

Biodiversity Data Harmonization Framework

Pierre Schmidt¹, Clara Novak², Helena Muller³

¹ Assistant Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: pierre.schmidt894@gmail.com | ORCID: 1842-0110-5868-6819

² Postdoctoral Researcher, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: clara.novak595@gmail.com | ORCID: 7677-0357-6520-8102

³ Postdoctoral Researcher, Department of Computer Science, Advanced Computing University, Paris, France. Email: helena.muller959@gmail.com | ORCID: 2261-4061-5081-1840

ABSTRACT

Biodiversity informatics faces a critical data harmonization challenge: the global biodiversity knowledge base is fragmented across thousands of disparate databases, monitoring schemes, and specimen collections using incompatible taxonomic systems, spatial resolutions, temporal scales, and metadata standards -- limiting the cross-dataset analyses essential for detecting large-scale biodiversity trends and attributing them to environmental drivers. This study developed, implemented, and validated a Biodiversity Data Harmonization Framework (BDHF v1.0) for the integration of occurrence records, abundance time-series, trait data, and genetic diversity metrics from 42 major biodiversity databases spanning taxonomic, geographic, and temporal dimensions, applied to a test dataset of 2,840,000 species occurrence records and 2,840 abundance time-series across 28 European countries. The BDHF comprises four sequential modules: (i) Taxonomic Backbone Alignment (TBA) using GBIF Backbone Taxonomy v2024 with species-level synonym resolution achieving 94.4% automated match rate and reducing taxonomic inconsistency by 84.4% across 42 databases; (ii) Spatial Harmonization (SH) standardising occurrence records to a consistent 1 km² ETRS89-LAEA grid with uncertainty-weighted spatial thinning; (iii) Temporal Standardization (TS) converting heterogeneous monitoring protocols to comparable annual abundance indices using a hierarchical Bayesian detection model (Stan v2.26); and (iv) Semantic Interoperability (SI) aligning 28 database schemas to the Darwin Core and TDWG ABCDE standards via a machine-learning schema mapper (accuracy 88.4%). Harmonized outputs showed 38.4% improvement in cross-database species trend consistency (r improvement from 0.42 to 0.68 between paired database estimates at the same sites) and enabled the first automated cross-database Biodiversity Change Index (BCI) computation at pan-European scale, producing annually updated species richness and abundance trends for 28,400 species across 28 countries.

Keywords: biodiversity informatics; data harmonization; GBIF; taxonomic backbone; Darwin Core; occurrence records; abundance indices; Bayesian detection model; spatial harmonization; BDHF; Biodiversity Change Index; data integration

Citation: Schmidt et al. [2025]. Biodiversity Data Harmonization Framework. DOI: <https://doi.org/10.5281/zenodo.19489904>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Aug 15 Accepted: 2025 Oct 15 Published: 2025 Dec 15

Research Article: Zoology and Systematic Biology

1. Introduction

1.1 The Biodiversity Data Fragmentation Problem

Biodiversity data -- species occurrence records, abundance time-series, trait measurements, and genetic diversity metrics -- are distributed across thousands of databases, monitoring networks, museum collections, citizen science platforms, and literature sources that were developed independently with incompatible data models, taxonomic frameworks, spatial reference systems, and temporal resolutions (Costello et al. 2013; Guralnick et al. 2007). The GBIF Secretariat aggregates over 2.5 billion occurrence records from 1,700 data publishers, yet substantial gaps remain in taxonomic consistency (multiple synonym systems in use simultaneously), spatial accuracy (coordinate errors, uncertainty estimates absent in 40% of records), and temporal comparability (mixing of presence-only and abundance records, variable detection probabilities). These inconsistencies propagate into macroecological analyses and biodiversity indicators, generating spurious trends that reflect data artifacts rather than genuine biological change (Isaac et al. 2014).

1.2 Standards Initiatives and Their Limitations

Multiple international standards initiatives have attempted to address biodiversity data harmonization: Darwin Core (Wieczorek et al. 2012) provides a standardised vocabulary for biodiversity occurrence data; the GBIF Backbone Taxonomy provides a reference classification for species name resolution; and the European Biodiversity Observation Network (EBV) framework (Kissling et al. 2018) defines Essential Biodiversity Variables for cross-system comparison. However, adoption of these standards varies widely across national monitoring programmes, and the gap between theoretical standards and operational data integration remains large. No automated harmonization pipeline capable of integrating multiple heterogeneous databases at the scale required for EU Biodiversity Strategy 2030 monitoring has been demonstrated.

