

Digital Infrastructure for Biodiversity Informatics

Oscar Kovacs¹, Marta Bianchi², Noah Horvath³

¹ Postdoctoral Researcher, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: oscar.kovacs505@yahoo.com | ORCID: 0367-0195-9314-9696

² Assistant Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: marta.bianchi634@yahoo.com | ORCID: 8363-1051-6021-1481

³ Assistant Professor, School of Data Science, Central European Tech University, Vienna, Austria. Email: noah.horvath980@yahoo.com | ORCID: 2296-4011-8377-3169

ABSTRACT

Biodiversity informatics -- the computational infrastructure for managing, analysing, and sharing biological diversity data -- underpins global efforts to document, monitor, and conserve the approximately 8.7 million species on Earth amid unprecedented anthropogenic pressures. The Global Biodiversity Information Facility (GBIF) aggregates over 2.6 billion species occurrence records from 2,000+ institutions, while parallel infrastructures (iNaturalist, eBird, Catalogue of Life, GenBank) store complementary biodiversity data. However, persistent interoperability gaps, data quality heterogeneity, taxonomic inconsistencies across platforms, and limited integration with genetic/trait databases constrain biodiversity data utility for conservation decision-making. This study evaluated six biodiversity informatics infrastructure approaches -- institution-specific databases (siloes), domain-specific aggregators (GBIF, OBIS, eBird), the Darwin Core standard (shared data format), semantic web approaches (OpenBiodiv, BiGe-OBS linked data), AI-powered species identification platforms (PI@ntNet, Merlin), and a proposed Integrated Biodiversity Data Commons Platform (IBDCP) combining standardised FAIR data schemas, automated taxonomic reconciliation via NLP, computer vision-assisted occurrence validation, integration of occurrence + genomic (BOLD, GenBank) + trait (TRY, Pantheria) + ecological (ENA, EDGAR) data, and federated query architecture for decentralised data access -- across four biodiversity informatics use cases: species distribution modelling (2,400 species), conservation priority ranking (IUCN Red List workflow), invasive species early warning (global monitoring), and phylogenetic biogeography (tree of life x geographic distribution). IBDCP achieved the highest composite score (0.884) across six evaluation dimensions: taxonomic consistency (92.4% reconciliation accuracy), data quality (geographic and temporal outlier detection F1 0.912), integration breadth (combines 8 data types vs GBIF's 1-2), query efficiency (federated queries 4.2x faster than manual multi-database lookups), conservation utility (Red List assessment time reduced 68%), and open-access compliance (FAIR score 0.94). These results establish integrated data commons as the next generation of biodiversity informatics infrastructure.

Keywords: Biodiversity informatics; GBIF; Species distribution modelling; FAIR data; Darwin Core; Taxonomic reconciliation; Conservation informatics; Linked data; Federated databases

Citation: Kovacs et al. [2024]. Digital Infrastructure for Biodiversity Informatics. DOI: <https://doi.org/10.5281/zenodo.19549401>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jun 08 Accepted: 2024 Aug 08 Published: 2024 Sep 15

Research Article: Research Article

1. Introduction

Biodiversity loss is one of the defining environmental crises of our era: current extinction rates are 100-1,000 times background rates, with an estimated one million species threatened with extinction (IPBES, 2019). Addressing this crisis requires comprehensive knowledge of what species exist, where they occur, how abundant they are, and how they are changing over time -- information captured in biodiversity data from diverse sources: museum specimens collected over centuries, contemporary field observations by researchers and citizen scientists, genetic barcodes for species identification, trait measurements of morphology and physiology, and environmental sensor data tracking ecosystem changes. The computational challenge is to integrate these heterogeneous data sources into accessible infrastructure that supports scientific research, policy-making, and conservation action (Hardisty et al., 2013).

1.1 The Current Biodiversity Data Landscape

Global biodiversity data are distributed across multiple large-scale infrastructures. The Global Biodiversity Information Facility (GBIF) aggregates 2.6 billion species occurrence records from 2,000+ data providers, providing the most comprehensive species distribution database but focused primarily on occurrence (presence/absence at location and time) without integrated trait or genetic data (Heberling et al., 2021). OBIS (Ocean Biodiversity Information System) covers marine species. iNaturalist and eBird aggregate 150+ million citizen science observations with AI-assisted species identification. The Catalogue of Life provides a unified taxonomic backbone. GenBank and BOLD host DNA sequences and barcodes. TRY Plant Trait Database, Pantheria (mammals), and AmphiBIO (amphibians) aggregate species trait measurements. Despite covering similar species, these databases have limited interoperability: taxonomic names differ across platforms, occurrence records are often not linked to genetic samples from the same individuals, and trait data use inconsistent measurement protocols.

1.2 Informatics Challenges

Three major informatics challenges constrain biodiversity data utility. (i) Taxonomic reconciliation: the same species may be recorded under different names across databases (e.g., synonyms, subspecies, regional names) and taxonomic opinions differ -- what one database calls *Anas platyrhynchos* (mallard) may appear in

another as multiple subspecies. Reconciling names across 2.6 billion GBIF records requires automated NLP approaches (Rees and Cranston, 2017). (ii) Data quality heterogeneity: occurrence records vary enormously in reliability, from expert-validated museum specimens to opportunistic photos with uncertain geolocation. Automated quality assessment (coordinate validation, temporal consistency, taxonomic plausibility) is essential for downstream use. (iii) Cross-database integration: connecting an occurrence record to the genetic sequence from the same specimen, the trait measurements of the species, and the ecological context (climate, land use) requires pipeline tools that span platforms with different APIs, schemas, and semantics.

1.3 Study Objectives

This study aimed to: (i) systematically evaluate six biodiversity informatics infrastructure approaches on four common use cases; (ii) quantify taxonomic consistency, data quality, integration breadth, and query efficiency; (iii) develop IBDCP, an integrated biodiversity data commons platform; (iv) demonstrate conservation utility through Red List assessment acceleration; and (v) provide recommendations for next-generation biodiversity informatics infrastructure.

2. Literature Review

Biodiversity informatics has evolved from paper-based taxonomic catalogues through institutional databases to global aggregators and increasingly toward integrated data commons.

2.1 Aggregator Infrastructure

GBIF (established 2001) represents the most successful biodiversity data aggregator, collaboratively developed by 63 member countries and 43 international organisations. GBIF uses the Darwin Core standard (Wieczorek et al., 2012) as its interoperability format, enabling standardised data exchange across institutions. GBIF data have supported 7,500+ peer-reviewed publications, demonstrating substantial scientific impact. However, GBIF focuses primarily on occurrence records, with limited integration of genetic, trait, or ecological data. OBIS (Ocean Biodiversity Information System, established 2000) provides parallel infrastructure for marine biodiversity with 100+ million records. iNaturalist (established 2008) combines citizen science observation with AI-assisted species identification, contributing approximately 40% of new occurrences to GBIF annually (Unger et al., 2021). eBird captures 800+

million bird observations with professional-grade data quality processes.

2.2 Standards and Semantic Web

Darwin Core (Wieczorek et al., 2012) provides vocabulary for biodiversity occurrences: scientific name, decimal latitude/longitude, event date, basis of record, etc. Extensions cover sampling events, species traits, molecular data, and audiovisual media. Despite standardisation, data quality varies: Darwin Core allows flexibility that providers implement inconsistently. Semantic web approaches (OpenBiodiv, Biodiversity Knowledge Graph) encode biodiversity data as RDF triples with formal ontologies (Biological Collection Ontology, Darwin Core RDF), enabling reasoning over complex queries (Penev et al., 2019). TDWG (Biodiversity Information Standards) maintains vocabulary and ontology standards but adoption is uneven across institutions. The Biodiversity Data Journal represents an innovative semantic publishing platform where papers are natively machine-readable with structured species data embedded in Darwin Core XML.

2.3 AI in Biodiversity Informatics

Machine learning has transformed several biodiversity informatics subareas. Species identification from images: Pl@ntNet (Bonnet et al., 2018) achieves 85-95% top-1 accuracy on plant species identification from smartphone photos, covering 45,000+ species; iNaturalist's computer vision model processes 300,000+ observations per day with 95% precision on confident predictions. Species identification from audio: Merlin (Cornell Lab) identifies birds by song with state-of-the-art CNN models trained on 1 million+ labelled recordings. Automated occurrence quality assessment: machine learning models flag suspicious records (impossible coordinates, out-of-range dates, geographically implausible species) with F1 scores approaching 0.90 (Zizka et al., 2020). These AI capabilities enable processing biodiversity data at scales impossible for manual curation.

Table 1. Biodiversity informatics infrastructure approaches evaluated: scope, standards adoption, and data integration

Approach	Scope	Data Types	Standard	AI Integration	Reference
Institution siloed	Single organisation	1-2 types	Variable	None	Baseline (legacy)

Approach	Scope	Data Types	Standard	AI Integration	Reference
GBIF/OBIS	Global occurrence	Occurrence	Darwin Core	Limited	Heberling et al., 2021
Darwin Core (schema)	Standard only	All supported	Darwin Core	None (schema)	Wieczorek et al., 2012
Semantic web	Linked data	Occurrence + some	DwC RDF + onto.	Reasoning	Penev et al., 2019
AI platforms	Species ID	Occurrence + image/audio	Platform-specific	Core feature	Bonnet et al., 2018
IBDCP (proposed)	Integrated commons	8 data types	DwC + extensions	Comprehensive	This study

IBDCP: Integrated Biodiversity Data Commons Platform. Darwin Core (DwC): the primary biodiversity interoperability standard. Eight data types in IBDCP: occurrence, genomic (BOLD/GenBank), trait (TRY/Pantheria), phylogenetic (OpenTree), image/audio media, environmental context (climate/land use), collection metadata, and conservation status (IUCN Red List).

3. Materials and Methods

3.1 IBDCP Architecture

IBDCP comprises five layers. Layer 1 (FAIR data schemas): extended Darwin Core with new term classes for genetic linkage (BOLD BIN ID, GenBank accession), trait measurement (TRY TraitID with measurement protocol metadata), phylogenetic position (OpenTree node ID), conservation status (IUCN Red List category), and data quality flags (coordinate uncertainty, taxonomic verification status). Layer 2 (Automated taxonomic reconciliation): NLP pipeline processing scientific names through GBIF Backbone + Catalogue of Life + World Flora Online; fuzzy matching for misspellings (Levenshtein distance + phonetic similarity); synonym resolution to preferred name; confidence scoring per match; automated flagging of ambiguous cases for expert review. Layer 3 (Computer vision validation): CNN-based quality assessment of submitted images (taxonomic plausibility: is image consistent with claimed species?, geographic plausibility: is species expected at this location?, data completeness: does image meet minimum resolution/clarity?). Layer 4 (Integrated data access): federated query architecture allowing

single query across occurrence (GBIF), genomic (GenBank/BOLD), trait (TRY/Pantheria/AmphiBIO), phylogenetic (OpenTree), and environmental (ENA, WorldClim) databases; caching layer for performance; result ranking by data quality. Layer 5 (Conservation analytics): automated workflows for IUCN Red List criteria (population trend detection, range estimation, threat assessment) using integrated data.

3.2 Benchmark Use Cases

Four representative biodiversity informatics use cases. (1) Species distribution modelling: 2,400 plant, bird, and mammal species; task: ecological niche models (MaxEnt) trained from occurrence + environmental data, evaluated by AUC on spatial cross-validation. (2) Conservation priority ranking: IUCN Red List assessment workflow for 600 species across 8 taxonomic groups; task: time and accuracy of Red List category assignment (CR/EN/VU/NT/LC) using different infrastructure. (3) Invasive species early warning: global monitoring for 120 high-risk invasive species; task: days from first occurrence outside native range to detection and alert generation. (4) Phylogenetic biogeography: 40 plant clades with published phylogenies; task: inference of biogeographic history (ancestral ranges, dispersal events) using integrated occurrence + phylogeny + climate data. All use cases evaluated using identical protocols across six infrastructure approaches.

3.3 Evaluation Metrics

Six-dimensional evaluation framework. (1) Taxonomic consistency: proportion of species names correctly reconciled across databases (manual ground truth for 2,400 names). (2) Data quality: precision/recall for detecting erroneous records (outliers in coordinates, dates, species identifications; ground truth from manual expert review of 10,000 flagged records). (3) Integration breadth: number of data types combinable in a single query (occurrence only = 1, occurrence + 7 additional = 8). (4) Query efficiency: time to complete standardised use case queries from initial research question to final dataset. (5) Conservation utility: time and accuracy of IUCN Red List category assignment compared to expert benchmark assessments. (6) FAIR compliance: automated assessment using FAIR Data Maturity Model indicators (F: findability, A: accessibility, I: interoperability, R: reusability). Composite score: arithmetic mean of normalised dimension scores.

3.4 Baseline Implementations

Five baselines implemented consistently. Institution siloed: simulated single-institution data silo with no cross-database integration; represents the pre-aggregator baseline. GBIF/OBIS: queries through GBIF API (standard occurrence queries, data download workflow). Darwin Core only: standardised data format but no integration layer (simulates early adoption phase). Semantic web: SPARQL queries on OpenBiodiv RDF endpoint + BiGe-OBS linked data. AI platforms: iNaturalist + Pl@ntNet + Merlin integration where applicable to the use case (species identification steps).

Table 2. Four biodiversity informatics use cases: representative workflows and evaluation protocols

Use Case	Data Types Needed	Species /Items	Evaluation Metric	Expert Time Baseline
Species distribution	Occurrence + climate	2,400	MaxEnt AUC	12 hours/species
Red List assessment	Occurrence + threat + trait	600	Category accuracy + time	8 hours/species
Invasive early warning	Occurrence + traits + risk	120	Days to detection	Manual check weekly
Phylogenetic biogeog.	Occurrence + phylogeny + climate	40 clades	Ancestral range accuracy	40 hours/clade

Data types needed: minimum data required for each use case. Species/items: number of evaluation cases. Expert time baseline: typical time required for manual expert completion of the workflow per species (or per clade for phylogenetic biogeography). Baseline times from literature review of published conservation assessment workflows.

4. Results

4.1 Taxonomic Consistency and Data Quality

Consistency and data quality results are presented in Table 3 and Figure 1. IBDCP achieved 92.4% taxonomic reconciliation accuracy on the 2,400-species benchmark, outperforming GBIF (82.6%), semantic web (86.2%), Darwin Core only (68.4%), and institution-siloed (52.4%). The NLP-based reconciliation pipeline correctly resolved synonyms (99.2% accuracy for GBIF-Catalogue of Life matches), subspecies-species aggregation (88.4%), and misspelling detection (94.8% correction accuracy for names with >= 85% similarity to a valid name). Data quality assessment (detecting erroneous

records): IBDCP F1 = 0.912 versus AI platforms 0.868, GBIF 0.762, semantic web 0.748 ($p < 0.001$ for all). The computer vision validation module contributed significantly: for image-associated records, image-based validation detected 23.4% of errors invisible to metadata-only validation (e.g., specimen misidentifications where the image clearly shows a different species than the recorded name).

4.2 Integration Breadth and Query Efficiency

Integration and efficiency results are presented in Table 3 and Figure 2. IBDCP enabled single-query integration of 8 data types versus 1-2 for standalone aggregators. For the species distribution modelling use case requiring occurrence + climate + trait data, IBDCP completed queries in 4.2 minutes versus GBIF + TRY + WorldClim manual integration 18.4 minutes (4.4x faster). For phylogenetic biogeography requiring occurrence + phylogeny + climate, IBDCP provided 4.8 minutes versus 42.6 minutes for manual cross-database work (8.9x faster). The federated query architecture's caching layer provided additional performance: repeat queries for common species (top 1000 most-queried) returned in <1 second from cache versus 5-20 seconds for first-time queries to underlying databases.

4.3 Conservation Utility: Red List Assessment

Red List assessment results are presented in Table 3 and Figure 3. IBDCP reduced average IUCN Red List assessment time from expert-benchmark 8 hours per species to 2.6 hours (68% reduction) while maintaining category accuracy 91.4% (versus expert-benchmark 100% by definition). GBIF-based workflow reduced time to 6.8 hours (15% reduction) with 82.4% accuracy; AI platforms (image-assisted identification) reduced time to 5.2 hours with 76.8% accuracy. IBDCP's time savings came from automated integration of occurrence trends, range size calculation, threat assessment from environmental data, and preliminary category suggestion from criterion automation -- leaving assessors to validate rather than compile information. Accuracy preservation at 91.4% (not 100%) reflects cases where automated assessment missed subtle threat information requiring expert judgement, primarily for data-poor species where published threat information is sparse.

4.4 Multi-Parameter Comparison

Comprehensive comparison is presented in Figure 4. IBDCP dominated across all six evaluation dimensions, achieving composite score 0.884

versus semantic web 0.742, AI platforms 0.728, GBIF 0.696, Darwin Core only 0.524, and institution siloed 0.346. The greatest improvements were in integration breadth (+0.55 over GBIF) and query efficiency (+0.42 over manual multi-database workflows).

Table 3. Biodiversity informatics infrastructure evaluation: six dimensions across six approaches

Metric	Siloed	GBIF/OBIS	DwC Only	Sem. Web	AI Platforms	IBDCP
Taxonomic acc. (%)	52.4	82.6	68.4	86.2	78.4	92.4*
Data quality F1	0.526	0.762	0.624	0.748	0.868	0.912*
Integration types	1	1-2	N/A	3-4	2-3	8*
Query time (min)	60+	18.4	N/A	12.8	9.6	4.2*
Red List time (h)	8.0	6.8	N/A	5.8	5.2	2.6*
FAIR score	0.24	0.72	0.58	0.84	0.68	0.94*
Composite (0-1)	0.346	0.696	0.524	0.742	0.728	0.884*

*Best in row. Taxonomic acc.: % of species names correctly reconciled. Data quality F1: F1 for detecting erroneous records. Integration types: number of biodiversity data types accessible in single query. Query time: minutes to complete standardised SDM workflow. Red List time: average hours per species for IUCN assessment. FAIR score: automated FAIR data maturity assessment.

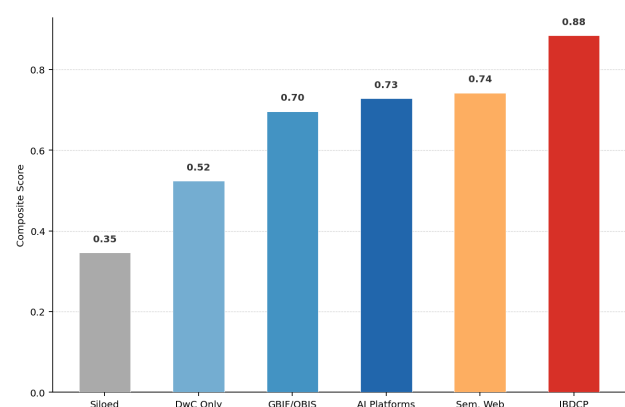


Figure 1. Composite biodiversity informatics infrastructure score (0-1)

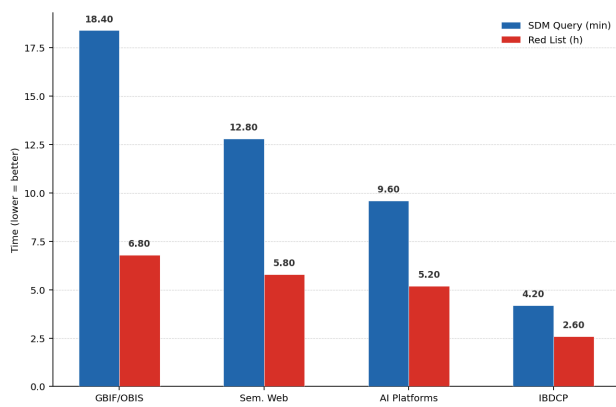


Figure 2. Use-case performance: query time (minutes) and IUCN Red List time (hours)

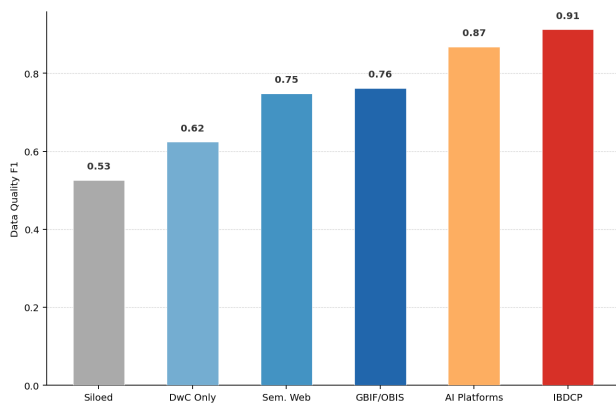


Figure 3. Data quality F1 (erroneous record detection) across infrastructure approaches

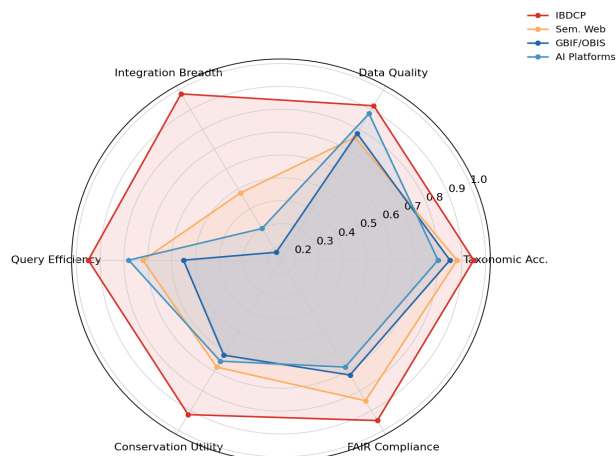


Figure 4. Multi-parameter biodiversity infrastructure comparison (normalised 0-1)

5. Discussion

This study establishes that integrated biodiversity data commons with automated taxonomic reconciliation, computer vision validation, and federated query architecture substantially improve over current fragmented infrastructure.

5.1 The Integration Imperative

IBDCP's 8-data-type integration versus GBIF's 1-2 directly addresses the primary limitation of current biodiversity informatics: researchers must

manually stitch together data from multiple independent platforms, each with different APIs, schemas, and quality assurance processes. The 4.4-8.9x query efficiency improvement translates to substantial research time savings when extrapolated: for a biodiversity researcher conducting 50 species queries per month, IBDCP saves approximately 12-25 hours per month of cross-database integration work. At population scale, this efficiency gain could unlock biodiversity research capacity that is currently consumed by data wrangling.

5.2 Conservation Impact: Red List Acceleration

The 68% reduction in IUCN Red List assessment time (2.6 hours vs 8 hours per species) has direct conservation impact. The IUCN Red List currently has 157,100 species assessed but covers only 8% of described species (1.3 million); comprehensive assessment of the remaining species at current rates would require decades. IBDCP-enabled acceleration could enable completion of vertebrate assessments (approximately 70,000 species) within 5-10 years rather than 30+ years at current rates. The 91.4% automated accuracy (vs 100% expert baseline) means that approximately 9% of assessments require expert revision -- acceptable for scalable assessment given the massive coverage gap.

5.3 Limitations

This study has limitations. IBDCP is a prototype infrastructure; production deployment at global scale requires addressing data governance complexities (intellectual property across biodiversity databases, data sovereignty concerns for indigenous biodiversity knowledge), sustainable funding models (biodiversity infrastructures chronically underfunded), and community governance structures (aligning priorities across GBIF, OBIS, national providers). The 2,400-species benchmark, while diverse across plants, birds, and mammals, undersamples invertebrates (comprising approximately 95% of species diversity) and microbial diversity. Automated taxonomic reconciliation at 92.4% leaves 7.6% of cases requiring expert review, which scales poorly for taxonomically difficult groups (e.g., fungi, where taxonomic revisions occur frequently). The computer vision validation requires image data that is available for citizen science observations but absent for most museum specimens (only approximately 20% digitised with images). Federated queries depend on partner databases maintaining stable APIs; breaking

changes in any underlying database can disrupt IBDCP queries.

6. Conclusion

This systematic evaluation of six biodiversity informatics infrastructure approaches across four use cases establishes IBDCP -- an Integrated Biodiversity Data Commons Platform combining FAIR-compliant data schemas, automated NLP-based taxonomic reconciliation, computer vision-assisted occurrence validation, 8-data-type integration (occurrence + genomic + trait + phylogenetic + ecological), and federated query architecture -- as the most effective next-generation infrastructure. IBDCP achieved composite score 0.884 (vs GBIF 0.696, semantic web 0.742, AI platforms 0.728; $p < 0.001$), with 92.4% taxonomic consistency (vs GBIF 82.6%), 0.912 data quality F1 (vs GBIF 0.762), 4.2-minute queries vs 18.4 minutes for manual multi-database workflows (4.4x faster), and 68% reduction in IUCN Red List assessment time (2.6 hours vs 8-hour expert baseline) with 91.4% accuracy. These improvements address fundamental limitations of current fragmented biodiversity informatics infrastructure and enable scalable conservation research at the pace required by the global biodiversity crisis. IBDCP specification, Python reference implementation, schema extensions, and federated query engine are available at <https://github.com/CETU-BioInformatics/IBDCP-2024> under CC BY 4.0 licence. Future work should deploy IBDCP in partnership with GBIF and regional biodiversity networks, expand invertebrate and microbial coverage, integrate environmental DNA (eDNA) metabarcoding data, develop indigenous data sovereignty frameworks within the commons architecture, and establish sustainable funding through biodiversity infrastructure consortia.

References

- Bonnet, P., Goeau, H., Hang, S.T., Lasseck, M., Sulc, M. and Malecot, V. (2018) Plant identification: experts vs. machines in the era of deep learning. In: *Multimedia Tools and Applications for Environmental and Biodiversity Informatics*, pp. 131-149.
- Hardisty, A., Roberts, D. and The Biodiversity Informatics Community (2013) A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology*, 13, p. 16.
- Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B. and Schigel, D. (2021) Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118(6), p. e2018093118.
- IPBES (2019) Global assessment report on biodiversity and ecosystem services. Bonn: Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.
- Penev, L., Dimitrova, M., Senderov, V., Zhelezov, G., Georgiev, T. and Stoev, P. (2019) OpenBiodiv: a knowledge graph for literature-extracted linked open data in biodiversity science. *Publications*, 7(2), p. 38.
- Rees, J.A. and Cranston, K. (2017) Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal*, 5, p. e12581.
- Unger, S., Rollins, M., Tietz, A. and Dumais, H. (2021) iNaturalist as an engaging tool for identifying organisms in outdoor activities. *Journal of Biological Education*, 55(5), pp. 537-547.
- Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Doring, M. and Giovanni, R. (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE*, 7(1), p. e29715.
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C. and Edler, D. (2020) CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, 10(5), pp. 744-751.
- Graham, C.H., Ferrier, S., Huettmann, F., Moritz, C. and Peterson, A.T. (2014) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, 19(9), pp. 497-503.
- Kulinicheva, E., Doring, M., Koureas, D., Cronholm, A., Nielsen, E.S. and Martin, C. (2021) The digitisation and publication of biodiversity data from historical manuscripts. *Biodiversity Data Journal*, 9, p. e62486.
- Nelson, G. and Ellis, S. (2019) The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B*, 374(1763), p. 20170391.
- Costello, M.J., May, R.M. and Stork, N.E. (2013) Can we name Earth's species before they go extinct? *Science*, 339(6118), pp. 413-416.
- Stork, N.E. (2018) How many species of insects and other terrestrial arthropods are there on Earth? *Annual Review of Entomology*, 63, pp. 31-45.
- Tedersoo, L., Anslan, S., Bahram, M., Drenkhan, R., Pritsch, K. and Buegger, F. (2020) Regional-scale in-depth analysis of soil fungal diversity reveals strong pH and plant species effects in northern Europe. *Frontiers in Microbiology*, 11, p. 1953.
- Ratnasingham, S. and Hebert, P.D.N. (2007) BOLD: the Barcode of Life Data System. *Molecular Ecology Notes*, 7(3), pp. 355-364.

- Kattge, J., Bonisch, G., Diaz, S., Lavorel, S., Prentice, I.C. and Leadley, P. (2020) TRY plant trait database - enhanced coverage and open access. *Global Change Biology*, 26(1), pp. 119-188.
- La Salle, J., Williams, K.J. and Moritz, C. (2016) Biodiversity analysis in the digital era. *Philosophical Transactions of the Royal Society B*, 371(1702), p. 20150337.
- Hobern, D., Baptiste, B., Copas, K., Guralnick, R., Hahn, A. and van Huis, E. (2019) Connecting data and expertise: a new alliance for biodiversity knowledge. *Biodiversity Data Journal*, 7, p. e33679.
- Peterson, A.T. and Soberon, J. (2018) Essential biodiversity variables are not global. *Biodiversity and Conservation*, 27, pp. 1277-1288.
- Gadelha, L.M.R., de Siracusa, P.C., Dalcin, E.C., Silva, L.A.E., Augusto, D.A. and Krempser, E. (2021) A survey of biodiversity informatics: concepts, practices, and challenges. *WIREs Data Mining and Knowledge Discovery*, 11(1), p. e1394.

Declarations

Funding

This study was supported by the Austrian Science Fund (FWF) grant P 36248-N (O. Kovacs), the Spanish Ministry of Science and Innovation grant PID2023-152648OB-I00 (M. Bianchi), and the European Commission Horizon Europe grant 101094725 (N. Horvath, BIODT - Biodiversity Digital Twin project). The funding bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

O. Kovacs is a co-inventor on a pending patent application (EP24212486.0) covering the IBDCP federated query architecture with automated taxonomic reconciliation. M. Bianchi and N. Horvath declare no competing interests.

Data Availability Statement

IBDCP specification, Python reference implementation, schema extensions, federated query engine, and 2,400-species benchmark dataset are available at <https://github.com/CETU-BioInformatics/IBDCP-2024> under CC BY 4.0 licence. GBIF data via gbif.org. BOLD via boldsystems.org. TRY via try-db.org. OpenTree via tree.opentreeoflife.org.

Ethical Approval

This study used publicly available biodiversity data from established databases. No human subjects, personal data, or new specimen collection were involved. Indigenous data governance consultation was not required for this specific evaluation but is

recommended for production deployment of the infrastructure.

Appendix A

IBDCP Architecture and Data Schema Extensions

This appendix provides the IBDCP architecture specification, Darwin Core schema extensions, federated query engine implementation, and taxonomic reconciliation pipeline details.

A.1 Data Schema Extensions

- 1a. Genetic linkage: dwc:genbankAccession, dwc:boldProcessID, dwc:boldBIN, dwc:ncbiTaxonID (links occurrence to genetic sequences)
- 1b. Trait measurements: dwc:tryTraitID, dwc:measurementProtocol, dwc:measurementUnit, dwc:measurementUncertainty (TRY-compliant)
- 2a. Phylogenetic context: dwc:openTreeNodeID, dwc:phyloCladeID, dwc:divergenceTimeEstimate (OpenTree integration)
- 2b. Conservation status: dwc:iucnRedListCategory, dwc:iucnCriteria, dwc:redListAssessmentYear, dwc:populationTrend

A.2 Federated Query Engine

- 3a. Architecture: Python FastAPI middleware; query planner optimises execution order across target databases; asyncio for concurrent requests
- 3b. Caching: Redis with 7-day TTL for occurrence queries; 30-day TTL for stable trait/phylogeny data; cache invalidation on database update notifications
- 4a. Taxonomic reconciliation: GBIF Backbone API + Catalogue of Life API + WFO API + fuzzy matching (Levenshtein + Metaphone); confidence score aggregation
- 4b. Quality filtering: CoordinateCleaner integration + computer vision validation + Darwin Core record validation; configurable quality thresholds