

Global Biodiversity Data Integration Frameworks

Lukas Jensen¹

¹ Professor, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: lukas.jensen827@yahoo.com | ORCID: 9647-1296-7974-7374

ABSTRACT

Global biodiversity data -- species occurrence records, population abundance time series, functional trait measurements, remote sensing habitat products, and phylogenetic trees -- are distributed across hundreds of repositories, databases, and monitoring programmes with heterogeneous formats, taxonomies, coordinate reference systems, and access policies that impede the integrative analyses needed for global biodiversity assessment. The fragmentation of biodiversity data infrastructure reflects the historically siloed development of ecological databases by research groups, national programmes, and international organisations with different scientific communities, funding streams, and technical standards. This review evaluates the current global biodiversity data infrastructure -- covering 18 major data repositories across five data types -- against FAIR (Findable, Accessible, Interoperable, Reusable) data principles, and proposes a federated integration framework capable of enabling cross-database analyses at the scale required for KM-GBF monitoring and IPBES global assessments. GBIF achieved the highest FAIR compliance score (87.4/100) followed by TRY (78.4) and OBIS (74.7). The 247-node biodiversity data network shows a mean integration completeness of only 38.4% across FAIR dimensions -- meaning that 61.6% of cross-database integration potential is unrealised. A federated architecture using persistent identifiers, standardised APIs, and shared taxonomic backbone would increase integration completeness to an estimated 74.7% at implementation cost of EUR 28.4 million over 5 years.

Keywords: biodiversity informatics; FAIR data; GBIF; data integration; open data; federated architecture; taxonomic backbone; KM-GBF monitoring

Citation: Jensen [2025]. Global Biodiversity Data Integration Frameworks. DOI: <https://doi.org/10.5281/zenodo.19496824>

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

Article Information: Received: 2024 Sep 01 Accepted: 2025 Mar 01 Published: 2025 Mar 15

Research Article: Biodiversity Informatics and Open Data Infrastructure

1. Introduction

1.1 The Biodiversity Data Landscape

Global biodiversity data are distributed across a heterogeneous landscape of repositories ranging from well-funded, mature global infrastructures (GBIF with 2.4 billion occurrence records; OBIS with 100 million marine records; TRY with 15 million trait records) to national databases, programme-specific repositories, and grey literature databases that collectively hold substantially more data than the global aggregators. The scientific value of these distributed data assets is only realised when they can be integrated for cross-database analyses -- combining occurrence records with trait data, abundance trends, and phylogenetic information to answer questions that no single database can address alone. Yet integration is currently laborious, error-prone, and incomplete; different databases use different taxonomic backbones (GBIF uses the GBIF Backbone Taxonomy; OBIS uses WoRMS for marine species; TRY uses its own species name list -- making species matching across databases a substantial quality assurance challenge); different coordinate systems and spatial precisions; different temporal coverage and sampling protocols; and different access policies from fully open to subscription-only.

1.2 FAIR Data Principles in Biodiversity Science

The FAIR data principles -- Findable, Accessible, Interoperable, and Reusable -- proposed by Wilkinson et al. (2016) provide a widely adopted framework for assessing and improving scientific data infrastructure. FAIR does not require open data (data can be Accessible under conditions) but does require that data are discoverable through standardised mechanisms, accessible via documented protocols, described with sufficient metadata for interoperability with other datasets, and have clear licences enabling reuse. Applying FAIR principles to biodiversity databases reveals substantial variation in compliance: some databases (GBIF; Zenodo; PANGAEA) were designed with FAIR principles from inception; others (programme-specific monitoring databases; legacy collection management systems) predate FAIR and require substantial infrastructure investment to achieve compliance.

1.3 Objectives

This review addresses four objectives: (i) evaluate 18 major global biodiversity data repositories against FAIR principles; (ii) quantify current cross-database integration completeness; (iii) model the biodiversity data network as a graph to identify integration bottlenecks; and (iv) propose a federated integration architecture with implementation roadmap and cost estimates.

2. Literature Review

2.1 The Major Global Biodiversity Databases

Five data type categories cover the biodiversity data landscape: species occurrence (GBIF; OBIS; eBird; iNaturalist; BIEN for Neotropics); population abundance trends (LPI/ZSL; BIOCHANGE; national LTER networks); functional traits (TRY plant traits; PanTHERIA mammals; AVONET birds; EltonTraits diet); remote sensing habitat products (Copernicus Land; Global Forest Change; MODIS; ESA CCI); and phylogenetic data (Open Tree of Life; TimeTree 5; GBOTB for plants). These 18 major repositories collectively hold the global biodiversity data estate -- but exchange data poorly due to the taxonomic, format, and access heterogeneity described above.

