polling the queue.

6.4 Ingester

The processing of a job is performed as a configurable sequence of handlers. The canonical sequence of handler steps is:

NOTE All pathnames used below are defined relative to a staging area specific to the submission batch and
job, "ingest_home/queue/batch/job". An arbitrary submission package is represented by "package"; an arbitrary package file or component is represented by "component".

1. Initialize. [Mandatory]

(1) Create a Merritt ingest metadata file, "mrt-ingest.txt", in ANVL syntax, documenting the parameters of the Submit-object method invocation and relevant properties of the controlling submission profile in the staging area "system" directory.

(2) Create a Merritt membership file, "mrt-membership.txt", containing the collection identifiers retrieved from the submission profile, in the staging area "system" directory.

(3) Create a Merritt object model (MOM) metadata file [MOM], "mrt-mom.txt", in ANVL syntax, specifying the object’s primary identifier, if known, type, role, and optionally, aggregate and local identifiers, if known, in the starting area "system" directory.

(4) Create a stub Merritt object resource map, "mrt-object-map.ttl", in Turtle syntax, in the staging area "system" directory. The resource map defines the object initially as an aggregation ("ore:aggregates") of the ingest, membership, MOM, and owner metadata files as well as the resource map itself. Each metadata file is associated to the object with an "mrt:hasMetadata" relationship, and is the subject of "mrt:metadataSchema" and "mrt:mimeType" relationships.

(5) Create a Merritt owner file, "mrt-owner.txt", containing the owner identifier retrieved from the submission profile, in the staging area "system" directory.

```
./system/mrt-ingest.txt

ingest: name
submissionDate: yyyy-mm-ddThh:mm:ss±zz:zz
batch: batch
job: job
userAgent: user
file: filename
type: type
profile: profile-identifier
queuePriority: priority
storageService: service
storageNode: node
notification: email[; ...]
suppliedIdentifier: primary | (:unas)
digestType: type | (:unas)
digestValue: value | (:unas)
creator: creator | (:unas)
title: title | (:unas)
date: date | (:unas)
localIdentifier: local[: local[: ...]] | (:unas)
note: note | (:unas)
```

NOTE Multiple notification email addresses are separated by semicolons (";") to conform to ANVL syntax.

```
./system/mrt-membership.txt
```
NOTE The primary and local identifiers are the two most significant properties of an ingested object. The primary identifier is the basis for all manipulation and reporting for the object. The local identifier is the mechanism for associating an object in a Merritt curation environment with external discovery and administrative metadata. The primary and local identifiers MUST be documented in the MOM file at this stage only if they are known. If the identifiers were not specified as Submit or Submit-object request parameters, the “primaryIdentifier” and/or “localIdentifier” elements MUST not be written. The primary identifier will always be available by the end of processing of Step 8 [Mint], below.

NOTE Multiple local identifiers MUST be expressed as a semicolon-separated list. Any semicolon embedded in an identifier MUST be represented in the standard ERC escape notation “%sc”.

```turtle
@prefix mrt: <http://uc3.cdlib.org/ontology/mom#>. @prefix ore: <http://www.openarchives.org/ore/terms/>.
<object-url>
  ore:aggregates
  <ingest-url>,
  <membership-url>,
  <mom-url>,
  <owner-url>,
  <resource-map-url> ;
  mrt:hasMetadata
  <ingest-url>,
  <membership-url>,
  <mom-url>,
  <owner-url>,
  <resource-map-url> .
<ingest-metadata-url>
  mrt:metadataSchema
  "MRT-ingest";
  mrt:mimeType
  "text/anvl".
  <membership-url>
```
2. **Accept.** [Mandatory]

(1) Copy the submission package to the staging area "producer" directory.

**NOTE** Merritt submission packages *always* take the form of a single file, which may be a single component object, a container file encapsulating an arbitrary number of object components, or a manifest file referencing an arbitrary number of object components.

