GFBIO DATA TRANSFORMATION SERVICE (DTS)

David Fichtmüller, Jörg Holetschek, Katja Luther, Anton Güntsch
Center for Biodiversity Informatics and Collection Data Integration - Freie Universität Berlin - Botanic Garden and Botanical Museum Berlin

MOTIVATION

Occurrence and taxon information are floating around in the cosmos of biodiversity networks and repositories in a multitude of standards, ABCD and DarwinCore archives being the best-known. Transforming data between the different formats is possible for some of them, albeit lossy due to different scopes and granularities, but since most data standards are evolving over time, conversion rules and implementations are in constant flow. Facilitating data transformations in one place will make them easier to maintain and can reduce the efforts for implementing new services or portals.

DATA TRANSFORMATION SERVICE

The DTS implemented by GFBio and to be continued by NFDI aims at filling this gap by providing a central service for converting data between different standards and can be used by GFBio or other constituents of the biodiversity community. It supports versioning, meaning changes in a given transformation will result in a new version and won't affect the existing implementation, thus providing the stability required for use in productive environments and reproducibility as the basis for good scientific practice.

The API of the DTS is divided into three parts:

- **Discovery** of transformations allows a potential user to list available transformations (see box at the center), their respective input parameters and versioning information.
- **Invoking** a transformation job can be done with a single GET call that includes the transformation to be executed (plus an optional version), a URL of the input file and potential additional parameters required by a given transformation.
- Once a job has finished, the result file is available for **Download** for 7 days.

https://transformation.gfbio.org

#1 XSLT

ABCD > PANGAEA PanSimple

Transforms a single ABCD document into a PanSimple document, which is used for harvesting purposes in the GFBio project.

ABCD > HTML Landing Page

#2 XSLT

Transforms a single ABCD document into a human-readable description of the dataset stored in the document. The page generated contains the dataset's metadata (such as title, description, contacts, taxonomic and geographic scope) and lists the individual records with their catalog number and scientific identification result. If the ABCD field RecordURI is filled, the detailed record pages are linked.

#3

ABCD (archive) > DarwinCore Archive

This transformation will create a DarwinCore archive for an ABCD dataset. The source document can be a single ABCD file storing one dataset or an ABCD archive containing multiple documents.

#4

CDM Light > PANGAEA PanSimple (Oct 2021)

Transforms a zipped CDM Light file into PanSimple documents, which are used for harvesting purposes in the GFBio project.

TRANSFORMATION ENGINES

The transformation engine implements the actual data conversion process and is decoupled from auxiliary steps such as unzipping input files or preparing the result files for download. This allows different technologies to be used, so the tool or programming language best suitable can be chosen.

Currently, DTS uses two different engines:

eXtensible Stylesheet Transformation is a well-established way of transforming XML contents between different schemas. DTS uses this engine for converting ABCD into PANGAEA PanSimple documents or HTML landing pages (transformations 1 & 2).

Pentaho Data Integration (PDI) is an extensive community framework for ETL purposes – extracting, transforming and loading data. Java based, it allows integrating code of almost any programming language or external web services. It is used for converting ABCD documents into DarwinCore archives and CDM-Light into PanSimple (transformations 3 & 4).

CONTACT & REFERENCES

Questions & further information: j.holetschek@bgbm.org
DTS website: https://transformation.gfbio.org
DTS API: https://transformation.gfbio.org/api
Concept (internal Wiki): https://gfbio.biowikifarm.net/internal/DTS_Concept

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation