OperationsProcedures
Operations and Procedures - an in-depth look

Updated Nov 3, 2009 by kyle.br...@gmail.com

Different operations correspond to different types of BioDatasources. In the HIT, there are two different classes of BioDatasources:

1. Metadata Updater BioDatasources
2. Operator BioDatasources

---

1. Metadata Updater (BioDatasources)

Metadata updaters gather information about the number of resources behind a given access point and create a new (or update an existing) Operator for each one. There is only a single class of operations that correspond to Metadata Updaters: Metadata Collection Operations.

Metadata Collection Operations

- Metadata Update

2. Operator (BioDatasources)

An Operator, representing a single resource, is used to perform harvesting and indexing operations against that resource. There are two classes of operations that correspond to Operators: Harvesting Operations and Indexing Operations:

Harvesting Operations

Harvesting has been broken down into 4 major operations. Keep in mind that often an operation depends on the output of its predecessor. These are as follows:

- Harvesting
  - Inventory
  - Process_Inventoried
    - Creating Name Ranges
  - Search
  - Process Harvested

Indexing Operations

Indexing procedures can vary considerably depending on the sources that are being harvested and the data structure that you are synchronising with. As a reference, the indexing procedures as carried out by GBIF are presented here and are broken down as such:

- Indexing
  - Synchronising
  - Extracting

Metadata Update

For some protocols, it is enough to issue a single metadata request in order to be able to collect all the information. For others, a series of different requests need to be made. The way in which the metadata update is performed for each protocol is presented here:

- metadata update - DiGIR
- metadata update - BioCASe
- metadata update - TAPIR (coming soon)
- Archive metadata update - DwC Archive (coming soon)

Metadata Update - DiGIR

The information we want:

1. How many resources are located behind that access point? Then for each resource, try to find answers to the following:
2. What is the resource's code?
3. What is the conceptual schema supported?
4. How many records are being served?
5. What is the maximum number of records returned in an inventory request?
6. What is the maximum number of records returned in a search request?
7. When was the resource last updated?
8. What is the minimum query length?
9. What is the version of the DiGIR protocol being used?

To answer these questions, a metadata request is issued. An example response can be seen below. Notice how it contains answers to all our questions:

```
<resources>
 <resource>
  <name>UAM Insects Specimens</name>
  <code>uam_ento</code>
  .
  .
  <conceptualSchema schemaLocation='http://bnhm.berkeley.edu/manis/DwC/darwin2jrw030315.xsd'>http://digir.net/schema/cor
  .
  .
  <recordBasis>PreservedSpecimen</recordBasis>
  <numberOfRecords>10892</numberOfRecords>
  <dateLastUpdated>2009-05-28T04:00:00-08:00</dateLastUpdated>
  <minQueryTermLength>1</minQueryTermLength>
  <maxSearchResponseRecords>25000</maxSearchResponseRecords>
  <maxInventoryResponseRecords>25000</maxInventoryResponseRecords>
 </resource>
 <resource>
  <name>UAM Herbarium Specimens</name>
  .
  .
  .
 </resource>
</resources>
```

The answer from question 3 (the conceptual schema) is then referenced against two different properties files:

- conceptualMapping.properties: contains conceptual schema to index-mapping file mappings
- protocolMapping.properties: contains conceptual schema to protocol version mappings

This helps us identify the answers to two remaining questions for each resource:

1. What index-mapping file should be used?
2. What version of the protocol is being used?

Afterward, each resource and its associated metadata is used to construct a new Operator (BioDatasources). Each Operator now contains all the information necessary to be able to start harvesting and indexing itself!

Metadata Update - BioCASe

Unlike DiGIR, BioCASE access points only have a single resource behind it. Encoded using ABCD, however, there can actually be several different Datasets wrapped up into one. In the HIT, Datasets are each divided into different Operator (BioDatasources), just like we do with resources behind a DiGIR access point.

Take for example the following extract from a response to a BioCASe search request (filtering by scientificName and encoded using ABCD 2.06):

```
<DataSets xmlns='http://www.tdwg.org/schemas/abcd/2.06'>
 <DataSet>
  <Metadata>
   <Description>
    <Representation>
     <Title>Herbarium collection</Title>
    </Representation>
   </Description>
  </Metadata>
  .
  .
  <Units>
   .
   .
   .
  </Units>
 </DataSet>
 <DataSet>
  <Metadata>
   <Description>
    <Representation>
     <Title>Insect collection</Title>
    </Representation>
   </Description>
  </Metadata>
  .
  .
  <Units>
   .
   .
   .
  </Units>
 </DataSet>
</DataSets>
```

In order to be able to treat Datasets individually (that is to filter inventory and search requests by Dataset title), however, the provider must ensure that Dataset title is a supported searcheable term. Therefore the first question we need to ask is:

1. Is Dataset title a supported searcheable term?
    - If NO: We have to treat all Datasets collectively as a single Operator.
    - If YES: We want to know how many different Datasets are located behind that access point, and convert each one into a separate Operator.

So first thing first, we need to first perform a capabilities request. The following is an extract from an actual response:

```
<biocase:capabilities>
 <biocase:SupportedSchemas namespace="http://www.tdwg.org/schemas/abcd/2.06" request="true" response="true">
  .
  .
  <biocase:Concept datatype="normalizedString" searchable="1">/DataSets/DataSet/Metadata/Description/Representation/Tit
 </biocase:SupportedSchemas>
</biocase:capabilities>
```

As you can see, this directly answers our 1st question. In addition, we are also able to answer:

2. What is the most recent content namespace (conceptual schema) that's supported?

The content namespace is then referenced against two different properties files:

- conceptualMapping.properties: contains conceptual schema to index-mapping file mappings
- protocolMapping.properties: contains conceptual schema to protocol version mappings

This helps us identify the answers to 2 other questions:

3. What index-mapping file should be used?
4. What version of the protocol is being used?

Now that answers have been gathered for all our preliminary questions, we can begin to think about creating new (or updating existing) Operators. The next step depends on how we answered question #1. If Dataset title is in fact a supported searcheable term, the next step is to perform a BioCASe scan request of all Dataset titles. This will give us the answer to the following question:

5. What are the names of all different Datasets?

The following is part of an actual response:

```
<biocase:content recordCount="2" recordDropped="0" recordStart="0" totalSearchHits="2">
    <biocase:scan>
      <biocase:value>Herbarium Specimens</biocase:value>
      <biocase:value>Insect Specimens</biocase:value>
    </biocase:scan>
</biocase:content>
```

This allows us to collect the names for the different Datasets. Afterward, we want to know:

6. How many records does each Dataset have?

In order to answer this question, we need to send a BioCASe search request, filtering by Dataset name and setting count equal to true. The following is part of an actual response that would be generated:

```
<biocase:content recordCount="0" recordDropped="0" recordStart="0" totalSearchHits="0">
 <biocase:count>42</biocase:count>
</biocase:content>
```

Each Operator now contains all the information necessary to be able to start harvesting and indexing itself!

Metadata Update - TAPIR

coming soon

Metadata Update - DwC Archive

coming soon

Harvesting

Inventory

Process Inventoried

Creating Name Ranges

Search

Process Harvested

Indexing

Synchronising Records

Extracting Records