
Event Agenda

TDWG 2026 Conference

Monday, September 21, 2026

CHECK-IN

8:30 AM – 9:00 AM | Location: Aula

Welcome

9:00 AM – 9:30 AM | Location: Aula

Address from TDWG Chair

9:30 AM – 9:45 AM | Location: Aula

Speaker



David Bloom

University of North Carolina at Greensboro

AI for Nature at Large Scale and High Resolution

9:45 AM – 10:30 AM | Location: Aula

AI AND ROBOT READY

Computation has fundamentally changed the way we study nature. New data collection technologies, such as GPS, high-definition cameras, autonomous vehicles under water, on the ground, and in the air, genotyping, acoustic sensors, and crowdsourcing, are generating data about life on the planet that are orders of magnitude richer than any previously collected. Yet, our ability to extract insight from these data lags substantially behind our ability to collect it.

The talk will discuss both recent advances and potential opportunities in AI that can turn these data into high resolution information source about living organisms, enabling scientific inquiry, conservation, and policy decisions. We will also address the challenges in scaling biodiversity research and action with AI, particularly in AI-ready cyber- and data infrastructures, as well as complex multimodal, multi-sensor, multi-scale data analysis, modeling, and visualizations.

Speaker



Tanya Berger-Wolf

Professor, Computer Science and Engineering | The Ohio State University

BREAK

10:30 AM – 11:00 AM | Location: Outside

Sámi perspectives on biodiversity Data Governance and more inclusive Data Standards

11:00 AM – 11:45 AM | Location: Aula

AI AND ROBOT READY

As more biodiversity data become available through global platforms such as GBIF, there is a growing need to consider how Indigenous rights, traditional knowledge, and community interests are reflected in these systems. For Sámi communities, biodiversity data are often closely connected to Sámi knowledge, traditional land use, livelihoods, and relationships with nature. This presentation introduces ideas behind a proposed Sámi-hosted biodiversity data portal and discusses how Local Contexts tags and labels might be used to make Sámi interests and connections to data more visible. Could these tools help provide information about provenance, cultural significance, appropriate use, and community expectations when or whether data are shared and reused?

The presentation will explore whether the CARE Data Maturity Model can provide a useful framework for assessing and guiding Sámi biodiversity data governance. Do their indicators related to collective benefit, control, ownership, and ethics reflect Sámi realities, or is there a need for additional approaches? We apply the most updated and transdisciplinary Sámi ethical guidelines as guiding principles for the project.

Finally, the presentation will discuss data repatriation and the importance of building Sámi-controlled archives, repositories, and data infrastructures. Returning data to communities is not only about access, but also about creating the capacity and governance needed to manage and use data in ways that benefit Saami communities.

The aim is to start a conversation about how these tools can support Sámi data governance in practice.

Speaker



Máret J. Heatta

Assistant professor | Samisk Høgskole/Sámi Allaskuvla

From biodiversity data to knowledge in the age of AI: Where do ontologies and standards fit?

11:46 AM – 12:30 PM | Location: Aula

AI AND ROBOT READY

Data – specimens and collections, human- and sensor-mediated observations, images, sequences, traits, and environmental measurements – provide the raw material for biodiversity informatics and broader life sciences, yet they come with questions: What do our data mean? Which concepts are equivalent, related, or distinct? How do we represent context, provenance, uncertainty, and

evidence? As new computational systems are applied to biodiversity data, a new question becomes increasingly visible and urgent: How do we support discovery and reuse without allowing fluent answers or plausible connections to obscure ambiguity, erase complexity, or reinforce errors? For four decades, TDWG has helped build the standards, ontologies, identifiers, metadata practices, and community agreements that make biodiversity data valuable. These are the same tools we need to create AI-ready biodiversity data.

Working across plant science, ontology development, and biomedical data management has reinforced one enduring lesson: the hardest data problems are rarely technical. They are problems of meaning, context, trust, and governance. AI does not eliminate these challenges—it magnifies them. Ontologies, standards, and community governance provide the shared semantic foundation that enables both people and machines to interpret data responsibly, transparently, and at scale. In an era when machines can generate answers faster than we can evaluate them, our greatest opportunity is not simply to build smarter AI, but to build richer, more trustworthy knowledge. The future of biodiversity informatics will depend not only on the intelligence of our algorithms, but on the quality of the shared knowledge infrastructure we create.

Speaker



Ramona Walls

Director Data Collaboration Center | Critical Path Institute

LUNCH

12:30 PM – 2:00 PM | Location: Outside

Contributed Oral (Plenary)

2:00 PM – 4:00 PM | Location: Aula

CONTRIBUTED ORALS

Selected presentations that give a taste of the week ahead

Moderator



Ely Wallis

Engagement Team Lead | Atlas of Living Australia, CSIRO

Building a New Foundation for the Biodiversity Heritage Library

2:00 PM – 2:15 PM

CONTRIBUTED ORALS

In 2025, the [Biodiversity Heritage Library](#) (BHL), one of the world's most critical resources of biodiversity knowledge, was placed at risk when its long-term host, the Smithsonian Institution, announced it would cease to host BHL's administrative and technical operations. This triggered a rapid, international response and a transition to a distributed, autonomous, community-supported global consortium.

In 2026, BHL is celebrating both its 20th anniversary and its first year operating independently. This presentation provides an overview of that transition, highlighting new governance and funding approaches, diversified technical and institutional support, and the role of transparent, community-driven communication in maintaining trust and engagement during a period of significant uncertainty.

With this new foundation in place, BHL is now well positioned to advance its role within the biodiversity informatics ecosystem as core infrastructure linking literature with taxa, specimens, institutions, and datasets across centuries. This creates new opportunities to enhance the accessibility, discoverability, and interoperability of biodiversity knowledge, supporting large-scale text and image mining, knowledge graph development, persistent identifier integration, and emerging AI-driven applications.

While BHL's long-term sustainability is not yet fully secured, it is rapidly filling with opportunities. As the opening presentation for this symposium, this talk sets the stage for the presentations that follow, framing BHL's evolution and the possibilities ahead, and inviting participants to engage with how biodiversity literature can be transformed into data, connections, and new forms of knowledge.

Speaker



Nicole Kearney

Communications Director | Biodiversity Heritage Library

Reducing the lag from observation to action

2:15 PM – 2:30 PM

CONTRIBUTED ORALS

Early detection and rapid response are among the most effective strategies for reducing the impacts of invasive alien species (IAS). Yet there is often a substantial lag between observing a species in the field and that information becoming actionable.

These delays arise throughout the biodiversity data lifecycle. Observations may remain in notebooks, local databases or disconnected systems. Identification can take time, particularly for difficult taxa. Data often require manual cleaning, standardisation and validation before publication. Even after publication, fragmented infrastructures and limited interoperability can slow the translation of biodiversity data into decision-ready information.

This presentation explores the concept of "lag" in biodiversity information systems and its implications for IAS management. Using examples from IAS monitoring and alert workflows, we discuss how delays propagate from field observation to policy uptake, and how these delays weaken early-warning capacity and reduce opportunities for rapid intervention.

Addressing these problems requires action on many fronts. Shared data standards, FAIR data mobilisation, interoperable infrastructures and automated publication workflows can shorten the path from detection to visibility. At the same time, progress also depends on strengthening taxonomic capacity, improving feedback, supporting near-real-time data publication, and creating incentives that reward timely sharing rather than delayed release.

Reducing lag is not only a technical challenge, but also an institutional and cultural one. Accelerating the movement of biodiversity observations into shared and reusable systems may be one of the most practical ways biodiversity informatics can support more responsive environmental governance.

Speaker



Quentin Groom

Sevior Researcher | Meise Botanic Garden

Keeping Colliders in Physics: Rethinking the Construction of Taxonomic Backbones and Names-matching in Large Data Aggregators

2:30 PM – 2:45 PM

CONTRIBUTED ORALS

A taxonomic backbone is essential to any biodiversity database displaying data in a systematic way for users. Perversely, one of the biggest challenges facing biodiversity databases is constructing, matching and displaying data using a taxonomic backbone. The Atlas of Living Australia (ALA) is a collaborative, digital, open infrastructure that pulls together biodiversity data from multiple sources. As with many countries, the taxonomy widely used in Australia is not always consistent with global views. This requires a national view of species names that government, researchers and industry can recognise. For many years, the ALA has constructed its taxonomic backbone using a process colloquially known as the "Large Taxon Collider" (LTC) that automatically combines taxonomic checklists from different sources. For the past 2 years, the ALA has done an overhaul of its taxonomic backbone, seeking to remove erroneous combinations, repetitions and align with nationally recognised taxonomic checklists centred upon the Australian Biological Resources' National Species List. This reduced the number of "accepted names" (not including synonyms) from over 500,000 to 233,000 and introduced governance. We found the LTC was valuable in combining different sources yet automation caused structural issues partly caused by the contributing taxonomic checklists themselves. This enabled us to disentangle these from issues arising from matching data to the taxonomic backbone using names-matching. The Catalogue of Life's ChecklistBank is now an essential part of our process, as are matching metrics. We have re-considered our approach to taxonomy, names, matching homonyms and how we measure success. We will discuss these issues, responses the ALA has developed so far, as well as next steps.

Speaker



Cameron Slatyer

Atlas of Living Australia CSIRO

Authors

Cam Slatyer (Atlas of Living Australia, CSIRO), Simon Sherrin (Atlas of Living Australia, CSIRO)

DwC-DP Integration at GBIF

2:45 PM – 3:00 PM

CONTRIBUTED ORALS

Following the public review of the Darwin Core Data Package (DwC-DP) standard, the GBIF Secretariat has committed to integrating DwC-DP into its existing data publishing infrastructure and ingestion pipeline. This talk presents recent engineering progress across four key areas:

IPT Extensions: The Integrated Publishing Toolkit (IPT) is being extended to handle data packages, enabling publishers to map and publish DwC-DP datasets through a familiar interface.

Data Validation: GBIF's validation services are being expanded to support DwC-DP, offering publishers meaningful quality metrics, including column population statistics, primary key uniqueness checks, referential integrity validation across schemas, and data type conformance reporting.

Registry and Metadata Synchronization: The GBIF Registry will automatically synchronize dataset metadata from Data Package descriptors, ensuring records remain accurate and up to date without manual intervention.

Ingestion Pipeline: The core ingestion pipelines are being updated to index occurrence and sampling-event records, together with their related extensions, published via DwC-DP. This work lays the foundation for broader community adoption of DwC-DP.

Together, these developments mark a significant step toward full DwC-DP support across the GBIF data publishing lifecycle, from authoring and validation to registration and indexing.

Speakers



Federico Mendez

Head of Engineering | GBIF Secretariat



Mikhail Podolskiy

GBIF Secretariat

Authors

Mikhail Podolskiy, Ahmet Hakan Uzun, Nikolaj Schaldemose Reibke

Using automation to consolidate biodiversity occurrence data from multiple sources to assist conservation decision making in South African National Parks.

3:00 PM – 3:15 PM

CONTRIBUTED ORALS

South African National Parks manages 21 National Parks within South Africa that includes more than 4 million hectares spanning both the Marine and Terrestrial environments. To support evidence-based decision making at park level data from multiple sources need to be consolidated allowing users and park managers to easily access this consolidated data through dashboards. Automation and AI allows consolidation of open access data from sources outside of the organisation, as well as open access data from parks like collection data and sensitive park specific data to be viewed and used together by decision makers. APIs allow automated harvests of data from platforms like GBIF as well as information around conservation statuses of species from IUCN sites which is required for managing species at park level. Additional features allow integration of the data using selection features like selecting a certain park/area, taxonomic group, time frame or conservation status without any prior knowledge of GIS or data integration needed. Further features allow a user to download climate data for a particular area, create species lists and obtain additional information and graphs on the conservation status of taxonomic groups in these areas and how these have changed over time. This presentation will showcase this system and highlight how it uses AI and automation to generate easily accessible evidence.

Speaker



Judith Botha

Senior Manager: Science Support | South African National Parks

Extending Darwin Core to Represent Sociobiodiversity and Traditional Knowledge

3:15 PM – 3:30 PM

CONTRIBUTED ORALS

Biodiversity data standards increasingly support automation, artificial intelligence, and large-scale interoperability. However, current standards still provide limited mechanisms to represent sociobiodiversity data, particularly traditional ecological knowledge, species uses, and the communities that generate and maintain this knowledge. As a result, much community-based biodiversity knowledge remains difficult to integrate into interoperable and machine-actionable infrastructures.

This contribution presents the ongoing development of a Darwin Core (DwC) extension for sociobiodiversity data, coordinated by the Brazilian Biodiversity Information System (SiBBR) and collaborators. The proposed extension aims to improve the semantic representation of biocultural information associated with species uses and traditional knowledge systems. Proposed terms include species use categories, plant parts used, preparation methods, medicinal and cultural applications, associated communities, territorial context, and provenance information.

The initiative recognizes that sociobiodiversity data are often generated through participatory and ethnobiological practices that require ethical and inclusive governance approaches. The work aligns not only with FAIR principles, but also with concerns regarding data sovereignty, attribution, and responsible reuse of traditional knowledge.

The proposal is being developed collaboratively through open discussions, use-case analyses, and iterative vocabulary design. Initial case studies involve ethnobotanical and historical datasets related to medicinal plants and traditional ecological knowledge from Brazilian coastal restingas and other socioecological contexts in the Global South.

Speaker



Clara Baringo Fonseca

Biodiversity Data Specialist at SiBBR | SiBBR (Brazilian Biodiversity Information System - RNP)

Authors

Keila Juarez

What if biodiversity data systems were designed to behave like ecosystems?

3:30 PM – 3:45 PM

CONTRIBUTED ORALS

Despite advances in global infrastructures such as the GlobalBiodiversity Information Facility (GBIF), a major gap remains between remote sensing data and biodiversity records. Most satellite and airborne observations begin as raw digital numbers and move through rigid, fragmented pipelines that rarely translate cleanly into interoperable, GBIF-ready data.

This talk reframes that problem through a systems lens inspired by ecological cascades. In ecosystems, small changes propagate through interconnected components, reshaping behavior, and structure over time. I argue that biodiversity data should function similarly and not as static products, but as events that trigger downstream processing and integration.

Building on this idea, I will introduce an event-driven, microservices-based approach to biodiversity data pipelines. Modular services handle key transformations (from radiance to reflectance, to derived products, to biodiversity integration) while responding dynamically to new data. This enables a continuous, traceable pathway from sensor-level measurements (L0/L1) to analysis-ready outputs (L2/L3) and ultimately to GBIF-ready records.

By shifting from static workflows to adaptive systems, this approach better supports reproducibility, interoperability, and cross-institutional collaboration. Rather than reducing complexity, it embraces it by designing biodiversity infrastructures that can evolve, respond, and scale alongside the ecosystems they represent.

Speaker



Kit Lewers

PhD Candidate, FINESST Fellow | CU Boulder, NASA

Mainstreaming high-throughput biodiversity monitoring using AI

3:45 PM – 4:00 PM

CONTRIBUTED ORALS

The rapid development of AI image and sound recognition technology in combination with reduced costs of camera traps and passive acoustic monitoring (PAM) devices opens opportunities to mainstream data driven species monitoring. Several user portals have been developed over the past years that bring these AI technologies to new user communities like ecologists and managers of natural resources. ARISE is the acronym for “Authoritative and Rapid Identification System for Essential biodiversity information” and is one such initiative that was launched in November 2025. ARISE provides end-users with a data management system for camera-trap, insect camera and PAM monitoring data and a library with different AI algorithms that can be used to analyse the data. At present 10 AI algorithms are available ranging from MegaDetector to remove images with humans or vehicles to SpeciesNet, DeepFaune and Trapper AI for camera-trap images and AvesEcho-v1 for PAM sound recordings. The number of available AIs is rapidly increasing.

The ‘Biodiversity Meets Data’ or BMD project aims to make these technologies available to monitor the state of conservation of species and habitats that are protected under the EU Nature Directives that are implemented in the Natura 2000 and Emerald Networks, together covering more than 30,000 protected sites across Europe. The managers of natural resources are BMD’s main stakeholder community that are provided with these new technologies. Directing our focus to the feasibility of wider adoption, we discuss the performance metrics of these automated pipelines alongside an assessment of their financial and computational costs—including model training and cloud operation —providing an example for scalable, economically viable AI-driven nature conservation.

Speaker



Niels Raes

Programme Manager Biodiversity Informatics | Naturalis Biodiversity Center

Authors

Freek de Bruijn, Khadija Ezzaki-Chokri

WELCOME RECEPTION

5:00 PM – 7:00 PM | Location: Oslo City Hall

Tuesday, September 22, 2026

CHECK-IN

8:00 AM – 8:30 AM | Location: 1st Floor

SYM30A - AI for Biodiversity Data

8:30 AM – 10:30 AM | Location: SAL A

AI AND ROBOT READY

Biodiversity researchers today are offered access to vast troves of data: billions of images and data points in online repositories, citizen scientists contributing millions of new observations every year, countless specimens in museum collections around the world, and an endless supply of studies, reports, and expert knowledge in articles and books stored in libraries or published online.

But data alone is not enough. To have any meaning it must be collated, organised, standardised, evaluated and interpreted. With such quantities of information, machine learning and artificial intelligence tools are vital to our ability to parse and understand the data available to us, and essential if we are to scale our research to meet the urgent biodiversity challenges the world faces today.

This session invites talks on cutting-edge applications of AI technology in automating the collection, preparation and interpretation of biodiversity data. We welcome contributions on topics such as novel applications of AI in data collection and mobilisation, large-scale studies enabled by AI, automation to extract data and new insights from historical collections and archives, and systems that improve our ability to arrange and interpret large biodiversity datasets.

We strongly encourage speakers to include the financial and computational costs of model training and operation, in addition to performance metrics, in discussing their work to allow assessment of its feasibility in wider use.

Moderators



David Williamson

Postdoctoral Researcher | NTNU University Museum



Wouter Koch

Senior adviser biodiversity informatics | Norwegian Biodiversity Information Centre

Session organizers



Arianna Salili-James

Natural History Museum UK



Rukaya Johaadien

GBIF Norway



Kit Lewers

PhD Candidate, FINESST Fellow | CU Boulder, NASA



Michal Torma

Head engeneer | University of Oslo



Maarten Trekels

Biodiversity Data Scientist | Meise Botanic Garden

Introduction

8:30 AM – 8:45 AM

AI AND ROBOT READY

Speaker



David Williamson

Postdoctoral Researcher | NTNU University Museum

AI Isn't Magic: Staying scientific in biodiversity research

8:45 AM – 8:57 AM

AI AND ROBOT READY

Recent developments in artificial intelligence have produced a flood of publications as researchers apply new tools for organising and understanding knowledge to their own fields of interest. In biodiversity research, advocates of LLMs and related technologies promise to convert museum collections and citizen science observations into annotated datasets at a mouse click; digitise written records faster and better than ever before; and automate synthesising the sum of scientific knowledge into answers to our most pressing questions. Critics say that such claims amount to marketing hype: the inherent susceptibility of such models to hallucinations, the biases in their training and use, and their inability to distinguish fact from fiction preclude them from being a reliable research tool.

How do we reconcile the obvious promise of such models with their clear deficiencies? Where generative AI is sometimes treated as an oracular "black box" from which truths appear, I argue that like previous developments in machine learning its workings are understandable, its applicability is specific and narrow, its capabilities both knowable and quantifiable. We have a responsibility to measure the reliability and accuracy of our models and to report on their costs and limitations. Such reporting should include definitions for ground truthing; broadly accepted accuracy metrics; the financial, computational and environmental costs of our models; and transparency and accountability around the training data that goes into them. By reference to established norms in other areas of machine learning I present an outline for practical reporting practices, and aim to catalyse discussion around incorporating these promising technologies into our research without compromising scientific integrity.

Speaker



David Williamson

Postdoctoral Researcher | NTNU University Museum

Delivering AI for biodiversity data in the public sector: data handling, security and humans-in-the-loop when small teams build at pace

8:57 AM – 9:09 AM

AI AND ROBOT READY

AI is reshaping how biodiversity data is collected, prepared and interpreted. It is also reshaping how the tools that do this work get built: practices loosely described as “vibe coding” or AI-assisted programming have changed what small teams can deliver, and how quickly. For public-sector bodies such as natural history museums, the opportunity is significant but the constraints are distinctive. Data governance, security baselines, accountability to research users, and the long-term stewardship of biodiversity records all shape what is possible.

This talk offers a practitioner's perspective from the Natural History Museum, London, on applying AI in the development of biodiversity platforms in a public-sector context. It draws on experience building the AWS Data Ecosystem, the Museum's cloud-based platform for UK biodiversity data. The platform ingests biodiversity records gathered over time and streamed continuously from a range of sources, and serves research scientists as its primary user community.

Two areas will be in focus. The first is data handling and security, including the questions raised when sensitive or pre-publication biodiversity data meets AI services in a public-sector environment. The second is human-in-the-loop design, and how to structure workflows so research scientists retain scientific authority over AI-assisted outputs while still benefiting from the productivity AI offers. The talk will also consider the financial and computational costs of AI-enabled approaches, particularly where cloud-based services are used operationally rather than experimentally.

The talk is practice-oriented. It aims to share what is working, what is not, and where shared guidance from the TDWG community would help most.

Speaker



Rachel Wiles

Technical Product Manager | Natural History Museum, London

Scalable Edge AI for Real-Time Biodiversity Monitoring: A Case Study on Tracking Invasive Species in Roadside Imagery

9:09 AM – 9:21 AM

AI AND ROBOT READY

Automated biodiversity monitoring is essential for tracking changes in species distributions and abundance, yet deploying cutting-edge AI on high-resolution imagery to detect diverse taxa remains challenging. While classical object detection models (e.g., YOLO) require exhaustive localized annotations that are infeasible to scale across massive taxonomies encompassing hundreds of thousands of plant species, global multi-label models offer a powerful alternative for complex species identification in a single forward pass.

In this work, we present a case study on continuous roadside monitoring to track invasive plants in near real-time. To capture localized invasions, vehicle-mounted cameras operate at high frequencies to guarantee dense spatial sampling. To develop an edge-deployable pipeline, we initially employed a multi-scale tiling framework with a Pl@ntNet model trained on global flora to automatically annotate complex roadside scenes. Through a process called knowledge distillation, these annotations are used to transfer the recognition capabilities of the heavy tiling framework (acting as a 'teacher') into a series of compact, fast models (the 'students'). These edge-deployed student models achieve performance comparable to the computationally demanding baseline across several datasets from different European countries.

We evaluate these student models on both species identification metrics and computational costs, highlighting critical operational trade-offs between image resolution, model parameters, and real-time latency. Ultimately, this framework paves the way for scalable, cost-effective ecological monitoring adaptable to diverse high-throughput data sources, including UAV surveys, plot imagery, and other extensive field campaigns.

Speaker



Giulio Martellucci

INRAE, UMR LISAH

Authors

Giulio Martellucci, Guillaume Mougeot, Fabrice Vinatier, Pierre Bonnet, Toke Thomas Høye, Alexis Joly, Hervé Goëau

AI-vision for integrated management of pest insects and their natural enemies

9:21 AM – 9:33 AM

AI AND ROBOT READY

Precision farming is being adopted by many farmers, but the number of reliable tools for controlling weed, pest invertebrates and disease in arable cropping system is still quite limited. Most decision support models for plant protection against invertebrate pests consider the pest invertebrate, the plant and abiotic factors affecting them. Natural enemies of invertebrate pests are seldom considered as a factor in the decision support models. Recent advancements in precision technologies used in weed management rely on machine vision models and algorithms where weeds are identified from the cultivated plant and the chemical herbicides are applied only where weeds are identified, without broadcast uniform spraying an entire field, resulting in a 40-97% reduction in the use of chemical herbicides. Our focus in this project is therefore to reveal whether similar techniques could be adapted for precise control of pest insects while protecting and conserving their natural enemies. Our aim is to develop an Artificial Intelligence (AI) workflow that utilizes identification of pest insects and their natural enemies identification for automatic crop management systems. This work begins with the creation of a labelled and annotated image dataset for selected crops with the key pests and natural enemies. Thereafter, the dataset will be used to identify the best modelling toolbox for AI-based machine vision model for automatic species identification of pests and their natural enemies. The result will serve as basis toward designing a framework for pest monitoring and conservation biological control in the selected crops at field level.

Speaker



Ritter Guimapi

Research Scientist | NIBIO

Authors

Arianna Salili-James, Bjørn Ingeberg Fesche, Aleksander Sendrowicz, Vegard Line, Ingeborg Klingen

Soundscape China - Listening to the Sounds of Nature with AI

9:33 AM – 9:45 AM

AI AND ROBOT READY

With the rapid advancement of artificial intelligence and acoustic sensing technologies, the use of sound for biodiversity monitoring has evolved from localized experiments to widespread application. The construction of a national intelligent monitoring network based on natural sounds has become a key research direction in the field of ecological monitoring. The "Soundscape China" project is precisely centered on the goal of intelligent natural soundscape monitoring. It addresses multiple critical components—ranging from sensor development and intelligent species identification modeling to the construction of big data platforms—and is dedicated to creating an integrated technological chain that encompasses "acquisition, recognition, and management." The project strives to provide a standardized, replicable, and scalable model for the industry.

To date, Phase I of the "Soundscape China" project has deployed over 120 acoustic monitoring devices nationwide, which have automatically collected and aggregated more than 2 million valid recordings. The monitoring coverage spans diverse ecosystem types, including forests, wetlands, farmlands, grasslands, and deserts, thereby providing robust data support for national biodiversity monitoring. The project has been officially selected for inclusion in the United Nations' "Decade of Science" program. In the future, it aims to further expand the intelligent natural soundscape monitoring network, while also deepening cross-disciplinary practices such as citizen science education and sound art. This will promote multi-disciplinary synergy and foster an innovative framework of positive interaction among technology, ecology, and society.

Speaker



Congtian Lin

Director of Scientific Data Center | Institute of Zoology, Chinese Academy of Sciences

Cross-Modal Coordination for Biodiversity Monitoring at Scale: Lessons from SmartWilds

9:45 AM – 9:57 AM

AI AND ROBOT READY

No single sensing modality is sufficient to monitor biodiversity at scale. Camera traps miss vocalizing birds, acoustic recorders miss silent species, drones struggle with species ID from altitude, and citizen science is opportunistic. Each modality has its own mature AI ecosystem (BirdNET, BioCLIP, MegaDetector, foundation models for drone imagery, iNaturalist), but automating biodiversity data collection, preparation, and interpretation at scale requires moving beyond single-modality pipelines toward cross-modal coordination, with smart sensors making decisions at the edge.

Coordinating across modalities is fundamentally a multi-scalar problem across both time and space. Standardizing these streams into datasets that support multimodal AI requires AI itself: foundation models for cross-modal alignment, LLM-assisted metadata extraction, and shared schemas like Darwin Core that make the result machine-actionable. We present SmartWilds, a co-located multimodal deployment at The Wilds Conservation Center, Ohio, combining camera traps, autonomous recording units, drones, citizen science, and field text across 220 hectares, with 380K Darwin Core occurrences and complete sensor configuration metadata.

We evaluate where each modality-specific AI pipeline is mature and where it is not, with drone species ID from altitude showing the largest gap, bird vocalization classification being robust but limited to vocal taxa, and camera-trap individual re-identification at scale remaining unsolved. We report annotation effort, computational cost, and known failure modes alongside performance, supporting honest assessment of which components are ready for edge deployment and which require centralized compute or human review.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Jenna Kline

Postdoctoral Associate | Massachusetts Institute of Technology

Authors

Anirudh Potlappally

Trapper 2.0: scalable, open-source platform for managing camera trapping projects with integrated AI pipelines

9:57 AM – 10:09 AM

AI AND ROBOT READY

Camera trapping has become a powerful tool for wildlife monitoring, but the resulting multimedia datasets, often comprising millions of images and videos, present significant analytical challenges. Manual annotation is time-consuming, expensive, and difficult to scale, making AI integration essential for modern biodiversity monitoring workflows.