(2) If "Type" is *file* (as opposed to *container* or *manifest*), define an "ore:aggregates" relationship between the object and the single file component, i.e., the submission package file, in the system object resource map "mrt-object-map.ttl"

```
./producer/package # Submission package
./system/mrt-ingest.txt # Ingest metadata
./system/mrt-membership.txt # Membership metadata
./system/mrt-mom.txt # MOM metadata
./system/mrt-object-map.ttl # Object resource map

...<object-url>
  ore:aggregates
  <file-uri> .
```

```
./system/mrt-owner.txt # Owner metadata
```
3. **Verify.** [Optional, if “DigestType” and “DigestValue” are available]

   (1) Compute and compare the message digest for the submission package, adding the results (“verified” or “failed”) to the ingest metadata component.

   ```
   ./producer/package
   ./system/mrt-ingest.txt
   ...
   packageIntegrity: verified | failed
   ./system/mrt-membership.txt # Membership metadata
   ./system/mrt-mom.txt # MOM metadata
   ./system/mrt-object-map.ttl # Object resource map
   ./system/mrt-owner.txt # Owner metadata
   ```

4. **Disaggregate.** [Conditional, if “Type” = `container`]

   (1) Validate the container according to its format, adding the result (“valid” or “invalid”) to the ingest metadata component.

   (2) Disaggregate the contents of the container in the staging area “producer” directory

   (3) Define an “ore:aggregates” relationship between the object and each disaggregated file component in the system object resource map “mrt-object-map.ttl”.

   (4) Delete the container file.

   ```
   [ ./producer/container ] # Deleted after processing
   [ ./producer/mrt-erc.txt ] # Optional ERC metadata
   [ ./producer/mrt-manifest.txt ] # Optional Checkm manifest
   ./producer/component # Disaggregated file components
   ...
   ./system/mrt-ingest.txt # Ingest metadata
   ...
   containerValidity: valid | invalid
   ./system/mrt-membership.txt # Membership metadata
   ./system/mrt-mom.txt # MOM metadata
   ./system/mrt-object-map.ttl # Object resource map
   ...
   <object-url>
   ore:aggregates
   <file-uri>,
   ...
   <file-uri>.
   ./system/mrt-owner.txt # Owner metadata
   ```

**NOTE** Container-based submission packages MAY contain a Checkm manifest component with the
reserved filename “mrt-manifest.txt” and/or an ERC component with the reserved filename “mrt-erc.txt”.

5. **Retrieve.** [Conditional, if “Type” = object-manifest]

   (1) Validate the manifest according to its format, adding the result (“valid” or “invalid”) to the ingest metadata component.

   (2) Read the manifest and perform a (possibly parallel) retrieval of all referenced file components, placing them in the staging area “producer” directory.

   (3) Define an “ore:aggregates” relationship between the object and each referenced file component in the system object resource map “mrt-object-map.ttl”.

