

Service Genre: Harvest

Intellectual Property Rights Statement: This Service Genre is a derivative work. This work was created by the Learning Systems Architecture Lab.

The Service Genre is derived from work created as part of the Evolving the JADL Integrated Prototype Architecture: Alignment with the e-Framework project within the Workforce ADL Co-Lab. The JADL IPA project was funded in part by the Joint ADL Co-Lab under contract N61339-06-C-0082. Any opinions, findings, conclusions or recommendations expressed herein are those of the author(s) and do not reflect the views of the U.S. Government, the University of Memphis or other project sponsors.

The JADL IPA Service Genre is derived from work created as part of the Federated Repositories for Education (FRED) Project within the Australian ADL Partnership Laboratory. The FRED project is sponsored by the Australian Commonwealth Department of Education, Science and Training under the Framework for Open Learning Programme.

The template structure and format of this document are based on the Service Genre Description document from the Evolving the JADL Integrated Prototype Architecture: Alignment with the e-Framework Technical Report. This work was funded in part by the Joint ADL Co-Lab under contract N61339-06-C-0082.

The template structure and format of this document are based on e-Framework documentation templates and guidelines, which are governed by the e-Framework *Intellectual Property Rights Statement* [<http://www.e-framework.org/Default.aspx?tabid=738>].

The template structure and format of this document are also based on Federated Repositories for Education (FRED) project documentation templates and guidelines, which were created as part of the FRED Project within the Australian ADL Partnership Laboratory. The FRED project is sponsored by the Australian Commonwealth Department of Education, Science and Training under the Framework for Open Learning Programme.

e-Framework work © Copyright 2007, e-Framework Partners. e-Framework work licensed under the *Creative Commons Attribution-ShareAlike 2.5 Australia License* [<http://creativecommons.org/licenses/by-sa/2.5/au/>].

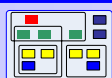
FRED project work © Copyright 2007, University of Southern Queensland and University of Memphis. FRED project work licensed under the *Creative Commons Attribution-ShareAlike 2.5 Australia License* [<http://creativecommons.org/licenses/by-sa/2.5/au/>].

JADL IPA project work © Copyright 2007, Workforce ADL Co-Lab. JADL IPA project work licensed under the *Creative Commons Attribution-ShareAlike 2.5 Australia License* [<http://creativecommons.org/licenses/by-sa/2.5/au/>].

Template structure and format © Copyright 2008, Learning Systems Architecture Lab. The template structure and format may be used under the *Creative Commons Attribution-ShareAlike 2.5 Australia License* [<http://creativecommons.org/licenses/by-sa/2.5/au/>].

Service Genre © Copyright 2008, Learning Systems Architecture Lab. The Service Genre may be used under the *Creative Commons Attribution-ShareAlike 2.5 Australia License* [<http://creativecommons.org/licenses/by-sa/2.5/au/>].

The appropriate attribution for a derivative of this work is: "This document is derived from work created by the Learning Systems Architecture Lab. © Copyright 2008, Learning Systems Architecture Lab." and should be followed by all of the attributions for this Service Genre as documented herein.



LSAL

Information Architecture & Design, Learning Technologies, Training

Introduction

The Introduction provides a brief, standalone overview of the Service Genre. It is for a non technical reader. It may duplicate other material in the Service Genre Description.

The harvest service genre defines an abstract service end point for harvesting metadata objects from the constituent repositories within the repository federation, e.g., a repository federated in the ADL-Registry (ADL-R) for the JADL IPA.

This harvest service genre is a derivative of the Federated Repositories for Education (FRED) harvest service genre. This derivative is technically equivalent to the source work; changes are editorial and in presentation only.

Harvest is the process by which an application collects or gathers (harvests) information about objects that are managed by a resource. A harvest service interface for the resource provides a mechanism for the external agent (harvester) to contact the resource (harvestee) to harvest the data. The harvested data is then used by the harvester to provide other services (e.g., to aggregate the data from multiple resources). While harvesting is generally focused on gathering metadata from the objects in a collection, any data, as defined by the service interface and resource, may be harvested.

This is a general description of a harvest service genre, independent of application end point, resource, data collected, harvesting protocol or underlying communications protocols and service models. The service genre does not include a mechanism to authorize clients. How to discover service implementations that support harvesting is out of scope.

Service Genre Description

The Service Genre Description is the complete, formal documentation of the Service Genre.