Trapper is an open-source, web-based platform designed to manage camera trap data at any scale, from individual research projects to large citizen science initiatives. It allows users to build automated AI analysis pipelines consisting of a variety of open-source AI models used for object detection, tracking objects in videos, species classification and distance and movement speed detection. Trapper's architecture allows for easy integration of new models as they are being developed.

The results of AI analysis can be reviewed and corrected by human experts. This human-in-the-loop approach maximises annotation efficiency without sacrificing data quality. The recently released version, Trapper 2.0, features support for processing videos as well as images. Trapper is compatible with the Camtrap DP data standard, ensuring interoperability with GBIF and broader biodiversity data infrastructures.

An important feature of Trapper's design is deployment flexibility. The entire Trapper stack, including the AI processing components, can be run on a personal computer, a local server or a cloud server, as well as on microservers like Raspberry Pi. AI processing can be done using CPU alone, with GPU acceleration or using Hailo-based processing, enabling computation on edge devices. This helps to significantly decrease computation costs, making AI-assisted biodiversity monitoring more accessible even with limited resources.

Speaker



Karolina Kuczkowska

Software developer, open data specialist | Open Science Conservation Fund

Authors

Jakub Witold Bubnicki, Izabela Stachowicz

Discussion

10:09 AM – 10:30 AM

AI AND ROBOT READY

Discussion for SYM30A

SYM23 - To see or not to see - that is the question. Managing restricted access data (RASD)

8:30 AM – 10:30 AM | Location: SAL B

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

This session seeks to explore the issues around solutions that data providers and conservation managers have encountered in handling Restricted Access Species Data (RASD) as well as providing an overview of the process of moving towards the ratification of RASD metadata extensions. Sensitive data (also known as Restricted Access Species Data - RASD) has been a long-running issue for data managers and is challenging for conservation decision-makers. The TDWG Sensitive Data Extension SubGroup has been working for the past 12 months on developing vocabularies for RASD treatments and reasons and a DWC RASD extension and DWC DP extension. The Task Group has proposed moving towards using RASD as a term (rather than sensitive species data) to escape many of the pre-conceived notions of what sensitive means. The intention of the extension is to improve the visibility of the reasons why data has been modified and to standardise the reasons and generalisations that have been applied to data, in alignment with FAIR principles. Session attendees will gain knowledge of a range of approaches for addressing challenging RASD decisions and will gain greater understanding of the vocabulary development efforts.

Session organisers will consider abstracts on the themes of issues around solutions that data providers and conservation managers have encountered in handling RASD and those providing an overview of the process of moving towards the ratification of RASD metadata extensions.

Moderators



Tania Laity

Atlas of Living Australia (CSIRO)



Cameron Slatyer

Atlas of Living Australia CSIRO

Session organizers



Simon Sherrin

Data Analyst Developer | Atlas of Living Australia



Martin Kaehrle

PhD Student | University of Wisconsin-Madison

Authors

Kristin Eschenfelder

Role-Based Access Control (RBAC) for RASD: The Flemish Biodiversity Portal Approach

8:30 AM – 8:45 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Managing Restricted Access Species Data (RASD) becomes complex when regional portals must balance local sensitivity requirements with global data sharing through aggregators like GBIF. The Flemish Biodiversity Portal (VBP), built on the Atlas of Living Australia (ALA), addresses this by utilizing a "twin dataset" architecture, ensuring that high-resolution data remains secure while a more generalized version is made publicly available. This strategy prevents the exposure of precise coordinates on global platforms while the dataset structure, metadata and most of the identifiers stay the same.

In this presentation, we detail how VBP manages these twins using Role-Based Access Control (RBAC) logic embedded within the `dynamicProperties` field. By utilizing machine-readable flags—such as `"rbac":true`, `"rbac_allowed":"HIGHRES"`—the portal dynamically determines data visibility. Users with authorized roles access the high-resolution "twin," while the general public is directed to the low-resolution version.

This dual-dataset approach ensures that "not seeing" high-resolution data is a deliberate, governed choice rather than a technical limitation. We will discuss the synchronization challenges inherent in maintaining twins and how this workflow aligns with the TDWG RASD extension's goal of standardizing the reasons and treatments for data restriction. By providing a transparent metadata trail for why a dataset is restricted or generalized, VBP demonstrates a scalable model for regional nodes to contribute to the global biodiversity data commons without compromising local conservation priorities.

Speaker



Dimitri Brosens

Coordinator Flemish Biodiversity Portal | Research Institute for nature and forest

Authors

Stefan Van Dyck

Restricted access species data in Finland

8:45 AM – 9:00 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

The Finnish Act on the Openness of Government Activities (621/1999) necessitates that species data in Finland is not made publicly available if it would endanger the protection of the species or area in question. The Finnish Biodiversity Information Facility (FinBIF) is a service for researchers, government and the public. The data harmonised in FinBIF's data warehouse includes the results of national monitoring programs, scientific collections and citizen science projects. FinBIF is the body primarily responsible for managing Finnish restricted access species data (RASD). FinBIF coordinates an expert panel, the Public Authorities Working Group, that maintains a list of taxa that require some level of data restriction. The Public Authorities Working Group determines the level of restriction and circumstances under which it should be applied. Each taxon included on the list must be subject to a concrete threat: persecution, harvesting, or disturbance. The list was first compiled in 2015, and a comprehensive update was carried out in 2025. Minor updates are made as needed. There are currently 155 listed taxa. Species data that is mobilised by FinBIF has restrictions applied on being harmonised in the data warehouse, before being made publicly available. The unrestricted data is made available in a parallel data warehouse that can only be accessed by registered users from statutory public authorities. Other data users may make a request for access to restricted data. Requests are fulfilled via a data access platform on a case by case basis with the consent of the data providers or relevant proxies. We outline the system for handling RASD in Finland and discuss how current and proposed biodiversity data standards may aid data providers and users.

Speaker



Will Morris

GBIF Finland Node Manager | Finnish Museum of Natural History

Authors

Eija-Leena Laiho

Restricted Access Species Data: The Swedish Model

9:00 AM – 9:15 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

For some 20 years, Sweden has used a central system, UserAdmin, for managing users and their roles. Certain roles decide which restricted access species data (RASD) a user is authorized to view. Out of the more than 158 million records in the Species Observation System (SOS) about 1.5 million are protected, belonging to 111 different species.

Records are automatically obscured if a species belongs to any of classes 3–5 on the national protection list. Additionally, a reporter can choose to manually hide any of their observations up until a future date of their choice. These will be handled as if they belong to class 3.

A protected observation is still visible to the reporter and the observers. They can also be available to authorized verifiers, authorities and agencies.

How can RASD be utilized?

- 1) A user can ask authorized staff in the local region if there is anything of concern in a certain area. They can give advice without revealing any sensitive data.
- 2) Upon authorization, some authorities/agencies can use Signal Search that answers yes/no if anything can be found that matches a broader taxonomic and geographic selection.
- 3) Restricted access can only be granted a person, who will then be screened in accordance with Swedish law and must sign a non-disclosure contract. This person is either allowed to do their own searches or to receive specific exports made by authorized staff. Receiving such exports is free of charge for official authorities but comes with a fee for commercial requests.

All observations in SOS can be accessed by logging in to either the web application Fynddata (fynddata.artdatabanken.se), maintained by SLU Swedish Species Information Centre, or Artsök (www.arkoll.se/artsok), a GIS based system managed by the County Administrative Board.

Speaker



Johan Liljeblad

Taxonomist | SLU Swedish Species Information Centre

Authors

Marit Persson Rådén, Mora Aronsson, Debora Arlt

Toward RASD Implementation in Brazil: Integrating Conservation, Open Data, and Traditional Knowledge Safeguards

9:15 AM – 9:30 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Brazil, a megadiverse country with extraordinary biological and sociocultural diversity, faces complex challenges in managing and publishing Restricted Access Species Data (RASD). Open-access platforms strengthen science, conservation, and policy, but also amplify risks of misuse of sensitive information.

Sensitivity varies across taxonomic groups and use contexts. Key concerns include illegal hunting, wildlife trafficking, predatory collection of commercially valuable species, and pressure on threatened, endemic, or low-density populations.

A further dimension involves species linked to associated traditional knowledge (ATK) of Indigenous peoples and local communities. Frameworks such as the CBD, Nagoya Protocol, ILO Convention 169, and Brazil's Law 13.123/2015 require free, prior, and informed consent for use of these resources—rarely met in practice. They provide important safeguards but also reveal limits: fragmented implementation, weak traceability, and uneven recognition of community protocols.

Paradoxically, traditional territories remain a critical gap in national repositories, despite their recognized role in conservation. Precise occurrence data are essential for monitoring, enforcement, and policy, producing tension between FAIR and CARE principles and effective protection.

We discuss challenges faced by Brazilian data providers in handling RASD and present approaches for differentiated access, including spatial generalization, multi-level sensitivity classification, and controlled access to precise coordinates. The work aligns with TDWG efforts on vocabularies and metadata extensions for RASD, advancing interoperable solutions that balance transparency, conservation, ATK protection, and visibility of traditional communities in biodiversity science.

Speaker



Clara Baringo Fonseca

Biodiversity Data Specialist at SiBBR | SiBBR (Brazilian Biodiversity Information System - RNP)

Authors

Keila Juarez, Lais Brasileiro

The Edge: Insights into problems that arise from implementing data solutions for restricted access (sensitive) species data

9:30 AM – 9:45 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

With the rising of Artificial Intelligence and Machine Learning, organisations and individuals are increasingly concerned about data on species that they wish to restrict for a wide variety of reasons including everything from the terms and conditions of data licence agreements through to concern about the poaching of attractive species. These so-called sensitive species or restricted access species data (RASD), present problems for the public display, dissemination and use of biodiversity data. Common treatment approaches to RASD involve generalising data through truncation of latitude and longitude, using grids to display data, or withholding points altogether. Many implementations use a list of identified sensitive species as the basis of generalisation. The Atlas of Living Australia (ALA) is a collaborative, digital, open infrastructure displaying biodiversity data from multiple sources, making it accessible and reusable. Data is presented in a consistent, current form, taking into account RASD, for scientists, policy makers, environmental planners, land managers, and industry. To do this, the ALA complies with 8 different approaches to RASD across governments jurisdictions as well as the requirements of diverse entities, from iNaturalist to Indigenous communities. Australia developed an agreed national framework for the management and sharing of RASD in 2023 (www.rasd.ord.au). In developing governance mechanisms to implement the framework, we have identified a range of surprising issues surrounding RASD arising from the complexities of taxonomy, geography and data coming from different sources with different approaches. We will discuss these issues and what responses the ALA has developed to each of them in order to represent data as systematically as possible.

Speaker



Cameron Slatyer

Atlas of Living Australia CSIRO

Authors

Simon Sherrin, Tania Laity

Using wildlife trade data to support proactive RASD decision-making

9:45 AM – 10:00 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

The poaching of wildlife is a growing biodiversity threat, with online trade increasingly facilitating exploitation. Wildlife traders monitor publications, biodiversity repositories, and social media to identify desirable species, and trade activity reveals species being poached prior to formal conservation assessment. While biodiversity data are fundamental to conservation and research it can also increase exploitation of vulnerable taxa. A challenge for RASD frameworks is the time lag between species discovery, conservation assessment, and data restriction. During this period, biodiversity databases may expose species to collection pressure before threats are formally recognised. Current approaches to sensitive species data management are often reactive, relying on formal conservation status or legal protection. This issue is particularly relevant for terrestrial invertebrates, many of which are newly described, range-restricted, data deficient, and increasingly targeted in wildlife trade. Public occurrence datasets are often the only available locality information for these taxa. In Australia, the recent restriction of occurrence data for the trapdoor spider *Moggridgea rainbowi* following a suspected poaching incident demonstrates adaptive data governance and underscores the value of extending precautionary approaches to additional vulnerable wildlife taxa. We propose that wildlife trade data can function as an early-warning indicator for identifying species subject to poaching and exploitation, and for determining when occurrence data warrant precautionary restriction, prior to formal conservation assessment. Incorporating wildlife trade data into RASD frameworks supports proactive biodiversity data governance and can reduce exploitation risk from data accessibility.

Speaker



Charlotte Lassaline

PhD Candidate | Wildlife Crime Research Hub, Adelaide University, Adelaide SA 5000

Authors

Bruno Alves Buzatto, Jacob Maher, Jessica R. Marsh, Phillip Cassey

A new standard for restricted access (sensitive) species data

10:00 AM – 10:15 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

There is growing concern among many organisations and researchers about so-called sensitive species or restricted access species data (RASD), that is, data that needs to be modified or withheld from the general public. This data is intrinsically difficult to share or display and the existing Darwin Core fields no longer give data managers the flexibility to deal with the diverse array of emerging issues that managers encounter. Following a workshop in 2023, a Darwin Core Task Group was established on sensitive species data. The Task Group has developed a suite of reasons and treatments and an underlying vocabulary that enables data managers to communicate what has happened to restricted data and why in line with FAIR principles.

The data extension is intended to handle a wide range of reasons for restricted access and a similarly wide range of treatment options that can be applied to the data. This proposed data extension paths the way for institutions, regions or countries to develop a consistent approach to the management of restricted data.

We will outline and discuss the new restricted access species data extension, including solutions for both Darwin Core Native and Darwin Core Data Packages.

Speaker



Tania Laity

Atlas of Living Australia (CSIRO)

Authors

Simon Sherrin, Mieke Strong

Discussion

10:15 AM – 10:30 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Discussion for SYM23

Speaker



Cameron Slatyer

Atlas of Living Australia CSIRO

Authors

Cameron Slatyer

SYM06 - Rocks, Minerals, and Metadata: Integrating Earth Sciences into the TDWG Community

8:30 AM – 10:30 AM | Location: SAL C

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Natural history collections extend well beyond biological specimens to encompass minerals, rocks, meteorites, fossils, and other geological materials fundamental to understanding Earth systems. Yet data standards development has historically prioritized biodiversity, leaving geological collections underserved and often isolated from broader aggregation and discovery infrastructures.

This symposium examines the current landscape and future directions for the standardization of earth sciences collection data. We will present progress on the Mineralogy Extension to Darwin Core, now approaching public review after two years of community development, and explore how existing standards can accommodate the distinct requirements of geological specimens.

Presentations will address practical implementation challenges, conceptual models for geological collections, and opportunities for advancement. We invite contributions from collection managers, curators, data architects, and researchers across the earth sciences collection community.

Moderators



Ben Norton

Yale University



Rachel Walcott

Principal Curator | National Museums Scotland

Session organizers



Adam Mansur

Smithsonian Natural Museum of Natural History



Gary Motz

Head of Computer Systems | Yale Peabody Museum

Introduction

8:30 AM – 8:35 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Speaker



Ben Norton

Yale University

Developing a community-curated controlled vocabulary for fossil collections using the term `dwc:materialEntityType`

8:35 AM – 8:50 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Data about fossil specimens are not always easy to distinguish from that related to modern material, especially when aggregated in broadly scoped platforms such as the Global Biodiversity Information Facility. While fossil specimens have traditionally relied on the `dwc:basisOfRecord` term to distinguish themselves, this needs to be reconsidered given the recent ratification of a conceptual model for Darwin Core, and the associated Darwin Core Data Package. With these, the terms `dwc:materialEntityCategory` and `dwc:materialEntityType` are expected to functionally replace `dwc:basisOfRecord` in the evolution of publishing data using the Darwin Core standard. The term `dwc:materialEntityCategory` aims to define a broad category identifying the nature of a `dwc:MaterialEntity` with a tightly controlled vocabulary including the value "fossil" or "fossilized." `dwc:materialEntityType` is a narrower concept for which we expect controlled vocabularies to be more discipline-specific.

We have been coordinating members of the paleontological research and collections communities to build consensus around a controlled vocabulary that fossil collections can use with `dwc:materialEntityType`. This vocabulary would streamline research requests and facilitate collections management workflows by enabling users to gain a better understanding of the physical material represented by the digital record. For example, looking at two current records of fossil insects without images, it would not be obvious if one was preserved in amber and another was a compression fossil on a slab. We propose an initial controlled vocabulary focused on a relatively small set of concepts, potentially to be expanded to sub-concepts in a future proposal. Here we present the process and our preliminary proposed vocabulary.

Speaker



Holly Little

National Museum of Natural History

Corresponding author



Erica Krimmel

Biodiversity Information Scientist | independent

Authors

Holly Little, Talia Karim, Amanda Millhouse, Lindsay Walker, Erica Krimmel

Extending Specify 7 for Geological Collections

8:50 AM – 9:05 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

For 25 years, Specify has provided biological research museums and genetic repositories with software and technical services for curating, integrating, and publishing collections information. With funding from the Swiss Academy of Sciences (SwissCollNet initiative) and in collaboration with four Swiss natural history museums (Basel, Bern, Lausanne, Geneva), the Specify Collections Consortium (SCC) extended Specify to support geological collections.

Geocollections database requirements are similar in many ways to those of biological collections, however their unique properties and cataloging protocols challenged us to extend Specify concepts and capabilities. In addition to adding new tables, such as Absolute and Relative Age, and a new tree, Tectonic Unit, Specify added support for new protocols, most notably, Collection Object Groups (COGs), Components, and Collection Object Types (COTs).

In this presentation, SCC staff will discuss and demonstrate the new tables, tree, and features added to the Specify schema in order to support Geology Collections and their unique needs. These new features were designed to also support functionality required by Natural History Collections, the backbone of our membership.

The SCC is an open-source, collections community collaborative governed, funded, and sustained by its member institutions. Our software and technical support priorities address member data and system infrastructure needs. With Specify's extension into the earth sciences, natural history museums will now be able to curate, integrate, and publish data from a robustly engineered single platform for both biological and geological collections.

Speaker



Theresa Miller

Membership and Fulfillment Officer | Specify Collections Consortium

Improving Access to Earth Science Collections Data via GeoCAsE and the European Open Science Cloud

9:05 AM – 9:20 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Earth Science collections curated by natural history museums and research institutions contain critical records documenting past climates, environmental change, natural hazards, biodiversity dynamics, and geological processes across deep time. Despite their considerable scientific and societal value, these data have often remained fragmented, inconsistently described, and difficult to integrate with other Earth observation datasets. The CETAF Earth Science Group (ESG) is working to improve the discoverability, interoperability, and reuse of Earth Science collections data across Europe. ESG is leading the further development of GeoCAsE (<https://geocase.eu>), a data network, aggregation service, and web portal for geoscientific collections data. The recently launched EOSC-Gravity project “Geodera” aims to integrate geological specimen- and sample-level data within the European Open Science Cloud (EOSC) through their Data Terra thematic node. Geodera will use representative case studies from different branches of geology to analyse and identify researchers’ data service needs, and to evaluate the usefulness of collections-based records for scientific applications and cross-domain research. Particular attention will be given to metadata harmonisation with external research platforms and datasets, including Earth observation data services within EOSC and Data Terra. This work includes assessing existing TDWG community standards, such as ABCD and Darwin Core, together with their geoscience extensions, as well as emerging concepts such as digital specimens. The project will identify gaps in current metadata schemas, provide training and practical guidelines to improve data quality, expand the GeoCAsE consortium, and strengthen interoperability across the Earth Sciences domain.

Speaker



Patricia Mergen

Liaison officer | Meise Botanic Garden/Royal Museum for Central Africa/ Belgian Science Policy Office

Speaker (VIRTUAL)



Laura Tilley

Consortium of European Taxonomic Facilities

Authors

Olle Hints, Patricia Mergen, Emily Veltjen, Alessandro Rizzo, Marion Massol

Improving geospecimen provenance and findability in the National Museums Scotland Scottish mineral collection

9:20 AM – 9:35 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Geospecimen metadata is only as valuable as it is findable. Specimens can be located through a variety of properties - for example by species, donor, location in the museum or provenance in the field. Unfortunately, place names are not always unique and historic descriptions are often ambiguous, inconsistent, and difficult to georeference at scale. Therefore, grid references are better but how many curators have time to look this up manually?

We present a curatorial workflow using the National Museums Scotland collection of Scottish minerals - about 18000 specimens collected over 200 years, with variable quality and certainty of provenance. It reconciles legacy mineral names, standardises locality information, and where possible, links specimens to structured localities and coordinates from external resources such as Mindat API.

Provenance is treated not as a single point on a map, but as a spectrum of certainty and scale, ranging from a specific named mine shaft to a whole district or county. Museum documentation can therefore preserve uncertainty and locality granularity that are often flattened in external resources. Linking collection records with Mindat can strengthen locality data through grid references and occurrence context, while the museum record retains curatorial detail.

This process improves discovery by enabling specimens to be found by searching at any scale, from broad region to specific locality. Museums can not only draw on external data but also use internal collection records to identify missing or faulty data. Moreover, minerals in the collection not yet recorded from Scotland in Mindat are readily identified. This enriches both information sources, and demonstrates geocollections as active contributors to data infrastructure development.

Speaker



Sarah Stewart

PhD candidate | University of Edinburgh

Authors

Rachel Walcott

Bridging Domains: How SESAR Supports Earth and Biodiversity Digital Extended Specimens

9:35 AM – 9:50 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

SESAR, the System for Earth and Extraterrestrial Sample Registration has served as a repository of Earth science specimen metadata since 2004. Scientists and collection managers “register” their specimen with SESAR, and SESAR provides them with a stable landing page containing metadata describing their sample, and a persistent identifier (an IGSN, or International Generic Sample Number, now managed by DataCite). Though originally focused in the Earth and geological science community, SESAR has gradually expanded to additional domains such as archaeology, paleontology, and environmental sciences. SESAR currently supports 1000+ individual researchers and dozens of collections, globally, in making their sample records publicly accessible, providing an open, global catalog and discovery platform for researchers to find samples and specimens. In this presentation, we provide an overview of SESAR and present the outcomes of a five-year database migration and modernization project, focusing on the ways in which we have integrated and formalized our adoption of broader community vocabularies and standards, such as GeoSciML, and our ongoing efforts to support Darwin Core. We share ways the new database enables greater flexibility in incorporating new community vocabularies and in better supporting the unique needs of organizations and curated collections. We conclude by sharing how SESAR supports the community vision of a “digital extended specimen.”

VIRTUAL PRESENTATION

Corresponding author



Natalie Raia

University of Arizona, College of Information Science

Speaker (VIRTUAL)



Andrea Thomer

University of Arizona, College of Information Science

Authors

Andrea Thomer, Sean Cao, Steve Richard, Kerstin Lehnert, Saebyul Choe

Developing and implementing relational mineral taxonomy model in EMu for Museums Victoria's Mineral Collection

9:50 AM – 10:05 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Mineral identification and classification data in Museum's Victoria's (MV's) EMu CMS have until recently been stored within individual specimens' catalogue records and managed as part of a series of shared hierarchical look-up-lists. This is unlike MV's biological taxonomic data, which have made use of a one-to-many relational structure between taxonomy and catalogue records at least since the implementation of EMu in the late 1990's. Disadvantages of the model previously applied to mineral data include difficulties making changes following nomenclature revisions, an inability to capture incomplete taxonomy (eg. cataloguing specimens only identified to the level of a group or solid solution series), and a limited ability to query the database on mineral properties, chemistry or classification schemes.

I report on the development and implementation of a relational mineral taxonomy model in EMu over 2025-26. This model captures a hierarchy of mineral taxonomy; subgroup, group and supergroup as formalised by the International Mineralogical Association (IMA) in 2009, the Dana and Strunz classification systems, ideal chemical formula and elements, and IMA approval status. It also allows use of some existing functionality for zoological taxonomy, capturing the history of identification and linking taxonomy records to publications in which species were described or revised. Data to populate 6,125 taxonomy records for IMA approved mineral species were extracted from Mindat.org via their API. A further 162 taxonomy records for invalid, as yet undescribed, or incompletely classified specimens were created and manually populated. These 6,287 taxonomy records have been linked to the 56,792 catalogue records containing mineral identification data in MV's database.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Oskar Lindenmayer
Museums Victoria

Understanding Geological Specimen Use for Effective Digitisation Strategies

10:05 AM – 10:20 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Without representation on major international platforms like GBIF, geological collections are currently much less digitally accessible to user communities than digitised biological specimens. Improving the access of users requires the digitisation of more geological and mineralogical specimens and the appropriate representation of their associated data. Forthcoming standards like the TDWG Mineralogy Extension and Minimum Information for a Digital Specimen (MIDS) will provide a framework with new fields that will enable the sharing of expressive top-level data, but there is a clear need for a better understanding what the audiences for these collections need from this data as we plan how to digitize and share collections data.

Here we present the preliminary longitudinal study of the users and use cases of geological and mineralogy specimens in museums, especially in three long-established well-known collections – the Smithsonian Institution, Natural History Museum of London, and National Museums Scotland. We analyze use of our specimens recorded in the published literature (including journals), loan forms, audience demographics, and web search. These data enable us to better understand the granularity of kind of data our audiences need and to develop more effective strategies about what and how to digitise.

Speakers



Rachel Walcott
Principal Curator | National Museums Scotland



Adam Mansur
Smithsonian Natural Museum of Natural History

Authors

Mike Rumsey, Sylvia Orli, Emily Brown

Discussion

10:20 AM – 10:30 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Discussion for SYM06

Moderator



Ben Norton
Yale University

Geologic Units in the Geological Collection Conceptual Data Model: A Framework for Stratigraphic Representation in Natural History Collections

10:20 AM – 10:20 AM

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Most geological specimens are fragments of larger rock bodies, and knowledge of their broader geological context is essential for research. Existing natural science collection systems inadequately represent provenance metadata, causing information loss during data mobilization. The Geological Collection Conceptual Data Model (GCDM; Norton et al., in prep.) introduces a dedicated Stratigraphy class model to address this gap.

A foundational decision distinguishes GCDM: adopting *geologic unit* over *stratigraphic unit*. The latter implies systematic layering—imprecise for unstratified bodies such as igneous intrusions and metamorphic complexes—while *geologic unit* covers both stratified and unstratified rock bodies.

The Geologic Unit model separates two orthogonal dimensions routinely conflated elsewhere. Geologic Unit Type captures the basis of definition (litho-, bio-, chrono-, magneto-, or tectonostratigraphic); Geologic Unit Rank captures hierarchical position within the type's framework (e.g., Formation, Member; Biozone; Eonothem, Erathem). An *isSpecificTo* relationship encodes that rank hierarchies vary by type.

GCDM supports five unit types across three criteria classes: material, temporal, and structural. Multiple units from any combination of types may be associated with a single specimen, reflecting that a rock body belongs concurrently to a formation, stage, and zone defined by different criteria. Nomenclatural control is achieved through linkage to authoritative stratigraphic lexicons.

This talk presents the Geologic Unit model within the broader GCDM, illustrates the architecture with worked examples, and shows how the framework enables interoperability with the Darwin Core GeologicalContext class.

VIRTUAL ONLY

Speaker



Ben Norton
Yale University

SYM18 - Building resilient data infrastructures for biodiversity science

8:30 AM – 10:30 AM | Location: ODIN

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Data infrastructures in science are facing increasing challenges. Cyberattacks, aggressive data scraping, loss of technical expertise, sudden funding cuts, political decisions limiting data access, and environmental hazards are just a few of the potential problems. Therefore, data repositories need strategies to build resilience as part of their daily operations and have plans in place for risk mitigation.

Biodiversity science is especially affected by these challenges, as it relies on very different data, which are hosted by a multiplicity of small and large repositories located all over the world. Some of the smaller repositories, like [DiscoverLife.org](https://discoverlife.org), are reporting AI bot crawlers as an "existential threat" for their knowledge-sharing mission, established repositories like the [Biodiversity Heritage Library](https://biodiversityheritagelibrary.org) are not immune from governance and funding challenges, and datasets used in [environmental and climate research](#) have been decommissioned and are not available anymore due to a political agenda.

The symposium is an opportunity to present how data infrastructures in biodiversity science are addressing resilience and risk mitigation in the face of current challenges. It is organised as an open session to invite contributions from all biodiversity science organisations—academic institutions, charities promoting conservation initiatives, citizen science projects, etc.—interested in the topic.