   ```
   ./producer/manifest # Object manifest
   [ ./producer/mrt-erc.txt ] # Optional ERC metadata
   ./producer/component # Referenced file components
   ...
   ./system/mrt-ingest.txt # Ingest metadata
   ...
   manifestValidity: valid | invalid
   ...
   ./system/mrt-membership.txt # Membership metadata
   ./system/mrt-mom.txt # MOM metadata
   ./system/mrt-object-map.ttl # Object resource map
   ...
   <object-url>
   ore:aggregates
   <file-uri>,
   ...
   <file-uri>.
   ...
   ./system/mrt-owner.txt # Owner metadata
   ``

   **NOTE** Manifest-based submission packages MAY contain an ERC component with the reserved filename “mrt-erc.txt”.

6. **Corroborate.** [Conditional, if “Type” = container and a manifest component (mrt-manifest.txt) is available]

   (1) Validate the manifest according to its format, adding the results (“valid” or “invalid”) to the ingest metadata component.

   (2) Read the manifest and compute and compare file size and digest values, adding the results (“verified” or “failed”) to the ingest metadata component. All files in the manifest MUST be found in the staging area; all files in the staging area, except the manifest itself, MUST be referenced in the manifest.
[ ./producer/manifest ] # Submission manifest
[ ./producer/mrt-erc.txt ] # Optional ERC metadata
[ ./producer/mrt-manifest.txt ] # Optional container manifest
./producer/component # File components
...
./system/mrt-ingest.txt # Ingest metadata

manifestFile: manifest
manifestValidity: valid | invalid
manifestIntegrity: verified | failed

./system/mrt-membership.txt # Membership metadata
./system/mrt-mom.txt # MOM metadata
./system/mrt-object-map.ttl # Object resource map
./system/mrt-owner.txt # Owner metadata

NOTE For manifest-based submission packages a manifest will always be available as the package file; for container-based packages a Checkm manifest, “mrt-manifest.txt”, may be available.

7. Characterize. [Optional]

(1) Invoke JHOVE2 against each component file in the staging area “producer” directory, copying the results “mrt-jhove2.xml” to the staging area “system” directory.

(2) Define an “mrt:hasMetadata” relationship between the component file and its JHOVE2 output, an “mrt:metadataSchema” relationship between the JHOVE2 output and “JHOVE2”, and a “mrt:mimeType” relationship between the JHOVE2 output and “text/xml” in the system object resource map “mrt-object-map.ttl”.

[ ./producer/manifest ] # Submission manifest
[ ./producer/mrt-erc.txt ] # Optional ERC metadata
[ ./producer/mrt-manifest.txt ] # Optional container manifest
./producer/component # File components
...
./system/mrt-ingest.txt # Ingest metadata
./system/mrt-jhove2.xml # JHOVE2 metadata
./system/mrt-membership.txt # Membership metadata
./system/mrt-mom.txt # MOM metadata
./system/mrt-object-map.ttl # Object resource map

...<component-url>
mrt:hasMetadata
  <jhove2-url> .
</jhove2-url>
</component-url>
mrt:metadataSchema
  ”JHOVE2” ;
mrt:mimeType
  “text/xml” ;
8. **Mint.** [Mandatory]

   (1) Define the Dublin Kernel “who”, “what”, “when”, “where” elements by concatenating the values from the optional ERC metadata component “mrt-erc.txt” in the staging area “producer” directory or the batch manifest, and the optional Creator, Title, Date, and LocalIdentifier values submitted as part of the Submit or Submit-object request.

   **NOTE** The Kernel “who”, “what”, “when”, and “where” elements MUST be defined; if no value is available, the string “(:unas)” MUST be used. The “why” element SHALL be defined only if a value was explicitly supplied in the ERC component, the batch manifest, or as a Submit or Submit-object request parameter. Multiple values for a single element MUST be concatenated using a semicolon (“;”) as the separator. Any semicolon embedded in a value MUST be represented in the standard ERC escape notation “%sc”.

   **NOTE** An object primary identifier **cannot** be specified in the producer directory ERC component. Any “where” element found in the ERC component MUST be assumed to define a local identifier list.

   (2) If the primary identifier is null and the local identifier list is not null:

   i. Retrieve the storage node from the submission profile.

   ii. Iteratively retrieve the object primary identifier associated with the profile and each local identifier in the Storage service, if defined.

   ```
   UA: GET /primary/node/profile/identifier HTTP/1.x
   UA: Host: store.cdlib.org
   UA: 
   OS: HTTP/1.x 200 OK
   OS: Content-type: text/plain