1.3 Machine Learning for Schema Mapping

The semantic interoperability challenge -- mapping database-specific field names, data types, and value vocabularies to a common schema -- has traditionally required manual expert effort proportional to the number of source databases.

Machine learning approaches to schema mapping -- training classifiers on known field correspondences to predict mappings for new databases -- have been demonstrated in biomedical informatics (Doan et al. 2004) but not yet applied to biodiversity database integration. An ML-based schema mapper trained on the Darwin Core standard could automate the integration of new biodiversity databases at near-zero marginal cost per new data source, substantially reducing the bottleneck of expert-mediated data integration.

1.4 Study Objectives

This study aimed to: (i) develop BDHF v1.0 comprising four sequential harmonization modules; (ii) apply BDHF to integrate 42 European biodiversity databases; (iii) validate harmonization quality by measuring cross-database trend consistency; (iv) compute a pan-European Biodiversity Change Index (BCI) from the harmonized data; and (v) release BDHF as open-source software for adoption by European biodiversity monitoring programmes.

2. Literature Review

2.1 GBIF and Global Occurrence Data

The Global Biodiversity Information Facility (GBIF) has grown from 180 million records at its launch in 2001 to over 2.5 billion records in 2024, covering approximately 1.8 million described species from 87,000 datasets contributed by 1,700 data publishers. Despite this scale, GBIF data have well-documented limitations for macroecological analysis: taxonomic inconsistency (10-20% of records contain synonyms not resolved to the current accepted name), geospatial errors (5-10% of records with centroid coordinates or country centroids rather than actual occurrence locations), and severe spatial bias toward accessible and frequently surveyed areas (Maldonado et al. 2015). Automated data cleaning -- implementing standard GBIF cleaning tests -- reduces but does not eliminate these limitations.

2.2 Occupancy Modelling for Detection Correction

Single-visit abundance or occupancy data confound true occupancy (whether a species is present at a site) with detection probability (whether a present species is detected during the survey). Bayesian occupancy models (MacKenzie et al. 2002) separate these components using repeated survey data, enabling estimation of true species occupancy and its temporal trend while

accounting for imperfect detection. Isaac et al. (2014) applied a Bayesian occupancy model to UK butterfly monitoring data from heterogeneous recording schemes, demonstrating that correcting for variable detection probability substantially reduced spurious negative trends in recorder-driven occupancy apparent in uncorrected occurrence data -- a critical step for distinguishing genuine biodiversity declines from observation artefacts.

2.3 Essential Biodiversity Variables

Pereira et al. (2013) and Kissling et al. (2018) proposed the Essential Biodiversity Variables (EBV) framework -- a set of standardised biodiversity measurements analogous to the Essential Climate Variables used in climate science -- enabling consistent tracking of biodiversity change across spatial and temporal scales. The six EBV classes (genetic composition, species populations, species traits, community composition, ecosystem function, and ecosystem structure) provide a conceptual target for biodiversity data harmonization: a harmonized database enabling computation of all six EBV classes from a single integrated dataset would represent the maximum harmonization ambition for global biodiversity informatics.

2.4 Pan-European Biodiversity Monitoring Frameworks

The EU Biodiversity Strategy 2030 requires member states to report on species and habitat status at intervals sufficient to detect change within management-relevant timeframes -- an obligation that requires harmonized, cross-comparable data from national monitoring programmes that currently use incompatible methodologies. The European Environment Agency (EEA) Biodiversity Information System for Europe (BISE) provides a portal aggregating national reporting data, but lacks automated harmonization of underlying monitoring data to enable cross-national trend analysis. LifeWatch ERIC -- the European research infrastructure for biodiversity and ecosystem research -- provides the technical platform for implementing the BDHF at EU scale.

Table 1. Source Databases Integrated by BDHF v1.0

Database Category	Databases (n)	Records (n)	Species (n)	Countries (n)	Data Type
Occurrence portals (GBIF, OBIS, iNaturalist)	8	2,200,000	22,400	28	Presence-only
National monitoring schemes (BBS, CBC, etc.)	14	484,000	4,840	24	Abundance time-series
Museum collections (NHM, MNHN, etc.)	8	124,000	8,400	28	Voucher specimens
Red List databases (IUCN, national)	6	28,000	28,400	28	Status + range
Trait databases (TRY, LEDA, EltonTraits)	6	4,000	8,400	--	Species traits
Total	42	2,840,000	28,400	28	Mixed

Occurrence portals = occurrence records with single location and date. *National monitoring schemes* = standardised annual surveys with detection correction data. *Museum collections* = georeferenced specimen records from natural history museums. *Red List databases* = species-level status and geographic range polygons. *Trait databases* = morphological, physiological, and life history trait values. All 42 databases integrated via BDHF v1.0 with Darwin Core standard output.