Moderators



Giuditta Parolini
Museum für Naturkunde Berlin - Leibniz Institute for Evolution and Biodiversity Science



Eva Alonso
DiSSCo Head of Governance | Naturalis Biodiversity Center

Introduction

8:30 AM – 8:35 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Moderators



Giuditta Parolini

Museum für Naturkunde Berlin - Leibniz Institute for Evolution and Biodiversity Science



Eva Alonso

DiSSCo Head of Governance | Naturalis Biodiversity Center

Governance as a Foundation for Resilient Biodiversity Data Infrastructures: Lessons from DiSSCo

8:35 AM – 8:47 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Governance as a Foundation for Resilient Biodiversity Data Infrastructures: Lessons from DiSSCo RI

Building resilient biodiversity data infrastructures depends equally on governance systems that sustain coordination, legitimacy, adaptability, and long-term service provision. As biodiversity science increasingly relies on distributed research infrastructures connecting institutions, collections, and communities across national and international scales, governance becomes a critical enabling component of resilience.

Resilient infrastructures require support from multiple stakeholder groups, including funding authorities, service providers, and strategic partners. Effective governance provides mechanisms through which these actors contribute to strategic direction, operational implementation, and continuous evolution of infrastructure services, ensuring alignment between scientific priorities, user needs, and sustainability objectives.

Here we explore governance as a foundational pillar through the case of DiSSCo (Distributed System of Scientific Collections). DiSSCo applies a multi-level governance model distinguishing decision-making, executive, implementation, and advisory functions, connecting a central coordination structure with a distributed network of national and institutional actors. This approach strengthens transparency, accountability, and responsiveness while supporting sustained service delivery and innovation.

The DiSSCo experience distills transferable lessons, including mechanisms for cross-layer communication, integration of independent advisory functions, and alignment with shared strategic priorities. Governance is not a supporting component of research infrastructures; it is a fundamental condition for their resilience and enduring scientific impact.

Speaker



Eva Alonso

DiSSCo Head of Governance | Naturalis Biodiversity Center

Resilience of Research Infrastructures through knowledge and engagement

8:47 AM – 8:59 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Data and research infrastructures provide a constantly growing variety of high-quality data, yet the potential of this data often remains unexploited, sometimes to the point of challenging the capacity of the infrastructure to remain relevant beyond the lifecycle of projects or similar initiatives.

For data infrastructures to be sustainable and resilient their services must be utilized. This requires the knowledge of existing datasets and their benefits by potential users, as well as data literacy skills to search for, retrieve, and combine them in a targeted manner. By the same token, long-term usability of data infrastructures calls not only for technical and financial support but also for an active engagement effort that reaches out to create alignment with other initiatives, boosts uptake of new developments for the widest possible audience, and makes data visible and attractive over time. Hence, communication and a community of practice is crucial.

Education, training, communication, and engagement are, therefore, decisive elements to make the best of existing data infrastructures and turn them into true spaces for exchange and networking where users can discuss methods and use cases and propose new research questions.

The presentation discusses challenges and potential solutions for the sustainable use of research infrastructures and emphasizes in particular the importance of community building as well as targeted education and training initiatives.

Speaker



Mareike Petersen

Lead Science Programme Collection Future | Museum für Naturkunde Berlin

Authors

José Alonso

Adopting Environmental Sustainability as a Strategy to Build Resilient Data Infrastructures in Biodiversity Science

8:59 AM – 9:11 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

In the aftermath of the COVID pandemic and the Russian invasion of Ukraine, ESFRI (European Strategy Forum for Research Infrastructures) [reported](#) on the challenges posed to research by high energy costs and shortages of critical resources required by scientific computing and large experimental facilities. As a response to the difficult situation, ESFRI recommended research infrastructures to increase the environmental sustainability of their operations by decreasing energy use, promoting eco-friendly processes, and increasing reuse and recycling, especially for critical resources in short supply. Since the report was published, new conflicts and geopolitical tensions have further destabilised energy markets and jeopardised the supply chain of key elements for computer hardware, such as rare earth elements. ESFRI's recommendation is now more relevant than ever.

The paper considers how data infrastructures in biodiversity science can become more resilient by adopting environmental sustainability as a guiding principle. Considering the growing amount of research data in biodiversity science and the increasing use of resource-intensive technologies like generative AI as part of research, biodiversity science can greatly benefit from understanding the environmental footprint of its data infrastructures and implement strategies to lower it. This is not only a required contribution to address global environmental challenges, such as climate change, but also a risk management strategy to reduce operational and strategic risks due to scarcity and/or high costs of energy and computing hardware, and to be prepared to comply with more stringent environmental reporting obligations that are likely to involve research infrastructures over time, at least in the European Union.

Speaker



Giuditta Parolini

Museum für Naturkunde Berlin - Leibniz Institute for Evolution and Biodiversity Science

Supporting resilient data infrastructure in a changing environment

9:11 AM – 9:23 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Recent concurrent losses of funding, people, and institutional support for U.S. environmental data infrastructure has forced a reckoning with the vulnerabilities of our previous systems. U.S. environmental data infrastructures, both human and technical, are more fragile than had been widely understood, often lacking sustainable funding, robust policies, and programmatic protections. Initial triage responses by data users and professionals were fragmented and relied heavily on copying data and information to repositories perceived as stable, often housed outside the U.S. As data rescue efforts expanded, we increased coordination and collaboration, and reduced fragmentation as groups communicated to share practices and identify high-priority datasets and products. In the last several months, the focus has shifted to improving the usability of data and products, including recreating and expanding lost environmental data services on new hosts. Data professionals have also built tools for assisting repositories in assessing resilience and planning for improved sustainability. Beyond the data itself, there has been a growing recognition that preserving and sharing knowledge about data infrastructures—methodologies, processes, and people, as well as analyses and applications—is critical for making data useful in the future. We will share reflections from a range of environmental data professionals and organizations on how to design and construct new data ecosystems that are more resilient and prepared for future challenges, and that preserve expertise, institutional knowledge and access to environmental data and knowledge.

Speaker



Rachael Blake

Director of Data Science | Intertidal Agency

From Pixels to Policy: Closing the Gap between In Situ Monitoring, Earth Observation, and Policy Actionability

9:23 AM – 9:35 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Scaling biodiversity monitoring requires aligning multi-level observations - from field sensors to satellite imagery - with international policy targets such as the CBD Global Biodiversity Framework and the UNDRR Sendai Framework. As thematic architectures like the EU's Biodiversity Data Space are established, a process exemplified by the Biodiversity Meets Data (BMD) project, the focus is shifting from simple data mobilization toward robust knowledge governance. The primary challenge is ensuring long-term digital resilience and persistent legal actionability for stakeholders, from policy-makers to Natura 2000 managers.

We address this by applying FAIR Digital Objects (FDOs) as "Actionable Knowledge Units" to bridge the gap between field deployments and regulatory requirements. We employ the WildLIVE Monitoring Ontology (WLMO) to link sensor deployments with biological events. By integrating the European Legislation Identifier (ELI) and the Open Digital Rights Language (ODRL), observations are mapped to frameworks such as the EU's Biodiversity Strategy for 2030 and the Nature Restoration Law.

Converting static legal text into machine-interpretable policy objects enables automated governance over data processing and access.

This approach ensures that critical environmental data remains a stable, machine-actionable cornerstone for conservation. We demonstrate this through three implementations: the Socotra Data Portal, safeguarding socio-ecological records in Yemen; the WildLIVE platform, integrating edge sensors with Earth Observation; and the Movebank database, providing harmonization and a digital security layer for heterogeneous biodiversity data.

Speaker



Claus Weiland

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main

Authors

Aidin Niamir, Sarah C Davidson, Nashat Hamidan, Lena Perzlmaier, Rajapreethi Rajendran, Jonas Grieb, Thomas Winter

The GBIF data index: crawling and crawled

9:35 AM – 9:47 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

The GBIF data infrastructure indexes over 120,000 datasets from 2,700 institutions, and maintains a website and public API sharing over 3.7 billion species occurrences.

GBIF runs a crawler to retrieve datasets for indexing, and is accessed by crawlers. We will share how our approach to crawling data has changed over the last 20 years, the impact AI bots and AI users have made on the system, and ongoing challenges we face in maintaining a reliable service.

Speaker



Matthew Blissett

Software Developer | Global Biodiversity Information Facility

Resilient Attribution Infrastructures for Biodiversity Science: Archiving and Risk Mitigation in Bionomia

9:47 AM – 9:59 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

[Bionomia](#) is an independent biodiversity data infrastructure that links natural history specimens shared through GBIF to the people who collected and identified them. The platform focuses on creating durable, attributable connections between biodiversity records and human expertise, particularly for collectors whose contributions are often obscured in downstream aggregations.

This presentation explores how Bionomia is adopting resilience-oriented practices in response to increasing infrastructural instability, including governance uncertainty, AI-driven scraping pressure, dependence on centralized services, and sustainability risks facing small biodiversity informatics projects. Rather than functioning solely as a web application, Bionomia increasingly treats itself as a distributed archival workflow. Public claims and profile datasets are versioned and deposited to Zenodo, creating DOI-backed snapshots that remain citable, reproducible, and recoverable even if operational infrastructure changes or disappears.

The talk will highlight practical strategies for resilience, including automated archival deposition, integration with ORCID and Wikidata identifiers, open-source reconciliation tools, and interoperable data products that can be mirrored or reused by other infrastructures.

Bionomia offers a case study in both the vulnerabilities and strengths of independently maintained scientific infrastructure, demonstrating how small projects can experiment with preservation, attribution, and reuse models under uncertain technological and political conditions.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



David Shorthouse

independent

Discussion

9:59 AM – 10:30 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Discussion for SYM18

Moderators



Giuditta Parolini

Museum für Naturkunde Berlin - Leibniz Institute for Evolution and Biodiversity Science



Eva Alonso

DiSSCo Head of Governance | Naturalis Biodiversity Center

WKS34 - Liberating Data from Historical Biodiversity Literature with The Unified Corpus Explorer (UCE)

8:30 AM – 10:30 AM | Location: BALDER

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Biodiversity research relies on data that are deeply structured, semantically explicit and interoperable across systems. This workshop introduces the **Unified Corpus Explorer (UCE)** as a tool for transforming biodiversity literature into machine-actionable research data.

Hosted by the Specialised Information Service Biodiversity Research (BIOfid), the workshop demonstrates how data from historical and contemporary texts, with a strong focus on publications from the 19th and 20th centuries, can be systematically mobilised through digitisation, NLP- and LLM-based annotation, and semantic enrichment using domain-specific ontologies. These sources contain essential information on species distributions, ecosystems and spatio-temporal dynamics that are largely absent from modern databases, yet are critical for establishing robust baselines for research on biodiversity change.

Participants will explore how UCE enables interactive access to richly annotated text corpora, supports standardized representations of extracted entities and relations and facilitates their integration into interoperable data workflows. The workshop highlights how text-based biodiversity information can be aligned with community standards and made accessible not only for human interpretation, but also to improve interoperability and reusability.

The workshop is aimed at researchers, interested in unlocking biodiversity knowledge from literature and embedding it into scalable, standards-oriented data ecosystems.

Workshop leaders



Marius Böyng

University Library Johann Christian Senckenberg, Goethe University Frankfurt



Gerwin Kasperek

Curator of the Biology Collection | University Library Johann Christian Senckenberg, Goethe University Frankfurt

Authors

Katrin Peikert, Mevlüt Bağcı, Jana Hoffmann

WKS35 - From Skeletal to Scientific: A Hands-on Workshop for Measuring Digitization Depth with the MIDS Standard

8:30 AM – 10:30 AM | Location: Møterom 5 / Meeting room 5

COLLECTIONS DIGITISATION & MANAGEMENT

The digitization of natural history specimens renders their associated data more accessible, both to human and robot users. As collections accelerate their digital transition, the need for objective metrics to measure the "completeness" of a specimen's digitization has become paramount. The Minimum Information about a Digital Specimen (MIDS, <https://github.com/tdwg/mids>) standard provides a framework of four levels to categorize the depth of digitization, helping institutions prioritize workflows, track progress and report to funders and researchers.

This workshop provides a practical, hands-on introduction to the MIDS standard. Participants will learn how their own collection data (Darwin Core or ABCD formats) are mapped to MIDS information elements and use the MIDSCalculator tool to generate scores and visualizations. By the end of the session, attendees will understand how to use MIDS as a diagnostic tool for the digitization level of collections, including why certain levels were achieved.

Participants should bring a laptop and should install a local version of the MIDSCalculator tool in advance. Instructions can be found on <https://github.com/AgentschapPlantentuinMeise/MIDSCalculator> and any issues encountered can be flagged in that repository. Participants are also encouraged to acquire a sample Darwin Core Archive (DwC-A) or ABCD file for specimen data (biology, geology, paleontology), either from their own institution or from a third party.

Workshop leaders



Mathias Dillen

Biodiversity Data Scientist | Meise Botanic Garden



Elsbeth Haston

Deputy Herbarium Curator | Royal Botanic Garden Edinburgh



Cat Chapman

Biodiversity Informatics Coordinator (Data Wrangler) | iDigBio, Florida Museum of Natural History

BREAK

10:30 AM – 11:00 AM | Location: 1st Floor

SYM30B - AI for Biodiversity Data

11:00 AM – 12:30 PM | Location: SAL A

AI AND ROBOT READY

Biodiversity researchers today are offered access to vast troves of data: billions of images and data points in online repositories, citizen scientists contributing millions of new observations every year, countless specimens in museum collections around the world, and an endless supply of studies, reports, and expert knowledge in articles and books stored in libraries or published online.

But data alone is not enough. To have any meaning it must be collated, organised, standardised, evaluated and interpreted. With such quantities of information, machine learning and artificial intelligence tools are vital to our ability to parse and understand the data available to us, and essential if we are to scale our research to meet the urgent biodiversity challenges the world faces today.

This session invites talks on cutting-edge applications of AI technology in automating the collection, preparation and interpretation of biodiversity data. We welcome contributions on topics such as novel applications of AI in data collection and mobilisation, large-scale studies enabled by AI, automation to extract data and new insights from historical collections and archives, and systems that improve our ability to arrange and interpret large biodiversity datasets.

We strongly encourage speakers to include the financial and computational costs of model training and operation, in addition to performance metrics, in discussing their work to allow assessment of its feasibility in wider use.

Moderators



David Williamson

Postdoctoral Researcher | NTNU University Museum



Wouter Koch

Senior adviser biodiversity informatics | Norwegian Biodiversity Information Centre

Session organizers



Arianna Salili-James

Natural History Museum UK



Rukaya Johaadien

GBIF Norway



Michal Torma

Head engeneer | University of Oslo



Maarten Trekels

Biodiversity Data Scientist | Meise Botanic Garden



Kit Lewers

PhD Candidate, FINESST Fellow | CU Boulder, NASA

Welcome back

11:00 AM – 11:03 AM

AI AND ROBOT READY

Assessing the Readiness of Plant Image Datasets for AI-enabled Spatially Explicit Ecological Monitoring: A Systematic Review of Bias, Representativeness, and Annotation Quality

11:03 AM – 11:15 AM

AI AND ROBOT READY

Despite rapid advances in AI models, the applicability of ecological monitoring remains constrained by limitations in dataset quantity and quality. While most plant image datasets focus on classification for species identification, object detection and segmentation datasets remain underdeveloped. These tasks are critical for spatially explicit ecological inference, including individual localisation, abundance estimation, and vegetation structure analysis.

This review systematically identified and analysed plant image datasets developed for ecological monitoring. An assessment framework inspired by clinical image datasets evaluation was adapted to ecological monitoring requirements and applied to 42 datasets. The framework evaluated dataset scale, ecological representativeness, annotation protocols, and data management practices.

The results revealed that most datasets lacked clearly documented annotation protocols and cross-annotator consistency controls, undermining the reliability of annotations as ground truth for AI-based ecological inference. Inconsistencies in taxonomic and geographic distribution further limit both model generalisability and the applicability of image-based monitoring at global scales. Agricultural weed detection dominated current applications, while broader biodiversity monitoring remained limited.

These findings highlight that while datasets are increasingly available, their design is often not aligned with the ecological levels required for biodiversity monitoring. Both annotation quality and ecological representativeness of available datasets fall short of the requirements for robust and generalisable image-based monitoring. Future work should prioritise ecologically representative and well-documented datasets for biodiversity monitoring.

Speaker



Yuchen Zhao

Newcastle University

Authors

Muhammad Shafiqul Hoq Akand, Dave George, Roy Sanderson, Ankush Prashar, Tong Xin

The WildLIVE Portal: Towards Scalable Wildlife Monitoring through the Integration of Citizen Science and Human-in-the-Loop AI Workflows

11:15 AM – 11:27 AM

AI AND ROBOT READY

Global biodiversity monitoring delivers an ever-increasing volume of recorded data, with even medium-sized projects gathering data in the order of hundreds of thousands of images annually. These vast amounts of data frequently overwhelm the manual review capacity of domain experts and small research teams. As a result, the scale of these observations creates significant bottlenecks that hinder timely data annotation and analysis for conservation efforts. AI-generated baseline annotations can reduce this workload, but they still require expert review before they can be treated as reliable scientific evidence. The primary challenge lies in establishing trust and quantifiable confidence in these annotations, regardless of their source.

To address this challenge, Senckenberg's [WildLIVE](#) portal combines AI workflows with Human-in-the-Loop (HITL) refinement. The platform supports multi-model object detection, including open-vocabulary approaches in which detections can be guided by textual category descriptions rather than being limited to a fixed label set. Through an active-learning infrastructure, the platform prioritises observations most in need of review, directing citizen scientists and domain experts toward images where labelling effort can have the greatest impact on model confidence. These evaluations are preserved as layered annotations linking back to AI4WildLIVE's FAIR Digital Object-based data model.

By integration of an AI-assisted annotation pipeline, active learning, AI-readiness of data and citizen science engagement, the WildLIVE portal establishes a shared, human and machine operable collaborative framework through which biodiversity monitoring data becomes a trusted, community-wide scientific resource.

Speaker



Rajapreethi Rajendran

Scientific Programmer | Senckenberg – Leibniz Institution for Biodiversity and Earth System Research

Authors

Jonas Grieb, Peleg Sapir, Simone Bianchi, Andreas Weinmann, Claus Weiland

HerbAudit: Validating AI-Driven Herbarium Transcriptions

11:27 AM – 11:39 AM

AI AND ROBOT READY

Millions of herbarium specimens are now digitized and accessible online, yet transforming specimen label images into structured, queryable data remains a major bottleneck in biodiversity informatics. While OCR and LLM technologies are increasingly applied to this challenge, the community lacks a shared framework to measure tool performance and identify data gaps.

To fill this gap, we developed HerbAudit, an open source tool combining herbarium label transcription with systematic quality evaluation, to be publicly available by September 2026. We also compiled a gold standard benchmark dataset of 90 typed and handwritten specimens from 63 institutions worldwide, each manually transcribed using Label Studio.

HerbAudit consists of two core components. The transcription component builds on VoucherVision logic, processing images via LLMs like Google Gemini and OpenAI to output Darwin Core formatted data. It includes a resolution adjustment submodule and reports estimated cost and carbon footprint metrics per image.

The evaluation component offers two distinct pathways. First, OCR tool outputs are compared directly against the gold standard manual annotations. Second, AI tool predictions in Darwin Core format are cross checked against GBIF and POWO, incorporating a built-in synonym check. Fields are scored using domain specific methods, such as geographic coordinates via the Haversine formula and author names via Character Error Rate. Users can apply transcription and AI evaluation to their own datasets, with the latter requiring records on GBIF to pinpoint errors and visualize where tools fall short.

All results are compiled into an interactive HTML report, giving developers and researchers a shared foundation to benchmark, compare, and improve digitization pipelines.

Speaker



Dilara Ağacık

PhD Candidate | IDLab, Ghent University

Authors

Maarten Trekels, Kenzo Milleville, Steven Verstockt, Kenneth Bauters, Quentin Groom

AI-assisted digitisation of printed identification keys: from book to structured data in minutes

11:39 AM – 11:51 AM

AI AND ROBOT READY

Printed identification keys are a major reservoir of curated biodiversity knowledge, but mobilising them into machine-readable form is hampered by the cost of expert transcription, taking days of work per key. We present a working pipeline that produces structured-key drafts from heterogeneous sources in minutes.

The pipeline runs in two stages. An extraction agent (local OCR plus targeted vision reads of figures) builds a taxa tree and atomic characters. A fresh-context audit agent then re-reads the draft against the source, catching character states that overlap or leave source-mentioned values uncovered, numeric characters without actual measurements, and described traits missing from the output - fixing each until a deterministic verifier signs off. Two transformations bridge dichotomous prose and matrix-form output: splitting bundled traits ("brown hairs, green stem, white dots") into separate characters, and merging synonymous states ("whitish yellow" / "pale yellow") across the file. All such decisions are documented for the taxonomist to consider; target output is an editable draft, not a finished publication.

We have run the pipeline on various inputs: dichotomous keys in scanned books, combined prose descriptions of species from multiple field guides, and online interactive keys. We also include a domain-aware language translation step which maintains context and user-provided terminology.

Costs depend on the input format and structure, but are limited to at most a few dollars per species, and will be presented in detail. The Clavis output format and its place in combined AI-plus-key identification workflows are addressed in a companion talk in "May the Data Be Structured" (SYM32).

Speaker



Wouter Koch

Senior adviser biodiversity informatics | Norwegian Biodiversity Information Centre

AI-assisted georeferentiation a-posteriori of herbarium specimens: a case study on Italian herbaria

11:51 AM – 12:03 PM

AI AND ROBOT READY

Italy's largest herbaria have recently undergone a large-scale digitization, making more than 4 million specimens available online through the national portal PlantData and the GBIF. Most specimens only had textual locality descriptions without geographic coordinates, limiting their use in many ecological analyses such as species distribution models. To address this limitation, we started a process of georeferentiation a-posteriori.

The data presented different challenges: locality labels were often in Latin or partially illegible, mixing locality names with habitat notes, altitude and collector annotations. Initial georeferencing approaches using national gazetteers and OpenStreetMap Nominatim achieved success rates below 10%, as geocoding services were frequently confused by excessive contextual details in locality strings. Workflows based on Google Maps APIs improved results (~50%) but remained highly sensitive to spelling errors and historical orthographic variants.

Thus, we developed pipelines that combine lightweight large language models (LLMs) with geocoding services. The models use the full label context to separate locality names from ancillary text and correct spelling inconsistencies before georeferencing. Small LLMs combined with Nominatim achieved ~71% successful georeferentiations at very low cost. Small LLMs paired with Google Maps improved this result to ~78%, while using Frontier LLMs reached up to 84% accuracy, but they remain too expensive for high volume data.

This approach shows that AI can assist and accelerate the process of georeferencing a-posteriori herbarium specimens. Small LLMs can achieve strong performance and, along with Nominatim, can be fully self-hosted, reducing reliance on external APIs and services.

Speaker



Matteo Conti

Authors

Stefano Martellos, Gianfranco Samuele, Michele Lussu, Alessandro Chiarucci

Towards Scalable Trait Measurements from Digitized Specimens

12:03 PM – 12:15 PM

AI AND ROBOT READY

The increasing digitization of natural history collections creates opportunities for large-scale extraction of quantitative specimen traits. Morphological traits such as specimen shape and size are important for research in ecology and biodiversity. To obtain reliable and comparable traits from images, pixel measurements are converted into metric units using a scale conversion factor (CF), typically calculated by measuring pixel distances between ruler tick marks.

We build upon our previous work on automated ruler detection and scale estimation for herbarium specimens by improving its robustness and scalability. The original approach relied solely on image keypoint matching for CF estimation, which performed well for conventional rulers but was unreliable for simpler ruler types. We extend the original method with a hierarchical framework that first identifies the ruler type and then selects the most appropriate method for CF estimation. These simple rulers, such as lines, blocks, and grids, are processed differently, using a combination of computer vision methods and LLMs.

To evaluate this new approach, our dataset has been expanded beyond herbaria to a broader range of domains, including zoological, entomological, geological, and microscopy images. Furthermore, we show how the estimated CF values can be stored as standardized metadata to support interoperability and how markers can simplify scale estimation in future digitization workflows. Finally, we show how the scale estimation method can be combined with segmentation models to automatically extract specimen traits.

Speaker



Kenzo Milleville

Postdoctoral researcher | IDLab, Ghent University

Authors

Maarten Trekels, Steven Verstockt, Quentin Groom, Dilara Ağacı

SYM13A - Responsible AI, Open Digital Curation, and Round-Tripping for Biodiversity Data

11:00 AM – 12:30 PM | Location: SAL B

AI AND ROBOT READY

We propose a session focused on **open digital curation** as a continuous, inclusive process across the biodiversity data lifecycle. Presentations would illustrate how human and AI assistance can be made transparent, attributable, versioned, and reusable during data mobilization and enrichment (e.g. annotations and linking, taxonomic determinations, georeferencing improvements, use of persistent unique identifiers).

Particular attention would be paid to **round-tripping**: ensuring that annotations and AI-generated and human enhancements flow back to source systems (e.g., collection management systems and institutional repositories), rather than remaining disconnected products. By emphasizing openness, provenance, and interoperability, the session would explore practical approaches for integrating human work and AI responsibly and sustainably while maintaining curatorial authority and trust.

With an emphasis on natural history data from museums and herbaria -- including specimens, images, and genomic resources -- the session presentations would feature case studies and methods that bridge institutional practice, community engagement, and **open science** infrastructure. Contributions would address challenges such as scaling AI-enabled curation, validating and crediting annotations, aligning with standards like Darwin Core, and **designing workflows** and **user interfaces (i.e. UI/UX)** that support

equitable participation and vetting by collection staff, researchers, and external contributors.

Moderators



Deborah Paul

Species File Group (INHS), University of Illinois



Siobhan Leachman

Wikimedian | Wikimedia Aotearoa New Zealand

Session organizer



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

Authors

Sabine von Mering

Open Digital Curation Needs: openness, interoperability, provenance

11:00 AM – 11:15 AM

AI AND ROBOT READY

Moving toward open digital curation to make round-tripping possible, consider the necessary interplay of three facets: openness, interoperability, and provenance. Via open application programming interfaces (APIs), data becomes accessible to humans and AI. Given openness, interoperability becomes possible and offers opportunities to contribute and collaborate. We share three examples here: [iNaturalist (iNat)](<https://www.inaturalist.org/>), the [Catalogue of Life (COL)](<https://www.catalogueoflife.org/>), and the [Global Biodiversity Information Facility (GBIF)](<https://www.gbif.org/>). With iNat's open APIs, our researchers using [TaxonWorks (TW)](<https://taxonworks.org/>) can pull images and related data into TW. In the other direction, iNat curators submit or point their users to the specific gitHub repository to log taxonomic inquiries directly to those curating these data. This reduces steps (e. g. emails) needed to get to round-tripping. From TW, our collections data go straight to GBIF. We grab the GBIF-related ingestion feedback directly into TW to again reduce the human mental data curation burden and imagine future automated validation. Our taxonomic data flow interoperably between TW to COL via [Checklist Bank](<https://www.checklistbank.org/>). To do all of the above, we must decide on and use identifiers (e. g. [ORCID](<https://orcid.org/>), [Wikidata Q numbers](<https://www.wikidata.org/wiki/Q43649390>)) and the related metadata that helps us ensure we have a provenance trail. Understanding our data through these breadcrumbs provides the foundation for round-tripping. In TW, imagine extracting data from literature using large language models (LLMs). Our communities need software and APIs to vet the resulting data, along with robust provenance. We find it's these decisions, usually human-mediated, that are difficult to capture in context, for example, with the user interface. When we connect via our APIs for data exchange, what must we persist? What can be temporary?

Speaker



Deborah Paul

Species File Group (INHS), University of Illinois

Authors

Matthew Yoder (University of Illinois, PRI, Illinois Natural History Survey, Species File Group)

Machine-Ready Botanical Data: Mapping Integrated Collections and Traits with Darwin Core Data Package

11:15 AM – 11:30 AM

AI AND ROBOT READY

The Darwin Core Data Package (DwC-DP) provides a powerful foundation for representing biodiversity data in ways that are both AI-ready and round-trip capable. We present a DwC-DP concept mapping developed for the Missouri Botanical Garden (MBG) that integrates herbarium specimens, living collections, and trait (extended specimen) data within a unified framework. To our knowledge, this represents the first detailed cross-collection DwC-DP concept map in the botanical domain, demonstrating how heterogeneous institutional data can be aligned into a coherent, machine-actionable structure.

