Indexing for Life

D6.1 - Integrated access to CoL in GBIF

Workpackage 6

Markus Döring

30th April 2013

Capacities Programme of Framework 7: EC e-Infrastructure Programme – Virtual Research Communities - INFRA-2010-2

Grant Agreement No:	261555
Project Co-ordinator:	Dr Alastair Culham
Project Homepage:	http://www.i4Life.eu
Duration of Project:	36 months
Start Date:	November 2010
End Date:	October 2013

Introduction

The Catalogue of Life has been included in various parts of the GBIF infrastructure to provide integrated access to the catalogue's taxonomy and names through the GBIF systems. An automatic GBIF workflow has been established that downloads and integrates the latest published catalogue on a regular basis.

GBIF Background

The GBIF infrastructure is made up of various distinct components of which the most important ones are quickly described here. The Catalogue of Life has a role in each of these.

Registry

A main role of the GBIF registry is to keep track of datasets and provide some basic metadata about them. It is meant to be a public service, but registered service URLs can be secured and don't need to be public. Currently there are 3 types of datasets that are managed:

- **Occurrence datasets**: Accessible species occurrence data such as specimens or observations. Currently there are roughly 11.800 registered occurrence datasets using various protocols such as DiGIR, TAPIR, BioCASE or Darwin Core Archives.
- **Checklists**: taxonomic, nomenclatural or other taxon oriented checklists published as Darwin Core Archives. This includes the Catalogue of Life.
- **External Metadata**: dataset descriptions originating from external metadata networks with the actual data being opaque to GBIF. Currently GBIF provides access to roughly 25.000 entries from the Knowledge Network for Biocomplexity (KNB) [1]

Checklistbank

All registered checklists are indexed by GBIF and kept in a single store called Checklist Bank (CLB). Checklist Bank offers uniform webservices across all checklists and uses the GBIF Backbone (see below) as a way to move between names in the different checklists. To do so, every checklist record, if possible, is mapped to a taxon from the GBIF backbone.

Taxonomic Backbone

The GBIF taxonomic backbone, often simply called the *nub*, is the reference taxonomy used by GBIF to link all occurrence and checklist records. It provides a standard classification most importantly used for browsing and calculating various metrics. It is a synthetic dataset which is updated from time to time (currently every couple of months) by a software which processes a prioritized selection of the checklists indexed in Checklist Bank. The backbone itself is also treated and stored as a regular checklist in CLB.

Occurrence index

Nearly 400 million occurrence records have been indexed by GBIF at this point and they are all linked to the backbone taxonomy if some taxonomic identification was provided. This is done by using a public taxonomic matching service which handles fuzzy name matching, partial classification matching and is homonym aware. It is important to stress that no taxonomic names found in occurrences are incorporated into the GBIF backbone.

[1] https://knb.ecoinformatics.org/

Portals

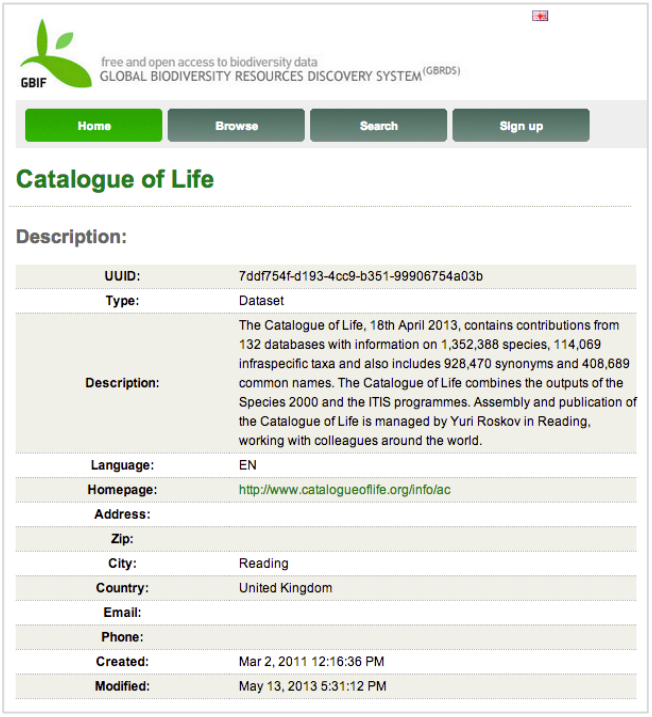
GBIF is about to release a new data portal later this year which provides a fully integrated access to the above listed components. As the current portal[2] is not maintained further, all work done for integrating the Catalogue of Life has focussed on the new portal. A first test preview will be available within the next weeks.

