



Biodiversity Informatics & KEYSTONE

Naturalis
Biodiversity
Center

Peter Schalk
Director Public & Markets Naturalis
Chair Governing Board GBIF
Executive Secretary Catalogue of Life

- **The domain of taxonomy and biodiversity informatics**
- **Community achievements (GBIF, COL, CETAF)**
- **Strategic plans**
- **The GBIF-COL-EOL-BHL-BOLD-ELODINS Summit**
- **Challenges for KEYSTONE**



My domain is about (y)our world

- Natural history collections are vouchers of observations in the past
- Networks of citizen scientist's who provide observation data
- Modelling to predict changes in the biosphere

Taxonomy and biodiversity science

- Robert May: "Without taxonomy to shape the bricks and systematics to put them together the house of biological sciences would be a meaningless jungle ..."
- Valid names are our common language. In science, in society.
- Problem: we keep changing definitions of the species concept.

Developments in natural history museums



33 members represent 57 institutions holding over half the world's biological collections

Big 6 CEO meetings



Joining forces (2016-2025):

- Single professional ICT environment for all collections
- Coordinated research agenda.
- Industrial approach to collection digitization (PPP)
- Amplifying international services (COL, GBIF, BOLD)



Employ 110 FTE IT staff to serve 4.000 FTE personnel = 2.75% (Market: 4-5%). *Undermanned.*

Cloud based CPU, storage and systems supported by a professional IT versus classic in-house technology, managed by “hard work”. *Room for professionalisation*

Redundancy in the community: similar IT systems for collection data management, species info systems, taxonomic systems. Limited collaboration and/or re-use of resources. *Room for more efficiency & cost reduction*

Catalogue of Life: tool for connecting data resources



Frank A. Bisby 1945 - 2011

1995. “Lets bring together the knowledge of the global taxonomic community in a single shared authoritative system on species concepts that can be used by all”. The resulting product is the **Catalogue of Life**.

2015

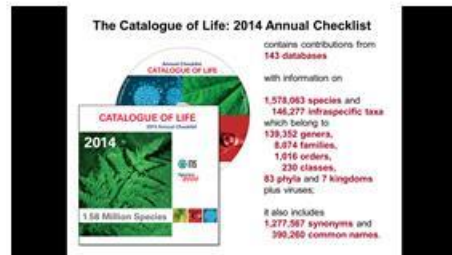
1.6 million species (82% of total)
160 databases from all over the world
Backbone to GBIF and BOLD
Millions of users on a monthly basis

An authoritative resource ...

- an electronic synonymic species checklist,
- a tightly integrated taxonomic hierarchy,
- intended for all 1.9 M extant known species.

.... constructed through international collaboration

- checklist and hierarchy constructed from sectors from 160 networked global species databases around the world
- and integrated using an international review panel of experts



COL Strategic Plan 2016-2020



Completing the CoL (quantity) 95%

Addressing gap areas in the species index. Define precise number of species (bottom up)

Extending the synonymic content towards completeness (federated approach with partners)

Amplifying quality

Addressing weak spots in CoL GSD coverage. Include indicator of completeness.

Mechanism for inviting user feedback on quality / errors (together with partners)

Improving infrastructure (through shared development)

Updating 4D4Life/i4Life CoL service e-infrastructure (security issues, performance)

Improve CoL construction route (QAW approach for part automation of updates)

Improving user services (fuzzy search, x-mapping)

Strengthening the CoL network of providers

Offering support GSD providers: hosting solutions (proto GSD) & migration services. Backups.

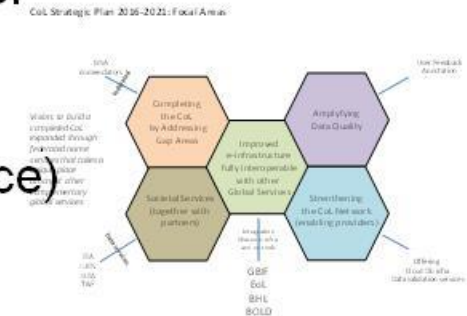
Providing feedback (statistics) on GSD and record level. Access to (literature) data. Funding!