We identify key areas in the mapping where controlled vocabularies are needed to further define the standard within the relevant communities. In particular, traits and character states constitute a critical class of information for AI applications, such as taxon identification and ecological modeling. While DwC-DP provides a structural basis for integrating such data, practical use requires alignment with external ontologies. We discuss an approach under development at MBG to integrate a trait ontology with DwC-DP, building on an internal pilot system.

Furthermore, we examine a gap in the current design relevant to AI-readiness around identifications: the need to capture not only the data that informs a determination, but the evidentiary and procedural basis for the decision. We consider how such information might be modeled externally to the core design.

This work highlights both the immediate utility of DwC-DP for institutional integration and the collaborative work needed, particularly around shared vocabularies, to advance interoperable and AI-ready botanical data systems.

Speaker



Nathan VanderKraats

Vice President, Biodiversity Informatics & Digital Infrastructure | Missouri Botanical Garden

Authors

Taylor AuBuchon-Elder (Missouri Botanical Garden), Matt Austin (Missouri Botanical Garden), Tony Mast (Missouri Botanical Garden), Molly Rooney (Missouri Botanical Garden), Paul Smock (Missouri Botanical Garden), Heather Stimmel (Missouri Botanical Garden), Rebecca Sucher (Missouri Botanical Garden), Jordan Teisher (Missouri Botanical Garden), Brian Yuan (Missouri Botanical Garden)

From Images to Structured Data: A Scalable AI Workflow for Natural History Museums

11:30 AM – 11:45 AM

AI AND ROBOT READY

Natural history museums worldwide face a major digitisation backlog. To unlock this critical bottleneck in research, the biodiversity informatics community is rapidly developing AI-based workflows to automate specimen label transcription and thus speed up the digitisation process. We present here a case study in which we describe an effective AI workflow for extracting specimen label data using Gemini 3.

The workflow combines batch processing with a two-stage Gemini large language model pipeline via API integrations. High-resolution specimen images are first processed to transcribe handwritten and printed labels. After post-processing to remove duplicate data, extract identifiers and fix common transcription errors, the cleaned transcriptions are resubmitted for data structuring using a predefined JSON schema aligned with collection management requirements.

A central design principle is provenance-aware, non-destructive integration of the data into the Finnish museum collection management system. Existing human-curated data are never overwritten. The workflow also records metadata describing which fields were generated by AI, while process-level provenance metadata are archived to support transparency, auditability, and reproducibility.

To support scalable quality assurance, the pipeline automatically generates validation reports from a random sample for human review. Our presentation discusses workflow architecture, provenance strategies, validation models, and lessons learned from deploying AI-assisted transcription at an institutional scale.

Speaker



Anne Koivunen

Head of Digitisation | Finnish Museum of Natural History

Authors

Mikko Heikkinen (Finnish Museum of Natural History)

Closing the Loop: The DiSSCo Annotation Validation Framework for Trustworthy Data Round-Tripping

11:45 AM – 12:00 PM

AI AND ROBOT READY

The Distributed System of Scientific Collections (DiSSCo) represents a paradigm shift in specimen data infrastructure, moving from a "custodian-only" model to a distributed, community-driven data lifecycle. Central to this is the ability for human experts and artificial intelligence (AI) agents to enrich Digital Specimens through annotations. However, as AI-driven workflows scale up, there is a risk of specimens becoming annotated with hundreds of low-quality annotations, which threatens synchronization with the source data. Without robust governance, collection managers will face "annotation fatigue," stalling "round-tripping"—the flow of data enrichments back to Collection Management Systems (CMS).

This presentation introduces the DiSSCo Annotation Validation Framework, which went through an international public consultation process in mid-2026. For responsible open digital curation, it treats annotations as first-class digital objects with full provenance, persistent identifiers (Handles) and attribution.

At its technical core, the framework differentiates by the scientific impact of the proposed data change. It aims to find a balance between low-impact corrections, which can be processed efficiently through automated services and high-impact modifications such as taxonomic reidentifications or stratigraphic updates, which should trigger a curated peer-review workflow.

By filtering and prioritizing incoming enrichments, the framework ensures that data round-tripping to source CMS remains high-quality, manageable, and highly trusted. This mechanism effectively closes the loop between external research communities, AI annotators, and physical collection institutions, ensuring Digital Specimens remain dynamic, accurate, and reflective of the latest scientific consensus.

Speaker



Wouter Addink

Manager Software Development | Naturalis Biodiversity Center

Geonomia: Metadata and Community for Georeferencing

12:00 PM – 12:15 PM

AI AND ROBOT READY

There has been considerable progress and investment into the digitisation of natural history collections, particularly among larger institutions that hold globally distributed specimen collections. These specimen records provide the "what, where, when" evidence used as the basis of much research. Data handling pipelines often include data cleaning steps which enhance our understanding of the core "what, where, when" and other attributes of the specimen record. Attributing these enhancements, along with acknowledging their source, to the digital specimen record is difficult due to the distributed nature of research tasks, many of which take place outside the holding institution of the specimen.

Distributed curation and management of digital specimen records are a necessary step in their use to support specific research interests and objectives, resulting in rich but fragmented, unusable data. Data are easily transferable, though trust in the data is essential for passive cross-institutional collaboration, particularly for subjective and ad-hoc tasks such as georeferencing.

With a focus on improving geographic coordinates or retrospectively adding new spatial data, attendees will be presented with the Geonomia project, which clusters GBIF-mediated specimen data to uncover field collecting patterns. The system includes a georeferencing application which will be demonstrated alongside community workshop outputs focussing on the digital curation of the georeferencing process and outputs. The aim of the project is to encourage the re-integration of comprehensive georeference metadata and continually improve the context of field collecting patterns.

Nicolson, N., Bachman, S. and Whittaker, A. (2026) Geonomia. Available at: <https://github.com/geonomia> (Accessed: 27/05/2026)

Speaker



Nicky Nicolson

Senior Research Leader, Biodiversity Informatics | Royal Botanic Gardens, Kew

Authors

Steven Bachman (Royal Botanic Gardens, Kew)

From satellite embeddings to Darwin Core: an open, ensemble-AI workflow for *Prosopis juliflora* mapping and reusable training-dataset generation

12:15 PM – 12:30 PM

AI AND ROBOT READY

Prosopis juliflora is among the most consequential invaders of global drylands, drawing down groundwater, depressing crop yields, and displacing grassland and pastoral commons. In Kutch, Gujarat, where *Prosopis* has overrun gauchar lands and Banni grassland fringes, classifier outputs guide biomass sourcing for a pyrolysis facility at Anjar, so every update carries operational consequence.

The pipeline fuses three openly licensed inputs on Google Earth Engine. AlphaEarth embeddings provide a learned per-pixel representation. Sentinel-2 contributes NDVI and EVI. SRTM derivatives (slope, roughness, TPI) separate *Prosopis* from similar canopy on rocky slopes. A four-model ensemble (two Random Forests, GTB, CART) is trained on versioned polygons tagged by source: presences, absences, false-positive corrections, and positive refinements. The correction loop turns model errors into named, attributable training records rather than silently retraining over them. Ensemble accuracy reaches ~91% on held-out test data.

Layered masks (hills, water, road buffers) exclude implausible pixels; each is a named, recorded operation with justification and source, auditable downstream.

Every positive pixel becomes a Darwin Core occurrence with basisOfRecord=MachineObservation, coordinateUncertaintyInMeters from pixel geometry, and identifiedBy naming both human curators and model version. Ensemble probability is carried as a MeasurementOrFact extension. The dataset, training polygons (CC BY 4.0), mask metadata, and model artefact are structured for GBIF publication, closing the round-trip from human curation to AI enrichment to reusable training material.

Speaker



Pranav Jha

Authors

Khyati Thacker (Climitra Carbon Private Limited), Shaurya Sharma (Climitra Carbon Private Limited), Aryaman Mihir Seth (Climitra Carbon Private Limited), Shubhankar Mihir Seth (Climitra Carbon Private Limited)

CO1 National & Global Taxonomic Backbones

11:00 AM – 12:30 PM | Location: SAL C

CONTRIBUTED ORALS

Moderators



Ely Wallis

Engagement Team Lead | Atlas of Living Australia, CSIRO



Elie Mario Saliba

Post-doc/TDWG Deputy Chair | Bernice Pauahi Bishop Museum

The United States Contribution to the Global Taxonomy of the "Catalogue of Life" in the First Quarter of the 21st Century

11:00 AM – 11:15 AM

CONTRIBUTED ORALS

In 1996, the Catalogue of Life (CoL) was planned as a comprehensive inventory of all valid species known to biological science globally. Creating this list required a coherent classification and representative synonymy. The first edition of CoL was published on CD in 2001. **The Integrated Taxonomic Information System (ITIS)** became a major data provider, contributing over half of the species in the first edition (incl. mammals, birds, and many other groups). Crucially, ITIS also provided the global hierarchy that served as the CoL Management Classification for many years. **The Species Files**, another big US contributor, joined CoL in 2004, gradually adding global checklists for Orthoptera, Blattodea, Phasmida, Aphidomorpha, Mantodea, Plecoptera, Embioptera, Coreoidea, Psocodea, Dermaptera, Lygaeoidea, Isoptera, etc. Over 25 CoL annual editions, other US-based global databases have participated in CoL: **Hexorallians of the world** (2002–2017); **ICTV** (2002–); **Moss TROPICOS** (2002–2022); **Systema Dipterorum** (2005–); **LecyPages** (2007–2017); **LDL Neuropterida** (2009–); **Brassicaceae** (2010–2023); **Early Land Plants Today** (2010–2022); **3i Auchenorrhyncha** (2011–); **Membracoidea of the World Database** (2011–2024); **Entiminae** (2016–); **PBI Plant Bug** (2016–); **The World List of Cycads** (2016–); **PaleoBioDB** (2017–); **Gymnodinium** (2018–); **Jewel Beetles** (2018–); **StaphBase** (2019–); **Sepidiini tribe** (2021–); **World Catalogue of Opiliones** (2021–); **Global Lepidoptera Index** (2022–); **Bryonames** (2023–); **LWS fleas** (2023–); **The World Odonata List** (2023–); **AntCat** (2025–). By 2025, US-based and expert-curated databases contributed 793,339 species, or 36% of the entire 2.2-million-species CoL base release. Also, US experts support CoL, making significant contributions to various biodiversity projects worldwide.

Speaker



Yury Roskov

Catalogue of Life Executive Editor | Species Files Group, INHS, University of Illinois at Urbana-Champaign

A Knowledge Organisation System for Taxonomic Data

11:15 AM – 11:30 AM

CONTRIBUTED ORALS

With the Taxon Concept Standard (TCS) now ratified, the community has a framework to better organize taxonomic data. While often designed for specific Floras, these models can serve as reference implementations for TCS and broader Knowledge Organisation Systems (KOS). This presentation introduces a model developed for VicFlora that is oriented around the Taxon Concept as the core of the system, with all other data layers acting as supporting components.

The model is structured across six key layers:

- **Semantics:** Taxon Concepts and mappings
- **Syntax:** Names, relations, usages, and typification
- **Governance:** Taxon trees and versioned nodes
- **Authority:** Provenance (references) and attribution (agents)
- **Narrative:** Profiles and descriptions
- **Identity:** External identifiers.

To manage complexity, we employ a "sidecar" pattern. Extension sidecars manage name types and reference roles, while functional sidecars allow Taxon Concepts to remain immutable, while their context can change. By decoupling these, we can update classifications (via Tree Nodes) or modify descriptive content (via Profiles) without redefining the concept itself. The Profile serves as the primary extension point, bridging internal models with other TDWWG standards like Plinian Core, Darwin Core, and SDD.

We distinguish between narrowly defined relationships and polymorphic ones—such as provenance and attribution—which require flexible handling. This architecture provides a blueprint for building taxonomic systems that are theoretically sound under TCS and functionally agile, demonstrating how a concept-centric KOS can effectively integrate keys, maps, and occurrences.

Speaker



Niels Klazenga

Biodiversity Informatics Developer | Royal Botanic Gardens Victoria

Managing observations when taxonomy changes: the case of Artportalen and Dyntaxa

11:30 AM – 11:45 AM

CONTRIBUTED ORALS

In the Swedish taxonomic database Dyntaxa (www.dyntaxa.se), updates are published continuously. This taxonomy is used by the Swedish citizen observatory Artportalen (www.artportalen.se). However, while about one million observations are added to Artportalen every two months, updates of taxonomy are only made once during this time.

In Dyntaxa, the taxon, with its corresponding taxonID, is the basis. The name for such a taxon can change without changing taxonID. Lumps are easy to deal with by pointing an unaccepted taxon (with its taxonID) to the accepted one. Technically, splits can also be dealt with automatically by keeping all observations on the old (sensu lato) taxon, even though it is no longer considered a species but essentially a collective taxon. However, this is usually unsatisfactory. When there is more information about the split you might be able to move observations to one or several of the new species. In Sweden, these cases are relatively few despite the volume of observations.

Another way is to inform the reporter of such observations of these changes. They will be presented with the option to move them if they now know what new species they observed.

The Dyntaxa data model can handle not only these relations between taxa, such as splits and different kinds of lumps, but also convey the reason for them and the statuses of the taxa involved. When an editor is working in Dyntaxa, lumping a species with another, an automated text is generated. This text summarizes which taxonIDs were involved and that, e.g., taxonID 25 was merged into taxonID 42. This is also stored in a machine-readable format in the database.

We show examples where automation works fine, and some where we currently assess them manually but could possibly automate more in the future.

Speaker



Johan Liljeblad

Taxonomist | SLU Swedish Species Information Centre

Authors

Mora Aronsson

Modernizing Brazil's Taxonomic Backbone: How ChecklistBank Supports Biodiversity and Genetic Heritage Systems

11:45 AM – 12:00 PM

CONTRIBUTED ORALS

Brazil, has made major advances in organizing and managing taxonomic information. It is one of the few countries with two official and continuously maintained national species catalogues: one for Flora and Funga and another for Fauna, curated by more than 800 taxonomists.

Both lists are published monthly using IPT and shared with GBIF and ChecklistBank. Taxonomy is the essential backbone for biodiversity information systems, since taxonomic updates require continuous reprocessing of species occurrence records.

Brazil operates two major systems dependent on an updated taxonomic backbone: SiBBr - Brazilian Biodiversity Information System), a Living Atlases platform and SIGGEN, the Brazilian Genetic Heritage Management System, where all research involving genetic heritage must legally register activities and outputs.

ChecklistBank has played a key role in strengthening Brazil's biodiversity taxonomic backbone. Initially, it supported detailed quality assessment of the official lists, identifying duplicated names, typographical errors, inconsistencies in Portuguese Darwin Core terminology, and other standardization issues. More recently, it enabled the construction of a unified Brazilian taxonomic backbone by integrating the two official catalogues, incorporating invasive species, generating identifiers for taxonomic tree construction, merging extensions, and producing a Darwin Core Archive (DwC-A) directly consumed by national systems.

Previously, this workflow in SiBBr required extensive technical effort to clean, harmonize, and generate updated taxonomic files. With ChecklistBank and the support of its technical team, the process became more efficient, transforming a complex operation into a routine update performed every six months.

Speaker



Clara Baringo Fonseca

Biodiversity Data Specialist at SiBBr | SiBBr (Brazilian Biodiversity Information System - RNP

Authors

Keila Juarez

Building a Taxonomic Reference Authority for Conservation: The Digital Catalogue of Biodiversity of Poland

12:00 PM – 12:15 PM

CONTRIBUTED ORALS

The Digital Catalogue of Biodiversity of Poland (DCBP) is a comprehensive inventory of all confirmed eukaryotic species and Cyanobacteria within Poland's administrative borders. The content is divided into taxonomic chapters managed by expert authors responsible for validating species composition and taxon attributes. These attributes cover occurrence status, threat categories, invasiveness, and other expandable characteristics. Each species entry is backed by a published reference or expert opinion documenting its presence in the country.

Currently, the DCBP comprises 100+ chapters published as individual GBIF checklist datasets and integrated into the Biodiversity Map (BioMap) portal, developed by the Polish GBIF Node. The total repository contains 73K species, including 36.7K animals, 19.9K fungi, 9.9K plants, 4.9K chromists, 0.5K protozoans, and 0.9K bacteria.

DCBP taxa belong to the taxonomic tree of the BioMap database, built on a PostgreSQL relational framework. This architecture connects checklists, species occurrences, datasets, collections, publications, and authors. The BioMap portal enables authors to manage their respective lists, edit or add new taxa, and update literature sources. Furthermore, the DCBP maps Catalogue of Life (CoL) taxon identifiers to facilitate external interoperability. We aim to establish the DCBP as the primary reference authority supporting Polish biodiversity research and conservation. Alongside other European national checklists, the Catalogue may serve as a valuable foundation for regional initiatives.

Speaker



Piotr Tykarski
University of Warsaw

National Biodiversity Information Action Campaign Production Process: A Case Study of Taiwan

12:15 PM – 12:30 PM

CONTRIBUTED ORALS

With the Kunming-Montreal Global Biodiversity Framework (K-MGBF), a robust biodiversity information infrastructure is essential for Taiwan's conservation goals. Since 2021, the Taiwan Biodiversity Information Alliance (TBIA) has implemented a national biodiversity data integration initiative through the 2021–2025 Action Campaign, focusing on species occurrence data, a Biodiversity Database Portal, inter-agency collaboration, and data standardization.

Building on this foundation, the 2026–2030 Campaign shifts toward long-term system resilience and operational sustainability. The Campaign was co-developed by an inter-agency working group, drawing on five years of experience while highlighting challenges in cross-domain coordination and staffing capacity.

Despite progress, key challenges remain in cross-domain governance, data circulation, integration of long-term monitoring needs such as Essential Biodiversity Variables (EBVs), and sustainable institutional and financial support.

In response, the Campaign defines five strategic directions: enhancing data openness and quality; strengthening interoperability via APIs and taxonomic checklists; building a comprehensive public-private network with stable funding; facilitating inter-agency technical exchange; and enhancing data usability through improved validation and citation practices. These strategies address governance, technical, data, and financial dimensions of infrastructure resilience.

This framework serves as an implementation-oriented synthesis, translating strategy into actionable guidance. This case study illustrates how biodiversity data infrastructures can evolve toward greater resilience, offering a reference for similar global initiatives.

Speaker



Chun-I Chang
Product Manager | Taiwan Biodiversity Information Facility, Biodiversity Research Centre, Academia Sinica Taoyuan City, Taiwan

Authors

Jerome Chie-Jen Ko, Melissa Jean-Yi Liu, Chia-Chi Wu, Chien-Wen Chen, Ping Huang, Jhu-Jyun Jhang, Ruey-Shing Lin, Li-Chung Lu, Chih-Yi Tsai

SYM11 - Data Quality: From Standard to Practice

11:00 AM – 12:30 PM | Location: ODIN

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

The utility of biodiversity data is defined by its fitness for use, a state achieved only when data is measured against rigorous, transparent benchmarks. The Biodiversity Data Quality (BDQ) standard is central to this work. BDQ provides the necessary foundation for consistent tests and assertions across the community. This session explores BDQ and related 'data quality' concepts and applications.

The session will provide an overview of BDQ by Lee Belbin, the BDQ ontology by Paul Morris, BDQ vocabularies by Arthur Chapman and BDQ-Darwin Core relationships by John Wiecezorek. Related presentations are encouraged.

A standard only gains value through its implementation and the subsequent communication of its results. Following the presentation session, a hands-on workshop led by **Paul Morris** will guide participants through the technical realities of implementing the BDQ

standard and interpreting its responses. This session will aim at ensuring data managers and researchers leave with the skills to apply these principles to their own datasets.

Moderators



Lee Belbin

Director | Blatant Fabrications Pty Ltd



Arthur Chapman

CEO | Australian Biodiversity Information Services

Introduction

11:00 AM – 11:05 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Speaker



Lee Belbin

Director | Blatant Fabrications Pty Ltd

An Introduction to the Biodiversity Data Quality Standard (BDQ)

11:05 AM – 11:20 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

The quality of biodiversity data is a critical determinant of its utility for science and policy. The **Biodiversity Data Quality (BDQ) standard** is a comprehensive, modular and flexible framework designed to assess the "fitness for use" of biodiversity data records.

The development of BDQ has been an 12-year journey. The need was realized at 2010 TDWG in Woods Hole. Work began in 2014 with the establishment of the BDQ Interest Group and its Task Groups. This work resulted in the 2020 publication "Developing Standards for Improved Data Quality" and the subsequent submission of BDQ in 2025.

The BDQ framework is anchored in the **Fitness for Use Framework Ontology (bdqffdo)** to provide the underlying semantics for data quality needs and solutions. Supporting this ontology are specialized vocabularies that define tests (e.g., Empty, ActedUpon), data quality dimensions (e.g., Completeness, Conformance), criteria (e.g., Found, Standard) and enhancements (e.g., Converted, Standardized).

110 core Single-record tests are the heart of BDQ. They are categorized into four **Test Types**: **Validation** - Factchecking (e.g., VALIDATION_LATITUDE_INRANGE), returning results like COMPLIANT or NOT_COMPLIANT; **Issue** - Flag potential problems (e.g., ISSUE_ANNOTATION_NOTEMPTY); **Amendment** - Proposes improvements (e.g., AMENDMENT_MONTH_STANDARDIZED) and **Measure** - Metrics that quantify quality (e.g., MEASURE_AMENDMENTS_PROPOSED).

The standard explicitly supports two primary data management workflows: **Quality Control (QC)**: Focuses on identifying and remedying errors to improve a dataset—the "run → analyze patterns → fix causes → re-run" loop. **Quality Assurance (QA)**: Focuses on filtering records to ensure only those fit for a specific use, such as species distribution modeling, are used in analysis.

Speaker



Lee Belbin

Director | Blatant Fabrications Pty Ltd

Authors

Arthur Chapman, Paul Morris, John Wiecezorek

GBIF Rule Based Annotations

11:20 AM – 11:34 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Biodiversity data quality is commonly run through automated tests and standardized assertions, yet many data quality issues are not captured by existing pipelines.

We present [rule-based annotations](#) as a practical approach for translating informal expert knowledge into structured, reusable data quality rules. Simple statements such as "there are no dragonflies in Antarctica" can be formalized as rules linking taxa, geography, and controlled vocabularies (e.g. suspicious, native, introduced), and applied to both current and future occurrence records.

Unlike traditional cleaning approaches or one-off feedback to data publishers, rule-based annotations are persistent, shareable, and extensible. Rules can incorporate additional constraints (e.g. basis of record, time, dataset) and are managed through an [experimental community-driven system](#). This enables a flexible, opinion-based layer of data quality assessment that complements existing standards and automated checks.

We demonstrate the approach using a case study on Odonata, where expert-defined rules identify a small number of outliers not captured by standard cleaning pipelines, yet with disproportionate effects on downstream species distribution models.

Just as filtering rules in an email inbox serve to filter and sort incoming emails, rule based annotations are meant to be future looking. Once a rule is created it is saved to a database, and is accessible to all GBIF users. Users can leave comments and upvote or downvote based on their perception of the general usefulness of the rule.

Speaker



John Waller

GBIF

Interpreting the Responses from BDQ Tests

11:34 AM – 11:48 AM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

BDQ (Biodiversity Data Quality) is a standard designed to facilitate the consistent development and use of a set of biodiversity data quality tests and responses from these tests.

The Standard takes Darwin Core (Wieczorek et al., 2011) encoded data, evaluates quality using data quality Tests and produces a report. The reporting from of BDQ Tests is the focus of this presentation (see [User's Guide to the Standard](#)).

Tests for data quality may serve two purposes: Quality Control and Quality Assurance.

- In Quality Assurance (QA), data are evaluated for fitness for use and are filtered down to those data that are fit for that use.
- In Quality Control (QC), tests are used to identify data that are not fit for a specific use. The goal is to improve data quality by proposing changes to the data.

A Data Quality Report, produced from running a BDQ Tests is expected to include a Response for each Test run.

In brief, each Response consists of the following elements:

- **Response.status** - Metadata describing the status of the Test run, including whether the Test was executed successfully. Values include RUN_HAS_RESULT, INTERNAL_PREREQUISITES_NOT_MET and EXTERNAL_PREREQUISITES_NOT_MET.
- **Response.result** - The result of the evaluation of the Test against the input data. Values include "COMPLIANT", "NOT_COMPLIANT", "COMPLETE", "NOT_COMPLETE", or data values depending on the type of Test.
- **Response.comment** - A human-readable comment providing additional context or information to assist users in the interpretation of the Test result.

These responses are dependant of the type of Test – Validation, Amendment, Issue or Measure.

The presentation will discuss these issues, and various pitfalls, especially with Amendment Tests, in using the data to improve databases of record.

Speaker



Arthur Chapman

CEO | Australian Biodiversity Information Services

Authors

Paul Morris, John Wieczorek, Lee Belbin

The Swedish experience on verification and how to improve quality on observations

11:48 AM – 12:02 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Quality on observations is always a challenge as we often know was needed but human and/or economic resources are lacking. Today technical validation is well implemented in most observations system (date is existing and relevant, coordinates is in a relevant area and more). AI has the last years been of great help to find anomalies in the reported observations, but it still a need for a human eye and brain to perform the last step. To get the final approved och not approved. For some observations (e.g. red listed species under debate by forestry or species under legal act as species listed in European regulation on invasive species) the verification by human is needed for companies and agencies to guarantee that the observation is correctly identified and reported from right place.

Since more than 2000 years Sweden has been evolving a quality assessment of observations, both by technical validation and human verification to reach as good quality as possible. The last five years have the verification system been improved to meet new legal requirements with more strict rules and, new tools and improved education of both the users and the persons doing verification. The Swedish system will can also be used as test for the BDQ-Standard.

The presentation will cover the system with verification system, categories, contract and education of people doing the verification and contract. It will also cover the results from the work with the "new" verification system in Sweden.

Speaker



Mora Aronsson

Senior Adviser | Swedish University of Agricultural Sciences

Authors

Marit Persson Rådén, Johan Liljebblad

Transitivity Failures in Taxonomic Name Resolution: Implications for Botanical Dataset Construction

12:02 PM – 12:16 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

A key decision when creating datasets about plants is whether to store the scientific names as originally recorded in a data source or to resolve them to currently accepted names. When a scientific name is resolved to an accepted name, and the original name is discarded, information is lost regarding the description and taxonomic classification used in the original identification. In practice, this may have limited impact if the original and accepted names consistently resolve to the same name in future taxonomic databases. However, this depends on the *transitivity* of name resolution across taxonomic database versions i.e. given names X, Y, Z and taxonomic database versions v1, v2 such that X resolves to Y in v1 and Y resolves to Z in v2, is it the case that X resolves to Z in v2? If this property does not hold in general then creating datasets by storing only resolved, accepted names may create unexpected outcomes with wide-ranging consequences for botany, and biological sciences more broadly, especially in an era where large scale data analyses are becoming increasingly prevalent in addressing many real-world challenges. Using the World Checklist of Vascular Plants and World Flora Online, we demonstrate that this transitivity fails for approximately 1% and 6% of plant names, respectively, and that the number of these discrepancies increases as new taxonomy versions are released. These results highlight the importance of storing verbatim names as found and resolving them to accepted names as close to the point of end-use as possible, using the most recent taxonomic database.

Speaker



Adam Richard-Bollans

Kew Research Fellow | Royal Botanic Gardens, Kew

Authors

Bob Allkin, Francesco Civita, Kristina Patmore, Rafaël Govaerts

How Darwin Core enables harmonizing Biodiversity Data via the OSCA Consortium

12:16 PM – 12:30 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Digitizing vast biological and geological collections presents a significant multi-disciplinary challenge, requiring the integration of domain expertise, information technology, and robust scientific infrastructure. The **Open Scientific Collections Austria (OSCA)** consortium (15 museums, universities, and scientific institutions) was created to digitize Austria's biodiversity and geological history. These collections serve as critical longitudinal baselines for researching global crises, including climate change and biodiversity loss.