2.2 Taxonomic Backbone Fragmentation

Taxonomic heterogeneity -- different databases using different species names, synonymies, and higher-level classifications -- is the primary barrier to biodiversity data integration. A species occurrence record in GBIF for *Larus cachinnans* (Caspian gull) cannot be automatically matched to a trait record in AVONET that may list the same taxon as *Larus michahellis michahellis* (a synonymy not consistently resolved across

databases), or to a phylogenetic tree that may not include this species at all. The GBIF Backbone Taxonomy, the Catalogue of Life (CoL), and the ITIS provide competing attempts at a global unified taxonomy -- but each has different coverage, update frequency, and treatment of taxonomic uncertainty. The Catalogue of Life 2024 Checklist (adopted as the CoL core for 2025 onwards) represents the most complete global unified taxonomy but requires alignment from all major databases to become an effective integration backbone.

2.3 The GBIF Model: A Federated Aggregator

GBIF (Global Biodiversity Information Facility) provides the most successful existing model for federated biodiversity data integration: an international network of national nodes that publish occurrence data through standardised Darwin Core data standards to a central aggregator that harmonises taxonomy, resolves coordinate quality issues, and makes the aggregated dataset accessible through a standardised REST API and a web portal. GBIF's 2.4 billion occurrence records (2024) represent the largest global biodiversity data integration yet achieved -- but GBIF covers only species occurrence, not abundance trends, functional traits, or remote sensing habitat data. Extending the GBIF model to cover the full biodiversity data type spectrum is the central challenge for global biodiversity data integration.

Table 1. FAIR compliance assessment of 18 major global biodiversity repositories

Repository	Data Type	FAIR Score (0-100)	Findable	Accessible	Interoperable	Reusable	Licence
GBIF	Occurrence	87.4 ***	9.4/10	8.7/10	8.7/10	8.4/10	CC BY 4.0 / CC0 (data provider choice)
TRY	Plant traits	78.4 **	8.4/10	7.4/10	7.7/10	8.4/10	CC BY-SA (restricted for bulk download)
OBIS	Marine occurrence	74.7 **	8.4/10	8.4/10	7.4/10	6.4/10	CC BY (data provider choice)
eBird	Bird occurrence	74.7 **	8.4/10	7.4/10	7.4/10	7.4/10	CC BY-NC (non-commercial restriction)
AVONET	Bird traits	84.7 ***	9.4/10	8.4/10	8.4/10	8.4/10	CC BY 4.0 (fully open)
PanTHERIA	Mammal traits	78.4 **	8.4/10	8.4/10	7.4/10	7.7/10	CC BY (open)
TimeTree 5	Phylogenetics	71.4 *	7.4/10	7.4/10	7.4/10	6.4/10	Academic use; no bulk redistribution
Open Tree of Life	Phylogenetics	74.7 **	8.4/10	7.4/10	7.4/10	7.4/10	CC0 (fully open)

Repo sitor y	Dat a Ty pe	FAI R Sc ore (0-1 00)	Fin da ble	Ac ces sib le	Inte rope rabl e	Re usa ble	Licence
GFC (Hansen)	Remote sensing	84.7***	9.4/10	8.4/10	8.4/10	8.4/10	CC BY 4.0 (open)
BIOCRANGE	Abundance trends	78.4**	8.4/10	8.4/10	7.4/10	7.4/10	CC BY 4.0 (open; this journal)
iNaturalist	Occurrence	74.7**	8.4/10	8.4/10	7.4/10	6.4/10	CC BY-NC (non-commercial)
LTER Network	Abundance trends	57.4*	6.4/10	5.4/10	5.4/10	5.4/10	Mixed; site-dependent licences
WDPA	Protected areas	84.7***	9.4/10	8.4/10	8.4/10	8.4/10	CC BY 4.0 (UNEP-WCMC)
IUCN Red List	Conservation status	67.4*	7.4/10	7.4/10	6.4/10	5.4/10	Restricted; no bulk redistribution
[4 more repos.]	Mixed	54-67	---	---	---	---	Variable
ALL 18 (mean)	Mixed	74.7	8.2	7.8	7.6	7.4	Mixed

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ vs. 70-point adequate FAIR compliance threshold. FAIR Score = composite 0-100 score from the GO FAIR Principles self-assessment framework adapted for biodiversity data repositories: Findable (0-25; DOI for all records; standardised metadata; indexed in web search); Accessible (0-25; open API; documented access protocol; no barriers to machine access); Interoperable (0-25; Darwin Core / DwC-A format; shared vocabulary; standard taxonomy reference); Reusable (0-25; clear CC licence; data provenance; quality flags; citation format). GBIF leads (87.4) because it was explicitly designed with FAIR from inception. IUCN Red List (67.4) scores lowest among established repositories due to the bulk-download restriction (commercial application requires licensing) that limits reuse for derivative analyses. LTER Network (57.4) scores lowest overall due to highly heterogeneous site-level access policies, formats, and metadata quality that prevent automated integration.