The harvest service genre defines content transfer from resource to resource as a pull mechanism, particularly in the context of repository federations. It specifies the minimum expected behaviors a harvest service is expected to provide, and the common design issues involved in implementing such a service.

Service Genre Metadata

The Service Genre Metadata contains basic labeling, classification and a version history for the Service Genre.

Name

- Service Genre Name: harvest
- LSAL ID: hdl:1870/2687306FA67643DE810A8C4789DDE0E0
- ADL Service Genre Name: harvest {collection registry}, harvest {competency resource}, harvest {CSDB}, harvest {knowledge base}, harvest {metadata registry}, harvest {repository}, harvest {repository registry}, harvest {rights license repository}, harvest {task list}, harvest {training catalog}, harvest {TSDB}, harvest {user profile repository}
- JADL IPA ID hdl:JADL-IPA-NA/E68E78264D634D3082FDBAE0B41D6924 (source for derivative)
- FRED Service Genre Name: harvest
- FRED ID hdl:FREDNA/2BF7810E0B434CE4B7B1124662A98B05 (source for derivative)

Classification

Classification Facets:

- Service Genre Status: Unapproved
- Domains: Repository
- Domain Coverage: Single
- Deployment Status: Prototype



- Deployment Scale: Isolated
- Maturity: Immature
- Composition: Individual
- Purpose: Exemplar, Application

Technical Facets:

- State Behavior: Stateful
- Transactional Behavior: Non Transactional
- Batch Behavior: Individual
- Time Constraint Behavior: None
- Service End Point: Provider
- Authentication / Authorization: Not Auth'ed
- Exposure: Public

Version

- LSAL Version: 1.0.0 [hdl:1870/2687306FA67643DE810A8C4789DDE0E0]
- JADL IPA Version: 1.0.0 [hdl:JADL-IPA-NA/E68E78264D634D3082FDBAE0B41D6924]
- FRED Version: 1.0.0 [hdl:FREDNA/2BF7810E0B434CE4B7B1124662A98B05]

Version History			
Version	Date	Author	Description
0.50	2007-07-31	DR	Initial JADL IPA version based on FRED harvest Service Genre version 1.0.0. hdl:FREDNA/2BF7810E0B434CE4B7B1124662A98B05 hdl:JADL-IPA-NA/E68E78264D634D3082FDBAE0B41D6924
0.51	2007-08-02	DR	Editorial review, consistency.
0.70	2007-08-02	DR	Draft for review.
0.90	2007-08-18	DR	Final editorial.
1.0.0	2007-08-31	DR	Final JADL IPA V1.0.0.
1.0.0	2008-06-18	DR	LSAL V1.0.0. Derivative from JADL IPA version.

Notation

The Notation element includes conventions used to describe the Service Genre.

The words MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, MAY, and OPTIONAL in this document are to be interpreted as described in [RFC 2119].

Notational conventions follow those given in the LSAL *Service Notation and Document Conventions*.

The identification and versioning scheme follows those given in the LSAL *Service Component Identification Scheme*.

The service classification scheme follows those given in the LSAL *Service Classification Scheme*.

The Service Genre Description follows those given in the LSAL *Service Genre Description Guidelines*.



Description

The Description element is an informal, standalone, non technical narrative description of the Service Genre (problem, process, business level capabilities and workflow).

The harvest service genre provides the mechanism to gather information from a resource. It is an example of a polling request-response process. (“Polling” is the process of information gathering that is initiated by the client, and recurs periodically.) The resource is assumed to be a managed collection of objects, each with harvestable data. The prime use is gathering information (metadata) about the individual elements or objects within the collection. Harvest is not normally used to gather information about the collection as a whole or to collect the resource itself, unless that information pertains to and is used to manage the harvesting process.

This service genre focuses on gathering information from repositories and other similar data collections of content objects. Typically the data gathered will be metadata about the objects in a repository. Typically this metadata is gathered for federation purposes. For example, harvesting can be used to gather metadata about the objects in the constituent repositories in the federation, to create a federation registry of metadata. The harvest service genre may also be used to gather the objects from the resource.

The service genre is specialized in service expressions to harvest a particular type of information about the objects in the collection, e.g., their metadata, by specifying data formats and communication protocols.