Central to this initiative was the adoption of the **Darwin Core (DwC)** data standard, which provided the foundational framework for cross-institutional interoperability.

We developed a centralized, cloud-based **Cache database** designed to aggregate DwC-compliant datasets from all consortium members. This infrastructure supports a coherent workflow that facilitates high-quality data production by both technical and non-technical researchers. To ensure global reach, the system automates the publication of records to international repositories, specifically **GBIF** and **DiSSCo**.

A dedicated **public Portal** was developed to enhance collaboration and provide open access to Austrian scientific data. The portal allows users to search, compare, and download records in DwC format. By enabling the cross-referencing of published data, the platform assists the scientific community in quantifying available information and identifying strategic gaps for future digitization. This integrated approach demonstrates how standardized data frameworks and collective infrastructure can overcome funding and technical barriers to unlock the potential of natural history collections.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Cristian-Dan Bara

Digital Product Owner (Scientific Collaborator) | Naturhistorisches Museum Wien

Authors

Frederike Korth, Heimo Rainer, Nedim Begic

WKS12 - Strengthening National Biodiversity Data Infrastructures through International Coordination

11:00 AM – 12:30 PM | Location: Møterom 5 / Meeting room 5

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Inspired by a symposium at the TDWG conference 2022 in Sofia, nine national biodiversity data infrastructures (NBDIs) from around the world have jointly developed and published a shared definition of their roles in relation to international services and initiatives. This work resulted in the identification of ten

essential functions that capture the core activities of national infrastructures and highlight their relevance for science, policy, and practice (<https://doi.org/10.1093/biosci/biae109>).

Building on this foundation, the proposed workshop aims to further consolidate collaboration among NBDIs by exploring the establishment of an NBDI Interest Group under the umbrella of TDWG Biodiversity Information Standards. Participants will discuss potential areas of activity, priorities, and modes of collaboration, with the goal of co-developing a draft charter for the Interest Group.

The workshop is designed as an interactive forum, providing ample space for open discussion and exchange of perspectives, while also offering opportunities for short presentations that showcase national experiences and strategic developments, ongoing initiatives, and emerging challenges. By bringing together NBDI coordinators, managers, technical experts, and other key stakeholders, the workshop will strengthen international collaboration and foster alignment between NBDIs and the broader TDWG community.

Workshop leaders



Anton Güntsch

Director, Center for Biodiversity Informatics | Freie Universität Berlin, Botanic Garden Berlin



David Fichtmueller

Computer Scientist | Botanic Garden and Botanical Museum (BGBM) Berlin

Authors

Tanja Melanie Weibulat, Victor Heijke

The Role of NFDI4Biodiversity within Germany's National Biodiversity Data Infrastructure

11:05 AM – 11:15 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Güntsch et al. (2025) define ten essential functions of National Biodiversity Data Infrastructures (NBDIs): providing base infrastructure; facilitating data mobilization, coordinated digitization, and the integration of regional biodiversity assessments and taxonomies; supporting users of international infrastructures; promoting standards; enabling citizen science; addressing legal issues; providing platforms for quality control, data integration, and analysis; and coordinating national funding. In this contribution, we discuss the role of NFDI4Biodiversity within Germany's NBDI.

Our analysis shows that the NFDI4Biodiversity consortium contributes to many, but not all, of the functions. It facilitates data mobilization, promotes standards, operates a helpdesk, integrates citizen science data, and develops cloud-based platforms for data integration. Its consortium partners provide base infrastructure and specialized services, while the NFDI framework provides legal expertise and supports strategic funding coordination.

Mandates for the ten functions, however, are distributed across a broader ecosystem. GBIF Germany and GFBio play key roles in data mobilization, the Federal Agency for Nature Conservation (BfN) coordinates national biodiversity monitoring, natural history museums lead digitization efforts, and specialized networks organize citizen science activities. NFDI4Biodiversity should therefore be understood not as the main provider of Germany's NBDI functions but as a coordinating and integrating actor within a heterogeneous infrastructure landscape.

Güntsch, A., et al. (2025) National biodiversity data infrastructures: ten essential functions for science, policy, and practice. *BioScience* 75.2

Speaker



Birgitta König-Ries

Friedrich Schiller University Jena/German Centre for Integrative Biodiversity Research (iDiv)

Authors

Barbara Ebert (GFBio - German Federation for Biological Data)

Assessing the maturity and sustainability of the LifeWatch Italy Data Portal using the Data Infrastructure Maturity Profile

11:15 AM – 11:25 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

LifeWatch Italy is the Italian Distributed Centre of LifeWatch ERIC, the European e-Science infrastructure for biodiversity and ecosystem research. It provides a full suite of services to support the scientific community through the entire research lifecycle, promoting best practices in data management. The Infrastructure aims to be a trusted and widely used hub for biodiversity research at the national level. However, assessing its maturity and keeping track of its status requires navigating a complex environment involving stakeholder coordination, financial security, human resources, and alignment with evolving technological standards. A newly developed comprehensive assessment framework, namely the Data Infrastructure Maturity Profile (DIMP), was applied to systematically assess the implementation of these critical components. It consists of more than 100 questions to document key characteristics, such as software and (meta)data management practices, governance, and sustainability. The DIMP was tested against the LifeWatch Italy Data Portal to identify strengths and areas for improvement. The assessment reveals that the LifeWatch Italy Data Portal is a FAIR-compliant platform that aligns well with community standards. It demonstrates powerful semantic capabilities and maintains high standards for data quality and validation. The analysis also identified several challenges: policies for data citation and long-term data preservation remain suboptimal, and the infrastructure's reliance on competitive funding creates sustainability risks. Furthermore, transparency and visibility towards stakeholders should be improved. Working to address such gaps represents a significant opportunity to consolidate LifeWatch Italy's role as a more resilient and trusted e-infrastructure.

Speaker



Andrea Tarallo

National Research Council of Italy (CNR), Institute of Research on Terrestrial Ecosystems (IRET)

Tackling the Transcription Bottleneck: A Multi-Agent AI Approach for Collections

11:00 AM – 12:30 PM | Location: BALDER

COLLECTIONS DIGITISATION & MANAGEMENT

Natural history collections hold vast research potential, but unlocking it requires transcribing specimen labels into structured data. Manual transcription is expensive and time-consuming, creating a bottleneck that leaves millions of digitized specimens without accessible data. This session presents our work on the Pixel Data Extraction Platform (PDEP), a multi-agent AI system we are developing to automate extraction, validation, and standardization of collection data from specimen images.

PDEP uses collaborative AI agents that work together on extraction tasks. A Project Manager Agent coordinates specialized agents for OCR/HTR processing of handwritten and printed labels, geographic normalization, collector name standardization, and taxonomic validation. Agents use Retrieval Augmented Generation (RAG) to validate and enrich extracted data against authoritative sources and curated reference lists. When high confidence is required, multiple agents can cross-validate findings through consensus mechanisms before a QA Agent performs final review.

The platform consumes image data from various sources, including Picturae's high-throughput conveyor systems, and produces Darwin Core-compliant output for integration into biodiversity data infrastructures.

We will discuss our architecture decisions, the challenges we are tackling around diverse label formats and languages, our approach to human-in-the-loop validation, and early lessons learned from development.

Workshop leader



Sebastiaan Bouten

VP of Sales EMEA / APAC | Picturae BV

Co-speakers



David Thornley

IT Architect | Picturae



Terence Doets

Picturae BV

LUNCH

12:30 PM – 2:00 PM | Location: 1st Floor

SYM08A - BHL at 20: A New Chapter for Biodiversity Literature and Data – Reimagining BHL: Foundations, Partnerships and Future Infrastructure

2:00 PM – 3:30 PM | Location: SAL A

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

In 2026, the [Biodiversity Heritage Library](#) (BHL) celebrates both its 20th anniversary and its first year as an independent consortium. Following its [transition](#) away from long-term hosting at the Smithsonian Institution, BHL is now operating from a stable foundation, with its people, expertise, and services firmly in place and

renewed momentum across its global network.

This session will explore where BHL stands today and what this new phase makes possible. Presentations will examine the outcomes of the transition, including updated governance structures, evolving funding approaches, strengthened partnerships, and technical developments that position BHL for sustained growth and innovation within the biodiversity data ecosystem.

Speakers will present emerging applications of BHL content and services, such as artificial intelligence and machine learning, large-scale text and image mining, knowledge graphs, annotation systems, and efforts to improve persistent identifier reliability and stewardship. Talks will demonstrate how BHL functions as essential infrastructure connecting literature with taxa, specimens, institutions, and datasets, from the 1100s to the current year.

The session will also reflect on two decades of BHL's impact, from mobilizing historical biodiversity knowledge to supporting modern data integration and reuse, inviting contributors and attendees to consider how BHL can continue to evolve as a core component of global biodiversity information infrastructure.

Moderators



Nicole Kearney

Communications Director | Biodiversity Heritage Library



David Iggulden

Head of Data & Digital (Library & Archives) | Royal Botanic Gardens, Kew

Introduction: Reimagining BHL

2:00 PM – 2:10 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Speaker



Nicole Kearney

Communications Director | Biodiversity Heritage Library

Reconsidering the original vision for a global biodiversity information facility

2:10 PM – 2:25 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

A global biodiversity information facility was recommended in 1999 by the OECD Biodiversity Informatics Subgroup of the Megascience Forum. In 2001, 25 years ago this month, GBIF, the Global Biodiversity Information Facility, embarked to meet the challenges outlined by the Forum. Soon thereafter the scope was narrowed due to limited funding to meet the challenge. The original vision was a more encompassing of data from literature, traits, and taxonomy. GBIF focused on species occurrences and community building. Critically GBIF continues to grow as an intergovernmental organization, not affiliated with the United Nations, which provides sustainability. Other important infrastructure projects such as BHL, COL, EOL (and others) filled the void but lacked GBIF's governance framework. What if the original vision had been funded and pursued? Would it have worked? Given the considerable stress on science and infrastructures today, could the original vision work now?

Speaker



Joe Miller

Executive Secretary | GBIF

For the Birds: Adapting Archives and Extracting Data for the Biodiversity Heritage Library

2:25 PM – 2:40 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

In December 2025, University of South Florida Libraries' was honored to join the Biodiversity Heritage Libraries as a member, as well as to provide redundancy support for the technical team. As part of this new step, the team at USF is curating materials from their more than 30 collections of Florida Environment and Natural History (FLENH) materials and their extensive rare books collections. Ornithological research, including published and unpublished works, form the core of these holdings. While mixed in their formatting, the materials are rich in data that remained layered behind human prose, handwritten notes, and inconsistent checklists. Making these materials computer readable and adaptable for scientists to process at scale requires a heavy human lift. From local Audubon Society sanctuary records to the vast holdings of the Searchable Ornithological Research Archive (SORA), USF is exploring existing and adapted methods for ingesting their content into BHL over the next few years. Along the way, the team is working with scientists and student interns in various fields to test the accuracy of human and AI-methods to extract data. The team is hopeful that its efforts will provide a fresh case study for loading archival materials that can build on the work of other BHL members while leveraging internal expertise in cataloging, copyright, digital preservation, GIS, and data management. This talk will showcase the possibilities and priorities for the more than 106,000 already digitized records across the FLENH collections.

Speakers



Sydney Jordan

Collections Coordinator, Florida Environment & Natural History Collections | University of South Florida Libraries



Amanda Boczar

Associate Director, Digital Initiatives | University of South Florida

BHL and the Planetary Knowledge Base: Structuring Biodiversity Literature for the Next Generation

2:40 PM – 2:55 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

As BHL enters its second decade, the Natural History Museum (NHM) is assisting in its evolution into a machine-actionable, FAIR-aligned biodiversity information infrastructure. This work is connected to NHM's Planetary Knowledge Base (PKB), which integrates fragmented natural history literature, collections, observations, genetic and environmental data into an explainable system for research, policy and action for nature.

The BHL work focuses on developing a biodiversity knowledge-extraction and discovery layer that operates alongside the BHL, linking users back to BHL pages, publications and metadata while enabling new forms of structured search and analysis. NHM has prototyped workflows that transform BHL's page-centric volumes into structured, machine-actionable scholarly resources through modern OCR, article-level segmentation, figure and table extraction, persistent identifier assignment, and entity extraction for taxa, people, institutions, localities, specimen identifiers, habitats and stratigraphy.

These outputs will support semantic discovery, programmatic reuse, knowledge graph construction, and alignment with biodiversity infrastructures such as GBIF, Catalogue of Life, Wikidata, GRSciColl, etc. All extracted entities and relationships retain provenance links to the original BHL pages and bibliographic records, with confidence and workflow metadata. The dataset also serves as a foundational evidentiary layer for the Planetary Knowledge Base (PKB), helping ground biodiversity knowledge in the underlying BHL literature.

These activities position BHL as a distributed, resilient, interoperable evidentiary layer for biodiversity research, connected to specimens, taxa, collections, observations, and external datasets.

Speaker



Qianqian Hiris Gu

Natural History Museum London

Authors

Ben Hartley, Vincent Smith

Possibilities to enhance the semantic navigation of BHL - further developing the "unit of work"

2:55 PM – 3:10 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

The BHL data resource includes page images, scanned OCR text and bibliographic metadata regarding its holdings. Data are ingested into BHL using the units of volumes, parts and pages. Later projects such as biostor have asserted more useful units of work such as articles within serial titles by allocating a series of pages to an article with a title and authorship. The BHL user interface has a long-standing feature enabling navigation via the results of taxonomic indexing, which looks for occurrences of scientific names within page texts.

This talk will discuss the possibilities for building on the article identification concept and creating more domain specific units of work from the examination of contents pages, standard body pages and index pages to determine the layout and organisation of the information held within the part. Units of work could comprise biographical entries organised alphabetically by person name, dictionary or glossary entries organised alphabetically, species entries in specimen catalogues, monographs and floristic works ordered taxonomically, or field work sites and gazetteer entries organised geographically. Identification of these would help drive wider use of BHL-mobilised data and build stronger links with user communities such as those working with biographical data about collectors and identifiers in the Bionomia project, those seeking historical geographic data for georeferencing and taxonomic projects such as the World Flora Online

Speaker



Nicky Nicolson

Senior Research Leader, Biodiversity Informatics | Royal Botanic Gardens, Kew

Discussion

3:10 PM – 3:30 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Moderators



Nicole Kearney

Communications Director | Biodiversity Heritage Library



David Iggulden

Head of Data & Digital (Library & Archives) | Royal Botanic Gardens, Kew

SYM13B - Responsible AI, Open Digital Curation, and Round-Tripping for Biodiversity Data

2:00 PM – 3:30 PM | Location: SAL B

AI AND ROBOT READY

We propose a session focused on **open digital curation** as a continuous, inclusive process across the biodiversity data lifecycle. Presentations would illustrate how human and AI assistance can be made transparent, attributable, versioned, and reusable during data mobilization and enrichment (e.g. annotations and linking, taxonomic determinations, georeferencing improvements, use of persistent unique identifiers).

Particular attention would be paid to **round-tripping**: ensuring that annotations and AI-generated and human enhancements flow back to source systems (e.g., collection management systems and institutional repositories), rather than remaining disconnected products. By emphasizing openness, provenance, and interoperability, the session would explore practical approaches for integrating human work and AI responsibly and sustainably while maintaining curatorial authority and trust.

With an emphasis on natural history data from museums and herbaria -- including specimens, images, and genomic resources -- the session presentations would feature case studies and methods that bridge institutional practice, community engagement, and **open science** infrastructure. Contributions would address challenges such as scaling AI-enabled curation, validating and crediting annotations, aligning with standards like Darwin Core, and **designing workflows** and **user interfaces (i.e. UI/UX)** that support equitable participation and vetting by collection staff, researchers, and external contributors.

Moderators



Deborah Paul

Species File Group (INHS), University of Illinois



Siobhan Leachman

Wikimedian | Wikimedia Aotearoa New Zealand

Session organizer



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

Authors

Sabine von Mering

Open Digital Curation and Round-Tripping Biodiversity Data Enhancements Across Collection Infrastructures: Six Years and 50 Million Links in Bionomia

2:00 PM – 2:15 PM

AI AND ROBOT READY

Throughout the past 6 years, [Bionomia](<https://bionomia.net>) has supported open digital curation through machine-readable annotations, bidirectional synchronization workflows, and contributor attribution that preserves curatorial authority while enabling broader participation. Case studies will illustrate how annotations involving specimen linkages can be generated by humans and redistributed back to source repositories rather than remaining isolated within downstream aggregators. Bionomia's data processing tools, services, and outputs enable lightweight integration with collection management systems while preserving institutional ownership and editorial control over incoming enhancements. Recently, the human name parsing tool was successfully ported to a high-performance Go package using Claude AI. The platform also exposes mechanisms for crediting contributors in ways that support professional recognition, reproducibility, and reuse. The continuous cycle of incremental enhancements has not been without its technical challenges, and these will be discussed in relation to Darwin Core and data exchange structures used by our community. By framing digital curation as an ongoing and collaborative process rather than a terminal publication event, Bionomia demonstrates how round-tripping infrastructures can strengthen resilience, trust, outreach, and long-term stewardship across biodiversity data ecosystems despite uncertainties about any one project's governance and sustainability.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



David Shorthouse

independent

Type specimens in Wikidata

2:15 PM – 2:30 PM

AI AND ROBOT READY

Type specimens play a vital role in ensuring the correct application of scientific names. Recently a type specimen data model has been created for Wikidata. This data model results from work undertaken by participants of the [Wikidata:WikiProject Natural History Specimen Data Model](https://www.wikidata.org/wiki/Wikidata:WikiProject_Natural_History_Specimen_Data_Model)!. The data model encourages the creation of type specimen Wikidata items that link to appropriately structured data. Several Wikidata properties were proposed and accepted by the Wikidata community to facilitate the formation of this data model. The data model also encourages the linking of institutional identifiers to appropriate type specimen items.

Documentation was produced to guide editors in the creation and linking of type specimen Wikidata items. This included an exemplar spreadsheet generated from an OpenRefine project creating Wikidata items for holotype specimens held in the New Zealand Arthropod Collection.

A data model for structured data on Wikimedian Commons was also proposed by the WikiProject. This is intended to apply to natural history specimen media files added to Wikimedia Commons. Again documentation was produced to assist editors to create or add appropriate structured data to natural history specimen image files.

One of the core aims of the WikiProject was to align specimen and type-material modelling across Wikimedia projects with existing biodiversity informatics standards. As a result crosswalk documentation was produced to illustrate the mapping between fields of the Wikidata type specimen data model and Darwin Core. Crosswalk documentation was also produced mapping Wikidata properties used to add structured data on Wikimedia Commons to media files and equivalent Darwin Core terms.

Speaker



Siobhan Leachman

Wikimedian | Wikimedia Aotearoa New Zealand

Here, There, and Everywhere: Modelling Research Expeditions in Wikidata

2:30 PM – 2:45 PM

AI AND ROBOT READY

Research expeditions are major sources of natural history collections and connect distributed cultural heritage across institutions and countries. However, expedition data remain fragmented, inconsistently modelled, and often lack stable identifiers. This hampers discoverability, interoperability, and provenance tracking, while also limiting efforts to contextualise collecting events and practices within their broader historical context. The TDWG Task Group "Modelling Research Expeditions" developed recommendations for modelling and linking of expedition data, working towards a shared terminology, a reusable data model and schema, and best practices for publishing expedition metadata as Linked Open Data.

The presentation proposes recommendations for an expedition data model connecting both historical and contemporary collecting events with specimens and other objects, participants, institutions, localities, itineraries, publications, and archival records. Particular emphasis is placed on interoperability with Darwin Core, the use of Wikidata identifiers, and collaborative community curation of expedition metadata as a complement or even replacement for the round-tripping of data back into collection databases. The recommendations aim to improve data quality, facilitate cross-institutional linking, annotation and validation workflows, and transparency regarding the historical context of collecting activities. By enabling machine-readable and multilingual representation of expedition data, the proposed model contributes to making collection data more FAIR and supports broader reuse in e.g., knowledge graphs, visualisation tools, and digital humanities research.

Speaker



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

Authors

Robert Cubey (Royal Botanic Garden Edinburgh), Annika Hendriksen (Naturalis Biodiversity Center, Leiden), Siobhan Leachman (Wikimedia Aotearoa New Zealand), Ben Norton (Yale University), Joaquim Santos (University of Coimbra), Dag Endresen (Natural History Museum, University of Oslo)

Digital curation across the biodiversity data lifecycle, at the node level (SiB Colombia), aggregator level (GBIF) and how to incorporate user feedback

2:45 PM – 3:00 PM

AI AND ROBOT READY

Digital curation should be a continuous process through the biodiversity data lifecycle. Within the GBIF ecosystem, it starts at the publishing level facilitated by GBIF nodes, continues with quality checks done automatically by GBIF Data Processing and incorporates user feedback provided through GBIF.org or when publishers are contacted directly.

Here we will explain how the data curation is done at SiB Colombia as one example of a GBIF node. Using a publishing support model called "acompañamientos", where a dedicated helpdesk team reviews some prioritized DwC elements and provides feedback to publishers before publication. With the main goal of facilitating data mobilization by ensuring the data is as complete as possible, meets quality requirements and is free of consistency errors.

The next step is the data processing done automatically by GBIF pipelines before ingestion, addressing some of the most common and relevant elements for users, such as dates, coordinates, controlled vocabularies, etc. These quality flags allow publishers and users to review and improve their data.

The next section covers how the user feedback is currently handled on GBIF, for taxonomic issues that are handled with a human/AI approach, where an AI agent generates JSON tags that are later reviewed by a human team that ensures results are correct and there are no hallucinations.

We also present an overview of user feedback related to occurrence records, covering how much feedback arrives at GBIF, how often publishers respond, and whether they are willing to correct their data.

Finally, we will explore how things can be improved on all fronts to have better data curation and round tripping in the long run for GBIF mediated data.

Speaker



Esteban Marentes Herrera

SiB Colombia

Authors

John Waller (Global Biodiversity Information Facility), Marie Grosjean (Global Biodiversity Information Facility), Matthew Blissett (Global Biodiversity Information Facility), Ricardo Ortiz (SiB Colombia)

Closing Gaps in Specimen Metadata: A Collaborative Annotation and Feedback Platform for Natural History Collections

3:00 PM – 3:15 PM

AI AND ROBOT READY

While the digitization of natural history collections is accelerating globally, GBIF data trends reveal a critical bottleneck: a persistent gap in the completeness of digitized specimen records. Many records still lack essential metadata, specifically accurate taxonomic classifications, precise geospatial coordinates, and explicit collection dates. Bridging these gaps requires a systematic, scalable annotation mechanism.

Inspired by collaborative frameworks such as the European DiSSCover initiative, we developed an annotation platform tailored for digital specimens. The platform ingests natural history collection data from the Taiwan Biodiversity Information Alliance (TBIA). The platform allows users to query target specimens by data completeness or holding institution through filtering functions. To optimize the workflow, it pairs an intuitive manual interface with AI-assisted data extraction, enabling users to verify and adjust AI-generated transcriptions to increase processing efficiency.

Aggregated annotations form a trackable feedback loop, allowing enriched records to be returned to the original data providers for review and merging into their primary databases. This continuous cycle improves the quality of open data while lowering the barrier to participation for smaller institutions. Collections from institutions without sufficient staff often fail to transcribe labels fast enough to keep pace with imaging; the platform lets them publish specimen images early and recruit public contributors to fill in the data, increasing the usability of their holdings and encouraging institutions with un-digitized collections to join the biodiversity informatics community.

Speaker



Szu-Hsien Lee

Taiwan Biodiversity Information Facility, TaiBIF

Authors

Kuo-Fang Chung (Biodiversity Research Center, Academia Sinica), Hsun-An Yang (Biodiversity Research Center, Academia Sinica), Jerome Chie-Jen Ko (Taiwan Biodiversity Information Facility, Biodiversity Research Center, Academia Sinica), Tsui-Ya Liu (Herbarium, Biodiversity Research Center, Biodiversity Research Museum, Academia Sinica)

Responsible and GreenAI: Does our community need Gigantic Data Centers to deliver?

3:15 PM – 3:30 PM

AI AND ROBOT READY

Both the public and private sectors invest billions in AI. Key players notably in the US, China and the EU try to outcompete each other by releasing the fastest and most powerful Large Language Models (LLMs) and building the largest data centers to cope with the increasing demand for computational power. AI Factories are end-to-end systems that continuously turn data into AI models and AI-powered services at scale. For example, in February 2025, the EU launched the InvestAI initiative to mobilise €200 billion in broader AI investment, including a dedicated €20 billion European fund for AI Gigafactories. In the US, figures around 500 billion US\$ invested in AI infrastructures are cited, with similar figures being estimated from China. Looking at the global scale, with the digitisation efforts and open science policies, billions of data points and media resources around biodiversity are made available and potential candidates for AI applications. One may argue that our discipline is actually in need of such large resources to deliver. When sizing down to specific taxon levels or smaller geographic areas, training datasets of several 100 000 to less than 1 Mio of records managed to deliver accurate AI models for automated species recognition based on images or sounds. When consulted, researchers often answer that they can run the algorithms on their local servers, even local computers, without being in need of heavy computing power from large data centers. Could Small Language Models (SMLs) dedicated to a specific discipline or question, including fact-checking, taking into account ethical and legal aspects, do the job and become a model for a greener and more rational use of AI?

Speaker



Patricia Mergen

Liaison officer | Meise Botanic Garden/Royal Museum for Central Africa/ Belgian Science Policy Office

CO2 Digitizing Collections & Legacy Data

2:00 PM – 3:30 PM | Location: SAL C

CONTRIBUTED ORALS

Moderators



Talia Karim

Collection Manager, Invertebrate Paleontology | University of Colorado Museum of Natural History



Shelley James

Collections Manager | Western Australian Herbarium - DBCA

Metadata Matters: Perceptions on the Implementation of Darwin Core and Other Considerations in the Biocollections Cyberinfrastructure

2:00 PM – 2:15 PM

CONTRIBUTED ORALS

Bio- and geo- collections inform scientific discovery, advance knowledge, influence policy decisions, and serve as educational resources. Aggregating biocollections data requires several steps from curation of physical items, data capture and digitization, and creating DarwinCore (DwC) compliant metadata to allow for discoverability in institutional, discipline-specific, and data repositories such as GBIF. With differing curation practices over time, technological and methodological advancements, and evolving funding priorities, the physical research objects, their associated metadata, and derived data and media create a complicated cyberinfrastructure. This survey of curators and collection managers was conducted to better understand their perspectives on the data curation of biocollections. The survey focused on the characteristics of collections professionals, their data storage, access, and reuse practices, challenges to DwC metadata implementation, and who their anticipated users are. Recruitment through TDWG, iDigBio, SPNHC, and discipline-specific listservs led to 60 usable responses. Half of respondents held doctoral degrees, had been in their roles for more than 10 years, and 75% worked at academic institutions. Many themes emerged in relation to perceived roadblocks when mapping to data standards, such as data complexity and representing linked or derived data, gaps in standards development and community implementation guidelines as well as discipline-specific data practices, capacity and technical challenges, and a need for continued support and education around implementing standards and communicating changes or new terminology. The results can inform the building of a robot-ready cyberinfrastructure to increase data access and reuse.

Speaker



Bradley Wade Bishop

University of Tennessee Knoxville

Authors

Kit Lewers, Talia Karim, Emily Braker

Pinned Insect Digitization Conveyor: image and data transcription workflows

2:15 PM – 2:30 PM

CONTRIBUTED ORALS

The Smithsonian National Museum of Natural History (NMNH) is currently digitizing pollinating insects using a Pinned Insect Digitization Conveyor developed by Picturae. The digital workflow, developed by the Smithsonian Office of Digital and Innovation, the NMNH Informatics and Data Science Center, and the Department of Entomology, are reviewed here. Insects and labels are placed on the conveyor and up to 1,500 specimens are processed daily. Four images are captured of each insect - one of labels and three focus-stacked images of the specimen. Conveyor images are run through a technical review process and are screened by Picturae before delivery to Smithsonian. Files are organized into a folder system which hold all specimen images from an insect drawer. A second technical check, using the Smithsonian system, Osprey, is followed by review of random images by an individual. Imaging errors may include issues related to automated focus stacking and specimens being captured at the wrong resolution for their size. When a drawer has passed QC, images are delivered to the institutional DAMS. A script then picks up the images for inclusion into the NMNH database of record, EMu. Stub records are created, images are attached, and the record is then published. Label images are submitted to a different vendor for data transcription. Transcribed data are run through Osprey where a review of randomly selected data is checked for accuracy. When a drawer has passed QC, data is transformed for bulk import back into EMu. Data is appended to the original stub records and images are made available for Open Access (CC0), if applicable. The current project goal is the digitization of 365,000 specimens of pinned insects of which the majority are pollinating insects.