Increase (societal) services.

Extend the number of service contracts and expert validation services

Country proto-checklist reports (with GBIF; hackathon output)

Address identifiers issue on name and taxon level (with partners)



GBIF: open access portal to all biodiversity data

The Global Biodiversity Information Facility (GBIF) is an international open data infrastructure, funded by governments.

Collections (past)



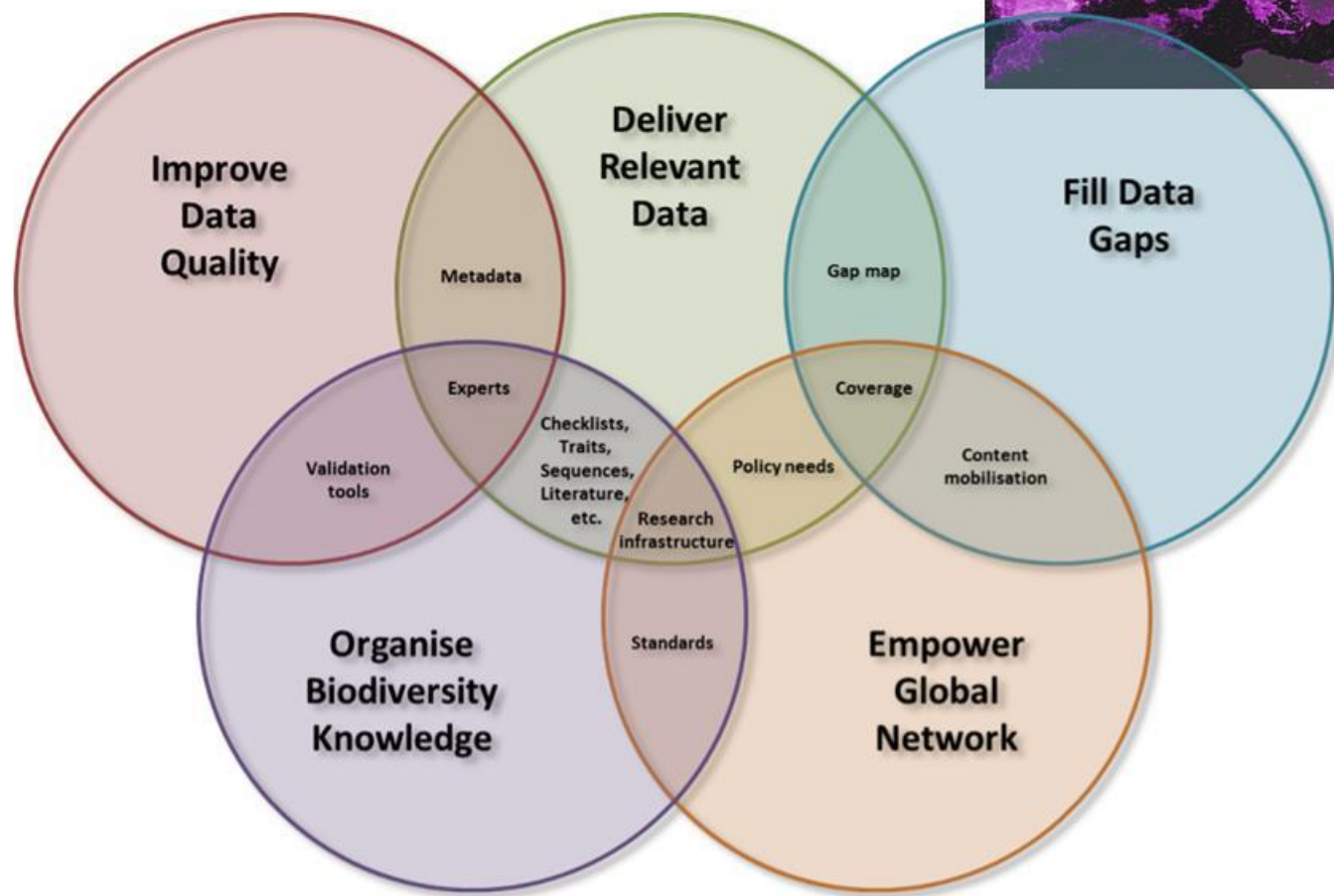
Observations (present)

GBIF serves millions of users on a monthly basis. It enables big data science in biodiversity never possible before.