Resources (collections) expose a “harvest” interface, defined in the service expressions that specialize this service genre. Clients may send requests to this interface to gather information stored or managed by the resource. This information could be metadata describing objects stored in the resource; or it could be the objects themselves. The service end point associated with the resource will respond to the request with a set of information, one (logical) item corresponding to each object managed by the resource. The client may request all information for a specified subset of the objects managed by the resource, or particular information about the objects. The service end point will determine what data (if any) to return in response to the harvest request. The details of the data model used to return results to the client are defined in the service expressions that specialize this service genre.

Note, as with other service genres, there is no expectation that all service implementations will be interoperable. Different service expressions MAY take different approaches to defining interfaces and data models.

As defined, the harvest service genre is not access controlled, i.e., any client may attempt to contact a harvest service end point. There are no authentication controls. The service end point for the resource is responsible for determining what results it will return.

Usage Scenarios

The Usage Scenarios element is an informal, non technical description of how the Service Genre is used. An illustration of process or problem workflows, expressed using services, is included. An illustration of an application using the components of the Service Genre may be included (but not a description of the design of the application). No critical or essential information required to understand the Service Genre should be included.

Enable–Harvest–Ingest

The typical workflow for harvesting involves three discrete pieces:

1. A (harvestee) service exposes a harvest *interface* to the resource managed by the harvest service. Developing this interface includes selecting which information about an object in the resource may or may not be requested through the harvest service. Often there is no differentiation made between information which can or cannot be harvested (i.e., the resource is open access).
2. A (harvester) external agent makes a request to collect information through the harvest service from the (harvestee) resource. The external agent “*harvests*” through this service genre. The resource, which is behind the service end point for the service instance of the service genre, “*exposes to harvest*” the available, harvestable information from the resource.
3. The (harvester) external agent adds the harvested information to a distinct, aggregating resource. This is done through an “*ingest*” service, which can be referred to in this context as an “*ingest (harvest)*” service.



Because the aggregating resource ingests the harvested data, it is often referred to as “harvesting” information from the harvestee resource. However, the harvest service genre described herein details only the delivery of information to a harvester. It makes no assumptions on what the harvester does with the information.

Harvest Metadata

A common scenario for harvesting, instantiating the workflow above, is to facilitate discovery of content objects across a set of data providers. This is done by harvesting metadata describing the objects from the various data providers (as opposed to harvesting the content objects themselves), by a single harvesting client. The harvesting client then builds a central resource (often called a registry), containing the harvested metadata. In the workflow outlined above, the harvestees are the data providers; the harvester is the harvesting client; and the central resource is the aggregating resource, i.e., the endpoint for “ingest (harvest)”.

Discovery of objects is performed against the central resource; not against the harvested repositories. Once the metadata has been harvested and collected in the central resource, discovery across the range of repositories involves only a single query on the central resource, searching through the harvested metadata, rather than launching a federated search across all of the data providers.

The metadata providers form a federation of repositories by having their metadata harvested and centralized. The federation can be ad hoc, with metadata harvested as found on open access repositories; or it can be controlled and formal, with policy guarantees underpinning it such as uniform metadata profiles and service-level agreements. The typical scenario is:

1. A constituent (source) repository in the federation exposes an interface to the harvest service.
2. Client identifies, through *harvest control functions* on the source repository, content objects with associated metadata records to be harvested, and metadata formats available for the harvest.
3. Client harvests a metadata instance for each content object from the source repository.
4. Client validates the harvested metadata against the registry schema.
5. Client ingests the harvested metadata into its registry.
6. End users may perform discovery using the harvested metadata on the registry.

Figure 1 illustrates this scenario as applied in an identifier-enabled repository:

- Content objects are identified and accessed through identifiers that resolved to the source repository-specific labels (Step 2);
- The harvestee resource provides a metadata record corresponding to the specified content object (Step 3);
- The harvested information is published to end users through the registry (Step 6).

The metadata record in the central resource is assumed to include a retrieval key for the associated content object, which can be used with an obtain request for that content object in the source repository (see *Obtain Resource* service genre). The retrieval key value may be populated before or after harvest, depending on the federation business model.

The metadata in the registry is expected to stay reasonably up to date. As a result, harvesting will usually be a periodically recurrent activity, and restricted only to those metadata records updated on the source repository since the last harvesting session.

A client may not be interested in ingesting all the metadata records available from the provider, but only those relevant to the registry subject matter. The provider needs to offer a mechanism of restricting harvesting to those records identified as relevant.

