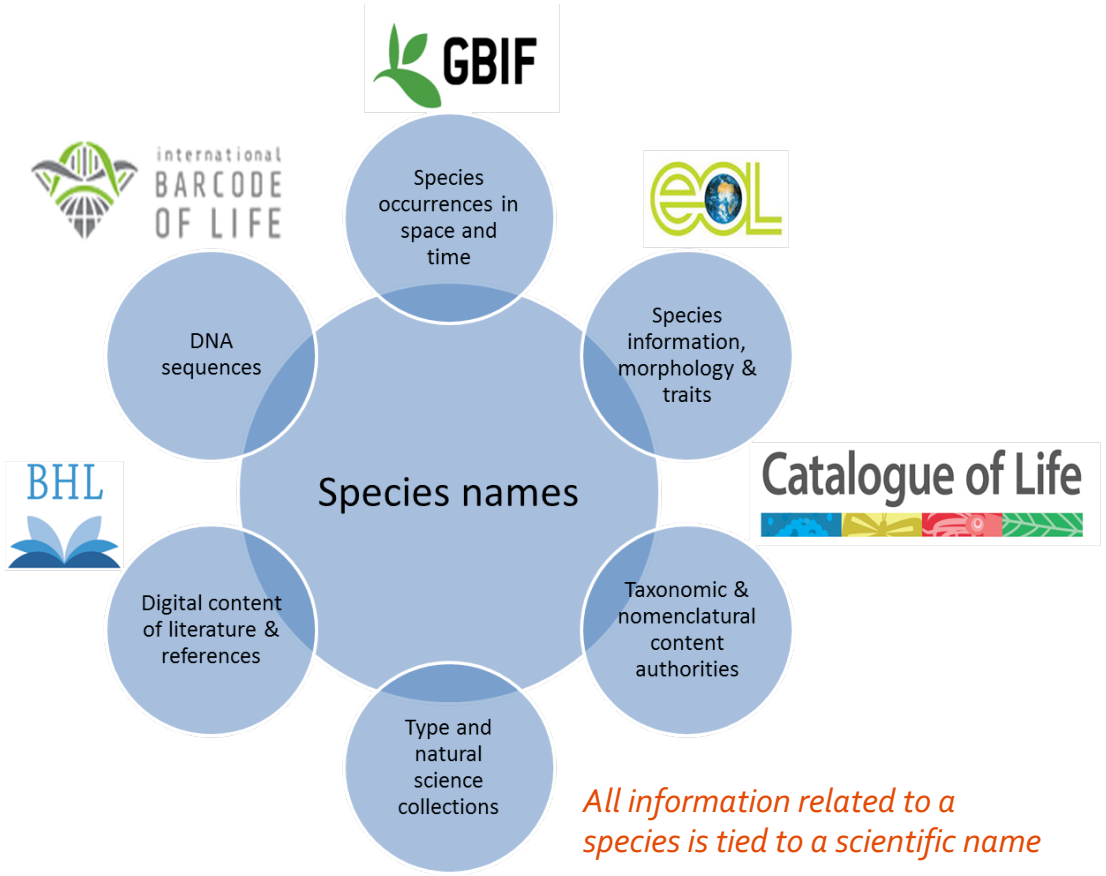
CATALOGUE OF LIFE PLUS

Innovating the Catalogue of life systems

This slide doc provides information about the Catalogue of Life plus project.

May 24, 2018, version 2.0

Prepared by: Olaf Bánki, Markus Döring, and Ayco Holleman



All information related to a species is tied to a scientific name

TABLE OF CONTENTS

The Catalogue of Life Plus initiative

We have set the goal in creating an open, shared, and sustainable consensus taxonomy and nomenclature foundation to serve the proper linking of data in the global biodiversity information initiatives.

In 2015 the global biodiversity information initiatives Biodiversity Heritage Library, Barcode of Life Data systems, Catalogue of Life, Encyclopedia of Life, and the Global Biodiversity Information Facility Secretariat took the first step to work on the idea for building a single shared authoritative taxonomic backbone that can be used to order and connect biodiversity data across various domains.

Each of these initiatives focus on the delivery of a consistent, normalised view of available data for a particular class of biodiversity information (GBIF - specimens and occurrence records, CoL - species names and concepts, EoL - species traits and species-level information resources, BHL - biodiversity publications, BoLD - barcode sequence records).

As a fundamental axis for organising their data, these global biodiversity information initiatives depend on the use of scientific names and the associated species concepts. Presently, there is no possibility to use the same foundation for names and taxonomy.

We continue building a consortium and a joined vision.

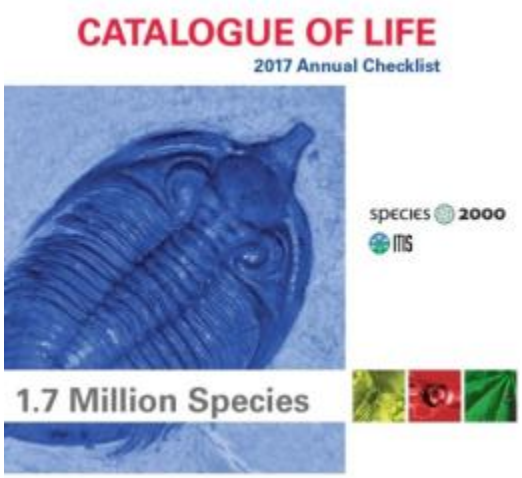
Following the initial meeting in 2015, the formation of the Catalogue of Life Plus initiative is gradually developing. The Catalogue of Life and Species2000 governance have adopted and endorsed the initiative. Also GBIF has formally embraced the initiative and incorporated it in their strategic and implementation plan for the period 2017-2021.

Catalogue of Life

status and issues

The Catalogue of Life is the most comprehensive and authoritative global index of species currently available. It consists of a single integrated species checklist and taxonomic hierarchy.

www.catalogueoflife.org



The 2017 annual checklist contains information from 156 taxonomic databases, resulting in 1.7 million accepted species names globally. These source databases are federated and delivered in different formats. Despite the global coverage in taxonomic groups, there are still gaps. Source databases do vary in completeness, curation, and in the use of nomenclators as pre-existing data foundation. The processes to build the Catalogue of Life are dated, over-reliant on manual intervention and suboptimal with respect to identifiers and the tracking of stable historical editions.

The Catalogue of Life is in need of an improved, stable, and performant IT infrastructure. This infrastructure should support and accelerate the editorial work for the Catalogue of Life. It should provide reliable identifiers for both names and taxonomic concepts. By integrating many more overlapping taxonomic and nomenclatural sources into a provisional Catalogue, extra information such as homotypic synonyms, literature references, and vernacular names can be offered for review in the existing, scrutinized taxonomic sectors. The use of the Catalogue of Life by others, including global biodiversity information initiatives, should be increased as well as metrics to monitor use.

GBIF Backbone Taxonomy

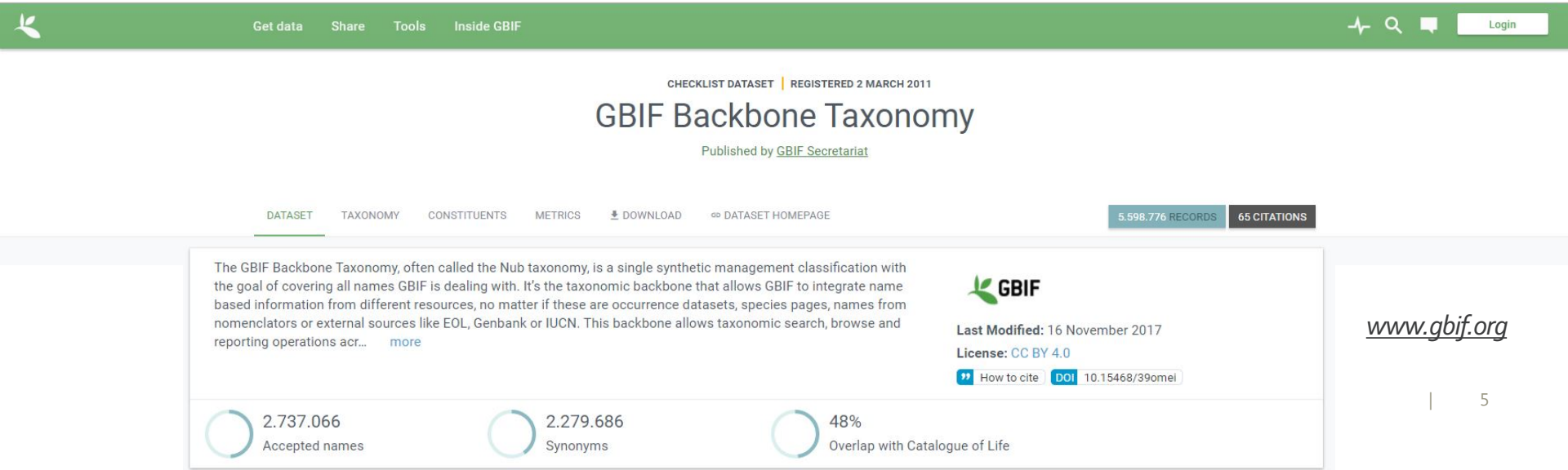
status and issues

The GBIF backbone allows taxonomic search, browse and reporting operations across all resources in a consistent way and to provide means to crosswalk scientific names from one source to another.

The GBIF Backbone Taxonomy is a single synthetic management classification with the goal of covering all names GBIF is dealing with. It's the taxonomic backbone that allows GBIF to integrate name based information from different resources, no matter if these are occurrence datasets, species pages, names from nomenclators or external sources like EOL, Genbank or IUCN. It is updated regulary through an automated process in which the Catalogue of Life acts as a starting point also providing the complete higher classification above families.

In addition 56 taxonomic sources have been used to assemble the GBIF backbone.

There is a need to build a backbone that is open for expert contribution instead of relying fully on an automated process. There is also a need for increasing the pool of scientific names to improve recall. In addition there is a need to address taxonomic gaps to improve the precision of the backbone.

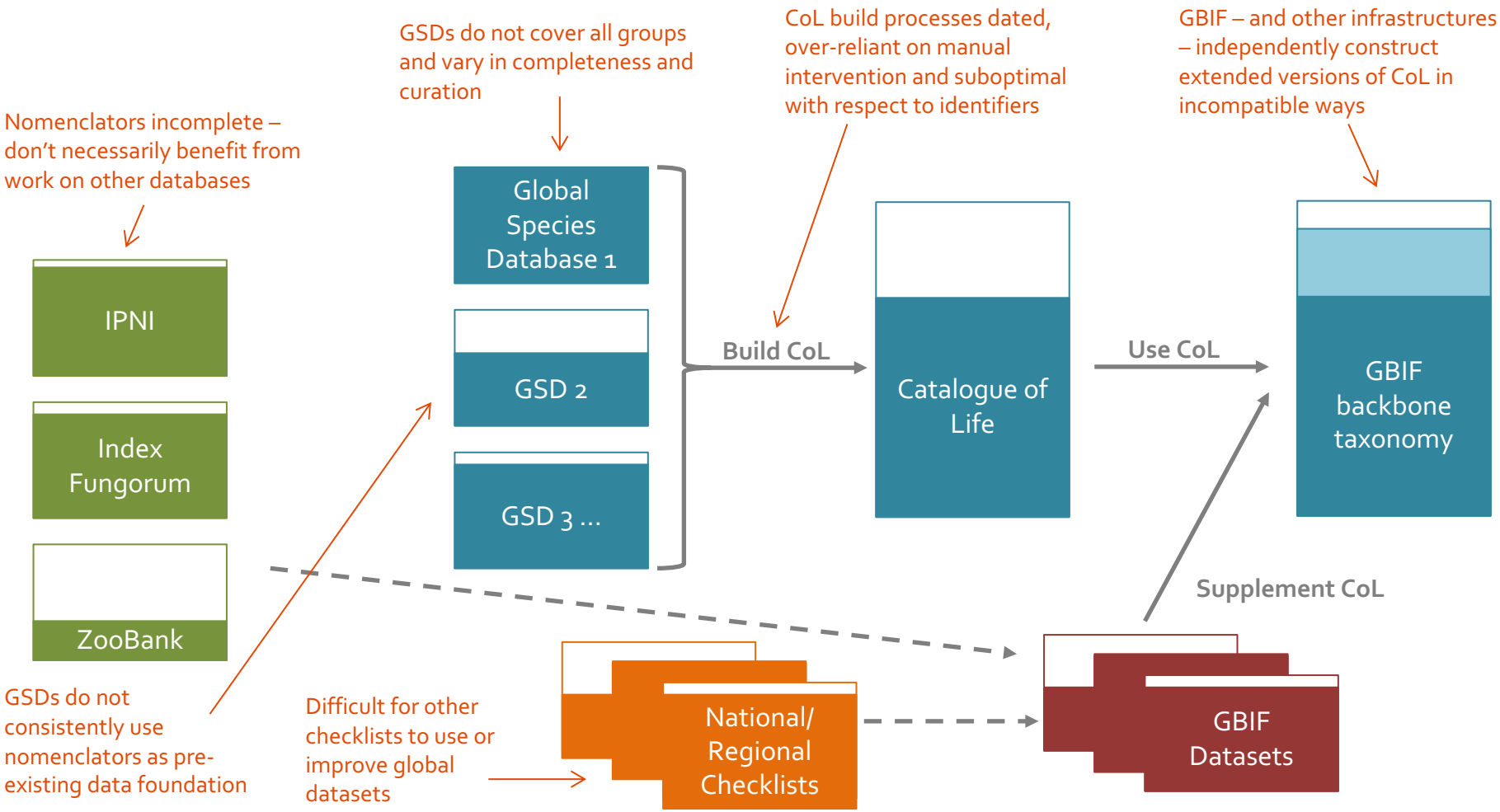


www.gbif.org

Current situation & issues GBIF & CoL

The Catalogue of Life and the GBIF Backbone Taxonomy are independently constructed.

It is unclear whether the same nomenclatural foundation is used. Global taxonomic gaps are not covered jointly.



Courtesy of D. Hobern

Catalogue of Life Plus project

Catalogue of Life Plus project

Project structure

The CoL+ project

2017 – 2019

As part of the GBIF global work programme activity 2b *deliver a names infrastructure*, the Netherlands government , through the Netherlands Biodiversity Information Facility, grants a funding of € 360K . This funding is to lay a fundament and kick-start the CoL+ initiative in the form of a project.

As initial project partners, consisting of the Global Biodiversity Information Facility Secretariat, Species 2000/Catalogue of Life, and Naturalis Biodiversity Center, we match the funding up to a total amount of 768K euro for the project for 2 years.

We identify project goals for

- Enabling a scrutinized (Catalogue of Life) and provisional taxonomic catalogue (GBIF Backbone Taxonomy)
- Separating fact (scientific name) from opinion (taxonomic concept)
- Providing (infrastructural) support to taxonomic and nomenclature content authorities
- Ensuring a sustainable, robust, and more dynamic IT infrastructure (hosted by GBIFS)

For the project we have set several specific objectives up to the end of April 2019.

Establishing a clearinghouse for nomenclature and taxonomy to reconcile sources

Establishing a partnership, governance, and roadmap for the infrastructure

We will start with the development of a clearinghouse infrastructure for names and taxonomy. Simultaneously we will strengthen the consortium and associated governance to ensure proper international embedding, sustainability and enhancement of these efforts after the project's end.

Link to CoL+ project proposal

Project structure

Building a common infrastructure for names and taxonomy through international collaboration.

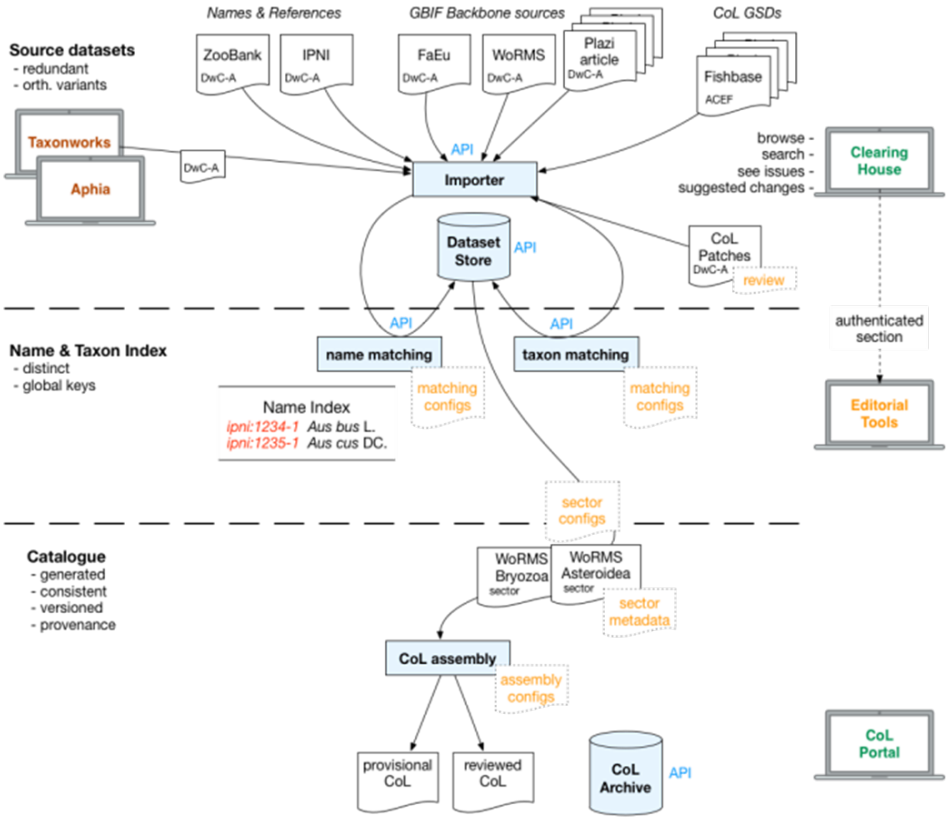
A steering committee is formed by the initial project partners Catalogue of Life (Species 2000 & ITIS), Naturalis Biodiversity Center, and the Global Biodiversity Information Facility Secretariat. The steering committee is complemented with representatives from the Encyclopedia of Life, Biodiversity Heritage Library, and the Barcode of Life data systems. Membership of the steering committee is open for those initiatives that substantially contribute to the development of the infrastructure and/or agree to make use of the backbone services of the clearinghouse for names and taxonomy.

A project team lead by Naturalis Biodiversity Center is established. It consists of developers from the Global Biodiversity Information Facility, Naturalis Biodiversity Center, Species File group – Illinois Natural History Survey, and representatives of the Catalogue of Life Editorial Board, Catalogue of Life information systems and taxonomic groups, Species 2000 secretariat as well as additional representatives from the Catalogue of Life and GBIF communities. It is expected that a front-end developer will be added to the project team in 2018.

Infrastructure development

- *Towards a clearinghouse for names and taxonomy*
- *Development schema*
- *(Provisional) Catalogue of Life*
- *Milestones*
- *Progress dataset store & importer milestones*

Towards a clearinghouse for names and taxonomy



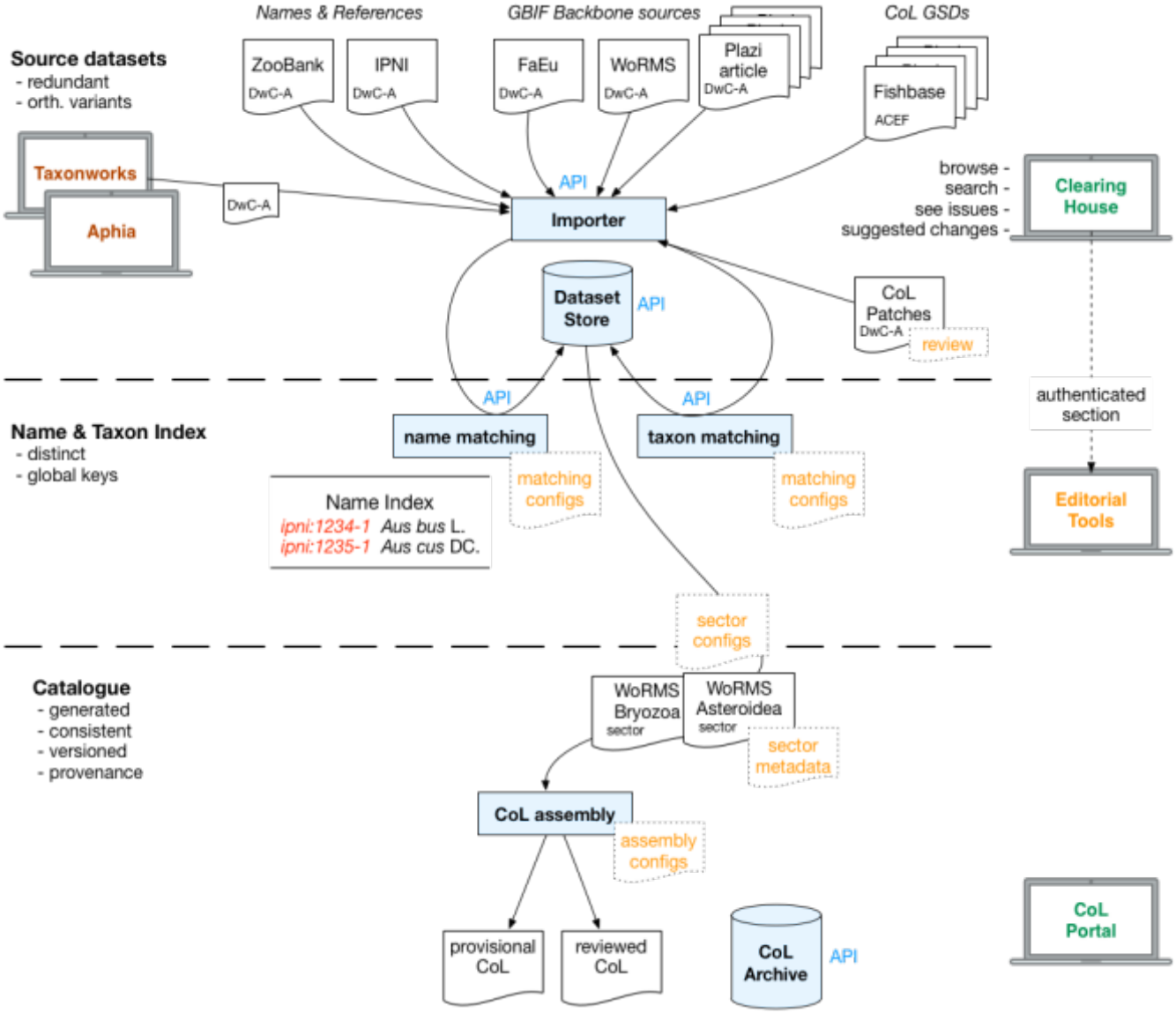
The specific objective is establishing a clearinghouse that covers scientific names across all life, allowing a reconciliation of nomenclatural and taxonomic data sources. It provides a single taxonomic view grounded in the consensus classification of the Catalogue of Life along with provisional taxonomic sources, shows differences between sources, and provides an avenue for feedback to content authorities and allowing a broader community to contribute.

The clearinghouse infrastructure both includes the necessary infrastructure to support the efficient functioning of the Catalogue of life, but also serves as the infrastructure to offer services on nomenclatural and taxonomic backbone to GBIF.

In the following set of slides the development roadmap of the IT infrastructure is further clarified.

Development schema

The development schema on the right shows the flow of data from the sources to the datastore, the names and taxon concept indexes and the assembly of the Catalogue of Life. On the outermost right hand side the planned front-end interfaces are shown. On the top the various data sources from where data originates. Editorial input is shown in orange colour.



(Provisional) Catalogue of Life

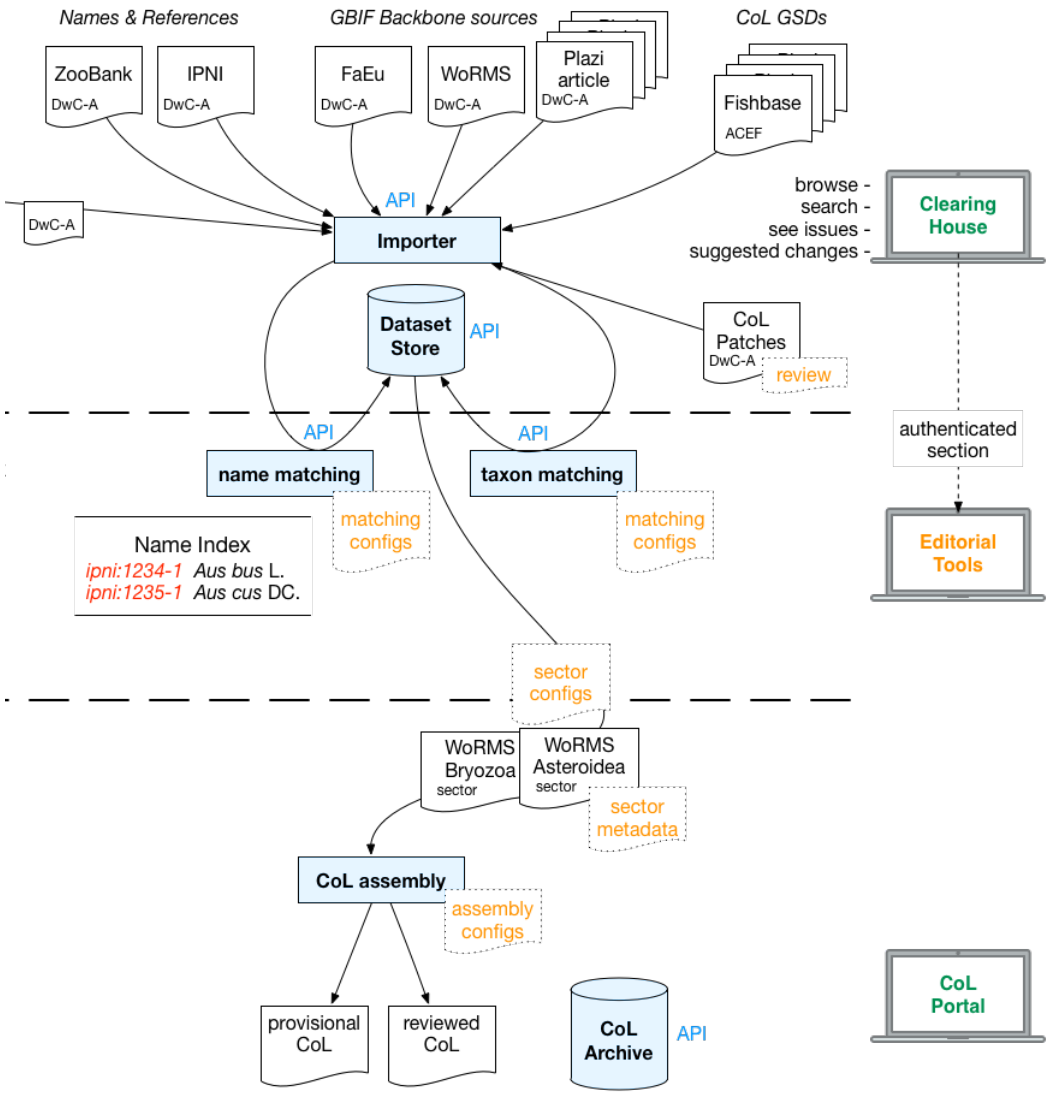
Innovating the Catalogue of Life systems through the CoL+ project should result into a new operational infrastructure for the current Catalogue of Life. This infrastructure should as well cover the data requirements for a GBIF taxonomic backbone service through a provisional Catalogue of Life.