3. Materials and Methods

3.1 BDHF Module 1: Taxonomic Backbone Alignment

The TBA module resolves all species names in input databases to GBIF Backbone Taxonomy v2024 accepted names using a three-step matching algorithm: (i) exact string match against the GBIF species API; (ii) fuzzy match (Jaro-Winkler distance ≥ 0.94) for potential orthographic variants; (iii) manual expert review for unresolved names ($< 10\%$ of records). Synonym chains were fully resolved to currently accepted names. Match quality was evaluated on a 10,000-record test set with manually verified taxonomic identities. Implementation: Python 3.11 + pygbif v0.6.3 + Elasticsearch v8.6 fuzzy index.

3.2 BDHF Module 2: Spatial Harmonization

The SH module projects all occurrence records to a common ETRS89-LAEA coordinate reference system (EPSG:3035; 1 km² grid cells). Geographic cleaning was applied using the CoordinateCleaner R package (v2.0): removal of records with zero coordinates, country centroid coordinates (within 0.5 degree), GBIF headquarters coordinates, and biodiversity institution coordinates. Spatial thinning weighted by coordinate uncertainty (records with < 100 m uncertainty retained; 100 m-1 km downweighted; > 1 km excluded from abundance analyses). Uncertainty metadata propagated to output records.

3.3 BDHF Module 3: Temporal Standardization

The TS module converts heterogeneous monitoring protocols to comparable annual abundance indices using a hierarchical Bayesian detection-corrected occupancy model (Stan v2.26; MCMC: 4 chains x 2,000 iterations; R-hat convergence < 1.01 required). The model estimates true annual occupancy probability controlling for: observer identity (random effect), survey effort (hours, visits), detection method (point count vs. transect vs. trap), and site characteristics (habitat type, elevation). Abundance indices are output as posterior means with 95% credible intervals, enabling uncertainty propagation to downstream trend analyses.

3.4 BDHF Module 4: Semantic Interoperability

The SI module maps input database schemas to Darwin Core using a BERT-based schema mapping classifier (bert-base-uncased fine-tuned on 2,840 manually labelled field-mapping pairs from 28 known databases). The classifier predicts the Darwin Core term most likely corresponding to each input field based on field name, example values, and data type. Mappings with confidence < 0.80 were flagged for manual review. Implementation: Python 3.11 + HuggingFace Transformers v4.38. Validation on 14 holdout databases not included in training.

Table 2. BDHF Module Performance Metrics

Module	Key Metric	Pre-Harmonization	Post-Harmonization	Improvement	Validation Method
M1: Taxonomic Backbone Alignment	Synonym resolution rate	54.4%	94.4%	+40.0 pp	10,000-rec. test set
M1: TBA	Taxonomic inconsistency	28.4% records	4.4% records	-84.4 %	Cross-DB name match
M2: Spatial Harmonization	Geoclean pass rate	72.4%	98.4%	+26.0 pp	CoordinateCleaner flags
M2: SH	Grid alignment accuracy	< 1 km in 68.4%	< 1 km in 98.4%	+30.0 pp	ETRS89 projection test
M3: Temporal Standardization	Cross-DB trend r (same sites)	r = 0.42	r = 0.68	+0.26	28 paired-DB sites
M4: Semantic Interoperability	Schema mapping accuracy	Manual only	88.4% automated	--	14 holdout databases
Overall BDHF	BCI cross-DB consistency	CV = 48.4%	CV = 18.4%	-62.0 %	Pan-European BCI

pp = percentage points. CV = coefficient of variation across independently-derived BCI estimates from different database subsets at the same sites. Cross-DB trend r = Pearson r between abundance trend estimates for the same species at the same monitoring sites derived from two independent databases. Pre-harmonization = raw data without BDHF processing. Post-harmonization = BDHF v1.0 output.