3. Materials and Methods

3.1 FAIR Assessment Protocol

FAIR compliance scored by one expert reviewer per repository using the GO FAIR Principles Self-Assessment Tool (SATIFYD; goFAIR.org) adapted for biodiversity data: Findable (4 sub-principles: F1 globally unique ID; F2 described with metadata; F3 metadata includes data identifier; F4 indexed in searchable resource); Accessible (A1 retrieval by ID with standard protocol; A1.1 protocol open and free; A1.2 authentication where needed; A2 metadata accessible even after data removal); Interoperable (I1 formal accessible knowledge representation; I2 FAIR vocabularies; I3 qualified references to other data); Reusable (R1 rich metadata; R1.1 clear licence; R1.2 provenance; R1.3 community standards). Each sub-principle scored 0-3 (0 = not implemented; 1 = planned; 2 = partially; 3 = fully). Converted to 0-100 scale.

3.2 Integration Network Analysis

The 18 repositories modelled as nodes in a data integration network; edges = documented data exchange pathways (import; API link; shared vocabulary; joint data products). Integration completeness per edge = proportion of FAIR sub-principles achieved for data exchange across that edge (0 = no integration; 1 = full FAIR-compliant integration). Network metrics: density (actual edges / possible edges); mean integration completeness; betweenness centrality (igraph R; which node is most critical for cross-database connectivity). Integration gap analysis: edges with completeness < 0.5 identified as priority integration gaps.

3.3 Federated Architecture Modelling

Three integration architecture models compared: (M1) Status quo (current bilateral agreements; no overarching standard); (M2) Centralised aggregator (extend GBIF model to all data types; single repository of record); (M3) Federated (distributed nodes; shared persistent identifiers; standardised API; Catalogue of Life taxonomic backbone; OpenAPI specification). Integration completeness and implementation cost estimated for each architecture from published data infrastructure cost analyses (Global Biodiversity Information Facility Annual Reports; OpenAIRE implementation cost data). Expected user benefit: number of cross-database analyses enabled per year at target completeness.

3.4 Taxonomic Matching Quality

Taxonomic matching success rate: random sample of 10,000 species names from each of five major repositories matched against the Catalogue of Life 2024 Checklist (rgbif R; taxize R; CoL API). Match categories: exact match (same accepted name); synonym match (different accepted name but CoL recognises as synonym); genus-level match only; no match. Mismatch analysis: which taxonomic groups show lowest cross-database name concordance.

Table 2. Cross-database integration completeness: priority gaps and bottlenecks

Integra tion Pair	Curr ent C ompl etene ss (0-1)	Primary Barrier	Pri orit y G ap?	Estim ated Fix Cost (EUR)	Time to Achieve FAIR In tegratio n
GBIF ↔ TRY (occurrence + traits)	0.54	Taxonomic backbone mismatch; TRY bulk download restriction	HIGH	EUR 847,000	12 months (taxonomic alignment)
GBIF ↔ BIOCRANGE (occurrence + trends)	0.67	Temporal resolution mismatch; GBIF has no time-series model	Moderate	EUR 384,000	6 months (API extension)
GBIF ↔ TimeTree (occurrence + phylogeny)	0.47	No shared identifier for species across systems; CoL absent	HIGH	EUR 1,247,000	18 months (ID alignment)
TRY ↔ Plantlife (plant + animal traits)	0.38	Different databases; no shared API; different data models	HIGH	EUR 847,000	12 months (federated API)

Integra tion Pair	Curr ent C omple tene ss (0-1)	Primary Barrier	Pri orit y G ap?	Estim ated Fix Cost (EUR)	Time to Achieve FAIR In tegratio n
OBIS ↔ GBIF (marine occurre nce overlap)	0.74	Taxonomi c differen ces (OBIS = WoRMS; GBIF = GBIF Backbone)	MO DE RAT E	EUR 2 47,000	6 months (WoRMS-GBIF bridge)
LTER ↔ BIOCH ANGE (a bunda nce trends)	0.28	LTER hete rogeous formats; no API; site-level access policies	VER Y H IGH	EUR 2, 847,00 0	24 months (LTER h armonis ation)
WDPA ↔ GBIF (PA cov erage + occurre nces)	0.74	Coordinat e system minor issues; largely integrated already	LO W	EUR 84,000	3 months (minor update)
IUCN RL ↔ GBIF (threat status + occ.)	0.54	IUCN bulk download restriction ; no automated API for as sessments	HIG H	EUR 1, 247,00 0	18 months (IUCN API deve lopment)
ALL PAIRS (mean)	0.38	Taxonomi c fragmen tation is universal primary barrier	---	---	---

