



# BOLD v5

Rutger Vos, Naturalis  
iBOL Europe community meeting  
27 January 2025

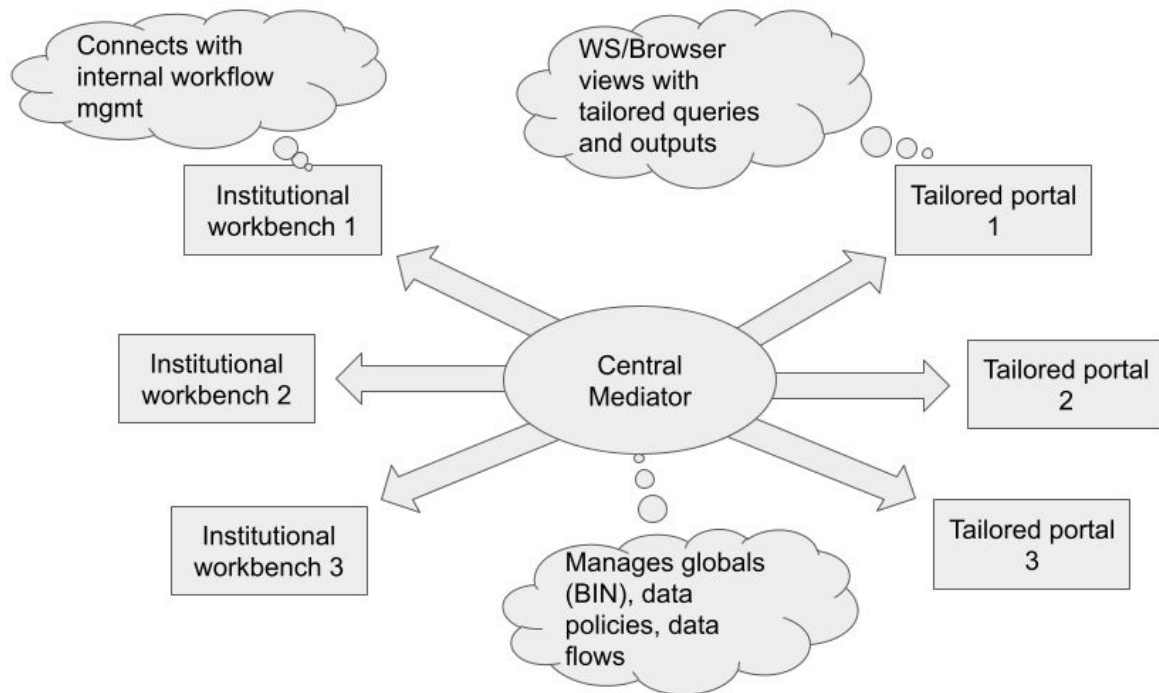
# What is BOLD?

- BOLD is a catch all term for a number of distinct functionalities:
  - Barcode (meta)data submission, validation and publication (the ‘workbench’)
  - Barcode (meta)data access via queries, APIs and data packages (the ‘data portal’)
  - Reverse taxonomy of unidentified sequences (‘ID services’)
  - Knowledge bases on DNA barcoding
- Up until v.4 this was implemented as a centralised, monolithic, closed-source web application that has become hard to maintain, hard to scale
- Also, the centralisation no longer meets all needs of an increasingly global community with disparate needs (trivial example: a need for data localisation)

# The BOLD v5 ecosystem

The vision for the future is for:

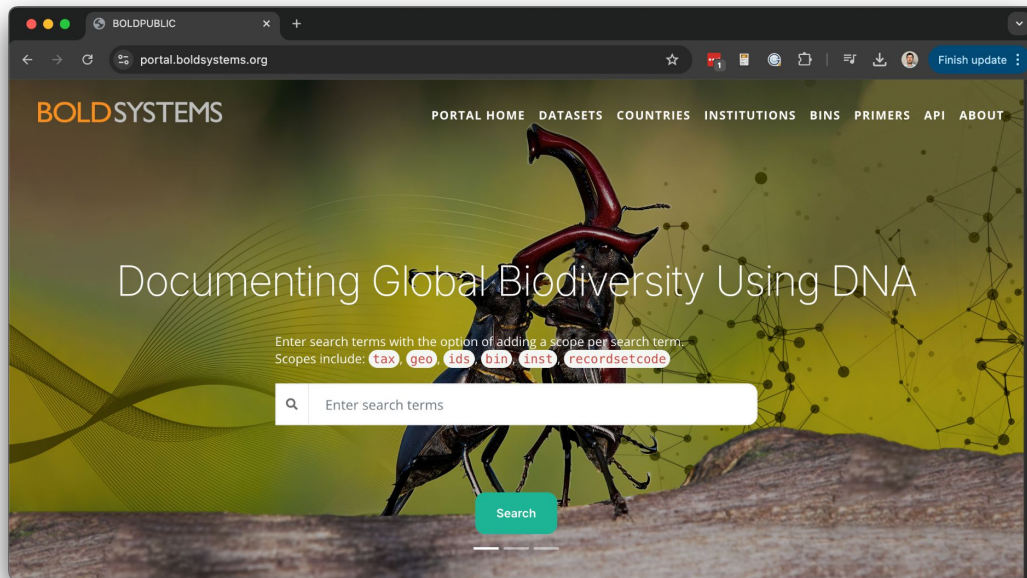
- Decomposition of portal and workbench
- Incremental open-sourcing of components
- Enabling of multiple instances of workbenches and portal
- Separate instances having custom extensions
- Overall referential integrity being maintained by a central mediator



# BOLD v5 data portal browser access

The new data portal component from a non-technical user perspective:

- Updated graphical interface with improved user experience
- More precise searching with more search predicates
- Explicit access to reference data sets with DOIs
- Barcode outputs aggregated at country and institutional level

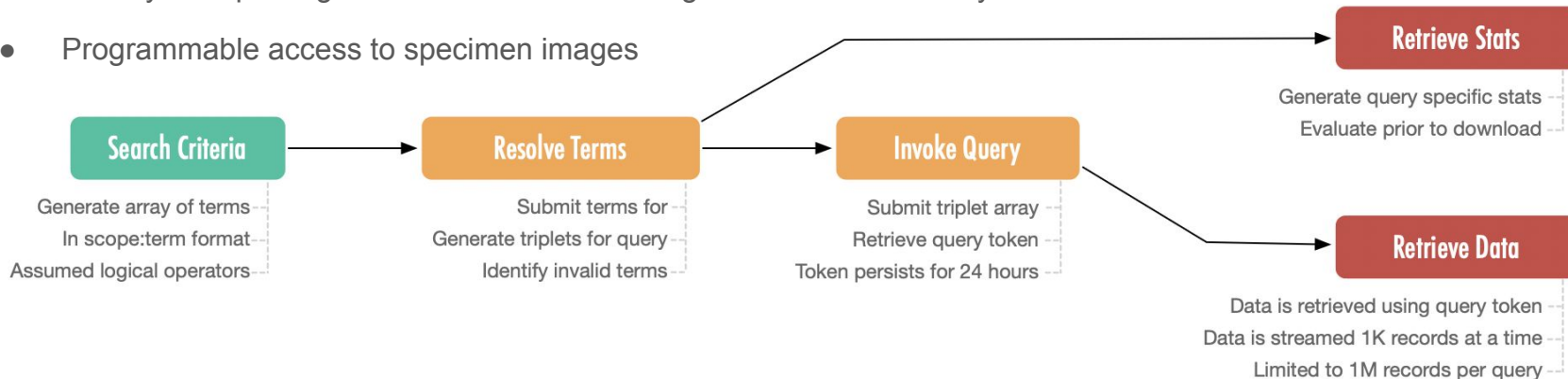


<https://portal.boldsystems.org/>

# BOLD v5 data portal programmable access

The new data portal component from a technical/power user perspective:

- More expressive and performant web service API
- An R-based toolkit for programmatic access to the API
- Weekly data packages with rich metadata using controlled vocabulary
- Programmable access to specimen images