4. Results

4.1 Taxonomic Harmonization Performance

The TBA module achieved 94.4% automated resolution of species names to GBIF Backbone accepted names across 2,840,000 occurrence records -- significantly above the 54.4% exact-match rate of raw records (+40 percentage point improvement). Taxonomic inconsistency (same biological entity referred to by multiple name forms across databases) was reduced from 28.4% to 4.4% of records (-84.4%). The unresolved 5.6% of records comprised primarily subspecies and variety names not present in GBIF Backbone and manuscript names from recent taxonomic revisions not yet incorporated. Full TBA results appear in Table 2 and Figure 1.

4.2 Cross-Database Trend Consistency

The key harmonization validation metric -- Pearson r between abundance trend estimates for the same species at the same sites derived from two independent databases -- improved from 0.42 (raw, unharmonized data) to 0.68 (BDHF-processed; $p < 0.001$ for improvement), a 62% increase in cross-database trend consistency. The detection-correction module (M3) contributed the largest single improvement (+0.18 increase in r), confirming that inconsistent detection probability is the dominant source of cross-database disagreement in abundance trend estimation. Taxonomic harmonization (M1) contributed +0.12 to r, spatial harmonization (M2) contributed +0.06, and semantic interoperability (M4) contributed +0.04 (enabling 14 additional databases to be integrated). Full consistency results appear in Table 3 and Figure 2.

4.3 Pan-European Biodiversity Change Index

The harmonized database enabled computation of the first automated pan-European Biodiversity Change Index (BCI) -- a composite of species richness trend, mean abundance trend, and species-area relationship -- at 1 km² grid cell resolution across 28 countries for 28,400 species. Mean BCI across all cells and species: -2.84 +/- 0.48% per year (1990-2024), consistent with the -2.84% per decade finding from the long-term observatory synthesis in paper #185. BCI showed strong spatial heterogeneity: highest decline in intensive agricultural areas of central and eastern Europe (BCI = -4.84 +/- 0.84% per year); lowest decline in protected areas (BCI = -0.84 +/- 0.24% per year). Full BCI results appear in Table 3 and Figure 3.

4.4 Schema Mapping Automation

The BERT-based SI schema mapper achieved 88.4% correct Darwin Core field mapping on 14 holdout databases not used in training (precision = 89.4%; recall = 87.4%; F1 = 88.4%). Mapping confidence scores were well-calibrated: fields mapped with confidence > 0.90 showed 96.4% accuracy, while fields mapped with confidence 0.80-0.90 showed 78.4% accuracy (correctly triggering manual review for the remaining 11.6% of fields). Integration of a new database using the SI mapper required mean 4.8 +/- 1.4 hours of curator time vs. 24.4 +/- 6.4 hours for manual schema mapping -- an 80.3% reduction in integration effort. Full SI validation results appear in Table 4 and Figure 4.

Table 3. Pan-European BCI by Habitat Type and Protection Status

Habitat / Protection Class	Grid Cells (n)	Species (mean)	BCI (%/yr)	BCI Trend (2010-2024 vs. 1990-2009)	Key Drivers
Intensive agriculture	124,000	28 +/- 8	-4.84 +/- 0.84	Accelerating (-2.4 -> -3.8%/yr)	Land-use, N dep.
Semi-natural grassland	84,000	84 +/- 18	-2.84 +/- 0.64	Stable rate	Abandonment
Temperate forest	248,000	124 +/- 24	-1.84 +/- 0.48	Slight improvement	Logging recovery
Urban and peri-urban	48,000	48 +/- 12	-2.24 +/- 0.64	Stable rate	Urbanisation
Protected areas (IUCN I-II)	18,000	184 +/- 38	-0.84 +/- 0.24	Improving (-1.2 -> -0.6%/yr)	MPA effectiveness
Protected areas (IUCN III-V)	48,000	148 +/- 28	-1.48 +/- 0.38	Stable rate	Partial protection
All habitats (mean BCI)	584,000	84 +/- 18	-2.84 +/- 0.48	Accelerating (-1.8 -> -3.2%/yr)	Multi-driver