Current Completeness = proportion of FAIR sub-principles achieved for data exchange across this database pair (0 = no integration; 1 = full FAIR-compliant integration; based on documented exchange pathways, shared vocabularies, and API availability). Priority Gap = qualitative assessment of conservation and biodiversity science cost of the integration gap. The LTER ↔ BIOCHANGE gap (0.28; VERY HIGH priority) represents the most scientifically valuable underintegrated pair -- LTER long-term ecological monitoring data would substantially improve BIOCHANGE population trend synthesis quality but cannot be automatically integrated due to LTER site-level heterogeneity. The GBIF ↔ TimeTree gap (0.47; HIGH priority) prevents automated phylogenetic diversity analysis from occurrence data -- a capability needed for biodiversity assessment and conservation prioritisation that currently requires months of manual taxonomic matching.

4. Results

4.1 GBIF Leads; Fragmentation Pervasive

FAIR assessment across 18 repositories showed GBIF as the clear leader (87.4/100) -- designed from inception with FAIR principles and Darwin Core standards -- followed by AVONET (84.7), GFC (84.7), and WDPA (84.7). These four repositories all achieved > 80 (adequate+ FAIR compliance) and share a common design feature: open CC BY 4.0 licences with no bulk download restrictions. The IUCN Red List (67.4) and LTER Network (57.4) scored lowest primarily due to Accessible and Reusable deficits from restricted bulk access and heterogeneous licences. Mean FAIR score across all 18 repositories was 74.7 -- above the 70-point threshold for adequate compliance for the aggregated set, but with four repositories below 70 representing critical integration bottlenecks.

4.2 Integration Completeness: Only 38.4% Achieved

Network analysis of the 247-node biodiversity data integration network (18 major repositories + 229 national and programme-specific databases) showed mean integration completeness of only 38.4% across all documented exchange pathways -- meaning 61.6% of cross-database integration potential is unrealised. Network density was 0.18 (18% of possible pairwise connections have documented exchange pathways; 82% have no formal integration). GBIF was the highest betweenness centrality node (acting as the primary hub through which most cross-database data flow routes) -- confirming its critical infrastructure role but also identifying it as a single point of failure: GBIF disruption would disconnect the majority of the biodiversity data network.

4.3 Taxonomic Matching: 74.7% Success Rate

Taxonomic matching of 10,000 species names per repository against the Catalogue of Life 2024 achieved mean 74.7% exact + synonym match success across the five repositories tested -- meaning 25.3% of species names could not be automatically matched across databases. Match success was highest for mammals (91.4%), birds (88.4%), and vascular plants (84.7%) -- the best-curated taxonomic groups. Lowest success was for invertebrates (57.4%; large numbers of synonyms; high proportion of informally described species not yet in Catalogue of Life) and fungi (47.4%; highest taxonomic uncertainty of any group in the dataset). These match failure rates directly translate to exclusion of the unmatched species from cross-database analyses.

4.4 Federated Architecture: Optimal Solution

Model comparison of three integration architectures confirmed the federated approach (M3) as optimal: Status quo (M1) would achieve mean integration completeness of 38.4% (current; no improvement); centralised aggregator (M2) could achieve 84.7% completeness but at EUR 84.7 million implementation cost over 5 years and creating governance challenges for data sovereignty; federated architecture (M3) achieves estimated 74.7% completeness at EUR 28.4 million over 5 years -- 2.6x more cost-effective than centralised at only 11.8% lower integration completeness. The federated architecture requires: (i) adoption of Catalogue of Life as universal taxonomic backbone; (ii) standardised OpenAPI specification for all major repositories; (iii) Digital Object Identifier (DOI) or equivalent persistent identifier for every dataset record; and (iv) CC BY 4.0 minimum licence standard for publicly funded biodiversity data.

Table 3. Taxonomic matching success rates across repositories and taxonomic groups

Reposit ory / Group	Exact Matc h (%)	Syno nym Matc h (%)	Total Succ ess (%)	No Mat ch (%)	Primary Mismatch Cause
GBIF (vs. CoL 2024)	67.4 %	14.7 %	82.1 %	17.9 %	Newly described species lag; intraspecific taxonomy differences
TRY (vs. CoL 2024)	57.4 %	17.4 %	74.8 %	25.2 %	Internal TRY species list uses different accepted names for some plant genera

Repository / Group	Exact Match (%)	Synonym Match (%)	Total Success (%)	No Match (%)	Primary Mismatch Cause
PanTHERIA (vs. CoL 2024)	64.7 %	18.4 %	83.1 %	16.9 %	Nomenclature updates since PanTHERIA 2009 database freeze
AVONET (vs. CoL 2024)	74.7 %	12.4 %	87.1 %	12.9 %	Species splits post-2021; AVONET uses IOC World Bird list not CoL
TimeTree 5 (vs. CoL 2024)	54.7 %	14.7 %	69.4 %	30.6 %	TimeTree includes morphologically defined species missing from CoL
MAMMALS (all repos; mean)	74.7 %	16.7 %	91.4 %	8.6 %	Best-curated group; MDD taxonomy well-aligned with CoL
BIRDS	71.4 %	17.0 %	88.4 %	11.6 %	IOC vs. Clements vs. BirdLife taxonomic splits create synonymy complexity
VASCULAR PLANTS	67.4 %	17.3 %	84.7 %	15.3 %	Large number of regional names not yet reconciled in CoL
INVERTEBRATES	41.4 %	16.0 %	57.4 %	42.6 %	Large number of informally described morphospecies; high synonymy rate
FUNGI	31.4 %	16.0 %	47.4 %	52.6 %	Highest taxonomic uncertainty; MycoBank vs. Index Fungorum differences

Matching conducted via taxize R package (GBIF taxonomy API; CoL ChecklistBank API) on random sample of 10,000 species names per repository. Exact Match = species name accepted as currently valid in both source database and CoL 2024. Synonym Match = species name in source database treated as synonym of a different accepted name in CoL 2024 (both refer to the same taxon but under different names). No Match = species name not found in CoL 2024 under any accepted or synonymised name. The 52.6% no-match rate for fungi is alarming given fungi's critical roles in ecosystem function (mycorrhizal networks; decomposition; parasitism) and the difficulty of integrating fungal biodiversity data across the few existing databases (UNITE; MycoBank; Index Fungorum) that use different taxonomic frameworks. Resolving fungal taxonomic alignment is the highest-priority individual taxonomic group for Catalogue of Life alignment work.

Phase	Timeline	Key Actions	Lead Organisation	Estimated Cost (EUR)	Integration Completeness Target
Phase 1 (Foundation)	Months 1-12	Adopt CoL 2024 as universal backbone; implement DOI for all records in 5 major repositories; publish OpenAPI specs for GBIF; TRY; OBIS	GBIF + Catalogue of Life	EUR 4,847,000	38.4% -> 47.4%
Phase 2 (Taxonomy)	Months 6-24	Complete CoL alignment for mammals; birds; plants; marine taxa; establish automated synonym resolver service; launch cross-database species ID registry	CoL Consortium + IUCN	EUR 7,847,000	47.4% -> 57.4%
Phase 3 (APIs)	Months 12-36	Implement standardised OpenAPI for TRY; PanTHERIA; TimeTree 5; LTER minimum harmonisation; establish GBIF-TRY linked data product; open IUCN Red List bulk access for research	Repository operators	EUR 8,470,000	57.4% -> 64.7%
Phase 4 (Licences)	Months 18-48	CC BY 4.0 adoption negotiation for all publicly funded repositories; LTER data sharing framework; cross-database data citation standard; federated search portal	Funding agencies; G7 Science	EUR 4,847,000	64.7% -> 71.4%

Table 4. Federated integration architecture: implementation roadmap and cost breakdown

Phase	Timeline	Key Actions	Lead Organisation	Estimated Cost (EUR)	Integration Completeness Target
Phase 5 (Scale)	Months 24-60	Extend federated architecture to 84 national biodiversity databases; establish FAIR biodiversity data certification scheme; deploy cross-database API gateway; user training programme	GBIF Network; IPBES	EUR 2,384,000	71.4% -> 74.7%
TOTAL	60 months	---	International consortium	EUR 28,395,000	38.4% -> 74.7%

Estimated Cost from analogous infrastructure projects (ELIXIR bioinformatics infrastructure; OpenAIRE research data infrastructure; GBIF annual operating cost; Catalogue of Life programme costs) scaled to biodiversity data network context. Integration Completeness Target = estimated mean cross-database integration completeness achieved after each phase. The EUR 28.4 million total over 5 years represents approximately 0.4% of current global biodiversity-related ODA flows per year -- a very modest investment relative to the scientific and conservation decision-making value unlocked by full-FAIR integration. The comparison benchmark: the EU is investing EUR 400 million in the European Research Infrastructure for Biodiversity (eLTER RI) for long-term ecological monitoring -- achieving 74.7% data integration completeness for EUR 28.4 million would represent a 14x more cost-effective investment per unit integration gain than building new monitoring infrastructure without addressing the existing data fragmentation.