   OS: identifier
   ```

   iii. This will result in either: (a) no primary identifier associated with any local identifier; (b) a single primary identifier associated with all or some local identifiers; or (c) multiple primary identifiers associated with various local identifiers. This last condition (c) MUST be considered an error and SHALL terminate further ingest processing.

   (3) If the primary identifier is null and is **not** retrievable via a local identifier as described in Step (2):

   i. Retrieve the identifier scheme, namespace, minter, and context from the submission profile.

   ii. Invoke the minter, passing the namespace and context to mint a new object primary identifier in the identifier scheme and namespace and associated with the context, adding the newly minted primary identifier as the value of the “assignedIdentifier” element in the ingest metadata component “mrt-ingest.txt” and as the value of the “primaryIdentifier” element in the MOM metadata component “mrt-mom.txt”, both in the staging area “system” directory.

   (a) If the ARK scheme is specified in the profile, a single ARK identifier is returned.
(b) If the DOI scheme is specified, both a DOI and an ARK identifier are returned. The DOI MUST be subsequently used only as a local identifier (in addition to any other local identifiers introduced through the Kernel “where” element); the object primary identifier MUST be the ARK.

UA: PUT /ezid/shoulerd/ark%3a%2f13030%2fm5 HTTP/1.x
UA: Host: n2t.net
UA: Content-type: text/plain
UA:
UA: _group: context

OS: HTTP/1.x 201 Created
OS: Content-type: text/plain
OS:
OS: success: ark

(4) If the primary identifier is not null or is retrievable via a local identifier as described in Step (2):

i. Invoke EZID, passing the object’s URL in the Merritt UI and the Dublin Kernel “who”, “what”, “when”, and optionally, “why” elements to update the metadata associated with the primary identifier. If the primary identifier was retrieved (and not originally supplied as a Submit or Submit-object request parameter), add it as the value of the “retrievedIdentifier” element in the ingest metadata component “mrt-ingest.txt” and as the value of the “primaryIdentifier” element in the MOM metadata component “mrt-mom.txt”, both in the staging area “system” directory.

UA: POST /ezid/id/primary-identifier HTTP/1.x
UA: Host: minter
UA: Content-type: text/plain
UA:
UA: _target: http://merritt.cdlib.org/?object=primary-identifier
UA: _erc: who: who%0awhat: title%0awhen: when%0a where where[%0a
why: why%0a ]

OS: HTTP/1.x 200 OK
OS: Content-type: text/plain
OS:
OS: success: primary-identifier

[ ./producer/manifest ] # Submission manifest
[ ./producer/mrt-erc.txt ] # Optional ERC metadata
[ ./producer/mrt-manifest.txt ] # Optional container manifest
./producer/component # File components
9. **Describe.** [Mandatory]

   (1) Copy the Dublin Kernel elements Creator, Title, Date, LocalIdentifier, PrimaryIdentifier, whether supplied, retrieved, or newly minted in Step 8(2), and optionally, Note, to an ERC metadata component “mrt-erc.txt”, in ANVL syntax, in the staging area “system” directory. Elements with no submitter-assigned value MUST be specified with the Dublin Kernel “(:unas)” code.

   (2) Define an “ore:aggregates” relationship between the object and the ERC metadata file, “mrt-erc.txt” in the “system” directly; an “mrt:hasMetadata” relationship between the object and the ERC metadata, an “mrt:metadataSchema” relationship between the ERC metadata and “ERC”, and an “mrt:mimeType” relationship between the ERC metadata and “text/anvl” in the system object resource map “mrt-object-map.ttl”.

   (3) If an ERC metadata component “mrt-erc.txt” exists in the “producer” directory, define an “mrt:hasMetadata” relationship between the object and the producer ERC metadata, an “mrt:metadataSchema” relationship between the ERC metadata and “ERC”, and an “mrt:mimeType” relationship between the ERC metadata and “text/anvl” in the system object resource map “mrt-object-map.ttl”.

   **NOTE** The primary identifier is written to the MOM metadata component only if it is not supplied as a Submit or Submit-object request parameter and is minted in Step 8(3)ii or retrieved using the localIdentifier in Step 8(4)i. If the primary identifier is supplied as a request parameter it will already have been written to the MOM component in Step 1(3). The local identifier is always written to the MOM component in Step 1(3).
[ ./system/mrt-erc.txt ]  # System ERC metadata

```
erc:
who: creator | (:unas)
what: title | (:unas)
when: date | (:unas)
where: primary
where: local[; local[; ...]] | (:unas)
[ why: exposition ]
```

./system/mrt-ingest.txt  # Ingest metadata
./system/mrt-jhove2.xml  # JHOVE2 metadata
./system/mrt-membership.txt  # Membership metadata
./system/mrt-mom.txt  # MOM metadata
./system/mrt-object-map.ttl  # Object resource map

```
<object-url>
  ore:aggregates
  <system-erc-metadata-url> ;
  mrt:hasMetadata
  <system-erc-metadata-url> .
  <system-erc-metadata-url>
  mrt:metadataSchema
  "ERC" ;
  mrt:mimeType
  "text/anvl" .
  [ <object-url>
    mrt:hasMetadata
    <producer-erc-metadata-url> .
    <producer-erc-metadata-url>
    mrt:metadataSchema
    "ERC" ;
    mrt:mimeType
    "text/anvl" . ]
```

./system/mrt-owner.txt  # Owner metadata

**NOTE**  It is possible for a single submission to define ERC metadata both within the package itself (i.e. "./producer/mrt-erc.txt") and by method arguments (i.e., Creator, Title, Date, LocalIdentifier, Note, used to define "/system/mrt-ingest.txt").

10. Member node.  [Optional, if interoperation with DataONE [DataONE] is desired]

   (1) If the Merritt DateONE manifest file “mrt-dataone-manifest.txt” does exist in the staging area “producer” directory:

   i.  Define a “mrt-hasMetadata” relationship between the object and the producer DataONE manifest file, a “mrt:metadataSchema” relationship between the producer DataONE manifest file and “DataONE-manifest”, and a “mrt:mimeType” relationship between the producer DataONE manifest file and
“text/anvl”, in the system object resource map, “mrt-object-map.ttl”.

The DataONE manifest file conforms to the general format:

```
#%dataonem_0.1
#%profile | http://uc3.cdlib.org/registry/ingest/manifest/
            mrt-dataone-manifest
#%prefix   | dom: | http://uc3.cdlib.org/ontology/dataonem#
#%prefix   | mrt: | http://uc3.cdlib.org/ontology/mom#
#%fields   | dom:scienceMetadataFile | dom:scienceMetadataFormat | dom:scienceDataFile | mrt:mimeType
metadata-file | metadata-format | data-file | mime/type ...
```

The list of defined DataONE science metadata formats is available at <http://mule1.dataone.org/ArchitectureDocs-current/design/WhatIsData.html#metadata-types>.

(2) If the file “mrt-dataone-manifest.txt” does not exist in the staging area “producer” directory:

i. Create the file “mrt-dataone-manifest.txt” in the staging area “system” directory, conforming to the Merritt DataONE manifest format, with a separate line associating the system ERC file “mrt-erc.txt” with each file in the “producer” directory:

```
#%dataonem_0.1
#%profile | http://uc3.cdlib.org/registry/ingest/manifest/
            mrt-dataone-manifest
#%prefix   | dom: | http://uc3.cdlib.org/ontology/dataonem#
#%prefix   | mrt: | http://uc3.cdlib.org/ontology/mom#
#%fields   | dom:scienceMetadataFile | dom:scienceMetadataFormat | dom:scienceDataFile | mrt:mimeType
mrt-erc.txt | ERC | data-file | mime/type ...
```

ii. Define an “ore:aggregates” relationship between the object and the system DataONE manifest file, a “mrt:hasMetadata” relationship between the object and the system DataONE manifest file, a “mrt:metadataSchema” relationship between the system DataONE manifest file and “DataONE-manifest”, and a “mrt:mimeType” relationship between the system DataONE manifest file and “text/anvl”, in the system object resource map, “mrt-object-map.ttl”.

(3) Use the file “mrt-dataone-manifest.txt” found in the staging area “producer” directory, or the file newly created in the “system” directory in step 10(1)(i) to create a DataONE resource map [ORE] with the filename “mrt-dataone-map.xml” in the staging area “system” directory:

NOTE The abstract structure and the specific serialization format of the DataONE resource map are not yet well defined and stable, and are thus not defined here.

(4) Define an “ore:aggregates” relationship between the object and the DataONE
resource map, a “mrt:hasMetadata” relationship between the object and the DataONE resource map, a “mrt:metadataSchema” relationship between the DataONE resource map and “DataONE-map”, and a “mrt:mimeType” relationship between the DataONE resource map and “text/turtle”, in the system object resource map, “mrt-object-map.ttl”.

(5) Register the object with Metacat [Metacat], the central metadata store for DataONE member nodes.
   i. Invoke the Metacat API to register the object.
   ii. Add the registration status, “success” or “failure”, as the value of the “metacatRegistration” element of the ingest metadata component, “mrt-ingest.txt” in the staging area “system” directory.

   NOTE: The failure of Metacat registration does not imply the failure and rollback of the Merritt ingest. The success or failure of the two processes The success or failure of Merritt ingest is completely independent of the success or failure of Metacat registration.

[ ./producer/mrt-dataone-manifest.txt ] # Optional DataONE manifest
[ ./producer/manifest ] # Submission manifest
[ ./producer/mrt-erc.txt ] # Optional ERC metadata
[ ./producer/mrt-manifest.txt ] # Optional container manifest
./producer/component ...
[ ./system/mrt-dataone-manifest.txt ] # Optional DataONE manifest
./system/mrt-dataone-map.ttl # DataONE resource map
[ ./system/mrt-erc.txt ] # System ERC metadata
./system/mrt-ingest.txt # Ingest metadata

... metacatRegistration: success | failure

./system/mrt-jhove2.xml # JHOVE2 metadata
./system/mrt-membership.txt # Membership metadata
./system/mrt-mom.txt # MOM metadata
./system/mrt-object-map.ttl # Object resource map

...<object-url>
  ore: aggregates
  [ <dataone-manifest-url> , ] # System, not producer
  <dataone-resource-map-url> ;
  mrt:hasMetadata
  <dataone-manifest-url> ,
  <dataone-resource-map-url> .
  <dataone-manifest-url>
  mrt:metadataSchema
  “DataONE-manifest” ;
  mrt:mimeType
  “text/anvl” .
  <dataone-resource-map-url>
  mrt:metadataSchema

Merrit Ingest Service
NOTE The "ore:aggregates" relationship for the DataONE manifest file only applies if the manifest is created in Step 10(2)(i); if the manifest was supplied by the producer, it is already a member of the object aggregation by virtue of Steps 4(1), 5(3), or 6(3).

./system/mrt-owner.txt # Owner metadata

11. Document. [Mandatory]

(1) Document the sequence of handlers, including those in subsequent steps, in the ingest metadata component "mrt-ingest.txt" in the staging area "system" directory.

... 

Handlers: handler/version; ...

./system/mrt-jhove2.xml # JHOVE2 metadata
./system/mrt-membership.txt # Membership metadata
./system/mrt-mom.txt # MOM metadata
./system/mrt-object-map.ttl # Object resource map
./system/mrt-owner.txt # Owner metadata

12. Digest. [Mandatory]

(1) Create a Checkm manifest "mrt-manifest.txt" in the "system" directory that references the URL, SHA-256 digest type and value, file size, and target filename for all files in the "producer" and "system" directories, except the manifest itself.

(2) Compute an SHA-256 message digest for the manifest.

...
13. **Transfer.** [Mandatory]

(1) Retrieve the Storage service and storage node identifiers from the profile.

(2) Invoke the Storage server `Add-version` method passing the storage node identifier, primary identifier, manifest, manifest size, manifest SHA-256 message digest algorithm and value, and if supplied, the local identifier list and profile identifier, and add the result (success or failure) to the ingest metadata component “mrt-ingest.txt” in the staging area “system” directory.