Notify

An alternate use for harvesting is to notify consumers of new objects added to a repository. A periodic harvesting operation gathers descriptions of the objects added to the repository since the last harvesting cycle; these descriptions are exposed through harvested metadata records. The descriptions are then made available to consumers through some form of notification or syndication.



Harvest Content

Harvest may be used to harvest other data disseminations besides metadata, e.g., an abstracting service may harvest abstracts or summaries of articles from a set of repositories to build a central collection of abstracts exposed to discovery.

Mirror

Harvest may be used to harvest data objects from a source directly in support of providing a data mirror or archive. The harvesting client will periodically harvest the source to get changes and updates from the source and reflect these in the mirror or archive.



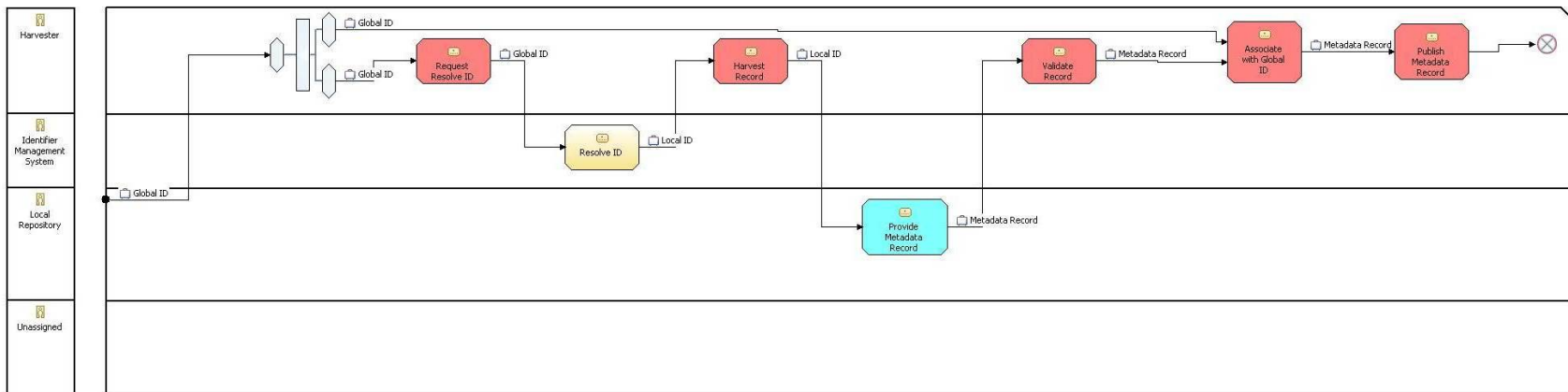


Figure 1: Harvest Metadata Workflow



Applicability

The Applicability element details when the Service Genre is used or not used. It represents specific constraints and assumptions on the use of the Service Genre. It is more specific and normative than the informal Usage Scenarios. No critical or essential information required to understand the Service Genre should be included.

As described in the *Usage Scenario* element, this service genre describes requests for information from a resource and the delivery of that information in the response. Any subsequent processing of that information, such as its ingestion in a second resource, is out of scope for this service genre.

As defined, the service genre is applicable for harvesting any defined data dissemination for any object within a resource. It is not restricted to harvesting metadata.

Behavior of the service genre is not defined when the resource requires authentication to permit harvest.

Behavior of the service genre is not defined when the resource requires authorization or access controls to permit harvest.

Behavior of the service genre is not defined when the harvest service end point filters results that are returned based on authorization policies or rules that need to be communicated to the resource through the harvest control interface.

Behavior of the service genre is not defined if communications need to be secure.

Functionality

The Functionality element details and illustrates the behaviors provided by the Service Genre, in terms of services, workflows, messages, resources, and data objects. It is not a technical description of the Service Genre, but it must provide sufficient information to develop the Requests & Behaviors of the Service Genre and to evaluate conformance of the Service Genre to the stated behaviors. It should not include implementation-specific information.

The harvest service genre supports two types of functions:

Harvesting Control Functions. These functions are used to get control and descriptive information about the resource (as a whole) being harvested. This information enables a client to successfully communicate with the service end point that is providing the interface to the resource. Control information gathered MAY include:

- Description (machine processible) of communications and transport protocols supported.
 - Protocol information SHOULD include version numbers.
 - Protocol and communications information MAY include overall flow and control parameters (e.g., synchronous, asynchronous, flow control, time outs).
- Description of information that can be gathered from or disseminated by the resource for the objects it manages (e.g., metadata formats). Schema and format information SHOULD include version numbers.
- Structural information about the collection of objects managed by the resource, used by the client to request particular information about an object.