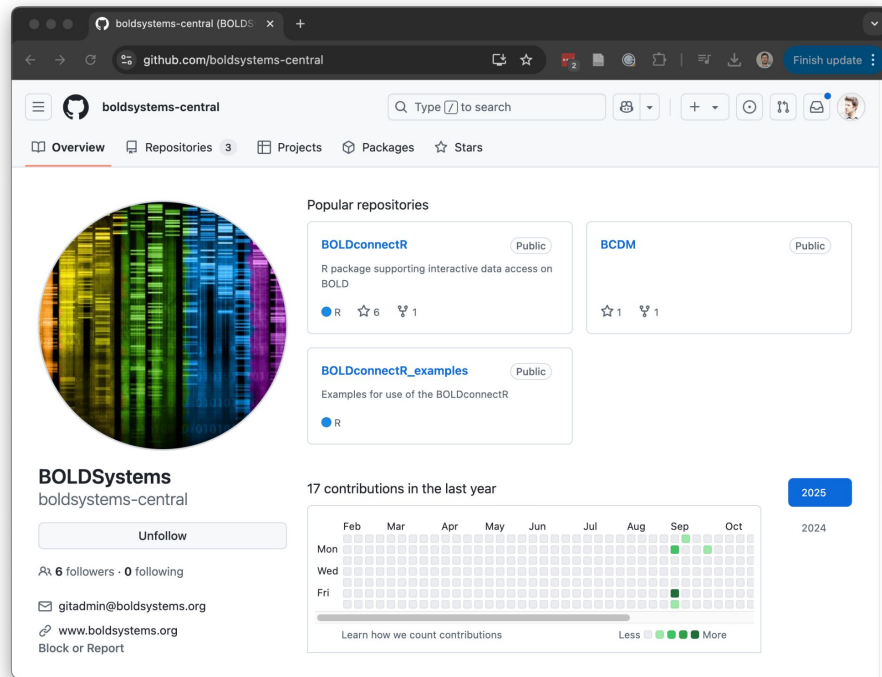
# The Barcode Core Data Model (BCDM)

- Barcode sequence data is only useful with sufficient metadata that needs to be predictable enough so that algorithmic operations (such as automated curation) can be performed on it.
- To meet this need, a controlled vocabulary is being developed, the Barcode Core Data Model (BCDM)
- The BCDM currently distinguishes 77 terms whose data type (e.g. string, date, number) and meaning (semantics) are defined clearly. Analogous to DarwinCore, for example. BCDM deals with the following concepts:
  - Taxonomic lineages at fixed ranks
  - Specimen identifiers and collecting institutes
  - Taxonomic identification of specimens
  - Collection events and localities
  - Sequencing process information
- BCDM data can be presented in multiple formats (tab-separated tables, JSON objects)
- Mappings between BCDM and other vocabularies have been developed, e.g.:
  - DarwinCore
  - ENA and UNITE records
  - MiXS minimum information standard

# Open source components

To promote power user access to BOLD data and the development of extensions to core functionality, BOLD code components are gradually becoming open source:

- BCDM reference version
- R library for API access
- (very near future) data portal code
- (more distant future) v5 workbench code



The screenshot shows the GitHub profile page for 'boldsystems-central'. The profile name is 'BOLDSYSTEMS' and the username is 'boldsystems-central'. The profile has 6 followers and 0 following. The bio includes the email 'gitadmin@boldsystems.org' and the website 'www.boldsystems.org'. The 'Popular repositories' section lists three repositories: 'BOLDconnectR' (R package supporting interactive data access on BOLD, 6 stars, 1 fork), 'BCDM' (BCDM reference version, 1 star, 1 fork), and 'BOLDconnectR\_examples' (Examples for use of the BOLDconnectR, 0 stars, 0 forks). The '17 contributions in the last year' section shows a calendar grid for 2024 with contributions in September and October.