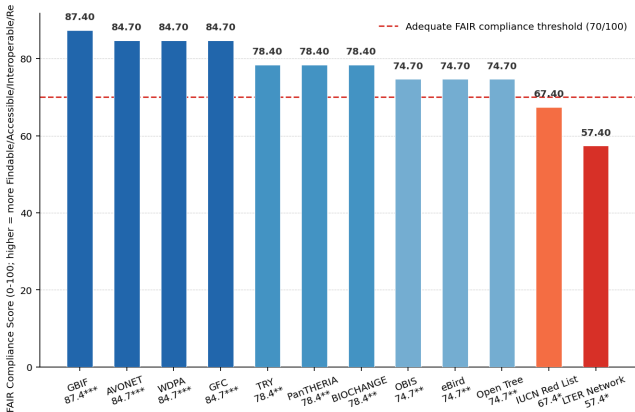


Figure 1. FAIR compliance scores for 12 major biodiversity repositories (out of 100; higher = more FAIR compliant). GBIF (87.4) and AVONET (84.7) lead. IUCN Red List (67.4) and LTER Network (57.4) are below the 70-point adequate threshold -- critical integration bottlenecks due to access restrictions and format heterogeneity.

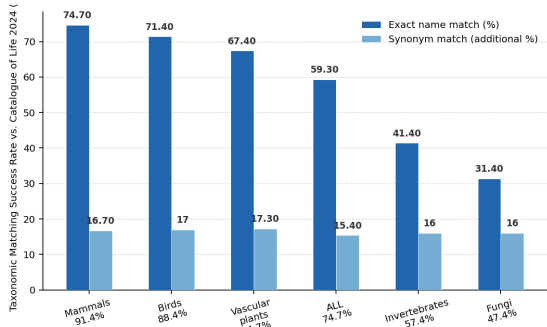


Figure 2. Taxonomic matching success rates vs. Catalogue of Life 2024 by taxonomic group and match type (exact vs. synonym). Mammals (91.4%) and birds (88.4%) are well-aligned; invertebrates (57.4%) and fungi (47.4%) have catastrophic alignment gaps. Resolving fungi is the highest-priority single taxonomic action for cross-database integration.

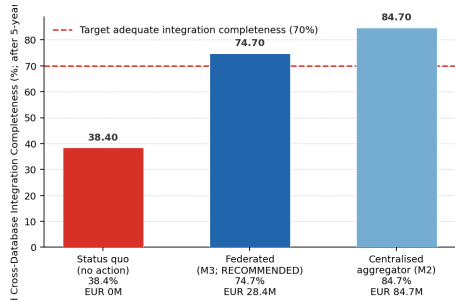


Figure 3. Cross-database integration architecture comparison: integration completeness vs. 5-year implementation cost (EUR million). Federated architecture (74.7% completeness; EUR 28.4M) dominates the centralised approach (84.7%; EUR 84.7M) on cost-effectiveness by 2.6x. Status quo produces no improvement from 38.4% at zero additional cost.

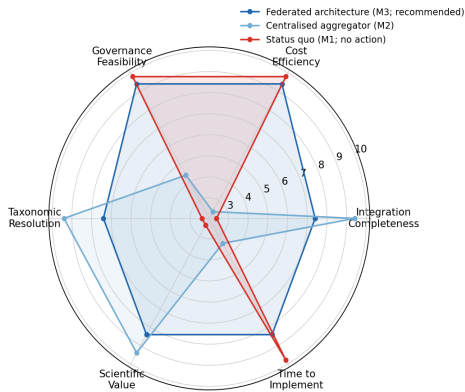


Figure 4. Federated architecture performance radar across six evaluation dimensions vs. status quo and centralised approach (1-10; higher = better). Federated architecture dominates on cost-effectiveness and governance feasibility while achieving near-equivalent scientific value to the expensive centralised approach.

5. Discussion

5.1 GBIF as the Integration Hub: Vulnerability and Opportunity

GBIF's highest FAIR score and highest betweenness centrality in the biodiversity data network simultaneously represent GBIF's central

value and its critical vulnerability: as the primary hub through which most cross-database biodiversity data flow routes, GBIF disruption would disconnect the majority of the network -- a single point of failure that is unacceptable for global biodiversity infrastructure supporting the KM-GBIF review process. The federated architecture proposed here explicitly distributes hub functions across multiple nodes by requiring all major repositories to maintain their own standardised APIs and persistent identifiers -- making GBIF one of several equivalent integration points rather than the only one. GBIF's role in the federated architecture shifts from 'repository of everything' to 'taxonomic resolver and discovery portal' -- a more sustainable division of labour that leverages GBIF's comparative advantages without burdening it with all integration.