Speaker



Jessica Bird

Smithsonian National Museum of Natural History, Department of Entomology

Authors

Sylvia Orli, Torsten Dikow, Erin Mazzei, Luis Villanueva, Corey DiPietro, Heather Herbstritt

Digitising heritage for science: The Medical Entomology Museum as a platform for entomological biodiversity research and vector-borne disease surveillance

2:30 PM – 2:45 PM

CONTRIBUTED ORALS

The Medical Entomology Museum (MEM) integrates heritage and science by maintaining one of Africa's significant collections of arthropods of medical importance. Established primarily to support malaria and vector-borne disease research, the museum houses specimens collected over the past century. These collections serve as an important resource for species identification, biodiversity studies, and monitoring ecological and environmental changes that influence public health. This presentation highlights the museum's role as both a historical archive and a dynamic research resource supporting FAIR (Findable, Accessible, Interoperable, and Reusable) biodiversity data standards. Ongoing digitisation efforts aim to convert legacy specimen records, including type material, archival collections, and historically under-curated or orphaned datasets, into structured, machine-readable formats suitable for integration into global biodiversity informatics systems. These outputs support biodiversity research and strengthen vector-borne disease surveillance through improved accessibility, interoperability, and analytical readiness. The digitised collections also provide baseline data for malaria and vector surveillance, enabling historical comparisons and trend analysis. Beyond its scientific contribution, the museum provides insight into the history of medically important insect research in southern Africa and the development of public health interventions. It also promotes public awareness by transforming preserved collections into accessible tools for learning and engagement. In this way, the MEM demonstrates how heritage institutions contribute to science, biodiversity data innovation, and stronger public health systems.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Eunice Jamesboy

Curator | VCRL/CEZPD/National Institute for Communicable Diseases and WRIM/ University of the Witwatersrand

Authors

Maureen Coetzee, Lizette Koekemoer, Basil Brooke

X-raying wild plant seeds: Preparing images for future AI-driven global biodiversity research

2:45 PM – 3:00 PM

CONTRIBUTED ORALS

X-ray imaging is an effective non-destructive method for assessing seed quality in seed banks, enabling the identification of full, empty or damaged seeds when estimating the proportion of potentially viable seeds within a sample. This method is routinely used at the Millennium Seed Bank, RBG, Kew, as an alternative to destructive cut-testing for evaluating seed lot quality. Over the past 25 years, more than 100,000 seed x-ray images have been captured, providing a valuable resource for documenting seed traits in wild plants (e.g., shape, size, length, width, and internal features). Here, we describe the processing of these images to generate a corresponding image with a structured grid overlay and standardised measuring scale, to support future AI-driven research to measure features. The image-processing script is Python-based and uses Pillow for standard image manipulation and pydicom for parsing medical-grade X-ray metadata. Script converts digital pixel data into a calibrated physical reference through a three-step mathematical transformation: (1) metadata extraction identifies image resolution from file headers by reading dots per inch (DPI); for standard JPEGs, this is obtained via Pillow, while for DICOM files, resolution is derived from pixel spacing or equivalent tags using pydicom; (2) pixel-to-metric conversion computes pixels per centimetre using standard conversions, enabling grid intervals according to image resolution; and (3) grid placement generates an evenly spaced coordinated sequence across image dimensions. Scale accuracy is validated against physical seed length measurements from a representative sample. This workflow standardises legacy and newly acquired x-ray images, enabling downstream computational and AI-driven global biodiversity research.

Speaker



Robert Turner

Digital Operations Team Lead | Royal Botanic Gardens, Kew

Corresponding author



Udayangani Liu

Science Data Resources Manager | Royal Botanic Gardens, Kew

Authors

Subair Porora, Sian McCabe, Robert Turner, Udayangani Liu (Royal Botanic Gardens, Kew)

Next-Generation Digitization: Integrating spectral reflectance into the online mobilization of herbaria

3:00 PM – 3:15 PM

CONTRIBUTED ORALS

Spectral reflectance is emerging as a prime candidate for next-generation herbarium digitization, as spectra provide non-destructive information on taxonomic identity and functional traits. We have operationalized digitization of spectral reflectance in the Missouri Botanical Garden Herbarium (MO) using a pipeline that runs in parallel to our specimen imaging and transcription workflow. Spectral measurements follow the protocol of the International Herbarium Spectral Digitization (IHerbSpec) working group – ensuring reproducibility and interoperability of spectra across the world’s herbaria – and are linked by keying on specific metadata with their corresponding specimen records in the collection management system, Tropicos. To date, our pipeline has resulted in >60,000 spectral measurements taken on >10,000 herbarium specimens (and counting), focusing on groups in which MO curators have taxonomic expertise. We are using these spectral measurements for research applications ranging from taxonomy to ecology, including, but not limited to, using spectra to detect misidentified specimens, predict ploidy, and evaluate how Amazonian trees have responded to climate change.

Speaker



Matthew Austin

Curator of Biodiversity Data | Missouri Botanical Garden

Authors

Gabrielle Rosa, Brian Yuan, Riley Fortier, Ryan Fuller, Nathan VanderKraats, Jordan Teisher

Experience of collaborative management of the World Auchenorrhyncha Database in TaxonWorks

3:15 PM – 3:30 PM

CONTRIBUTED ORALS

The World Auchenorrhyncha Database (WAD) initially began as a series of smaller projects managed using the 3i software platform (<http://dmitriev.speciesfile.org>). The 3i system was based on an MS Access database linked to custom web interfaces and was designed to provide detailed information on species taxonomy, nomenclature, distribution, biological associations, bibliography, and identification tools, including interactive and dichotomous keys.

Eight years ago, all Auchenorrhyncha-related projects were merged into a single database, and the data were migrated to TaxonWorks (<https://taxonworks.org>), an open-source biodiversity information workbench developed to replace three previously available platforms: 3i, mx, and SpeciesFiles. Migrating data from legacy databases into TaxonWorks created numerous opportunities for collaborative data entry and management.

Following the migration, the number of curators and users with extended access rights increased substantially, from only one or a few individuals to more than 65 contributors. In addition to implementing global nomenclatural changes within the group, curators now focus on more applied goals, such as adding species images, developing interactive and dichotomous identification keys, and refining distributional data to produce comprehensive regional species checklists.

Although many users are not directly involved in data editing, they continue to contribute significantly by commenting on nomenclatural and taxonomic issues and by sharing bibliographic resources and photographs. The global coverage of the WAD currently exceeds 48,150 valid species (<https://hoppers.speciesfile.org>), and its data are shared across major global biodiversity initiatives, including Catalogue of Life, GBIF, iNaturalist, Global Names.

Speaker



Dmitry Dmitriev

cybertaxonomist | University of Illinois at Urbana-Champaign

CO3 Semantic Interoperability and Standards Infrastructure

2:00 PM – 3:30 PM | Location: SAL D

CONTRIBUTED ORALS

Moderator



Ben Norton

Yale University

AURORA - A Shiny app designed to create Darwin Core Archives from messy biodiversity datasets

2:00 PM – 2:15 PM

CONTRIBUTED ORALS

Biodiversity monitoring data are commonly stored in heterogeneous tabular formats, hindering standardisation, validation, and publication through public repositories. Researchers often face technical challenges, such as taxonomic variations and multiple date and geospatial format inconsistencies, that must be addressed before datasets can be submitted to global repositories like GBIF. To address these challenges, we developed the AURORA Shiny application that provides an intuitive, end-to-end workflow for transforming raw datasets into publication-ready Darwin Core tables. The application supports data ingestion from diverse tabular formats, enabling users to convert raw datasets into tidy data format. An interactive mapping interface allows users to align user-specific fields with Darwin Core terms by combining manual selection with automated suggestions and mandatory-field validation. The tool automates essential data-cleaning tasks for taxonomic integrity by integrating API calls to the WoRMS and GBIF Backbone Taxonomies and facilitates the use of controlled vocabulary in the Extended Measurement or Fact table. Finally, a quality control module offers diagnostics on data completeness and identifies missing key fields, providing detailed information about issues and errors for tracing and resolution. Currently, the Occurrence, Event, and extended Measurement or Fact extension DwC tables are available for export. A metadata form that follows the GBIF IPT format is provided to allow prefilling and downloading, facilitating metadata review among dataset authors. The output comprises standards-compliant DwC tables ready for publication. By streamlining these time-consuming processes into a user-friendly interface, this application reduces the technical barriers to biodiversity data mobilization and global sharing. The application is accessible to the public via a web browser by accessing <https://bio-shiny.ua.pt/aurora/>

Speaker



Sara Araújo
University of Aveiro

Authors

Álvaro Magnenti, Heliana Teixeira, Ana Hilário, Sofia Ramalho, Fabio Matos, Eva Ramirez-Llodra

Rebuilding core functionalities of the GGBN system: taxonomic alignment with international research infrastructures

2:15 PM – 2:30 PM

CONTRIBUTED ORALS

The Global Genome Biodiversity Network (GGBN) is an international consortium of biodiversity biobanks and collections focussing on standards and best practices, providing access to information on > 7.1 million samples. GGBN informs researcher ensuring ABS and FAIR principles. GGBN assigns persistent identifiers (PIDs) to all object types (DNA, tissue, specimen, culture, environmental sample) and establishes relationships among these, enabling cross-linking with other infrastructures (e.g., ENA). Use of DwC-DP will further improve this.

Collection data incl. scientific names are currently submitted using TDWG standards (DwC, ABCD) but many received names do not exactly match with global taxonomic aggregators (e.g., Extended Catalogue of Life), or missing classification data causes ambiguity; reducing discoverability, semantic interoperability and gap analysis.

We aim at a machine-assisted curation pipeline that: (i) preprocesses verbatim names (AI-assisted), (ii) matches and links names to authoritative taxonomic aggregators, (iii) augments records with aggregator data (classification, synonyms, name-IDs), (iv) provides access to enriched records and their relationships via faceted, fuzzy and synonym-aware search (Portal and API) and (v) feeds back name processing results to data providers. We will draw on the experience gathered by other aggregators of collection data.

GGBN aligns with the DwC Conceptual Model by making relationships among material entities, organisms, and occurrences explicit and machine-actionable. We will present challenges encountered and expected benefits for CMS workflows, data consumers, and aggregators. We welcome discussion on how DwC-DP and PID practices can be adopted across CMSs to support interoperable, research-ready biodiversity data.

Speaker



Walter Berendsohn
Botanical Garden Berlin, Freie Universität Berlin

Authors

Gabriele Droege

Bridging eLTER Data Reporting Format: enabling interoperable biodiversity and ecological data publication across research infrastructures

2:30 PM – 2:45 PM

CONTRIBUTED ORALS

Environmental and biodiversity research infrastructures increasingly face the challenge of mobilising heterogeneous observational data across domain-specific systems, harmonizing it into FAIR data products or digital twins, while preserving semantic richness and machine-actionable interoperability. Within the eLTER Research Infrastructure, ecological and environmental measurements are distributed through the Data Reporting Format (DRF), a structured, multi-table specification designed to support multi-domain standard observational workflows through different entities.

To strengthen semantic interoperability within the broader informatics landscape, a controlled vocabulary has been developed that provides labels and definitions for DRF data specification terms and aligns them with the Darwin Core standard where conceptual equivalence exists. This mapping preserves the native expressiveness of eLTER observational structures while enabling alignment with established biodiversity data exchange mechanisms and improving discoverability across infrastructures and other TDWG-compliant systems.

Building on this, an R-based workflow has been implemented that transforms DRF datasets into Data Packages. This packaging layer facilitates FAIR and machine-actionable (“robot-ready”) data publication, while enabling downstream integration into biodiversity data infrastructures without requiring convergence toward a single monolithic data model.

This work demonstrates how infrastructure-specific reporting standards can remain fit-for-purpose while participating in a coordinated biodiversity data ecosystem through explicit semantic alignment and reusable transformation workflows. The approach provides a practical contribution to strengthening interoperability.

Speaker



Alessandro Oggioni

National Research Council of Italy - Institute of Marine Science (CNR-ISMAR)

Authors

Hanna Koivula (CSC - IT Center for Science, Espoo), Johannes Peterseil (EAA - Environment Agency Austria), Lisa Reiss (EAA - Environment Agency Austria), Sarah Venier (EAA - Environment Agency Austria)

Infrastructure-driven solutions to strengthen semantic interoperability in ecology

2:45 PM – 3:00 PM

CONTRIBUTED ORALS

Speaker



Cristina Di Muri

National Research Council (CNR), Research Institute on Terrestrial Ecosystems (IRET)

Authors

Martina Pulieri, Alexandra Nicoleta Muresan, Andrea Tarallo, Lucia Vaira, Alberto Basset, Ilaria Rosati

A³ctionable Understanding: Action units for post-FAIR biodiversity knowledge infrastructures

3:00 PM – 3:15 PM

CONTRIBUTED ORALS

The FAIR Principles have made biodiversity data findable, accessible, and interoperable, but not operationally usable. A critical layer is missing between semantic representation and context-sensitive action, one that distinguishes the structural capacity of a knowledge representation to support operations (actionability) from the epistemic reliability of using it in a specific context (applicability). Conflating these has contributed to well-documented failures in conservation practice.

We introduce Action Units as formal knowledge objects that extend plan specifications within the Semantic Units Framework by making applicability conditions explicit as first-class typed components. Three types correspond to three operation classes: epistemic action units ask '*What is the case here, and does this knowledge apply?*'; transformational action units ask '*How can available representations be processed into usable form?*'; and intervention action units ask '*What can be done in this situation to achieve a desired outcome?*' Conditional action units, i.e., executable IF–THEN rules linking SPARQL context queries to workflow triggers, enable knowledge graphs to serve as AI-ready decision-support systems.

Applied to biodiversity science, the framework addresses conservation failures and enables practitioners to identify where actionable knowledge is missing. We propose the TripleA Principle—Actionability, Applicability, and Auditability (tracing evidence and failure contexts)—as a post-FAIR guiding framework, and argue that A³ctionable Understanding—Actionability, Applicability, and Action Units—provides the representational foundation next-generation biodiversity knowledge infrastructures require.

This is a conceptual proposal; implementation is deferred to subsequent work.

Authors

Tim Alamenciak, Robert Fröhstückl

SFBorg: a frictionless biodiversity data exchange

3:15 PM – 3:30 PM

CONTRIBUTED ORALS

Background: Sharing biodiversity datasets across projects and organisations depends on common exchange formats e.g., Darwin Core Archive (DwCA), Darwin Core Data Package, or the Catalogue of Life Data Package (ColDP). These formats bundle multiple CSV, JSON, or XML files into a single compressed archive. While widely adopted, this approach has a fundamental limitation: the data is effectively inert until a recipient imports it into a database. Generation of a dataset's archive is error-prone, and introduced inconsistencies might create significant problems on the receiving end. Also, there is no standard mechanism to detect what has changed between two successive releases of the same dataset.

New information: We introduce SFBorg, a SQLite-based ecosystem for biodiversity datasets exchange centred on the Species File Group Archive (SFGA) schema. An SFGA single file is a self-contained SQLite database, allowing recipients to query and modify data immediately using standard SQL tools — no import step required. The ecosystem includes: sf, a universal converter between DwCA, ColDP, and other formats, which also computes data diffs between two SFGA archives to identify added, modified, or removed taxa, names, and synonyms; harvester, for ingesting non-standard or legacy sources; and gnldb, which loads SFGA archives directly into the GNverifier PostgreSQL database. Shared functionality is centralised in the sflib library, reducing duplication and lowering the cost of adding new tools. SFBorg is currently in production use across Species File Group projects. All code is open source and available on GitHub under the MIT licence.

Speaker



Dmitry Mozzherin

University of Illinois Urbana-Champaign

Authors

Geoff Ower

DS19 - Strengthening resilience in biodiversity data infrastructures: From practice to global coordination

2:00 PM – 3:30 PM | Location: BALDER

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Recent alignment between DiSSCo and GBIF shows how shared standards (Darwin Core Data Package, and DiSSCo's openDS specification) can enable interoperability without information loss. Specimen-level data can now move more seamlessly across systems when conceptual foundations are aligned. Building on this, a broader question emerges: how can collaboration move beyond bilateral technical integrations toward a truly coordinated biodiversity data ecosystem? Researchers increasingly need to discover and combine diverse data types — collections, observations, environmental measurements, molecular data — yet these remain siloed across specialised infrastructures like DiSSCo, GBIF, BHL, eLTER, ELIXIR, and LifeWatch. Using the DiSSCo-GBIF collaboration as a starting point, this session invites participants from technical, scientific, and policy communities to explore what it would take to enable findability, usability, and interoperability at scale. Where should alignment happen: conceptual models, identifiers, metadata, vocabularies, access mechanisms, or governance? How can infrastructures collaborate without converging into a single data model? And what practical steps can reduce fragmentation for users today?

The session combines short scene-setting inputs with open discussion focused on shared challenges, emerging opportunities, and actionable next steps — fostering a shift from project-level thinking toward a more resilient and connected biodiversity data landscape.

Moderator



Sharif Islam

Senior Data & System Architect | Naturalis Biodiversity Center

Panelists



Sam Leeflang

Lead Developer Collection Data Infrastructure and Access team | Naturalis Biodiversity Center



Hanna Koivula

Senior Research Data Management Specialist | CSC - IT Center for Science, Espoo



Joana Pauperio

European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK



Dmitry Schigel

Lead scientific officer | Global Biodiversity Information Facility (GBIF), Secretariat



Anne Fouilloux

CTO | Lifewatch-ERIC

Discussion leaders



Giuditta Parolini

Museum für Naturkunde Berlin - Leibniz Institute for Evolution and Biodiversity Science



Eva Alonso

DiSSCo Head of Governance | Naturalis Biodiversity Center



Wouter Addink

Manager Software Development | Naturalis Biodiversity Center



José Alonso

Senior Communications Officer (DiSSCo, BGE+) | Naturalis Biodiversity Center

WKS51 - Practical Biodiversity Data Quality (BDQ)

2:00 PM – 3:30 PM | Location: Møterom 5 / Meeting room 5

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Workshop leaders



Paul Morris

Museum of Comparative Zoology, Harvard University



Lee Belbin

Director | Blatant Fabrications Pty Ltd



Arthur Chapman

CEO | Australian Biodiversity Information Services

BREAK

3:30 PM – 4:00 PM | Location: 1st Floor

SYM08B - BHL at 20: A New Chapter for Biodiversity Literature and Data – Transforming BHL: Data Extraction, Semantic Navigation and Future Interfaces

4:00 PM – 5:00 PM | Location: SAL A

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

In 2026, the [Biodiversity Heritage Library](#) (BHL) celebrates both its 20th anniversary and its first year as an independent consortium. Following its [transition](#) away from long-term hosting at the Smithsonian Institution, BHL is now operating from a stable foundation, with its people, expertise, and services firmly in place and renewed momentum across its global network.

This session will explore where BHL stands today and what this new phase makes possible. Presentations will examine the outcomes of the transition, including updated governance structures, evolving funding approaches, strengthened partnerships, and technical developments that position BHL for sustained growth and innovation within the biodiversity data ecosystem.

Speakers will present emerging applications of BHL content and services, such as artificial intelligence and machine learning, large-scale text and image mining, knowledge graphs, annotation systems, and efforts to improve persistent identifier reliability and stewardship. Talks will demonstrate how BHL functions as essential infrastructure connecting literature with taxa, specimens, institutions, and datasets, from the 1100s to the current year.

The session will also reflect on two decades of BHL's impact, from mobilizing historical biodiversity knowledge to supporting modern data integration and reuse, inviting contributors and attendees to consider how BHL can continue to evolve as a core component of global biodiversity information infrastructure.

Moderators



Nicole Kearney

Communications Director | Biodiversity Heritage Library



David Iggulden

Head of Data & Digital (Library & Archives) | Royal Botanic Gardens, Kew

Introduction: Transforming BHL

4:00 PM – 4:05 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Speaker



Nicole Kearney

Communications Director | Biodiversity Heritage Library

Biological occurrence data from historic scientific correspondence: enabling automated approaches for structured data extraction from body text

4:05 PM – 4:15 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Historical species occurrence data are essential for incorporating the past into analyses of biodiversity change. In addition to sources specifically intended to document occurrences, such as surveys, occurrence data are also implicitly contained in many other historical documents, where they are embedded in diverse linguistic structures.

Here, prerequisites for scalable information extraction of biodiversity-relevant data from written sources were investigated using content from the scientific journals published by the Museum für Naturkunde Berlin: the *Deutsche Entomologische Zeitschrift* and *Zoosystematics and Evolution* and their predecessor periodicals accessible through the Biodiversity Heritage Library.

Focussing on information concerning the observation of biological species (occurrences), together with data on habitats and related preserved specimens, a data model was developed to align natural-language text elements with facets of occurrence instances. Based on this model, a text corpus derived from the journals was manually annotated to generate ground-truth text annotations and occurrence data. The data associated with the identified observation events were serialized as RDF (Resource Description Framework) graphs and aligned with the Darwin Core standard and the data schemas of Wikidata and the Global Biodiversity Information Facility.

We outline how the annotated corpus can be used to develop and evaluate AI-based information-extraction workflows for species occurrence data from natural-language prose. In addition, the project generated insights into the acquisition and use of copyright-protected sources for text- and data-mining approaches in contrast to open access models.

Speaker



Christian Bölling

Museum für Naturkunde Berlin - Leibniz Institute for Evolution and Biodiversity Science

Authors

Aron Marquart, Sandra Sánchez Páez, Stefanie Paß, Franziska Schuster

A closer look at steps needed to extract research-ready biodiversity data from legacy literature

4:15 PM – 4:25 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Research-ready biodiversity data must follow established standards. However, this is not easily achieved in the case of data extracted from legacy literature. The literature is unstructured and heterogeneous in format, content and style, which is why data extraction presents many challenges, some of which will be illustrated during this talk. One of the problems is synonymy, a problem which goes much further than mere taxonomic or nomenclatural synonymy. Another common phenomenon is coreference, this is, two or more expressions referring to the same object. Do we need to apply or develop computer-linguistic solutions for these problems... or can we just rely on LLMs to solve them for us? This talk will discuss approaches applied in the BIOfid text mining and annotation pipeline which is under development with a focus on a German-language corpus of biodiversity legacy literature, but we expect that this approach will also help solve problems with corpora in other languages.

Speaker



Gerwin Kasperek

Curator of the Biology Collection | University Library Johann Christian Senckenberg, Goethe University Frankfurt

Multimodality for Knowledge Extraction from Historical Entomology Literature

4:25 PM – 4:35 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Historical entomological publications combine textual descriptions and scientific illustrations that contain complementary biodiversity knowledge. While texts describe insect taxa, host plants, and ecological interactions, illustrations capture life cycles, morphological traits, and species interactions. Existing biodiversity knowledge extraction approaches mainly focus on text and rarely integrate visual information from historical illustrations.

We propose a multimodal knowledge graph for representing entities, traits, and ecological interactions extracted from historical entomological literature. As a case study, we analyze 50 articles and illustrations from Maria Sibylla Merian's *Der Raupen wunderbare Verwandlung, und sonderbare Blumennahrung* (1679). Our workflow combines OCR, semantic annotation, image-region linking, and knowledge graph construction to align textual and visual information. The workflow includes: (i) extracting text with OCR, (ii) defining an annotation ontology for concepts and events in illustrations, (iii) annotating articles with these concepts, and (iv) linking illustration regions to corresponding text passages.

The workflow generates semantic triples from text and illustrations, e.g., {(tulip, instance_of, petal), (petal, has_color, purple), (caterpillar, has_shape, elongated)}. Beyond visual traits, extracted relations represent biologically relevant information such as insect–host plant associations and developmental stages. The resulting knowledge graph supports semantic enrichment and automated caption generation for historical biodiversity collections.

Speaker



Jana Hoffmann

Senckenberg German Entomological Institute, Münchenberg

Authors

Sefika Efeoglu (Senckenberg – Leibniz Institution for Biodiversity and Earth System Research)

Future interfaces for BHL

4:35 PM – 4:45 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

As part of TDWG 2026 there is a workshop on future interfaces for the Biodiversity Heritage Library. The workshop will be a mostly "hands on" exploration of live demos, often based on a subset of BHL content.

This talk puts the workshop in context by surveying user interfaces employed by digital libraries in other fields. It explores different approaches to search, the use of maps, and navigating by terms or entities (such as people or taxa). It considers some of the new possibilities offered by AI-based approaches, such as searching for similar images, and wrapping complicated queries in easy to use natural language interfaces.

Having found what you are looking for, how do you want that item displayed? The talk considers various approaches, such as PDFs, such as custom-built viewers, or viewers based on the IIIF standard. Increasing adoption of IIIF has resulted in a flurry of new approaches to annotating content, including overlaying published maps on a shared map of the world, or using in-browser machine-learning tools to classify images.

The success (or otherwise) of the various demos will give insights into what interfaces will enable users to get the most value from BHL.

Speaker



Roderic Page

Discussion

4:45 PM – 5:00 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

LT17 - Bots, Bits, and Biodiversity

4:00 PM – 5:00 PM | Location: SAL B

AI AND ROBOT READY

This session will consist of a series of short, rapid presentations on robotic and automated approaches to biodiversity data acquisition. It will focus on the design and use of physical and hybrid systems that generate structured, machine-actionable datasets and their alignment with TDWG Biodiversity Information Standards.

The session explicitly welcomes contributions across the full development spectrum, including early-stage prototypes, experimental platforms, field-deployable systems and production infrastructures.

Relevant topics include, but are not limited to:

- robotic specimen handling and imaging
- automated sensing platforms in the laboratory or field
- high-throughput digitisation systems
- workflows for generating training and validation data for machine learning
- automated or semi-automated annotation and quality control
- feedback loops between data acquisition and model performance
- methods for structuring outputs in compliance with TDWG standards

By using a rapid-fire format, the session aims to maximise the number and diversity of case studies, enabling comparison of different technical approaches and maturity levels. The discussion will synthesise common challenges, identify opportunities for standardisation, and explore how emerging automated workflows can contribute to scalable and sustainable biodiversity data infrastructures.

Speaker



Jack Hollister

Senior Robotics Technology Engineer | The natural History Museum, London

Authors

Introduction

4:00 PM – 4:05 PM

AI AND ROBOT READY

Moderator



Jack Hollister

Senior Robotics Technology Engineer | The natural History Museum, London

Agentic AI Systems for Specimen Interaction and Camera-Based Observation

4:05 PM – 4:10 PM

AI AND ROBOT READY

The study, recording and interpretation of natural history specimens is a highly contextual activity, drawing on curatorial expertise, published literature, collections data and comparative knowledge across taxa. In many cases, working with specimens is also a physical process, requiring observation, annotation, transcription, imaging and careful handling. These activities are time-consuming and can limit the number of specimens that can be examined, recorded or reinterpreted at scale. This presentation explores the development of an agentic AI system capable of interacting with physical specimens through a dedicated imaging and specimen observation platform. Operating on local/on-device hardware, the system integrates large language models, computer vision, camera control and hardware control to support direct observation and interpretation of specimens within collections environments. The system uses a camera-based platform to capture visual data from specimens and incorporate this information into the model's reasoning process. By combining visual input with collections data and relevant literature, the agent can support tasks such as specimen inspection, metadata verification, annotation and morphological interpretation. Through the integration of RAG, the system links observed specimens to contextual knowledge, enabling more informed and interactive analysis. This establishes a foundation for embodied AI within museum settings, where models can move beyond static datasets and engage more directly with physical collections. Live demonstrations will show how such systems can support curatorial and research workflows, providing a new interface between researchers, AI systems and physical specimens while maintaining local control over data, models and processing.