```
UA: POST /content/node/object HTTP/1.1
UA: Host: store.cdlib.org
UA: Accept: response/form
UA: Content-type: multipart/form-data; boundary=boundary
UA:
UA: --boundary
UA: Content-disposition: form-data; name="manifest"
UA: Content-type: text/checkm
UA:
UA: manifest
UA: --boundary
UA: Content-disposition: form-data; name="size"
UA: Content-type: text/plain
UA:
UA: size
UA: --boundary
UA: Content-disposition: form-data; name="digest-type"
UA: Content-type: text/plain
UA:
UA: type
UA: --boundary
UA: Content-disposition: form-data; name="digest-value"
UA: Content-type: text/plain
```
14. **Fixity.** [Conditional, if Transfer was successful]

(1) Retrieve the Fixity service, owner, and collection identifiers from the profile.

(2) For each object component file, invoke the Fixity service *Add-Item* method, passing the component’s Storage service URL, size (in octets), SHA-256 message digest, and contexts for the object’s owner, collections, and primary identifier, and add the result (number of
successful adds, number of failed adds, URLs of the individual failed items) to the ingest 
metadata component “mrt-ingest.txt” in the staging area “system” directory.

UA: POST /add HTTP/1.x
UA: Host: fixity.cdlib.org
UA: Content-type: multipart/form-data; boundary=boundary
UA:
UA: --boundary
UA: Content-disposition: form-data; name="url"
UA:
UA: url
UA: --boundary
UA: Content-disposition: form-data; name="source"
UA:
UA: merritt
UA: --boundary
UA: Content-disposition: form-data; name="size"
UA:
UA: size
UA: --boundary
UA: Content-disposition: form-data; name="digest-type"
UA:
UA: sha-256
UA: --boundary
UA: Content-disposition: form-data; name="digest-value"
UA:
UA: value
UA: --boundary
UA: Content-disposition: form-data; name="context"
UA:
UA: owner
UA: collection
UA: ...
UA: identifier
UA: --boundary
UA: Content-disposition: form-data; name="response-form"
UA:
UA: xml
UA: --boundary

OS: HTTP/1.x 201 Created
OS: Content-type: response/form
OS: Location: http://fixity.cdlib.org/state/url
OS:
OS: state

(3) Add the counts of Add-item successes and failures to the the ingest metadata component 
“mrt-ingest.txt” in the staging area “system” directory.

```
[ ./producer/manifest ] # Submission manifest
[ ./producer/mrt-manifest.txt ] # Optional container manifest
[ ./producer/mrt-erc.txt ] # Optional ERC metadata
. ./producer/component # File components
...
[ ./system/mrt-erc.txt ] # System ERC metadata
```
15. **Notify.** [Mandatory]

   (1) Retrieve the notification email address(es) and format from the submission profile.

   (2) Email notification of Ingest results, formatted as indicated by the notification format in the submission profile. Only a single notification email is generated for each batch.

   a. The notification message MUST minimally include the following information for the entire submission (batch or single object):

      - Submitting user agent.
      - Submission package filename.
      - Submission date/timestamp.
      - Completion date/timestamp.
      - Ingest status.

   The notification message MUST minimally include the following information for each object processed in the submission:

      - Primary identifier.
      - Local identifier, if defined.
      - Version identifier.
      - Creator, title, and date, if defined (either as request parameters or in an included ERC metadata component, “mrt-erc.txt”).
      - Checksum algorithm and value.
      - Submission date/timestamp.
      - Completion date/timestamp.
      - Metacat registration status, if the Member Node handler was invoked in Step 10.
      - Merritt Ingest status.

   Additional information MAY also be included in the notification message.

   To aid in the automatic processing of the emailed notification, the subject line of the email message MUST conform to the Merritt template:

   ```
   Subject: service [instance]: status -- message: extra; ...
   ```
where service is “Ingest”; instance is “dev” or “stg” (or not provided, if production); status is “OK” or “fail”; message is “Submission processed”; and extra is the batch identifier.

Subject: Ingest: OK -- Submission processed: bid-batch-uuid

The body of the email message MUST be in ANVL format:

Submission ID: bid-batch-uuid
Job(s):

:Number of pending job(s): k
:Number of completed job(s): m
:Number of failed job(s): n

User agent: identity/name
Queue priority: nn
Submission date: yyyy-mm-ddThh:mm:ss-hh:mm
Completion date: yyyy-mm-ddThh:mm:ss-hh:mm
Status: COMPLETED

More detailed notification information is provided in an attached file using the notification format specified in the submission profile, which can be ANVL, CSV, JSON, RDF/Turtle, XHTML, or XML.

(3) If an error condition is raised, send email notification to the service administrative address.

16. Callback. [Optional, if callback URL specified in profile]

(1) Retrieve the callback URL from the submission profile.

(2) Format the notification message as specified in the profile.

(3) Send the notification to the callback URL in the body of an HTTP POST request:

    POST /callback/path HTTP/1.x
    Host: callback.host
    Content-type: application/x-www-form-url-encoded

    jobstate=notification

(4) Receive an HTTP response:

    HTTP/1.x 200 OK

17. Cleanup. [Mandatory]

(1) Delete the staging area.

6.5 File System Instantiation
The Ingest service is instantiated in a file system as:

```xml
<ingest_home>/
  0=ingest_0.21
  ingest-info.txt
  log/
  profiles.txt
  profiles/
  queue/
  stores.txt
```

Within the file system hierarchy rooted at an Ingest home directory, all file and directory names starting with “ingest”, “ing”, “merritt”, or “mrt”, on a case-insensitive basis, are reserved.

6.5.1 Namaste Tag (0=ingest_version)

The home directory MUST contain a file named “0=ingest_0.21” that is the service’s Namaste tag [Namaste]. The tag file MUST contain the Ingest service specification name and version:

```
Ingest/0.21
```

A Namaste tag fulfills the same function for a directory that a magic number does for a file.

6.5.2 Global Service Properties (ingest-info.txt)

The home directory MUST contain a file named “ingest-info.txt” that specifies the global properties of the service.

```plaintext
name: Ingest01
identifier: ingest.cdlib.org/ingest01
description: UC3 curation repository ingest service
serviceScheme: Ingest/0.21/0.5
baseURI: http://ingest.cdlib.org/
supportURI: mailto:merritt-support@cdlib.org
adminURI: mailto:merritt-admin@cdlib.org
```

Within an Ingest properties file all property names starting with “ingest”, “ing”, “merritt”, or “mrt”, on a case-insensitive basis, are reserved.

6.5.3 Profiles (profile.txt and profiles/)

The home directory MUST contain a file “profiles.txt”.

```
profiles.txt
```
This file contains a list of identifiers for all registered profiles.

```plaintext
profile
...
```

Each profile MUST be defined by file “profiles.txt” in a sub-directory named “profiles”.

```plaintext
profiles/
  profile.txt
  ...
```

Each profile is defined by the ANVL file “profile.txt”

```plaintext
identifier: identifier
description: description
type: type
role: role
[ aggregate: aggregate ]
owner: owner-identifier
collection: collection-identifier[; ...]
queuePriority: priority
storageService: service
storageNode: node
contentModel: model
identifierScheme: scheme
identifierNamespace: namespace
identifierMinter: minter
handler: handler[; ...]
...
notification: email[; ...]
notificationFormat: anvl | csv | json | rdf | xhtml | xml
[ callbackURL: url ]
```

Note: The Ingest service should determine which profiles are registered by reading the file “profiles.txt”, not by looking for files “profile.txt” in the “profiles/” sub-directory. This permits partially completed profile definitions to be present in the directory without making them actionable.

### 6.5.4 Queue (queue/)

The home directory MUST contain a sub-directory named “queue” that holds all batches and jobs in the queue.

```plaintext
queue/
  batch/
```
6.5.5 Storage Services (stores.txt)

The home directory MUST contain a file named "stores.txt" that defines the Storage and Access services known to the Ingest service in terms of their base URIs.

```
store.1: http://store1.cdlib.org/
access.1: http://access1.cdlib.org/
...
store.n: http://storen.cdlib.org/
access.n: http://accessn.cdlib.org/
```

The association between a Storage service and its Access service is made via the numeric suffix on the "store.n" and "access.n" property names.

**NOTE** There is always a one-to-one relationship between given Storage and Access service.

References


[DataONE] DataONE, “OAI-ORE.” Welcome to Data Observation Network for Earth (DataONE) <http://www.dataone.org/>.