Data Harvesting Functions. These functions are used to get information about the objects in the resource being harvested. Requests MAY specify:

- What objects to harvest: all or a subset, specified by object identifier, object association (e.g., specifying a metadata object through the identifier of the corresponding content object), or selection condition, e.g., when created, type of object, what part of a collection.
- What information to disseminate: all or a subset, specified by type (e.g., object versus metadata versus identifier) or format of what to disseminate (e.g., attributes, applied dissemination transformation).

Mechanisms MAY exist to provide flow control so that large results sets are returned in chunks.

No other functionality is defined. The functionality that is defined MAY be extended.



Requests & Behaviors

The Requests & Behaviors element details all of the behaviors exposed by the Service Genre. It lists functionality that can be used by applications or Service Implementations. The information must be sufficient to specialize the Service Genre to one or more Service Expressions.

The format and definitions for requests and responses SHALL BE defined by the service expressions that specialize the service genre. Requests and behaviors SHALL meet the following conditions:

- At least one defined data harvesting function SHALL be defined.
 - The data harvest function SHALL be capable of harvesting information for all objects managed by the resource exposed to harvest.
 - Restrictions on harvest SHALL be exposed through *Data Harvesting Functions*.
 - Restrictions on harvest SHALL NOT result from extrinsic factors such as size constraints.
 - The request SHOULD include mechanisms to specify or control the format or content of results sets.
 - The request SHOULD include mechanisms to specify the part of the results set to be returned.
 - Responses SHOULD include flow control information.
- At least one defined data harvest control function SHOULD be defined.
 - The response SHALL return basic metadata about the target resource.
 - The response SHOULD include protocol and result format information.
- Responses SHALL include error indicators or other needed control information.

Use & Interactions

The Use & Interactions element details how the how the Requests & Behaviors are combined to provide the stated functionality of the Service Genre. This is a precise technical description of how the Service Genre provides its capabilities.

The model for a client to interact with a service implementation SHALL BE defined by the service expression that specialize the service genre.

Structure

The Structure element provides a conceptual model of how the Service Genre manipulates data and state to provide results in response to requests. An illustration of the structure should be included. The structure information is used to specialize the Service Genre to one or more Service Expressions, but is not needed to understand how to use or interact with the Service Genre.

The resource is assumed to be a managed collection of objects. The harvestable data could be the content of the object, metadata describing the object, or both. This data can be expressed in one or more consistent data formats (including the possibility of on-the-fly transformations—“disseminations”). A request to harvest information specifies the objects from which data is to be harvested, and the formats in which the data will be presented. As a precondition to that request, the requestor must ascertain through the harvest control functions what the objects to be harvested are, and what the available data formats are.

The diagrams illustrate the flow for *Harvest Control Functions* and *Data Harvest Functions*.