Speaker



Jack Hollister

Senior Robotics Technology Engineer | The natural History Museum, London

Authors

Steen Dupont

Rediscovering Archives – Using LLM's for the Transcription and Extraction of Ecological Data from Historical Archives

4:10 PM – 4:15 PM

AI AND ROBOT READY

Historical archives held by museums and research institutions contain vast quantities of material that remain underutilised due to limitations in accessibility, readability, and searchability. Recent advances in multimodal large language models (LLMs) present new opportunities for improving access to these collections, particularly for complex historical documents that challenge conventional optical character recognition (OCR) and handwritten text recognition (HTR) approaches. This is especially relevant within natural history archives, where materials range from structured to unstructured, often containing handwritten and typed text alongside historical terminology.

This project investigates the use of multimodal LLMs for the transcription and interpretation of historical archival documents, with a focus on prompt engineering and schema-guided extraction to produce structured, machine-readable outputs. We focus on natural history collection materials such as those housed at Natural History Museum, UK, and Zoological Society London, as well as publicly available datasets such as the India Office botanical records from the British Library. Our study evaluates transcription performance – measured by Levenshtein distance, word error rate, and character error rate - across mixed-format, handwritten documents against ground truth transcriptions. By examining how multimodal LLMs perform on ecologically and historically relevant archival materials, this work contributes to emerging research on AI-assisted archival accessibility and explores how historical collections can be transformed into structured digital resources suitable for biodiversity and historical research.

Speaker



Phoebe Santos

MSc student | UCL/Natural History Museum

Authors

Robin Freeman (Institute of Zoology, Zoological Society of London), Arianna Salili-James (Natural History Museum, London), Santiago Martinez Balvanera (University College London), Daan Scheepens (University College London)

Semantic Governance and Knowledge Organization in AI-Augmented Biodiversity Infrastructures

4:15 PM – 4:20 PM

AI AND ROBOT READY

As artificial intelligence becomes increasingly integrated into biodiversity data infrastructures, AI-generated annotations, classifications, and metadata are reshaping how biodiversity knowledge is organized, interpreted, and governed. While large language models and other generative AI systems create new opportunities for metadata enrichment, semantic integration, and data discovery, they also introduce challenges related to interpretability, semantic consistency, provenance, and institutional trust. This presentation examines biodiversity infrastructures from a knowledge organization perspective, focusing on how probabilistic AI outputs interact with structured biodiversity standards, controlled vocabularies, and long-standing cataloguing practices. The talk will explore emerging challenges surrounding semantic ambiguity, machine-generated biodiversity knowledge, and the governance of AI-assisted curation workflows within biodiversity repositories. Particular attention will be given to the need for provenance-aware, semantically transparent, and interoperable approaches that align AI-generated outputs with community standards and reproducible biodiversity data practices. By framing AI integration as both a technical and governance challenge, this presentation highlights opportunities for TDWG to support the development of shared semantic frameworks and best practices for AI-augmented biodiversity infrastructures.

Keywords: knowledge organization; semantic governance; biodiversity informatics; large language models; provenance; metadata; semantic interoperability; AI-generated annotations; biodiversity standards; AI-assisted curation

Speaker



Andrew Jones

Assistant Curator | Tulane University

KakraCards: An AI-Assisted Pipeline for Liberating Six Decades of Seabird Heritage Data

4:20 PM – 4:25 PM

AI AND ROBOT READY

The continuous monitoring of the Common Gull (*Larus canus*) breeding colony on Kakrarahu islet in the Baltic Sea, Estonia, is one of the region's longest-running ecological research projects, dating back to 1962. With continuous breeding success data collected since 1968 and ringing and age data since 1971, this project holds a vast repository of longitudinal population dynamics data. Historically, these data were recorded in field notebooks, systematized on catalogue cards, and later partially transferred to spreadsheets. Modern data collection has now transitioned to a specialized mobile application for real-time field use.

This work-in-progress project outlines the ongoing effort to rescue and digitize this extensive legacy dataset using a hybrid approach of manual data entry and novel AI-based digitization techniques and store it together with the latest field data. Because the monitoring methodology has remained constant over the decades, the recovered data holds immense value for understanding long-term population dynamics.

The primary goal of this initiative is to make these unique ecological data widely available following FAIR principles. While multiple open-access publication approaches will ultimately be utilized, this presentation focuses on one particular pathway. By utilizing PlutoF, a comprehensive common data space for biodiversity developed by the University of Tartu, we establish a pipeline to structure, store, and publish the historical dataset to the Global Biodiversity Information Facility (GBIF). This workflow demonstrates an effective framework for rescuing legacy mass-monitoring data and ensuring over six decades of ecological observations become globally accessible.

Speaker



Kristjan Adojaan

volunteer, coordinator | Vaibla Bird Observatory

Authors

Kristjan Adojaan, Allan Zirk, Kessy Abarenkov, Urmas Kõljalg, Jeffrey Carbillet, Tuul Sepp, Lauri Laanisto, Hannah Métaireau, Richard Meitern, Lauri Saks, Jaanis Lodjak

Data Quality in Focus: Bridging the Biodiversity Data Gap for Sustainable Finance

4:25 PM – 4:30 PM

AI AND ROBOT READY

Sustainable finance increasingly demands biodiversity data and metrics, yet these are only credible when underlying data are traceable, quality-assessed, and explicit about uncertainty. Currently, the financial sector faces a "biodiversity data gap": existing datasets are sparse and taxonomically biased, and often lack the depth needed for reliable risk assessments. Without robust data standards, investors cannot reliably evaluate the local impact of their portfolios or meet emerging nature-positive targets.

Drawing on insights from the Mistra FinBio program, we identify critical roadblocks in biodiversity reporting and propose a key improvement: a "data quality hierarchy." By mapping data and model outcomes to quality tiers and standardized metadata, it would enable AI-ready data for financial risk models, turning observations into trustworthy, decision-ready inputs.

The TDWG community is invited to collaborate on closing this data gap, focusing on improving data quality standards. Biodiversity data should not just be collected, but made interoperable, machine-actionable, and ready for the next generation of sustainability reporting.

Speaker



Emma Granqvist

Director, Swedish Biodiversity Data Infrastructure | Swedish Museum of Natural History

Robots Rapidly Writing Records: Machine Annotations at Scale with DiSSCo

4:30 PM – 4:35 PM

AI AND ROBOT READY

In the face of the biodiversity crisis, digitization efforts in natural science collections need to rapidly increase. Scaling up these efforts requires the support of the community – and machines. DiSSCo, the Distributed System of Scientific Collections, enables this through annotation and harmonization of the data.

When natural science collection datasets are ingested into DiSSCo, specimens are transformed into Digital Specimens, assigned a Persistent Identifier, harmonized to a shared data model, and become and become fully actionable by machines as discrete knowledge units.

Through this process, DiSSCo enables community curation through annotations by humans and machines. With Digital Specimens harmonized to one data model across host institutions, DiSSCo is uniquely situated to host "plug and play" machine services to act on any specimen in the DiSSCo infrastructure. Instead of being confined to a single dataset they were designed for, Machine Annotation Services (MAS) adapted for DiSSCo can act on any Digital Specimen, breaking institutional silos and furthering collaboration.

For the first time, DiSSCo MASs have been tested at scale, running on all digital objects in a dataset as they are ingested into the DiSSCo infrastructure. These pilot services link all Digital Specimens in a given dataset to counterparts that can be found in GBIF, ENA, or BOLD. With the pilot successful, there is now an opportunity to run more complex services (such as label transcription using large language models) at scale on natural science collection data in DiSSCo.

This lightning talk gives an overview of the MAS-on-Ingestion feature in DiSSCo and what that means for transforming collection data.

Speaker



Soulaine Theocharides

Software Developer | Naturalis Biodiversity Center

Authors

Wouter Addink, Sam Leeflang

From Collection Drawers to AI – A One-Week Recipe

4:35 PM – 4:40 PM

AI AND ROBOT READY

Over 724 million entomological specimens exist worldwide. The collection at the Norwegian Institute of Bioeconomy Research (NIBIO) holds approximately 200,000 partly databased specimens, highlighting the urgent need for digitisation with high-quality images and structured metadata to support research and outreach.

We asked: what can be built in one week to initiate digitising a small collection, and how can AI enhance the process?

During a joint visit between NIBIO (Division of Biotechnology and Plant Health) and the Natural History Museum UK (AI & Innovation), we developed a functional, low-cost imaging setup using recycled office materials, simple LED lighting, and two borrowed cameras (Nikon Z7 and Z7 II). The aim was not perfection, but speed, accessibility, and replicability.

We focused on Scolytinae (bark beetles), a group central to NIBIO's research. Specimens of varying sizes were imaged in dorsal and lateral views and filmed in 360° rotation. In total, several hundred specimens were captured, producing a compact but versatile dataset.

This dataset was used in three AI pipelines: (1) CNNs trained for species classification; (2) 3D reconstruction from video to generate interactive models for outreach; and (3) combined CNN and LLM approaches to extract specimen label data directly from videos.

Together, these experiments demonstrate that rapid, low-cost digitisation can meaningfully support AI-driven classification, data extraction, and engagement, providing a scalable entry point for modernising biological collections.

Speaker



Arianna Salili-James

Natural History Museum UK

Imaging, AI and robotics for high-throughput insect digitisation at CSIRO's National Research Collections of Australia

4:40 PM – 4:45 PM

AI AND ROBOT READY

CSIRO's National Research Collections of Australia (NRCA) contain an estimated 15 million specimens, including approximately 12 million insects. To accelerate scalable specimen data acquisition, NRCA is collaborating with CSIRO technology science teams to develop integrated imaging, OCR, machine learning, and assisted-robotic workflows for high-throughput digitisation.

A key development is the Insect Recording and Imaging System (IRIS), a multi-angle imaging platform for pinned insect specimens that captures specimens and their associated labels without removing labels. IRIS reduces imaging time from more than 6 minutes to approximately 30 seconds per specimen, substantially increasing throughput while minimising specimen-handling risks. Computer vision pipelines segment and reconstruct label data from image sets, integrating with in-house OCR tools for automated transcription and structured metadata extraction designed for bulk upload into our Specify collections management system.

In addition, our robotics research teams are developing autonomous specimen manipulation, including robotic handling systems, gripper design, automated loading and unloading workflows, and human-in-the-loop approaches that prioritise specimen safety while improving scalability. Supplementary exploratory workflows include high-resolution drawer capture and automated drawer-level metadata generation to support inventory management and tracking.

This case study highlights the development of hybrid imaging and robotic workflows for generating machine-actionable specimen data at scale. We discuss throughput optimisation, workflow integration, automated metadata extraction, human-robot interaction and the challenges of balancing automation and specimen safety to digitise collections at scale.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Nicole Fisher

Digitisation manager | CSIRO, National Research Collections Australia

Authors

Corinna Paeper, Nicole Fisher, Pete Thrall, Lars Andersson, Ran Wei, Brandon Matthews, Chuong Nguyen, Hu Zhang, Ahalya Ravendran, Brendan Tidd

kabr-tools: A Modular Workflow for Automated Video-Based Behavioral Monitoring Across the Autonomy Spectrum

4:45 PM – 4:50 PM

AI AND ROBOT READY

Quantifying animal behavior from video footage is central to conservation and welfare assessment, but drone and camera-trap deployments now generate datasets that are too large to process manually. Existing tools support individual stages of video-to-behavior processing, yet ecologists lack practical guidance on which stages can be reliably automated and which require manual review. We present kabr-tools, an open-source modular workflow for video-based behavioral monitoring that supports a spectrum of autonomy, from fully manual annotation to fully automated processing. Components for detection, tracking, and classification are interchangeable and taxon-agnostic, producing structured outputs that align with biodiversity data standards. We evaluate the workflow against expert annotation at each pipeline stage and trace how errors propagate into downstream behavioral estimates, using drone footage of Grevy's zebras, Plains zebras, and reticulated giraffes at Mpala Research Centre, Kenya, and drone and camera trap footage of African wild dogs at The Wilds, Ohio. Detection is the primary bottleneck in both systems, since once individuals were reliably detected, tracking and classification added limited additional error for common behaviors. Automated classification achieved 91.3-F1 for the three most frequent ungulate behaviors, declining to 50.3-F1 for rare categories. A camera-source ablation confirmed that tail-class failure reflects insufficient training examples rather than a modality-specific limitation. This error characterization positions animal behavior studies on the autonomy spectrum, given their research question and required precision, and highlights where standardized, machine-actionable behavioral outputs depend on targeted human review.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Jenna Kline

Postdoctoral Associate | Massachusetts Institute of Technology

Authors

Jenna Kline, Elizabeth Campolongo

Discussion

4:50 PM – 5:00 PM

AI AND ROBOT READY

Moderator



Jack Hollister

Senior Robotics Technology Engineer | The natural History Museum, London

SYM22 - State of the Collections Description Interest Group. What's Happening and What's Not!

4:00 PM – 5:00 PM | Location: SAL C

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

This session will:

- 1) Bring together the active task groups under, and maintenance groups affiliated with the Collections Description Interest Group. Key topics will include Latimer Core, MIDS and Expeditions.
- 2) Invite talks from groups who are interested in the work of the Interest Group and have ideas and thoughts around new work that could be taken up.

Moderators



Sharon Grant

Director of Information Systems | Field Museum of Natural History



Ben Norton

Yale University

How is digitisation doing? Monitoring the increase in digitised collections using MIDS

4:00 PM – 4:15 PM

COLLECTIONS DIGITISATION & MANAGEMENT

The Minimum Information about a Digital Specimen (MIDS) standard has been under development by the TDWG MIDS task group for several years and entered expert review in early 2026. It provides a framework for automatically and reproducibly assessing the digitisation level achieved by a specimen, based on its provided data properties. Applying the MIDS level calculation to historical and recent GBIF specimen data snapshots, we shed light on the growth of digitised specimen collections in recent years and investigate patterns across the dimensions of taxonomy, preservation type, location, and age.

Speaker



Elsbeth Haston

Deputy Herbarium Curator | Royal Botanic Garden Edinburgh

Authors

Matthew Blissett, Cat Chapman, Ben Norton, Stefan Seifert, Mathias Dillen

COLLECTIONS DIGITISATION & MANAGEMENT

The Global Registry of Scientific Collections (GRSciColl) aims to catalogue physical scientific collections, their content, location, associated staff and institutions. GRSciColl was originally developed before the Latimer Core (LtC) Standard became ratified. It has its own model, terms and associated vocabularies. The information can be stored in different types of entities:

- Institution and collection entries, which contain a lot of textual descriptions.
- Specimen occurrence records shared on the Global Biodiversity Information Facility (GBIF) platform and linked to GRSciColl institution and collection entries.
- Semi-structured tables called “collection descriptor groups”. They are used to describe collection content and are associated with collection entries.

In this presentation, we will show where and how GRSciColl is integrating the LtC standard:

- GRSciColl entries can be accessed as well as uploaded in the LtC standard.
- GRSciColl collection descriptors groups use the LtC terms and suggested vocabularies. The objectClassificationName term is particularly useful when capturing information about subcollections corresponding to polyphyletic groups. It allows for example to find collections containing algae: <https://scientific-collections.gbif.org/collection/search?objectClassification=Algae>
- The use of LtC standard in training material for modelling of collection data.

We will also highlight some of the challenges of this integration.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Marie Grosjean

Data administrator | GBIF Secretariat

COLLECTIONS DIGITISATION & MANAGEMENT

Latimer Core (LtC) is the TDWG standard intended to support description and exchange of information about collections and their associated organizational, and preservation contexts. Over the past year, the Maintenance Group has focused on improving its conceptual model, refining documentation, and exploring practical serialization and interoperability patterns that help long-term sustainability and implementation across diverse collection-management environments.

This presentation will provide an overview of the current state of LtC, summarize the work undertaken during 2025–2026, and introduce some implementation examples, including draft JSON and RDF/XML exchange formats. Together, these examples aim to illustrate both the flexibility of the LtC conceptual model and the unresolved questions surrounding structure, identifiers, and semantic interoperability.

The presenters will invite discussion of larger questions that extend to other TDWG efforts: How do we make standards understandable and maintainable for people across different domains of expertise? How should TDWG approach controlled vocabularies and governance? How should we model relationships between terms, and structure their identifiers (IRIs)? How do we document dataset creation, generation, or AI-assisted workflows? What happens when our standards depend on concepts borrowed from outside the TDWG ecosystem?

This session embraces the messier reality of standards work: balancing flexibility and structure, interoperability and usability, technical rigor and human comprehension. The goal is to share where LtC stands today, and invite discussion. How can TDWG standards remain useful, maintainable, and adaptable as biodiversity data practices continue to evolve?

Speaker



Sharon Grant

Director of Information Systems | Field Museum of Natural History

Authors

Maarten Trekels, Ben Norton, Kate Webbink, Jutta Buschbom, Janeen Jones, Elie Saliba

Assessing record quality and completeness in geoscience collections

4:45 PM – 5:00 PM

COLLECTIONS DIGITISATION & MANAGEMENT

The Minimum Information about a Digital Specimen (MIDS) standard allows collections to assess digitization progress using a numeric, discipline-specific score. Here we describe efforts to assess this type of data quality score in collections of geological specimens at the Smithsonian National Museum of Natural History (NMNH). We examine MIDS scores for subsets of the collection and compare them to established, internal record quality scores from the Smithsonian's Collections and Digitization Reporting System (CDRS). Our goal is to determine how to best implement MIDS, while also reviewing potential updates to CDRS.

MIDS scores can range from 0 (Bare) to 3 (Extended). Most records from the geology collections score 1 (Basic) or 2 (Intermediate). No records score above 2. Low scores largely reflect 1) a lack of detailed metadata for older specimens and 2) incomplete data from ongoing digitization projects. Certain subcollections have more complete data but recently acquired collections score similarly to older collections.

CDRS scores are calculated in the collections database using discipline-specific rulesets. Like MIDS, CDRS assesses digitization progress, but it differs from MIDS in four key ways: 1) it includes elements for physical collections management, like storage location and accession status; 2) it excludes some properties that appear in MIDS; 3) it may include data quality checks, which were omitted from MIDS by design; and 4) inconsistent rulesets make comparisons between disciplines difficult.

We expect MIDS to be most useful for improving the completeness of published datasets and allowing comparisons between organizations, whereas CDRS scores will better capture internal data quality and collections management concerns.

Speaker



Adam Mansur

Smithsonian Natural Museum of Natural History

Authors

Holly Little

LT16 - Community-Powered, Research-Ready: Citizen Science for a Digital Future

4:00 PM – 5:00 PM | Location: ODIN

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Citizen science has evolved from a niche data-gathering tool into a powerhouse of global biodiversity monitoring. However, as the research community moves toward "Robot-ready" and machine-actionable data, a critical gap remains: how do we maintain the authentic, local, and human-centric nature of community-based work while meeting the rigorous requirements of modern informatics?

This session invites presenters to share their experiences, successes, and challenges in managing Citizen Science programs across all scales. We seek to explore the "bridge" between the volunteer in the field and the digital infrastructure that hosts their data. Topics for discussion include:

- * Program Design: How local initiatives are adapting to produce higher-quality, interoperable data.
- * The Volunteer Experience: Maintaining engagement and agency as data becomes more automated.
- * Bridging the Gap: Case studies on translating "unstructured" community observations into valuable research assets.
- * Inclusivity & Technology: Ensuring that the drive toward standardized, "Robot-ready" data does not exclude under-resourced communities or traditional knowledge holders.

By sharing diverse perspectives—from software developers to community organizers—this session aims to foster a dialogue on how we can empower humans and "robots" to work together for the future of biodiversity conservation.

Moderators



Kumiko Totsu

Data Specialist | National Institute for Environmental Studies



Tim Alamenciak

Session organizers



Ram Dayal Vaishnav

Head of Operations | The Naturalist School



Vijay Barve

Asia Support Team | GBIF

Authors

Ram Dayal Vaishnav, Vijay Barve

Capitalizing the Invisible Ledger: A Cost-of-Production and Shadow Pricing Methodology for Valuing Open-Source Biodiversity Assets in India

4:05 PM – 4:11 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Open-access biodiversity repositories are fundamental for Ecosystem Condition Accounts under the UN System of Environmental-Economic Accounting. However, because data from decentralized networks like iNaturalist and eBird relies on civic volunteerism, it carries a nominal market price of zero and is omitted from GDP. This leads to a public-good market failure, causing suboptimal resource allocation, spatial data gaps, taxonomic biases, and unverified backlogs.

This paper introduces a reproducible valuation methodology to calculate the capitalization of crowdsourced natural data infrastructure. Analyzing India's footprint of over 4.8 million iNaturalist observations against 62 million global occurrences on the Global Biodiversity Information Facility, we model a bottom-up framework to quantify true data production costs.

The methodology deconstructs the unit value of an observation into three non-overlapping cost layers: an infrastructure data-intake base from platform operational footprints (~₹11/observation), a field curation labor asset utilizing a shadow-wage proxy for specialized field survey hours (~₹39/observation), and a taxonomic expertise premium derived from technical consultancy rates (~₹50/observation).

Aggregating these baselines establishes a theoretical production cost of ₹100 per research-grade observation, capitalizing India's citizen science ledger at a proxy value of ₹48.02 crore. For context, the Government of India's official annual data collection budget across the Forest, Botanical, and Zoological Surveys stands at ₹392.58 crore. This methodology provides the baseline required to internalize data-production externalities, clear verification bottlenecks, and scale open-source biodiversity monitoring.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Ashwin A

Coordinator - Citizen Science and Research | The Naturalist School

African Ocean Biodiversity Atlas

4:11 PM – 4:17 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

The African Ocean Biodiversity Atlas addresses UN Ocean Challenges 8 and 2 by advancing digitalization and ecosystem restoration through innovative ocean applications. Led by Ocean Rock Base and expanded by the Africa Ocean Alliance, the project develops a comprehensive digital repository of marine biodiversity, including fisheries, mangroves, algae, and blue carbon ecosystems across 20 African nations. Key components include a Fisheries Atlas and Blue Carbon Biodiversity Outlook, supported by locally developed, affordable technologies that enhance data sharing among indigenous communities and policymakers. The initiative provides high-quality geospatial, photographic, and video data to inform the establishment of Marine Protected Areas, promote sustainable fisheries, and support blue economy investments. Pilot projects in Ghana demonstrate scalable models for mapping coastal ecosystems and fisheries, fostering regional collaboration and capacity building. By integrating community-driven data collection with scientific research, the project aligns with UN Decade goals to protect and restore marine biodiversity and ecosystems. Our efforts aim to empower policymakers, researchers, and private stakeholders with reliable, accessible information to drive sustainable ocean management, climate resilience, and blue carbon initiatives across Africa.

Speaker



Peter Teye Busumprah

Ocean Rock Base And Africa Ocean Alliance

From citizen observations to long-term ecological knowledge: integrating iNaturalist records into LTER framework through automated enrichment

4:17 PM – 4:23 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Citizen science observations have become major contributors to biodiversity knowledge, generating large volumes of high-quality data through widespread volunteer participation. Much of this information is shared via global open-access platforms such as iNaturalist; however, these observations often remain outside formal long-term ecological research workflows.

To address this gap, an automated R-based workflow has been developed to harvest “research-grade” iNaturalist observations occurring within sites of the Long-Term Ecosystem Research (LTER) network in Europe and transform them into enriched, research-ready data for long-term ecological analysis. The workflow identifies observations geographically located within LTER sites and classified as high quality on iNaturalist, and enriches them via automated semantic augmentation using authoritative external sources.

Each record is complemented with: (i) IUCN Red List conservation status, based on geographic relevance; (ii) nativeness information (Darwin Core “establishmentMeans”), filtered by country and sourced from iNaturalist; (iii) taxon protection status under the EU Habitats and Birds Directives, retrieved from the EUNIS database; and (iv) contribution to eLTER Standard Observations.

The resulting R package, “ReLTER.inatEnrich”, provides an innovative integration layer enabling citizen-generated biodiversity data to directly support structured long-term ecological research. This approach offers a dual benefit: it supplies LTER sites with a replicable and user-friendly protocol to complement existing monitoring and enables cross-site comparisons, while enhancing the scientific value of citizen science data through contextual enrichment and integration into standardized ecological frameworks.

Speaker



Alessandro Oggioni

National Research Council of Italy - Institute of Marine Science (CNR-ISMAR)

Authors

Alessandro Oggioni (CNR IREA - Institute for Electromagnetic Sensing of the Environment), Alice Lenzi (CREA - Council for Agricultural Research and Economics), Alessandro Campanaro (CREA - Council for Agricultural Research and Economics, Italy), Alexander Palummo (CNR IBE - Institute of Bioeconomy), Nicola Vuolo (CNR IBE - Institute of Bioeconomy), Caterina Bergami (CNR ISMAR - Institute of Marine Sciences)

An eDNA Platform for Community-Powered LA River Restoration

4:23 PM – 4:30 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

The Los Angeles River runs roughly 51 miles, confined for nearly its entire length within a concrete flood-control channel. Channelization severed the river from its floodplain, collapsing riparian and wetland habitats and biodiversity. Community organizations including Friends of the LA River and Metabolic Studio have worked to restore the river through cleanup, education, and rewilding segments. Revitalization at this scale requires a way to prioritize effort and measure success, in particular whether biodiversity is returning, and which sites will benefit. Environmental DNA (eDNA), the genetic material organisms shed into air, water, and soil, provides a snapshot of biodiversity across the tree of life from citizen scientist-collected samples, avoiding the time, labor, and specialized expertise that conventional field surveys demand. In collaboration with University of California’s CALeDNA program and community partners supported by Metabolic Studio, we implemented eDNA sampling along the LA River. To make the multi-marker, high-dimensional dataset usable, we built the eDNA Explorer platform. This platform identifies organisms from raw eDNA sequencing data, integrates ~400 remote-sensing geospatial features from Google Earth Engine, and presents results through an intuitive interface that community groups can interpret without bioinformatics expertise. GBIF occurrence data is integrated, allowing users to filter and visualize detections by geographic plausibility. Machine learning models identified environmental and site-level factors most associated with biodiversity outcomes, turning a complex dataset into actionable insights. These tools give community groups a shared evidence base to interpret restoration progress and prioritize future revitalization effort.

Speaker



Ryan Martin

Senior Data Scientist | eDNA Explorer

Authors

Ryan Martin, Jim Jeffers, Julie Stanford, Rachel Meyer

Ground-Truth from the Ground Up: Community-Led Occurrence Data for *Prosopis juliflora* Restoration in Kutch's Climate-Stressed Drylands

4:30 PM – 4:36 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

In landscapes where ecological knowledge lives in people rather than databases, citizen science cannot mean app-based observation submission alone. Kutch, Gujarat, is one such landscape, an arid dryland system where communities hold decades of firsthand knowledge about how *Prosopis juliflora* has moved across grasslands, salt flats, and grazing corridors. This talk presents how Climitra puts that knowledge to work as primary biodiversity data.

Community field teams operate alongside satellite-based land cover classification, not beneath it. They confirm, correct, and spatially extend the occurrence layer into zones that spectral indices routinely misclassify. Structured to Darwin Core standards and contributed to GBIF, their observations move from informal ground knowledge to citable, interoperable records usable by researchers, conservation practitioners, and automated monitoring pipelines alike.

The urgency is real. With Super El Niño conditions forecast for 2026-27, *juliflora* is positioned to expand precisely where drought stress weakens native grassland resistance. Community-generated occurrence data is not a supplement to remote sensing here, it is a frontline climate adaptation tool. Downstream restoration outcomes, biochar soil amendment, fuel briquette provision, and native species recovery are spatially targeted only because communities made the data research-ready.

We share protocols for structured field data collection without app dependency, quality assurance workflows, and practical pathways for making vernacular ecological knowledge interoperable with global biodiversity infrastructure.

VIRTUAL PRESENTATION

Speaker



Pranav Jha

Corresponding author



Udbhav Rana

Climitra Carbon Pvt. Ltd.

