Biocache-store

CLI tool for managing occurrence records

Manash Shah
GBIF Sweden
October 03, 2017

Biocache-Store

Naturhistoriska riksmuseet

- Command Line Interface tool
- Implemented in Scala
- Built with maven
- External property file for configuration
- Ansible scripts available for setup

```
/ #
/ # biocache
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/biocache/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/biocache/lib/logback-classic-1.1.3.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2017-10-01 16:45:45,871 INFO : [ConfigModule] - Using config file: /data/biocache/config/biocache-config.properties
2017-10-01 16:45:46,855 INFO : [Config] - Using the set of blacklisted media URLs defined in: /data/biocache/config/blacklistMediaUrls.txt
--------------------------------------------------------------------------------
 Biocache management tool
 Commit ID: 67cfb9c49aa546a7c447fa14970a2321a96fa4d0
 Build date: 2016-08-22T01:18:24Z
 For more detail run with --version or type 'version' at the prompt
--------------------------------------------------------------------------------

Please supply a command or hit ENTER to view command list.
biocache>
```

Functions / Features

- Manages occurrence records
  - Loading
  - Sampling
  - Processing
  - Indexing
- Additional support
  - Outlier detection
  - Duplicate detection
  - Identifying extra-limital outliers

Loading

- Download and extract resource (DarwinCore Archive)
- Retrieve metadata JSON from registry
- Construct a map of supplied field name and Index
- Check collectory(registry) for institutionCode and collectionCode
- Get related multimedia
- Load the data into the occurrence store (Cassandra db)

Sampling

- Get the distinct coordinates for this resource
- Build Location coordinates set
- Generate sampling
- Intersection with available layers
- Load the sampling into the loc table
- Load sampling to occurrence records

Processing

- Process a record
- Add metadata and records quality assertions
  - Map default values from the data resource configuration
  - Taxonomic Classification matching
  - Parse locality information

Processing (contd.)

- Point matching
  - Process geospatial details of the record
  - Parse latitude/longitude
  - Retrieve associated point mapping
- Type status normalization
  - GBIF's vocabulary

Processing (contd.)

- Event Date parsing
  - Date validation
  - Support for date ranges
- Collectory lookups for attribution
- Sensitive Data processing

Processing (contd.)

- Miscellaneous Assertions
- Process
  - Images
  - Interactions
  - EstablishmentMeans
  - Identification
  - Collectors
  - MiscOccurrence
  - OccurrenceStatus

Indexing

- Index the records (SOLR instance)
- Conform to the fields as defined in the schema
- Bulk-processor
  - Reprocessing the entire dataset
  - Resampling the entire dataset
  - Creating a new complete index offline

Outliers detection

- Checks for outliers
- Takes a taxon
- Intersects the corresponding occurrences for the input taxon with the environmental layers
- Flags the potential outliers
- Update the datasource

Duplicate detection

- Get a distinct list of species LSID and a distinct list of subspecies LSIDs (without species LSIDs) that have been matched
- Break down all the records into groups based on the year, its comprising months and subsequently date
- With the smallest grained group, group all the similar "collectors" together
- With the collector groups, determine which of the records have the same coordinates (ignoring differences in precision)

Extra-limital outliers

- Takes a taxon
- Intersects with a predefined expert distribution polygon for the given taxon
- Flags the potential outliers

Additional functions

- Create DarwinCore Archive
- Delete columns records or resource
- Download, Migrate media
- Import / Export images

https://github.com/AtlasOfLivingAustralia/biocache-store