In the provisional Catalogue of Life additional data about species originating from various sources is gathered that complement and extent the current content of the Catalogue of Life. The gathering of additional data may be targeted at increasing the coverage of scientific names, mostly coming from trusted nomenclatural sources. It may also contain links to references (e.g. from BHL), and/or adding new species names published in online journals. The provisional Catalogue of Life could also contain taxonomic data sources that do not have a global coverage with the purpose to fill taxonomic gaps or with the aim to add vernacular names.

Part of the additional data gathered will be offered to the taxonomists who provide species data for specific taxonomic sectors of the Catalogue of Life. This could for example be ithe form of additional synonyms (e.g. homotypic synonyms), or preferred nomenclatural spelling.

The provisional Catalogue of Life should in the future also respond to the specific requirements for taxonomic and nomenclatural backbone services for other biodiversity information initiatives (such as BoLD, EOL and BHL), including the European initiatives of DiSSCo and LifeWatch.

Milestones



A development roadmap for the infrastructure is drafted. It contains ten broadly formulated milestones (ordered in sequential steps of development):

1. **Dataset Store** √
   Caches entire datasets in PostgreSQL
2. **Importer** √
   Imports datasets into the dataset store
3. **Names Index**
   Matches names with index of unique names
4. **Editorial Tools**
   Manages the Catalogue of Life assembly
5. **Catalogue of Life Assembly**
   Assembles the scrutinized & provisional CoL
6. **Archive Store**
   Archives CoL versions & data provenance
7. **Taxon Index**
   Matches with index of unique taxon concepts
8. **User Comments**
   Enables user comments & authentication
9. **Review Queues**
   Shows data for review to publishers
10. **Final Deployment**
    Enables to switch off former infrastructure and releases the operation of the new infrastructure

√ indicates the milestone is completed

Progress

Dataset store & importer milestones

A milestone that lays the foundation for the entire infrastructure.

The first milestone of the CoL+ project delivers a dataset store and datasets importer for the clearinghouse infrastructure. This includes a back-end API installation and documentation: http://api.col.plus

A demonstration environment is currently hosted at the GBIF Secretariat. The datastore includes data sources coming from the global species databases (CoL GSDs) in ACEF format. It also includes data sources coming from the GBIF Backbone Taxonomy in DWC-A format.

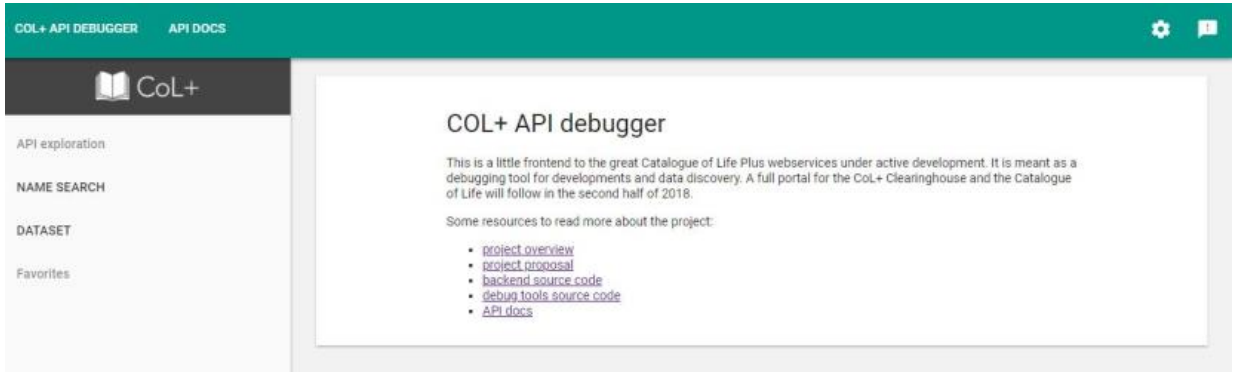
An API debugger has also been developed: http://tools.col.plus/ This tool allows to search for species names and datasets.