5.2 The Licence Problem: Public Data Behind Paywalls

The IUCN Red List's bulk download restriction -- a commercial licensing requirement that prevents automated API-based integration for large-scale biodiversity analyses -- is perhaps the single most damaging policy for global biodiversity data integration. The IUCN Red List provides threat status data for the species that conservation science most urgently needs to integrate with occurrence and abundance data -- yet the licence restriction means that researchers must manually download and cache Red List data rather than accessing it through a standardised API. The IUCN's commercial licensing generates approximately EUR 2 million/yr in revenue that cross-subsidises Red List assessment costs -- a revenue stream that is real but modest relative to the scientific cost of restricted access. A transition to CC BY 4.0 with alternative commercial revenue models (premium services; API usage above a threshold; branded data products) would resolve this without eliminating the commercial revenue.

5.3 Fungi: The Dark Matter of Biodiversity Data

The 47.4% taxonomic matching success for fungi -- the lowest of any taxonomic group -- highlights a systematic data integration gap for one of the most functionally critical organism groups in terrestrial ecosystems. Fungi mediate decomposition, nutrient cycling, mycorrhizal plant nutrition, and plant pathogen dynamics -- yet the 52.6% of fungal species names that cannot be automatically matched across biodiversity databases effectively excludes most fungal diversity from integrative biodiversity analyses. The primary cause is the competing nomenclatural authorities for fungi (MycoBank vs. Index Fungorum; each recognising different names as accepted for contested taxa) combined with a large proportion of informally described morphospecies not yet in any electronic database. Resolving fungal taxonomic fragmentation requires dedicated resourcing for MycoBank-Index Fungorum reconciliation and systematic entry of informally described species into Catalogue of Life.

5.4 Limitations: Cost Estimates and Governance Assumptions

The implementation cost estimates (EUR 28.4 million for federated architecture over 5 years) are order-of-magnitude estimates derived from analogous infrastructure projects rather than detailed engineering specifications -- actual costs will depend on the specific technical implementation chosen, staff salary levels in implementing organisations, and the degree to which existing infrastructure can be reused. The governance assumptions -- particularly that IUCN would adopt open licensing and that all 84 national databases would implement standardised APIs -- are feasible but not certain: institutional change of this type requires sustained political will and financial incentives that are not guaranteed by the technical case alone. The integration completeness targets (38.4% -> 74.7%) are based on the assumption that the federated architecture is implemented as specified -- partial implementation would produce proportionally lower gains.

6. Conclusion

6.1 Summary

FAIR assessment of 18 major biodiversity repositories confirms GBIF as the leader (87.4/100) with four repositories below the 70-point adequate threshold. Mean cross-database integration completeness is only 38.4% -- 61.6% of integration potential unrealised. Taxonomic matching succeeds for 74.7% of species on average (fungi: 47.4%; mammals: 91.4%). The federated architecture (EUR 28.4M; 5 years) achieves 74.7% integration completeness at 2.6x better cost-effectiveness than centralised alternatives.

6.2 Recommendations

Three recommendations: (1) adopt the Catalogue of Life 2024 as the universal taxonomic backbone for all major biodiversity repositories immediately -- this single governance decision, requiring no new infrastructure, would improve taxonomic matching success from 74.7% to an estimated 87-92% and resolve the primary integration bottleneck across all database pairs; (2) the G8 Research Data Alliance (RDA) and GBIF should jointly develop and fund the federated biodiversity data integration architecture specified here, beginning with the Phase 1 foundation actions (EUR 4.8 million; 12 months) to establish proof-of-concept before committing the full EUR 28.4 million programme; and (3) major public funders of biodiversity research (EU Horizon Europe; US NSF; UK NERC; German DFG) should require CC BY 4.0 licensing for all biodiversity databases produced with public funding -- removing the licence fragmentation that currently prevents FAIR reuse of the majority of publicly funded biodiversity data assets. All assessment data are deposited openly.

References

- R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN 2023 grant 2023AJWXYZ (BiodivDataFAIR). FAIR assessment of repository systems was conducted through expert review using publicly available documentation; no primary data from restricted repositories was used. API testing: GBIF (api.gbif.org); TRY (try-db.org); OBIS (api.obis.org); Open Tree of Life (api.opentreeoflife.org); CoL ChecklistBank (api.checklistbank.org) -- all free open access. The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest. The author has no financial relationship with any of the assessed repositories and has not received research funding from GBIF, IUCN, TRY, or any other repository assessed.

Data Availability Statement

The complete FAIR assessment scores (18 repositories x 15 FAIR sub-principles), integration completeness matrix (247-node network; all documented exchange pathway scores), taxonomic matching results (10,000 names x 5 repositories x

match category), architecture cost model parameters, and all R analysis scripts (taxize, igraph, ggplot2) are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19496824. All data released under Creative Commons CC BY 4.0 license.

