Data pipelines for biodiversity data

Tim Robertson, Matthew Blissett

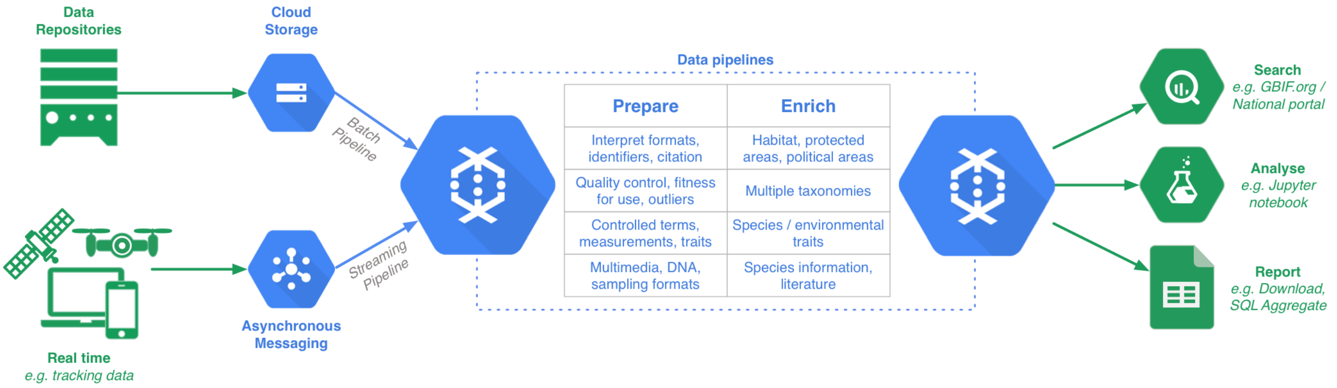
Version 1.0, 2019-05-21 09:39:56 UTC

Table of Contents

Data Pipelines

1. Process biodiversity data

Data pipelines provides components to integrate, structure, interpret and transform biodiversity data. Built for extensibility, portability and high performance, data pipelines powers services such as GBIF.org [https://www.gbif.org/].



Prepare	Enrich
Interpret formats, identifiers, citation	Habitat, protected areas, political areas
Quality control, fitness for use, outliers	Multiple taxonomies
Controlled terms, measurements, traits	Species / environmental traits
Multimedia, DNA, sampling formats	Species information, literature

1.1. Features

Some of the high-level capabilities and objectives of Data pipelines include:

- Support a variety of input formats (Darwin Core [https://www.tdwg.org/standards/dwc/], ABCD [https://www.tdwg.org/standards/abcd/], CSV files, Excel files, Shapefiles etc) with easy opportunity to include new connectors
- Support batch (e.g. a project CSV) and streaming inout (e.g. append-only tracking data)
- Align data to a standardized vocabularies, supporting multilingual data labelling
- Apply quality controls to flag errors, detect outliers and apply statements about the suitability of the data for a variety of uses (also known as fitness for use indicators)
- Enrich data by:
    - cross referencing with geospatial gazetteers for political boundaries (e.g. GADM.org [https://gadm.org/], EEZ [http://vliz.be/vmdcdata/marbound/], protected areas) and biogegraphic regions, landuse categorisation and environmental surfaces
    - organizating to multiple taxonomic classifications including the GBIF Backbone taxonomy [https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c], Catalogue of Life [http://www.catalogueoflife.org/] and national legislative taxonomies such as ITIS [https://www.itis.gov/]
- Allow consumers to easily understand the data preparation and enrichment process that has been applied (i.e. preserve and document data provenance).
- Provide clear documentation for all data transformations
- Support multiple runtime environments such as Apache Spark [https://spark.apache.org/], Google Dataflow [https://cloud.google.com/dataflow/], Amazon EMR [https://aws.amazon.com/emr/] or local

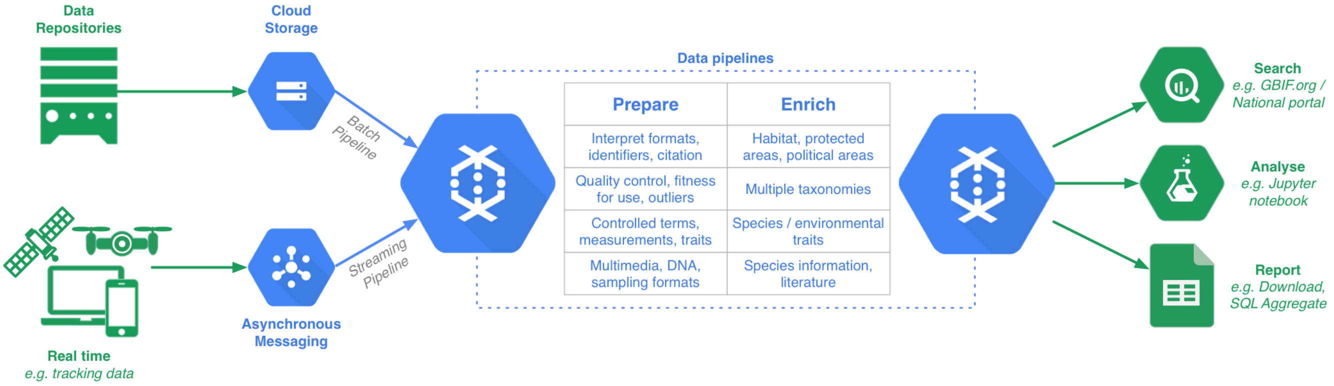
machine

- Ensure pipelines can be deployed in a high throughput environment. GBIF.org [https://www.gbif.org/] target the processing and indexing into Elasticsearch of 1 Billion records in under 12 hours

2. Architecture

The data pipeline project comprises of a collection of components which can be used as libraries in your own project and a set of runnable pipelines built around Apache Beam [https://beam.apache.org/]. Beam provides the ability to execute the pipelines on a variety of target environments, such as a local Apache Spark cluster or on Google Cloud Dataflow. Additionally Beam provides excellent IO adapters to make it easy to source and sink data into a variety of datastores (e.g. Apache Solr, Elasticsearch, Apache HBase, Apache Kafka etc) and file formats (e.g. Apache Avro).

2.1. Deployment example 1: GBIF



Prepare	Enrich
Interpret formats, identifiers, citation	Habitat, protected areas, political areas
Quality control, fitness for use, outliers	Multiple taxonomies
Controlled terms, measurements, traits	Species / environmental traits
Multimedia, DNA, sampling formats	Species information, literature

3. Roadmap

Data pipelines is a new project under active development.

3.1. Features

Sollicitudo / Pellentesi	consectetur	adipiscing	elit	arcu	sed
Vivamus a pharetra	yes	yes	yes	yes	yes
Ornare viverra ex	yes	yes	yes	yes	yes
Mauris a ullamcorper	yes	yes	partial	yes	yes
Nullam urna elit	yes	yes	yes	yes	yes

Sollicitudo / Pellentesi	consectetur	adipiscing	elit	arcu	sed
Malesuada eget finibus	yes	yes	yes	yes	yes
Ullamcorper	yes	yes	yes	yes	yes
Vestibulum sodales	yes	-	yes	-	yes
Pulvinar nisl	yes	yes	yes	-	-
Pharetra aliquet est	yes	yes	yes	yes	yes
Sed suscipit	yes	yes	yes	yes	yes
Orci non pretium	yes	partial	-	-	-

3.2. Deployments

3.3. Governance

Preparation

1. Overview

2. Input formats

3. Identifiers

4. Basic parsing

These operations are done on all terms:

1. Leading or trailing whitespace is removed.
2. Tabs and newlines are converted to spaces.
3. The exact value `null`, `\N`, `''` or `""` is removed

5. Controlled vocabularies

6. Taxonomy

7. Date/Time

8. Location

8.1. Interpretation

Darwin Core Terms: country, countryCode, decimalLatitude, decimalLongitude, geodeticDatum, coordinateUncertaintyInMeters, coordinatePrecision, verbatimCoordinates, verbatimLatitude, verbatimLongitude, verbatimCoordinateSystem, verbatimSRS

8.1.1. Interpret the datum

Parse the `geodeticDatum` using #. If it fails or is unspecified, add issues .

8.1.2. Interpret the coordinatePrecision

This records a value in degrees. The value should be in the range 0–1°, otherwise add an issue #.

8.1.3. Interpret the coordinateUncertaintyInMeters

This should be a number greater than zero (not equal to zero), less than (half the Earth's circumference). It should also be greater than the precision calculated by https://docs.gbif-uat.org/georeferencing-best-practices/1.0/en/#uncertainty-related-to-coordinate-precision, and greater than the uncertainty introduced by an unknown datum https://docs.gbif-uat.org/georeferencing-best-practices/1.0/en/#uncertainty-from-unknown-datum

8.1.4. Interpret the country

Use the uncertainty from above.

8.2. Grids

8.2.1. UTM

8.3. Political boundaries (GADM.org)

8.4. Habitats

8.5. Protected areas

Indexing

Formats

2. Field delimited formats

3. Data packages (DwC-A)

4. Geospatial formats

5. Essential biodiversity variables (EBVs)

6. Ad-hoc SQL

Community

1. GBIF Pipelines Specification

The GBIF …

1.1. Commands

Build the HTML and PDF documents:

```
docker run --rm --user $(id -u):$(id -g) -v $PWD:/documents/ gbif/asciidoctor-toolkit
```

1.2. Project layout

```
index.en.adoc   # The master document file.
*.en.adoc       # Other parts of the document
img/            # Images
en/             # Generated English document.
```

2. Tutorials

3. Contributor guide

4. Build and releasing

5. People

Part 1

1. Chapter 1

1.1. Chapter 1 point 1

2. Chapter 2

Part 2

1. Chapter 1

2. Chapter 2