The datastore milestone allows for studying the proposed datamodel of the clearinghouse. Feedback could be provided on the CoL+ github repo or by email to the project team.

The current work around the datastore is work in progress and frequent changes and updates are expected.

Please also note this milestone only contains datasets as content. Other milestones of the development roadmap are needed to actually built the Catalogue of Life.

The development of the names index is already fully in progress. The project team and steering committee are busy with arranging personnel for the front end development. A more accurate timeline for the development roadmap can be provided once this process is completed.



Partnership & engagement

Partnership, engagement & governance

Progress partnership & engagement

Partnership, engagement & governance

A governance model should ensure the responsibility for improving content remains with the respected nomenclature and taxonomy content authorities, including editorial boards (e.g. Catalogue of Life)

The specific objective of the CoL+ project is establishing a partnership and governance for the clearinghouse and its associated components that enables continuing commitment after the project's end.

One of the main priorities for the CoL+ project is to built the consortium of partners that are willing and able to contribute to the infrastructure development and to use the infrastructure once it is operational. Knowing and acknowledging key requirements, concerns, and benefits of partners is vital.

Engagement with nomenclature and taxonomy content authorities is another main priority. The resources of the CoL+ project will not allow for engaging with all relevant content authorities at once. The strategy is to carry out several pilots with content authorities to understand and identify mutual requirements, concerns, and benefits of the clearinghouse infrastructure.

The clearinghouse should support the work of editorial boards (e.g. Catalogue of Life). including setting-up editorial tooling for the provisional catalogue (e.g. in the case of taxonomic gaps).

The clearinghouse should be developed through documented user requirements for services.

To faciltate the work done by taxonomists the project will look into making direct connections between the clearinghouse infrastructure and existing taxonomic editing tools.

The CoL+ project will put measures in place to guide the transition to the new clearinghouse infrastructure once it is operational. This includes a governance model that should deal with how the infrastructure is maintained, hosted, and developed further. At the end of the project a roadmap should be available that guides and clarifies how the clearinghouse infrastructure fits into the wider landscape (for example in the context of the Global Biodiversity Informatics Outlook).

Progress

Partnership & engagement

Membership of the steering committee is open for those initiatives that substantially contribute to the development of the infrastructure and/or agree to make use of the backbone services of the clearinghouse for names and taxonomy.

The initial steering committee has been expanded with representatives from the Barcode of Life data systems, Biodiversity Heritage Library, Encyclopedia of Life, and ITIS.

From July 2017 onwards, various meetings took place with stakeholders. The CoL + project was presented several times in the GBIF community (including at the 14th global Nodes meeting and the 24th GBIF Governing Board) and at the 42nd and 43rd meeting of the Consortium of European Taxonomic Facilities.

With the following parties, agreements have been made about future cooperation: Kew / International Plant Names Index & Index Fungorum for connection with nomenclature information, Species File group - Illinois connection with taxonomic editing tool TaxonWorks, LifeWatch / Worms about taxonomic editing tool Aphia, World Flora Online about taxonomic plant information and editing tools. These partnerships will be further elaborated in 2018.

The CoL+ project participated in the strategic alignment group of the Distributed Systems of Scientific Collections (DiSSCo). This is a European research infrastructure initiative that seeks placement on the ESFRI roadmap.

The CoL+ project was involved in the creation of the Synthesis + proposal that is submitted to the European Commission in March 2018. Synthesis + is a direct contribution to the creation of the DiSSCo initiative.

In March 2018 the CoL+ project and development road map were discussed with the Catalogue of Life governance. The plans were fully endorsed by the governance.

Communication

Modes of communication

Modes of communication

For more information:

https://github.com/Sp2000/colplus

In the future:
https://col.plus

olaf.banki@naturalis.nl

This slide doc will be periodically updated, especially to communicate milestones and deliverables of the CoL+ project.

Periodically, the CoL+ project will organise webinars to communicate on the project plans and progress.

The project team communicates through scrum calls two times a week. These scrum calls are focused on supporting daily activity.

The CoL+ project is also represented at the monthly Species 2000 secretariat meetings.

The main information sources are:

- The CoL+ project proposal that has a DOI so it can be referred to: https://doi.org/10.5281/zenodo.1194825
- The CoL+ github repository that can be used to submit issues: https://github.com/Sp2000/colplus
- Back-end API documentation: http://api.col.plus
- An API debugging tool: http://tools.col.plus