GBIF STRATEGIC PLAN 2017-2021

Five draft priorities

- **Deliver Relevant Data**
 - Support supranational research and policy needs
 - Enhance relevance to taxonomy and climate change research
- **Fill Data Gaps**
 - Identify and prioritise spatial, temporal and taxonomic gaps
 - Engage countries and institutions with data to address gaps
- **Improve Data Quality**
 - Enhance automated and human curation of data
 - Develop clear quality indicators
- **Organise Biodiversity Knowledge**
 - Link data to other online biodiversity data sets
 - Classifications, genes, traits, literature, etc.
 - Improve stability and persistence of data
- **Empower Global Network**
 - Provide tools and best practice guidance
 - Promote and support engagement of countries and institutions everywhere within GBIF



The COL-GBIF-EOL-BHL-BOLD summit statement

Aligning strategic development goals to create a ‘super resource’

The global data aggregators COL-GBIF-EOL-BHL-BOLD **together meet essential needs in the biodiversity community and society**. In 2015 they agreed to align strategic developments and enter into a **mutually beneficial and durable relationship**.

First step: working towards a single shared taxonomic index

- Achieving shared all encompassing taxonomic backbone that serves BOLD, COL, GBIF, EOL, BHL, LifeWatch communities and other users
- Combine developing forces and funding resources to realize a jointly owned and maintained taxonomic index to organize biodiversity data