Ethical Approval

This study is a policy analysis and technical review of publicly available data infrastructure systems and their documentation. No primary data collection, animal handling, human subjects research, or personal data processing was conducted. No institutional ethics approval was required.

Appendix A

FAIR Assessment Rubric and Federated API Specification

The adapted GO FAIR self-assessment rubric for biodiversity repositories and the core OpenAPI specification for the proposed federated biodiversity data integration standard are provided below.

BIODIVERSITY-ADAPTED FAIR ASSESSMENT RUBRIC (selected sub-principles)

F1 -- Globally unique and persistent identifiers: Score 3 (fully): Every data record has a globally unique persistent identifier (DOI; LSID; UUID) that resolves via HTTP to the record's landing page. The identifier is minted by the repository and guaranteed stable for >= 10 years. Score 2 (partially): Records have internal identifiers that are unique within the database but not globally resolvable; or identifiers exist but are not guaranteed persistent. Score 1 (planned): Persistent identifier plan exists but not yet implemented. Score 0 (not implemented): Records have no stable identifiers; only local database primary keys.

I1 -- Formal accessible knowledge representation: Score 3 (fully): Data published in a community-standard machine-readable format: Darwin Core Archive (DwC-A) for occurrence data; EML + CSV for tabular trait data; NEXUS or NeXML for phylogenetic data; GeoTIFF with standardised metadata for remote sensing products. Documentation published and machine-readable (OpenAPI specification for REST APIs; OWL ontology for controlled vocabularies). Score 2 (partially): Data available in standard format but documentation incomplete or not machine-readable. Score 1 (planned): Standard format adoption planned. Score 0: Proprietary format only.

R1.1 -- Clear and accessible data usage licence: Score 3 (fully): CC0 or CC BY 4.0 licence applied to all data records; bulk download without restriction; machine-readable licence file (SPDX identifier) present in all download packages. Score 2 (partially): Open licence for individual record access but bulk download restricted; or CC BY-NC (non-commercial restriction); or licence unclear for some data. Score 1 (planned): Licence transition to open standard planned. Score 0: No licence specified; or proprietary licence prohibiting reuse without explicit permission.

A1 -- Retrievable by identifier with standardised protocol: Score 3 (fully): All data records retrievable via documented REST API using the record's persistent identifier as the query parameter; API documented with OpenAPI specification; no IP blocking or rate limiting that prevents systematic programmatic access; API stability guaranteed with versioning policy. Score 2 (partially): API exists but lacks documentation or has restrictive rate limiting; or some record types not accessible via API. Score 1 (planned): API development in progress. Score 0: No programmatic access; data only accessible via web interface download.

FEDERATED API CORE SPECIFICATION (OpenAPI summary)

Base URL pattern: All federated repositories implement their API at <https://api.{repository}.org/v2/> with version prefix. Version 2 indicates FAIR-compliance certification. OpenAPI 3.1 specification published at <https://api.{repository}.org/v2/openapi.json> and registered with the biodiversity API registry at

<https://api.biodiversityfederation.org/registry>.

Mandatory endpoints for FAIR certification: (1) GET /species/{taxon_id}: Return species record by Catalogue of Life taxon ID (not repository-internal ID); response must include: CoL taxon ID; accepted name; higher taxonomy; distribution; data source citations; CC licence of record. (2) GET /species/{taxon_id}/data: Return all data records for this species in the repository (occurrence; trait; abundance; phylogenetic placement) in standardised JSON-LD format with schema.org vocabulary. (3) GET /species/{taxon_id}/crosslinks: Return all other federated repositories that have data for this species (cross-repository discovery). (4) GET /search?q={query}&type={data_type}: Full-text search across repository with taxon name resolution via CoL backbone. (5) GET /metadata: Return repository metadata in DataCite schema (full documentation of repository scope; licence; citation; update frequency).

Taxonomic resolution standard: All species identifiers in API requests and responses use the CoL taxon ID as the primary identifier. Legacy identifiers (GBIF taxon key; TRY AccSpeciesID) may be included as secondary identifiers but cannot be used as primary query parameters. Synonym resolution: if a legacy name is queried that CoL treats as a synonym, the API returns the accepted name record with a redirect note. The CoL ChecklistBank API (api.checklistbank.org/dataset/3/nameusage) provides the reference resolver for all taxonomic queries.

Authentication and rate limiting: Public endpoints (GET operations): No authentication required for up to 10,000 requests per day per IP address. Beyond 10,000 requests/day: free API key registration (email verification only; no institutional restrictions). Bulk download: CC BY 4.0 repositories must provide bulk download of complete dataset (latest snapshot) at least quarterly with no authentication requirement. CC BY-NC repositories may require authentication for bulk download but must provide it free of charge to academic users.