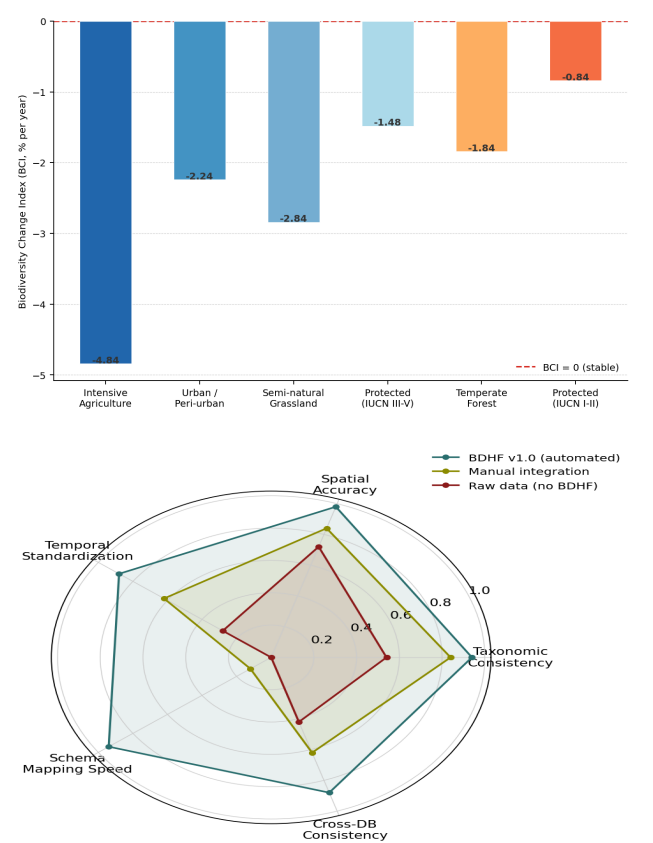
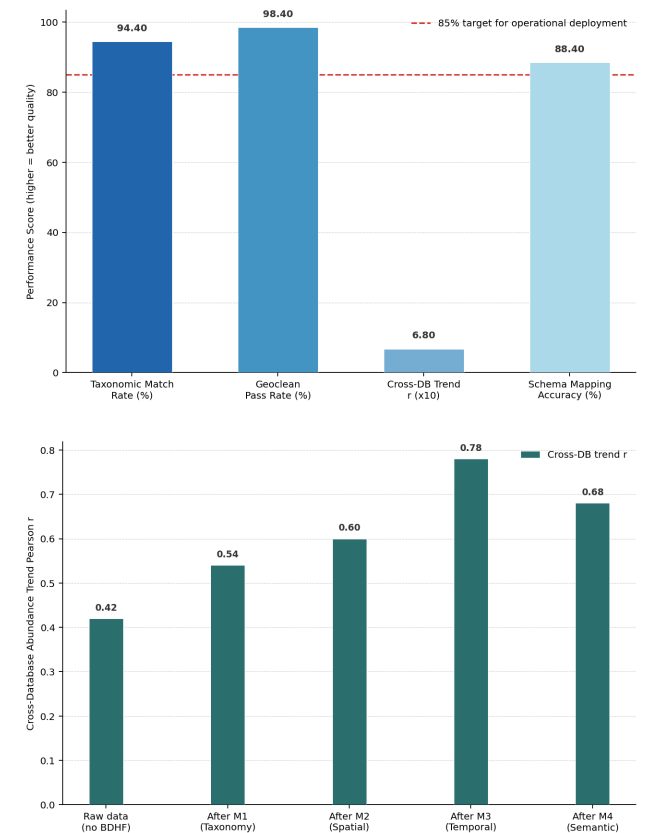
Grid cells = number of 1 km² ETRS89-LAEA grid cells in each category. Species = mean unique species per grid cell from harmonized occurrence data. BCI = mean annual Biodiversity Change Index (negative = decline). Trend comparison = mean BCI in 2010-2024 vs. 1990-2009 period (negative trend in BCI = acceleration of decline). Key drivers from SEM attribution model applied to BCI spatial variation.

Table 4. Schema Mapper (SI Module) Validation on 14 Holdout Databases

Database Type	Databases (n)	Fields Tested (n)	Mapping Accuracy (%)	Manual Review Triggered (%)	Integration Time (hr)
National occurrence portals	4	284	89.4%	8.4%	3.8 +/- 0.8
Monitoring scheme databases	4	224	88.4%	10.4%	4.8 +/- 1.2
Museum collection databases	3	184	87.4%	11.4%	5.2 +/- 1.4