*the
Life
partnership*



ELODINS

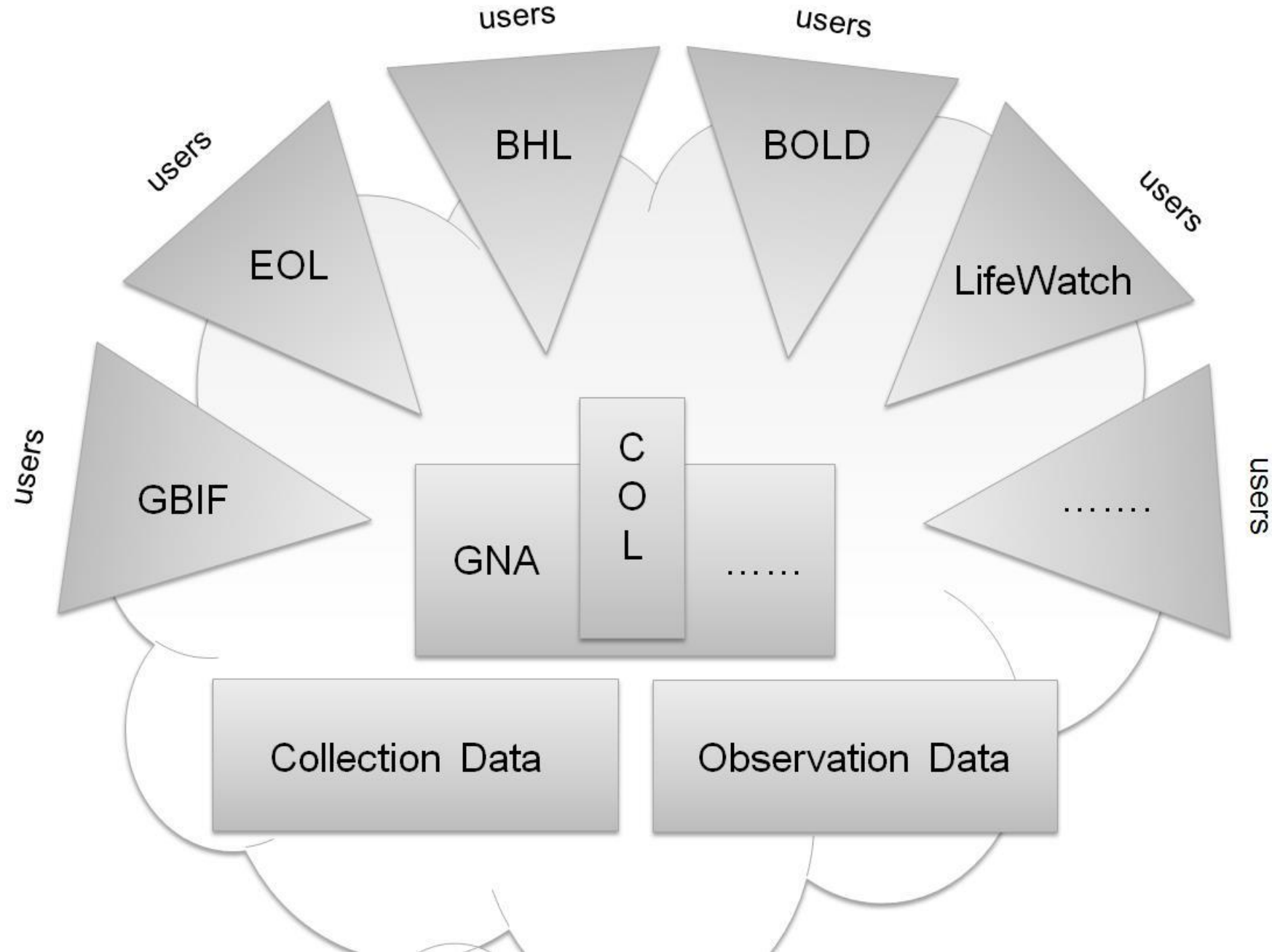
Catalogue of Life



International
BARCODE
OF LIFE



Towards a centrally shared and owned name index





Opportunities for KEYSTONE

Wouter Addink

Biodiversity data is often searched for by species name or common name, or group of species (taxon). Species can be subdivided in closely related lower taxa, like a subspecies or a variety. A species name is in latin and has 2 parts, example: *Ursus maritimus* (common name: Polar Bear in English). Animal names have different nomenclature rules from plants.

A scientific name represents an opinion on a taxon concept at a certain point in time, it is not a unique identifier for the taxon concept. A taxon concept is delimited by a description and described by specimen (individual objects collected in the field and stored in a museum). Specimen have identifications (a scientific name attached to the specimen at a certain point in time)



Challenges for KEYSTONE

Wouter Addink

1. We need to be able to extract scientific names (in latin) from text (sometimes latin, usually English or another language)
2. Similar for common names (can be different in different local areas or countries)
3. Since scientific names are not unique, we need to resolve them to the currently accepted name (or ideally we want to resolve to a unique taxon concept identifier)
4. We want to resolve a common name to a scientific name
5. We need a correct ranking of results: searching for (genus) *Accipiter* should first return documents with exact match *Accipiter*, not give *Accipiter accipiter* a higher ranking (because 'accipiter' occurs twice in that document)
6. In results there should be no difference between male and female endings *Pterostyrax hispida* and *Pterostyrax hispidus*
7. Descriptions contain characters of a species with a state or value (leaf:round). descriptions can be in different language. Search for 'round leaf' should return species or specimen with a round leaf.



Opportunities for KEYSTONE

Wouter Addink

8. Specimen have multimedia recordings, like an image, which could be analyzed to extract keywords for comparison with a written description.
9. Specimen should be searchable by distribution. Often a locality is recorded as verbatim text, like 'Coimbra'. There may be multiple locations with that name or the name may have changed over time.
10. We need to include temporal information in the search: a specimen found centuries ago on a location may not occur there at present.
higher level:
- 11 We want to predict if, given a certain habitat, a species is introduced, that will cause competition for native species and how likely it is that it will be introduced there by humans or other vectors.
- 12 we want to predict that given a temperature rise of 1 degree Celsius a species distribution will change or it will become extinct. What areas could it invade?
- 13 We want to find likely hosts for an insect given the distribution for that insect and other data.



The Beginning

... for the Keystone Community?

peter.schalk@naturalis.nl