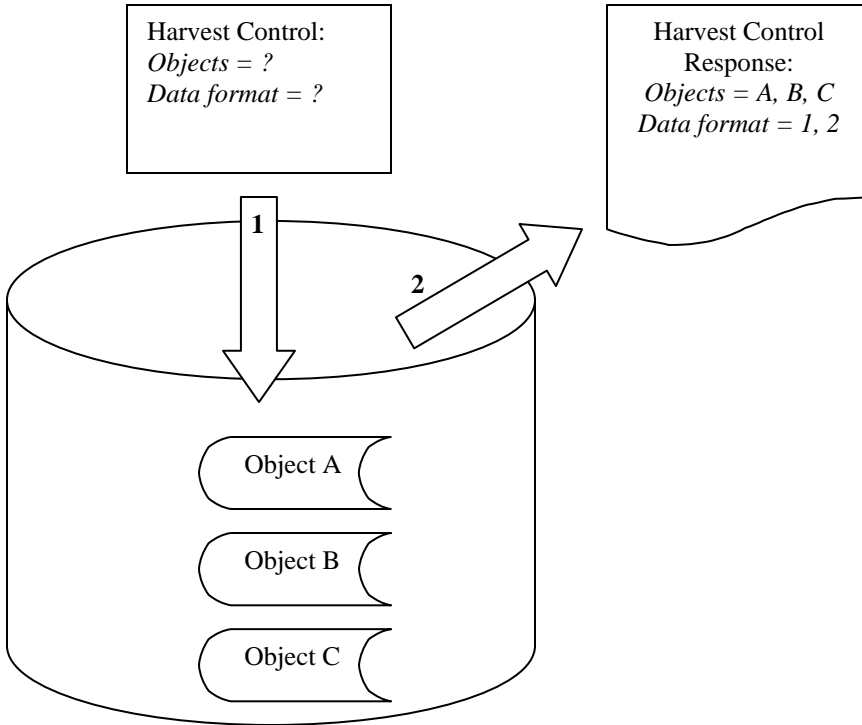


Figure 2: Harvest Control Function

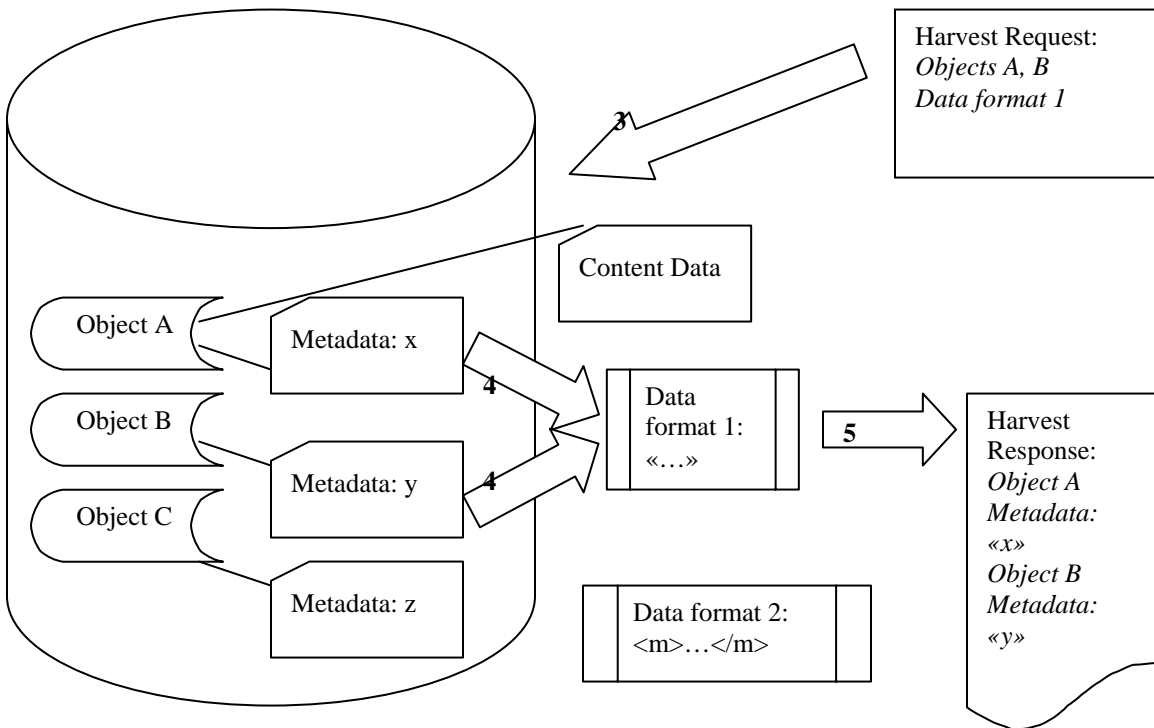


Figure 3: Data Harvesting Function



Design Decisions & Tradeoffs

The Design Decisions & Tradeoffs element documents overall choices, tradeoffs and their implications on the design of the Service Genre. It does not address the issues related to the internal details of a Service Implementation used to implement the Service Expression based on the Service Genre. No critical or essential information required to understand the Service Genre should be included.

Client authentication and authorization is not a part of this service genre.

Implementation Guide & Dependencies

The Implementation Guide & Dependencies element describes issues of concern in specializing the Service Genre to one or more Service Expressions and their corresponding Service Implementations. Resolution of issues discussed is deferred to the actual Service Implementation design. No critical or essential information required to understand the Service Genre should be included.

The following design decisions apply to the service expression that specializes the service genre.

Design:

- The service expression MAY include the specification of the communications protocol as part of its definition (e.g., as in OAI-PMH) or it MAY layer the functions on top of another communications protocol (e.g., using SQI as the communications and control protocol).
- The service expression SHOULD clearly and cleanly separate control functions from harvest functions, e.g., global control functions should not be part of individual harvest requests.

Consistency:

- The service implementation SHALL ensure that all objects and all disseminations managed by the resource are harvestable at all times, i.e., timing of updates and transactions on the resource do not impact harvest requests in a way that would omit objects from result sets.

Performance:

- A service implementation SHALL be capable of handling simultaneous requests from different clients.
- A service implementation SHOULD implement an indexing scheme or equivalent method to permit efficient harvesting by identified selection criteria.
- A service implementation SHOULD NOT implement harvest as a search of the objects in the resource if that imposes a severe performance penalty.
- Load balancing SHOULD be implemented for large resources or those which are harvested frequently (continuously).

Security and Privacy Considerations:

- Service implementations may be subject to denial-of-service attacks.
- Care should be taken to maintain privacy of any personal data or other records that may disclose usage patterns.
- There are no authorization or authentication controls. Care should be taken to maintain data privacy.

Applicable Standards

The Applicable Standards element lists domain-specific standards applicable to the Service Genre as a whole. Standards are described in terms of name, version and citation link. Conformance requirements and extensions should be noted. Standards used to implement applications are excluded. No critical or essential information required to understand the Service Genre should be included.

None. No standards are directly applicable to the service genre as a whole. The service expressions that specialize the service genre SHALL BE defined in terms of standards:

- Service expressions SHALL specify applicable harvesting standards and protocols.
- Service expressions SHALL specify applicable data formats and schemata for harvested data.



- Service expressions SHALL specify applicable communications and transport protocols.

The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0

<http://www.openarchives.org/OAI/openarchivesprotocol.html> is an example of a protocol that can be used to define a service expression that specializes the service genre.

Known Uses

The Known Uses element documents actual uses of the Service Genre in applications and systems, including how used, extensions, limits.

Actual: None as documented.

Potential: The service genre could be used in a service usage model for a repository federation containing a central federated metadata registry. The federated metadata registry would be the client that sends harvest requests to the service implementations providing harvest interfaces to the repositories in the federation. The requests would be used to gather the metadata used to populate the federation registry. The client would periodically harvest the repositories in the federation to obtain updates to the objects held in the repositories. The federated metadata registry could also provide a service implementation interface to allow other clients to harvest the metadata in the federation registry, building a federation of federations. The typical implementation would use an OAI-PMH-based service expression as a specialization of the service genre.

Service Genre Dependencies

The Service Genre Dependencies lists other Service Genres that this Service Genre is dependent upon. Dependent Service Genres are identified by name and version.

None.

Related Service Usage Models

The Related Service Usage Models element documents and illustrates how the Service Genre is used in Service Usage Models. Related Service Usage Models are identified by name and version. No critical or essential information required to understand the Service Genre should be included.

(FRED) repository federation: V1.0.0.

[\[link to service usage model\]](#)

Harvest (genre) is a part of the repository federation service usage model (genre based) and is used to gather metadata from the repositories and collections that participate in the federation to build the registry data used for discovery. The repository federation service usage model provides an integrated set of service genres used to populate and use the metadata registry that supports a repository federation. Functionality includes content management (creation and management of metadata objects within a repository federation), content discovery (discovery of content objects from a repository federation) and content delivery (retrieval of and access to content objects discovered through a repository federation).

(ADL-R) repository federation: V1.0.0.

[\[ADL-R repository-federation-sum-v100 / hdl:1870/E2FE4AD428A1468FA284E270245F72D7\]](#)

The (ADL-R) repository federation service usage model is a derivative subset of the (FRED) repository federation service usage model. For the purposes of this service genre they are interchangeable.

Related Service Patterns

The Related Service Patterns element documents and illustrates how the Service Genre is used in Service Patterns. Related Service Patterns are identified by name and version.

None.



References

The References element includes references and bibliographic citations to works needed to understand the Service Genre.

The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0.

Glossary & Terminology

The Glossary & Terminology element defines domain-specific terms used in documenting the Service Genre.

All repository domain terms in the *Domain-Specific Terms* section of the LSAL *Service Glossary* are applicable to this service genre.

Working Notes / Things To Do

The Working Notes element documents open issues in the development of the Service Genre and is for internal project use only. It should be deleted before the Service Genre is submitted for publication.

None.