Until the new portal becomes available GBIF still updates the data presented in the current portal, but it often takes several months until data changes at the source actually become visible. An improved latency down to a week or even a day was a key motivation for building the new portal.

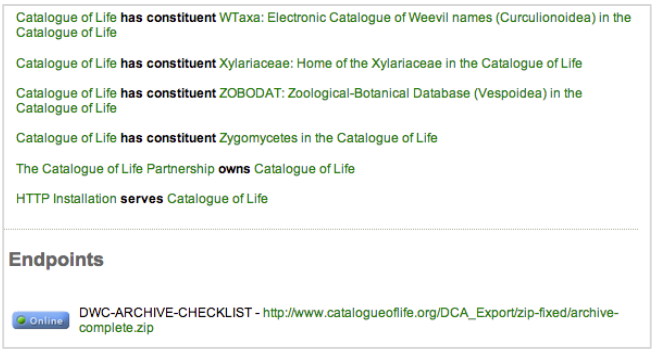
Description of GBIF CoL processing workflow

CoL in GBIF Registry

The entire catalogue of life is registered as a checklist with the dataset key 7ddf754f-d193-4cc9-b351-99906754a03b[3], with the URL of the Darwin Core Archive provided by the CoL download service being registered as a service endpoint. In the GBIF registry all stored metadata details such as the title, description and contacts are automatically updated based on metadata found in the registered endpoints. In the case of CoL this would be an EML file bundled with the Darwin Core Archive. Currently this metadata file is missing from the downloaded archive, so manual curation is needed which is error prone and often lacking behind the actual version used. Future versions of the download service will provide the needed metadata for fully automated updates.



To help citation, every GSD included in the CoL is also registered as a constituent dataset of the catalogue on its own. Again the information stored for a GSD comes from an EML metadata file per GSD which is included in the download archive already, but which is currently still missing some key information**Error! Bookmark not defined.**.