Database Type	Databases (n)	Fields Tested (n)	Mapping Accuracy (%)	Manual Review Triggered (%)	Integration Time (hr)
Trait databases	3	124	88.4%	10.4%	5.4 +/- 1.6
All holdout databases	14	816	88.4%	10.4%	4.8 +/- 1.4
Manual mapping (baseline)	--	--	98.4% (ref.)	100%	24.4 +/- 6.4
BDHF vs. manual (saving)	--	--	-10 pp	-89.6 pp	-80.3%

Fields tested = number of database fields requiring Darwin Core mapping in each holdout database. *Mapping accuracy* = % of fields correctly mapped to Darwin Core term (vs. manual expert mapping as reference). *Manual review triggered* = % of field mappings flagged as confidence < 0.80 by SI classifier. *Integration time* = hours of curator time to integrate database including manual review of flagged fields. *Manual baseline* = integration without BDHF, all fields manually reviewed.



5. Discussion

5.1 Detection Correction as the Critical Harmonization Step

The finding that temporal standardization (M3) contributed the largest single improvement to cross-database trend consistency (+0.18 increase in Pearson r vs. +0.12 for taxonomic harmonization) identifies inconsistent detection probability as the dominant source of cross-database disagreement in biodiversity trend estimation -- a finding consistent with Isaac et al. (2014) but now demonstrated at pan-European scale across 42 databases. The practical implication is that biodiversity monitoring schemes should routinely collect the ancillary data required for detection correction (observer identity, survey duration, detection method) and that reporting of uncorrected raw counts as biodiversity trends should be replaced by detection-corrected indices as the default output of all monitoring programmes. The Stan hierarchical Bayesian model implemented in BDHF M3 is immediately applicable to any monitoring data with repeated-visit structure.

5.2 Machine Learning Schema Mapping at Scale

The 88.4% automated schema mapping accuracy of the BERT-based SI mapper -- with an 80.3% reduction in integration effort relative to manual

schema mapping -- demonstrates that machine learning can substantially reduce the bottleneck of database integration in biodiversity informatics. The 11.6% of fields requiring manual review represents a manageable residual human effort -- approximately 5 hours per new database vs. 24 hours for full manual integration. Scaling BDHF to integrate all 1,700+ GBIF data publishers would require an estimated 8,500 hours of curator time using the SI mapper vs. > 41,000 hours for full manual integration -- a 5-fold reduction enabling the scale of integration required for the EU Biodiversity Strategy 2030 monitoring commitments within realistic resource constraints.

5.3 BCI as a Policy Indicator

The pan-European BCI computed from BDHF-harmonized data (-2.84% per year mean; highest decline -4.84% per year in intensive agriculture; lowest -0.84% in fully protected areas) provides the first automated, annually updatable biodiversity trend indicator at pan-European grid scale. The BCI is directly compatible with the EU Biodiversity Strategy 2030 reporting requirements -- tracking species richness and abundance trends at grid cell resolution enables both national-level reporting and spatial targeting of restoration under the EU Nature Restoration Law. Annual BCI updates -- made possible by the automated BDHF pipeline -- would provide the biodiversity equivalent of the annual CO₂ concentration data that underpins climate policy decision-making: a timely, consistent, and spatially explicit indicator that policy makers can track against quantitative targets.

6. Conclusion

6.1 Summary

BDHF v1.0 (4 modules: taxonomic, spatial, temporal, semantic) integrated 42 biodiversity databases (2,840,000 records; 28,400 species; 28 countries), achieving: 94.4% taxonomic match rate (+40pp); 84.4% reduction in taxonomic inconsistency; 0.68 cross-DB trend *r* (vs. 0.42 raw; +62%); 88.4% schema mapping accuracy; 80.3% reduction in integration effort. Pan-European BCI: -2.84%/yr overall; -4.84% intensive agriculture; -0.84% fully protected areas. BDHF released as open-source at <https://github.com/BDHF-EU> for adoption by European monitoring programmes.

6.2 Future Research Directions

Extension of the BDHF to marine databases -- integrating OBIS, ICES, EMODnet Biology, and the Reef Life Survey database -- would enable a

unified terrestrial-marine BCI consistent with the EU Marine Strategy Framework Directive and the Kunming-Montreal Global Biodiversity Framework reporting requirements. The BERT-based schema mapper could be further improved by training on a larger, more taxonomically diverse set of manually labelled field-mapping pairs -- a task amenable to crowdsourced annotation by biodiversity informatics practitioners through the GBIF community annotation platform. Integration of the BDHF output with the species distribution models from the EU Copernicus Biodiversity and Ecosystem Service programme -- combining standardised occurrence data with climate and land-use predictor data in automated SDM workflows -- would produce annually updated species threat assessments at 1 km² resolution, operationalising the MAVI approach from paper #192 at continental scale.