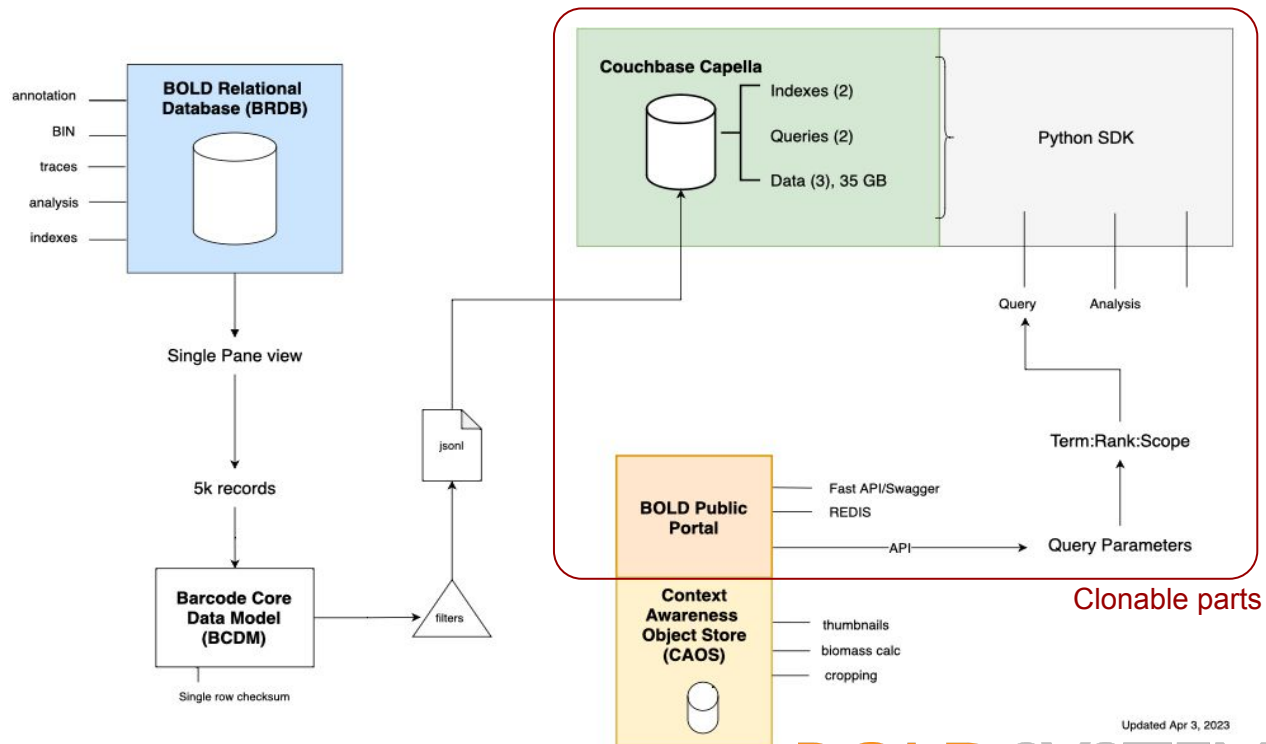
# BOLD v5 data portal architecture

Data portals host BOLD data transformed to comply with BCDM in a document store

The document store is accessed through web services via a Python software development kit

The portal presents the query results from various facets in a graphical user interface

The results are further enhanced with images and raw data from an object store





# Readiness levels and impact

- A first instance of the data portal has recently been unveiled by CBG and made available at [portal.boldsystems.org](https://portal.boldsystems.org)
- A beta version of European mirror has been circulated with the BGE consortium at [boldsystems.eu](https://boldsystems.eu)
- Upgrading the European mirror to the release version is slated for Q1/2025
- Central mediation among multiple workbenches and portals is active WIP
- The BCDM standard and CAOS access have simplified automated record curation



# Thank you for your attention

## Acknowledgements:

- BOLD v5 architecting and development is led by the Centre for Biodiversity Genomics, Guelph, Canada
- International collaboration on BOLD v5 is funded by Biodiversity Genomics Europe

**BOLD**SYSTEMS