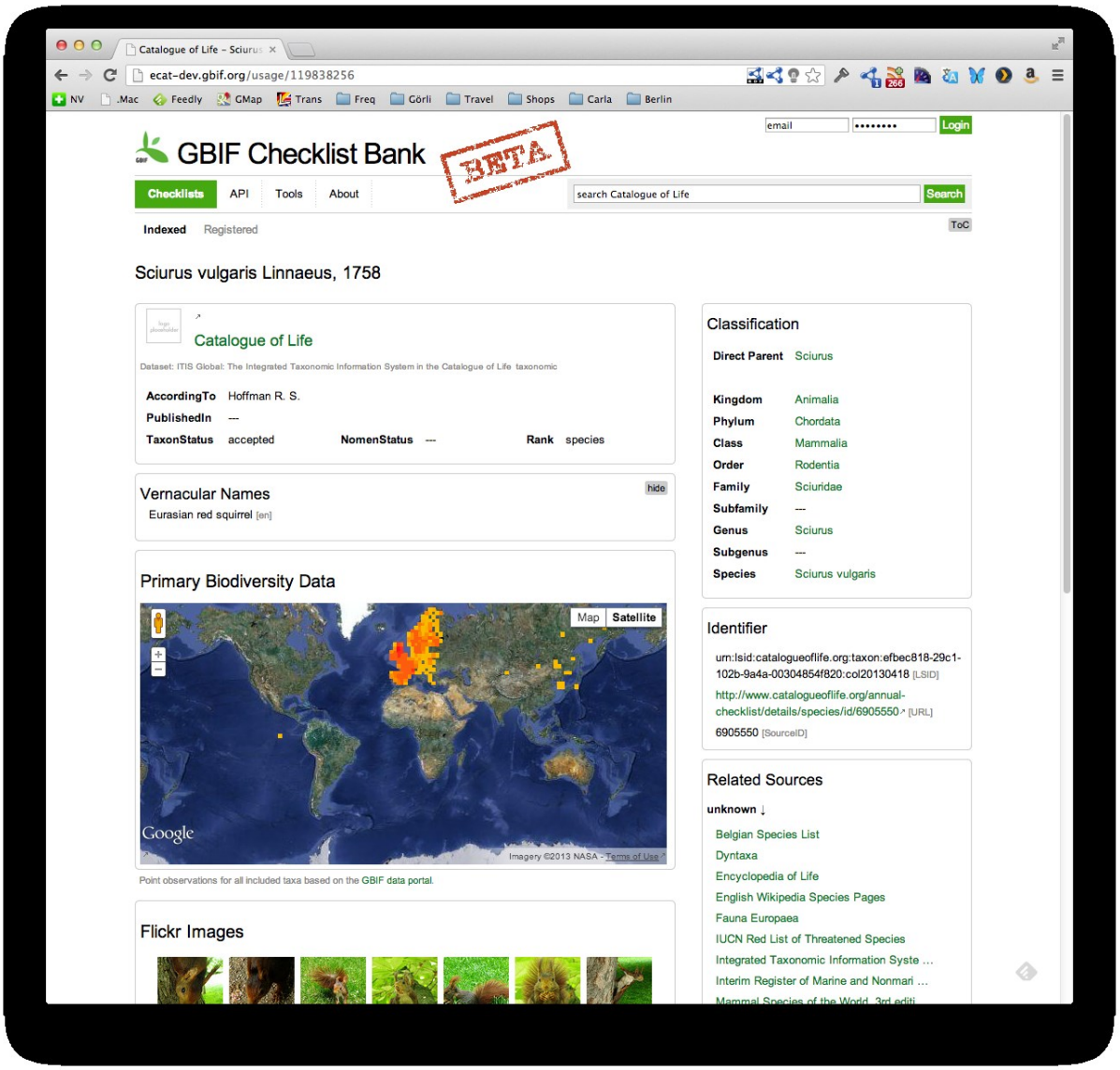


[2] http://data.gbif.org/

[3] http://gbrds.gbif.org/browse/agent?uuid=7ddf754f-d193-4cc9-b351-99906754a03b

CoL in the GBIF Checklistbank

The entire catalogue as provided in the download archive is inserted on a regular basis into checklist bank. Until the new portal becomes available Checklist Bank offers a standalone user interface which can be browsed[4] and used through webservices[5]. A taxon page with the latest scrutinizer (given as *according to*) is shown in this screenshot:



CoL and the GBIF Backbone

For GBIF by far the most important role of the Catalogue of Life is when building the GBIF backbone. The catalogue acts as the starting seed and also provides the entire higher classification above families as the only source. That means every species covered by the CoL will also be included one to one in the GBIF backbone. Because of this, GBIF does not make use of the i4Life crossmapping tools; every CoL taxon has its corresponding taxon in the GBIF taxonomy known in advance. It also means that it is crucial for GBIF that the CoL provides a comprehensive higher taxonomy, as GBIF does not augment the CoL taxonomy above family level. Currently CoL is contributes 56% of all taxa in the GBIF backbone and 40% of all accepted family names.

[4] http://ecat-dev.gbif.org/checklist/10
[5] http://ecat-dev.gbif.org/api/clb