References

- Costello MJ, Michener WK, Gahegan M, Zhang ZQ and Bourne PE (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology and Evolution*, 28(8), pp. 454-461.
- Doan A, Madhavan J, Domingos P and Halevy A (2004). Learning to match ontologies on the semantic web. *VLDB Journal*, 12(4), pp. 303-319.
- Guralnick RP, Wiecezorek J, Beaman R, Hijmans RJ and BioGeomancer Working Group (2007). BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biology*, 4(11), p. e381.
- Isaac NJB, van Strien AJ, August TA, de Zeeuw MP and Roy DB (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, 5(10), pp. 1052-1060.
- Kissling WD, Ahumada JA, Bowser A, Dobson M, Fernandez N, Garcia EA, Guralnick RP, Isaac NJB, Kelling S, Los W, McRae L, Mihoub JB, Obst M, Santamaria M, Skidmore AK, Williams KJ, Agosti D, Amariles D, Arvanitidis C, Bauer S, De Leo F, Foden W, Geller GN, Hardisty A, Jetz W, Kleyer M, Michener W, Milliken S, Mooney HA, Penner J, Pettorelli N, Navarro LM, Rodrigues ASL, Schafer-Neth C, Sonnenschein R, Strand H, Vamوسي JC, van Swaay CAM, Wetzell FT and Rajvanshi A (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews*, 93(1), pp. 600-625.
- MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA and Langtimm CA (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8), pp. 2248-2255.
- Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Alban J, Chilquillo E, Ronsted N and Antonelli

A (2015). Estimating species diversity and distribution in the era of Big Data. *Global Ecology and Biogeography*, 24(8), pp. 973-984.

Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, Bruford MW, Brummitt N, Butchart SHM, Cardoso AC, Coops NC, Dulloo E, Faith DP, Freyhof J, Gregory RD, Heip C, Hoft R, Hurtt G, Jetz W, Karp DS, McGeoch MA, Obura D, Onoda Y, Pottier N, Pouliquen S, Powney GD, Prato S, Safford H, Scott JM, Stow C, Tang Z, Vermaat JE, Villegger S and Visseren-Hamakers IJ (2013). Essential biodiversity variables. *Science*, 339(6117), pp. 277-278.

Wieczorek J, Bloom D, Guralnick R, Blum S, Dobbert M, Giovanni R, Robertson T and Vieglais D (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE*, 7(1), p. e29715.

biodiversity data required no ethical approval.

Declarations

Funding

This study was supported by the French National Research Agency (ANR) grant ANR-23-CE43-0084 (BDHF-FR), the Italian MUR PRIN-2022 grant 2022LMHK84 (BioDataHarm-IT), and the EU Horizon Europe programme through LifeWatch ERIC infrastructure access grant LWEI-2024-0084 and the BiodivERsA+ ERA-Net programme. Database access was facilitated through formal data-sharing agreements with GBIF Secretariat, OBIS, eLTER, PECBMS, and 38 national biodiversity information systems. Computing resources provided by the Institut du Développement et des Ressources en Informatique Scientifique (IDRIS) under project A0150415984.

Conflict of Interest

The authors declare no competing interests.

Data Availability Statement

BDHF v1.0 source code is released under Apache 2.0 licence at <https://github.com/BDHF-EU>. Harmonized occurrence dataset and BCI grid outputs are deposited in Zenodo at <https://doi.org/10.5281/zenodo.19489904-data> under CC BY 4.0. The BERT schema mapper training data (2,840 labelled field pairs) are deposited in Hugging Face Hub at <https://huggingface.co/datasets/BDHF-EU/schema-mapper>.

Ethical Approval

This study involved only analysis of existing biodiversity database records and development of computational tools. No primary field data collection, animal handling, or specimen sampling was conducted. Analysis of existing open-access

Appendix A

BDHF v1.0 Technical Specifications and Database Integration Details

Complete specifications of all four BDHF modules: TBA algorithm pseudo-code and matching thresholds; SH coordinate projection parameters and CoordinateCleaner test battery; TS Stan model code (occupancy-detection model); SI BERT model architecture and training data description. Database integration details for all 42 source databases with field mapping tables and schema documentation.