Authors

Udbhav Rana, Pranav Jha, Khyati Thacker

The Asian Nature Challenge as a Beacon of Citizen Science: Regional Contributions, Macroecological Data Networks, and Strategies for Optimizing Taxonomic Identification Efficiency in India

4:35 PM – 4:40 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

The Asian Nature Challenge (ANC) represents a major milestone in decentralized biodiversity monitoring, serving as a powerful engine for global open-access macroecological data networks. By evaluating massive crowd-sourced datasets—specifically `observations-740399.csv` and `observations-740433.csv`—this cross-study analyzes the mechanics of regional participation, data quality, and taxonomic validation patterns across Asian nations. Empirically, the ANC acts as a beacon of citizen science, mobilizing tens of thousands of participants to generate over 190,000 observations, mapping critical target species, endemics, and geographic range extensions. However, comparative analysis reveals stark disparities in verification efficiency: while Taiwan and Japan achieve high Research Grade conversion rates of 62.97% and 56.08% respectively, India exhibits an identification bottleneck, converting only 40.11% of its 47,346 observations. Granular taxonomic breakdowns for India indicate high proficiency in ornithology (89.88%), contrasted against systemic deficits in mega-diverse groups such as *Plantae* (27.18%), *Arachnida* (30.89%), and *Fungi* (4.04%). These backlogs stem from a structural deficit in expert identifier networks relative to active field observers, alongside limitations in local computer vision models for complex tropical flora and micro-fauna. To scale up utility for formal scientific research, India must transition from raw data acquisition to active identification curation. Lessons from successful regional networks show this can be achieved by engineering dedicated expert-curator incentives, hosting targeted digital 'ID-a-thons', and integrating localized AI-assisted taxonomic vetting pipelines to streamline the validation flow.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Ashwin A

Coordinator - Citizen Science and Research | The Naturalist School

Authors

Ram Dayal Vaishnav, Ashwin A, Lavanya Bhat, Adarsh Pradhan

Growing Together: Community-Driven Conservation of Plant Genetic Resources

4:36 PM – 4:42 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

The EU project INCREASE (Intelligent Collections of Food Legumes Genetic Resources for European Agrofood Systems) fosters agricultural biodiversity through participatory research. With 28 partners from 14 countries, it focuses on legumes (chickpea, bean, lentil, lupin). A central component is the Citizen Science Experiment (CSE) with common bean across Europe. CSE aims to: increase citizen awareness, generate phenotypic data of 1130 bean lines across European environments, and test decentralized conservation approaches for germplasm conservation and multiplication.

Beans are distributed with the help of the INCREASE CSA mobile app, integrating observation and teaching tools, and legal frameworks (SMTA via FAO Easy SMTA interface). The system employs digital infrastructure for seed exchange with DOI registration via the FAO Global Information System GLIS, ensuring machine-actionable data provenance. Each citizen receives a bag with 5 PGR lines plus one commercial check variety. Citizens are assigned to Randomization Units based on latitude, elevation, and 3 bioclimatic variables to ensure environmental diversity.

Dissemination prioritizes native languages through voluntary translators, with partners recruiting in their own countries using all available media. Engagement is fostered via picture galleries, questionnaires, contests, and private Facebook groups. We share experiences designing a program balancing human-centric citizen science with modern biodiversity informatics requirements. The CSE attracted over 26,000 participants, including 650+ schools, fostering public education on PGR topics from an early age. We discuss how the INCREASE CSA app bridges community observations and research assets, demonstrating a scalable decentralized conservation model.

Speaker



Markus Oppermann

Senior Scientist | Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

Authors

Markus Oppermann, Stephan Weise, Kerstin Neumann

Discussion

4:42 PM – 5:00 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

WM48 - IQAIR Task Group Working Meeting: Image Quality and AI-Readiness Metadata for Biodiversity Media

4:00 PM – 5:00 PM | Location: Møterom 5 / Meeting room 5

AI AND ROBOT READY

The TDWG Image Quality and AI-Readiness (IQAIR) Task Group focuses on developing metadata concepts and vocabulary needed to describe biodiversity media in ways that support automated processing and AI-based analysis. As digitization efforts continue to produce large volumes of specimen images and other multimedia resources, standardized descriptors for image quality, content, and machine usability are increasingly important for enabling computational workflows.

This working meeting will bring together members of the IQAIR Task Group and interested participants to advance ongoing work on defining image quality and AI-readiness metadata for biodiversity media. The meeting will review progress on candidate metadata terms, examine alignment with existing TDWG standards such as Audiovisual Core and Darwin Core, and discuss use cases involving automated trait extraction, image analysis, and machine-generated annotations.

Participants will discuss practical implementation issues, including how repositories and aggregators can incorporate image quality descriptors, automated metadata extraction, and machine-actionable annotations into existing data infrastructures.

The goal of the meeting is to refine vocabulary proposals, clarify use cases, and identify next steps for developing community recommendations that support AI-ready biodiversity media within the TDWG standards framework.

Workshop leaders



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute



Henry L. Bart

Professor Emeritus of Ecology and Evolutionary Biology & Curator Emeritus of the Royal D Sutkus Fish Collection Collecion | Tu...

DS52 - Discussion of Responsible AI for Open, Integrated Data Creation, Curation and Enrichment

4:00 PM – 5:00 PM | Location: BALDER

AI AND ROBOT READY

This discussion session is a companion session to *SYM13 Responsible AI, Open Digital Curation, and Round-Tripping for Biodiversity Data* and will explore the themes raised by talks in the two sessions held earlier in the day.

Discussion leaders



Deborah Paul

Species File Group (INHS), University of Illinois



Siobhan Leachman

Wikimedian | Wikimedia Aotearoa New Zealand



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

Wednesday, September 23, 2026

EXCURSIONS

9:00 AM – 5:00 PM

Conference Social

6:00 PM – 10:00 PM | Location: Munch Museum

The conference social will be held at [Munch Museum](#)

Two drinks and finger food will be offered.

There will be a guided tour of the Museum and the offer of a ticket to enter the museum at another time as well.

Thursday, September 24, 2026

CHECK-IN

8:00 AM – 8:30 AM | Location: 1st Floor

SYM29A - From Cabinets to Clouds – Bridging the gap between collection- and database management

8:30 AM – 10:30 AM | Location: SAL A

COLLECTIONS DIGITISATION & MANAGEMENT

Physical collections and digital databases are often managed in parallel by collection managers and data managers. As dedicated development teams become more common in institutions, opportunities for misalignment between those working in the collection cabinets and those in the data clouds arise.

Across institutions, this relationship is handled in different ways. A wide range of approaches for coordinating collection digitisation pipelines exist, reflecting differences in institutional structures and ongoing integration of new development roles into long-established collection management practices and workflows. In some cases, workflows are closely aligned across teams, while others are more loosely defined. These differences offer an opportunity to learn how institutions balance collection management expertise with emerging digital practices.

We aim to bring together a range of examples to learn from each other's institutional workflows to implement biodiversity data standards and contribute to global biodiversity knowledge. Topics could include experiences and best practices in facilitating functional relationships between collection and data managers; decision-making processes for data standards; having one expert in two domains vs. two experts; enrich

Moderators



Caitlin Thorn

Data Manager | Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science



Nora Lentge-Maaß

Head of Coordination Office for Digitization on Demand, Deputy Head of Department Services for Collection Future | Museum für Naturkunde Berlin

Session organizers



Frederik Berger

Museum für Naturkunde Berlin



Mareike Petersen

Lead Science Programme Collection Future | Museum für Naturkunde Berlin



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

Introduction

8:30 AM – 8:40 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Speakers



Nora Lentge-Maaß

Head of Coordination Office for Digitization on Demand, Deputy Head of Department Services for Collection Future | Museum für Naturkunde Berlin



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

Collaborative Development of an Institution-Wide Controlled Vocabulary

8:40 AM – 8:55 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Developing an institution-wide controlled vocabulary presents significant organisational and technical challenges. At the *Museum für Naturkunde Berlin*, a Controlled Vocabulary (CV) Working Group was established to harmonise terminology used across the institution. Recent digitisation efforts and migration to a shared Collection Management System (Specify) revealed inconsistencies in terminology and object description practices across subcollections, limiting interoperability and discoverability.

Over a two-year period, the CV Working Group developed an institution-wide controlled vocabulary through an iterative process. Initial focus groups identified requirements and consolidated existing terminology lists into a preliminary thesaurus. A key factor in the project's success was the structured coordination of stakeholder involvement. Internal presentations and regular communication with collection managers and other interested parties helped build institutional support for adopting the CV and updating collection records. Domain expertise from collection staff was essential to refine the vocabulary and ensure that it remained accurate and applicable across subcollections. Following the distribution of the preliminary vocabulary, participants were invited to review and refine term definitions.

The CV has since been integrated into institutional digitisation workflows, and its ongoing maintenance is now embedded in collection data management practices at the museum. This case study highlights the importance of collaboration and iterative review in the implementation of institution-wide vocabulary standards.

Speaker



Caitlin Thorn

Data Manager | Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science

Authors

Sabine von Mering, Nora Lentge-Maaß, David Ware, Jessica Zimmermann

From Silos to Standards – Advancing Standardised Collection State Documentation for Institutional Continuity and Coordination

8:55 AM – 9:10 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Collection-specific knowledge often lies exclusively with collection management staff. While collection managers understand the state of the collections they oversee, this information rarely reaches decision-makers or data managers in a standardised form. In South Africa, many institutions lack dedicated data managers, and those that do often lack access to collection-related contextual information needed to prioritise digitisation workflows and support data compliance.

Standardising collection state documentation addresses these challenges on three fronts. Operationally, institutions gain visibility into collection conditions and priorities, improving resource allocation. Institutionally, the system creates institutional memory that persists beyond staff turnover, preserving knowledge and collection history for future staff. Strategically, a shared standard introduces a common vocabulary that enables siloed institutions to move beyond isolated problem-solving toward coordinated, community-based approaches.

In 2025, we launched a prototype system across partner institutions using Google Apps Script and Google Data Studio. Although engagement varied, the pilot validated the concept. However, maintaining the system, onboarding collections, capturing data through Google Forms, and manually configuring dashboards proved labour-intensive and difficult to scale.

To address these limitations, the prototype was refactored into a standalone web application that streamlines collection state documentation and reporting workflows across institutions. The system supports institutional continuity while enabling a more standardised approach to collection management in South Africa.

Speaker



Jacobus Christiaan Steenkamp

Senior Specialist: Data Coordination | Natural Science Collections Facility (South African National Biodiversity Institute)

Bridging the Gap: Strategies for Integrating Person Identifiers into Collection Databases

9:10 AM – 9:25 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Unambiguous identifiers for person names are critical for research based on collection data. Every specimen in a collection is the result of human agency — through collection, preparation, description, and curation. Identifying these agents enhances data quality by referencing the source of intellectual authority, by resolving collector networks or by providing additional insights into the history of collections. While identifier management workflows have existed for decades, Wikidata has become a central broker for biodiversity-related identifiers. The introduction of the “collection items at” property (P11146) in 2022 enabled the seamless linking of collectors to holding institutions in Wikidata. In less than four years, over 47,000 items (i.e. unambiguously identified persons) have been assigned this property, showcasing a significant cross-disciplinary community effort.

Despite this progress, the integration of unique identifiers into primary collection databases remains slow. This process requires deep domain knowledge, identifier management expertise, and, most crucially, significant time. This creates a fundamental bottleneck: collection curators possess the necessary knowledge and access but lack the time to prioritize integration, whereas the broader community has the capacity for disambiguation but lacks access to institutional records.

This presentation will delve into the process of person name disambiguation, exploring possible solutions that can decouple the workflow and enable collections curators to benefit from community-curated databases like Wikidata. By understanding these challenges and implementing strategies for collaborative data enrichment, we can enhance our ability to leverage human agency within biodiversity collection data.

Speaker



Frederik Berger

Museum für Naturkunde Berlin

Authors

Sabine von Mering

Confronting legacy data for primary type specimens

9:25 AM – 9:40 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Often historic natural history collections pose great challenges such as tracking taxonomic changes between the database and the physical labels, properly recording specimen movements, and maintaining legacy databases. This is particularly true for the SI NMNH Entomology Department, which is home to ~150,000 primary types.

In 2007, data was migrated from INQUIRE, a database designed by the department, to EMu, a Collection Information System (CIS) used museum wide. Due to structural incompatibilities of INQUIRE with EMu, the data migration faced challenges and significant nuances of INQUIRE data were lost. All type records migrated from INQUIRE were categorized as “INQUIRE TYPES” in EMu and publicly flagged, while new type records were categorized as “Entomology Types”. From the time of the migration, the department tried to remedy the lost nuance by updating INQUIRE Type records whenever a type was handled.

Dubbed the Type Inventory Project for Entomology Specimens (TIPES) Project, Collections staff are working alongside Curators to reconcile primary type specimen records in EMu against registration files and the physical collection, while maintaining a list of what should be physically present. The project has created opportunities to redefine workflows and ensure proper data capture, physical specimen security, and legal accountability, while sharing up-to-date records online and on GBIF to mitigate future data loss.

The goal is to develop a streamlined workflow that removes communication barriers and prevents bottlenecks from forming on any one role. Additionally, the department is exploring AI-driven tools to identify gaps in the Primary Type collection via automated literature reviews for keywords, compared against the existing known inventory after its completion.

Speaker



Jessica Bird

Smithsonian National Museum of Natural History, Department of Entomology

Authors

Torsten Dikow, Cailin Meyer

From Legacy Database to Cloud: Challenges in Migrating Legacy Collection Data to a FAIR Collection Management System

9:40 AM – 9:55 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Natural science collections play a vital role in scientific research, yet they often rely on outdated legacy databases that reflect years of curatorial practices. At [Senckenberg German Entomological Institute](#), these collections are currently managed across Corel Paradox relational databases and additional spreadsheets, each with customized schemas and specific entomological terminology. However, only 22,470 of our estimated 3,800,000 specimens are accessible through the [GBIF Global Registry of Scientific Collections](#). The lack of standardization, interoperability and accessibility restricts the potential of our valuable collection data for biodiversity research.

We present a working example of our approach and processes to migrate heterogeneous data into the [DINA Collection Management System](#). We identify three key challenges in this migration process. First, we must reconcile inconsistent database schemas and curatorial terms with DINA's data model, which is closely aligned with the Darwin Core standard. Second, we must tackle data quality issues that arise from maintaining a legacy curation tool, for example, outdated taxonomic names and discrepancies with public taxonomic backbones, such as the [Catalog of Life](#). Finally, we must coordinate efforts across different institutes within Senckenberg, each with unique schemas, curation workflows and institutional cultures.

By completing the migration to DINA, we anticipate achieving FAIR-compliant, identifier-based collection data that enhances accessibility and usability through GBIF and for researchers worldwide. We aim to demonstrate how data managers can bridge collection management expertise and new digital infrastructure to implement data migration projects that meet FAIR standards.

Speaker



Marije Timmermans

Authors

Jana Hoffmann

A framework for aligning collections data with the global data ecosystem from data creation to extension in the Smithsonian National Museum of Natural History Department of Paleobiology

9:55 AM – 10:10 AM

COLLECTIONS DIGITISATION & MANAGEMENT

For many years, the biodiversity and geoscience informatics communities have worked together to establish a dynamic global data ecosystem that supports interdisciplinary and cutting-edge research. Natural history collections provide vital specimen-based data that are foundational for this ecosystem. While informatics communities expand data across this landscape by developing new concepts, data models, standards, and more, collections are challenged to move beyond baseline data practices and keep pace with evolving data needs.

The Smithsonian National Museum of Natural History Department of Paleobiology (NMNH Paleo) recently launched a multi-year program to create an integrated framework that addresses enhancements in our data capture and processing, ultimately enabling a more robust local data ecosystem that can adapt to global developments. To achieve this goal, we will evaluate the best methods for implementing results from work by the broader community throughout our processes, creating data that can be extended or integrated with new sources and made more actionable and open for broader use. This initiative will also synthesize components of disparate projects, leveraging practices refined over many years across the biodiversity informatics and paleontology collections communities, as well as through recent cyberinfrastructure evaluation projects. Importantly, our work will be supported by a cross-functional team, incorporating a variety of expertise and perspectives to inform this framework. While our primary focus is on the NMNH Paleo collections, this integrated approach is intended to be adaptable to other paleontology collections and likely other natural history collections.

Speaker



Holly Little

National Museum of Natural History

Authors

Amanda Millhouse, Erica Krimmel, Alexandra Lawrence

Discussion

10:10 AM – 10:30 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Moderator



Caitlin Thorn

Data Manager | Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science

Authors

Sabine von Mering, Nora Lentge-Maaß

SYM15 - Operationalizing Biodiversity Digital Twins within Data Space Ecosystems

8:30 AM – 10:30 AM | Location: SAL B

DIGITAL TWINS & EMERGING INFRASTRUCTURES

This session explores the transition from prototype to operation of Digital Twins (DTs) in biodiversity, building upon the foundational approaches elaborated in projects such as Biodiversity Digital Twin (BioDT), Integration of biodiversity monitoring data into the Digital Twin Ocean (DTO-BioFlow) and Biodiversity Meets Data (BMD). As the landscape of environmental monitoring matures, we examine how these technological advancements are being institutionalized within architectures like the Green Deal Data Space, the EU flagship initiative Destination Earth (DestinE), Research Infrastructures such as LifeWatch ERIC, eLTER, DiSSCo, and intergovernmental organizations involving GBIF and OBIS to support long-term environmental governance.

We invite contributions detailing the practicalities of this shift - specifically regarding data models, FAIR workflow and provenance preservation, metadata standardization, and the operational expansion of DT approaches. By fostering a dialogue between diverse DT initiatives, this session aims to define a roadmap for a sustainable digital ecosystem that empowers key policy frameworks like IPBES with scalable, evidence-based tools, ensuring that DT innovations evolve into permanent assets for European and global biodiversity decision-making.

Moderators



Claus Weiland

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main



Dag Endresen

GBIF Node Manager for Noarway | Natural History Museum, University of Oslo, Oslo

Session organizers



Gabriela Zuquim

Scientific Coordinator | CSC - IT Center for Science



Sharif Islam

Senior Data & System Architect | Naturalis Biodiversity Center



Hanna Koivula

Senior Research Data Management Specialist | CSC - IT Center for Science, Espoo



Niels Raes

Programme Manager Biodiversity Informatics | Naturalis Biodiversity Center

Authors

Tomas Martinovic, Taimur Khan

Introduction

8:30 AM – 8:35 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

Speakers



Claus Weiland

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main



Dag Endresen

GBIF Node Manager for Noarway | Natural History Museum, University of Oslo, Oslo

The BMD Cubing Engine: Harmonizing Biodiversity and Earth Observation Data

8:35 AM – 8:49 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

The Biodiversity Meets Data (BMD) project seeks to provide Natura 2000 site managers with data-driven tools for evidence-based environmental decision-making. Such tools depend on biotic and abiotic datasets to be rendered interoperable in a reproducible manner. Yet the transition from prototype biodiversity data products to operational workflows is hindered by fragmented data sources, inconsistent metadata practices and insufficient preservation of provenance.

To address these challenges, the BMD data cubing engine designs pipelines for producing spatially and temporally harmonized datasets, integrating biodiversity records with Earth Observation (EO) data. The engine generates analysis-ready products that can be reused in downstream modelling and decision-support workflows. A central design principle is provenance preservation: transformations, source datasets and harmonization steps are explicitly captured to support transparency and reproducibility.

This presentation will focus on the practical data model and metadata challenges encountered when aligning EO data with biodiversity observations. Particular attention will be given to metadata interoperability across heterogeneous inputs, in particular the representation of spatial, temporal, taxonomic and methodological context. By presenting the current BMD cubing architecture, we offer a practical contribution to the operational expansion of biodiversity Digital Twin approaches. The proposed design illustrates how FAIR workflows, standardized metadata, and traceable data transformations can support reusable ecological data products for integration into emerging European and global biodiversity data infrastructures.

Speaker



Mathias Dillen

Biodiversity Data Scientist | Meise Botanic Garden

Authors

Niels Billiet

Building the Foundations for a European Biodiversity Data Space

8:49 AM – 9:03 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

European Green Deal Data Spaces serve as catalysts for unlocking data-driven innovation by integrating open and private datasets to help Europe achieve climate neutrality by 2050. By fostering cross-domain collaboration in earth and environmental sciences, economics, and urban planning, these initiatives connect scattered systems into a single, interoperable ecosystem. In the domain of conservation, this is vital for supporting the EU's nature restoration targets and deforestation reduction efforts. Within this strategic context, the Biodiversity Meets Data (BMD) project is actively developing the foundations of a European Biodiversity Data Space.

To enable near-data processing, we are building on workflow.earth, a FAIR-compliant workflow engine engineered originally within the Destination Earth (DestinE) initiative, which is now actively applied to streamline the ecological monitoring of Natura 2000 sites. Leveraging Argo Workflows and declarative pipeline definitions, these workflows fully encapsulate the instructions required for portable, reproducible, and deterministic execution across any Argo-compatible Kubernetes cluster (current in operation at UFZ/JSC and LifeWatch ERIC), while offering researchers a clear, transparent, and readily visualizable view of the underlying scientific processes.

Data portability and workflow transparency are achieved by adopting the Workflow Run RO-Crate (WRROC) specifications. This enables the packaging of workflow definitions, runtime parameters, data outputs, and rich metadata into self-contained FAIR Digital Objects (FDOs). All encapsulated primary data are also preserved as linked data, facilitating the seamless discovery, mobilization, and replication across the broader Common European Data Spaces.

Speaker



Claus Weiland

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main

Speaker (VIRTUAL)



Lena Perzmaier

Senckenberg Gesellschaft für Naturforschung

Authors

Lena Perzmaier, Sharif Islam, Taimur Khan, Julian Oeser, Julio Paneque, Ayco Holleman

Beyond Data Integration: Ecological and Semantic Interoperability in Coastal Digital Twins

9:03 AM – 9:17 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

Coastal zones are highly complex environments that are increasingly affected by sea-level rise, erosion, extreme weather events, biodiversity loss, pollution, and intensifying human activity. The development of coastal zone digital twins brings together diverse data types stemming from multiple domains, including marine science, urban planning, ecology, shipping, fisheries, climate science, and disaster management. The perspectives of these different disciplines and their respective data governance regimes collide, resulting in data integration challenges.

The paper presents a map of coastal zone semantic interoperability challenges, situated in relation to international frameworks including the EIF - European Interoperability Framework. Examples are drawn from a demonstrator twin for Oslofjord, where data is combined from diverse cross-disciplinary observation systems and governance domains. We have organized our map into different classes of problem-types, along with suggestions for tooling and potential solutions. We provide a concrete, practical overview of issues encountered by twin developers, spanning from problematic definitions and ontology misalignment, to inconsistent classifications, divergent coordinate reference systems, fundamental differences in spatial representation, separate metadata standards, temporal mismatches, as well as the limitations of available ontologies. We further cover the non-technical challenges, such as those related to uneven adoption of standards.

The work lays the foundation towards our main goal, “How do we enable a new generation of environmental digital twins that seamlessly integrate biodiversity data between marine, coastal, and terrestrial ecosystems of twins?”

Speaker



Laura Slaughter

Senior Researcher | SINTEF Digital

Authors

Arne Jørgen Berre

From Buzzwords to Blueprints: What It Really Takes to Operationalise Biodiversity Digital Twins

9:17 AM – 9:31 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

The language of Digital Twins and Data Spaces has rapidly entered biodiversity monitoring and conservation but what does operational reality actually require? In this talk, I move past the buzzwords to examine the concrete components that determine how a Digital Twin transitions from prototype to sustained infrastructure, and where this fits into the emerging Data Space ecosystem.

Drawing on experience from BioDT, DTO-BioFlow, and Biodiversity Meets Data (BMD), I identify building blocks that make the difference. Operationalisation demands more than a working pipeline: it requires interoperable metadata patterns, robust provenance preservation, and honest engagement with real-world metadata quality problems. I illustrate this through concrete examples where metadata and crosswalks become decisive. For instance, harmonising Annex I habitat types across EEA Natura 2000 datasets to reconciling marine functional trait vocabularies. These are not edge cases; how we handle them determines whether data can actually flow between systems. We can also learn from the earth observation use cases, where the combination of Cloud-Optimized GeoTIFF (COG) and SpatioTemporal Asset Catalog (STAC), and cloud object storage represents something more fundamental than technical convenience: a set of contracts between producers and consumers governing access patterns, discoverability, and long-term availability. This architecture is what makes scalable services possible, and offers a model for operationalising biodiversity twins and connecting them into a data space.

I conclude by situating these technical foundations within the institutional architectures the session addresses and outline what a shared metadata and interoperability contract could look like.

Speaker



Sharif Islam

Senior Data & System Architect | Naturalis Biodiversity Center

Digital Twins as System-Optimization Tools: Provisioning Edge and Cloud Infrastructure for Biodiversity Monitoring at Scale

9:31 AM – 9:45 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

Monitoring dynamic ecosystems increasingly requires autonomous systems that make decisions at the far edge, where sensors meet the environment. Biodiversity digital twins, however, are typically positioned as integration and visualization layers over data streams produced by sensor networks whose hardware, models, and edge-cloud placement have already been fixed. This framing treats the cyber-physical system as fixed and external. We argue that monitoring biodiversity at landscape scale should be approached as a system-optimization problem, and operational DTs must model and help provision the cyberphysical system itself, including the edge AI stack where most decisions are made.

Autonomous ecosystem sensing must balance processing speed, energy use, network availability, and ecological goals in ways single-platform designs cannot resolve. Smart sensors, powered by edge AI, close a local loop at the edge; coordinating them across modalities, sites, and ecosystems is a provisioning decision with ecological consequences. Where to place a sensor or model, when to offload to the cloud, and how mission settings interact with accuracy are decisions a data-only DT cannot represent, let alone optimize.

Robotics-readiness requires embodied intelligence, where models, sensors, and platforms act on the world they observe. Extending DT architectures to support this requires co-simulating layers that data-only DTs treat as out of scope, combining ecological realism with edge-deployed AI models, smart sensors, and edge-to-cloud networks. We argue that modeling the cyberphysical system, including the dynamic ecosystems it observes, is what makes DTs operational and necessary infrastructure for biodiversity monitoring at scale.

Speaker



Tanya Berger-Wolf

Professor, Computer Science and Engineering | The Ohio State University

Authors

Jenna Kline

How to operationalize large-scale aerial biodiversity monitoring?

9:45 AM – 9:59 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

At the time of this conference, millions of songbirds are migrating south across Europe. Many do this at high altitudes and under the cover of darkness, making it hard to monitor them at scale with citizen science or bio-logging. Weather radars can detect these aerial movements, 24/7, through an existing network that covers most of continental Europe.

Over the last decade, the radar aeroecology community has developed methods to extract biological signals from weather radar data. Through the Aloft data portal (<https://aloftdata.eu>), these data are now available from 2012 onwards for 181 radars in 24 European countries and updated near real-time. The resulting biological data consists of vertical profiles time series (VPTS) with the speed, direction and density of animals aloft. The data are standardized following the VPTS CSV standard and can be accessed, processed and visualized through a number of open source software tools ([getRad](#), [bioRad](#), [CROW](#)).

While these data do not contain species-level information, they are an important quantitative source used to study biomass flows, migration timing and aerial habitat use. They have also enabled the development of migration prediction models, which are important for flight safety, collision prevention, disease spread modelling and public outreach.

To truly evolve to a digital twin of the sky, the Aloft data portal needs to address a number of key challenges. Funding for maintaining software and infrastructure is limited, access to unfiltered weather radar data remains an issue, and methodological work is needed to better separate bird and insect signals. By collaborating with other biodiversity infrastructures, we hope we can continue to provide and expand radar-derived data products for policy and research in Europe.

Speaker



Peter Desmet

Open data coordinator | Research Institute for Nature and Forest (INBO)

Authors

Pieter Huybrechts, Bart Kranstauber, Cecilia Nilsson

Operationalising biodiversity DT ecosystems through assessment-driven service mapping and reusable tools and workflow integration

9:59 AM – 10:13 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

Operationalising biodiversity Digital Twin (DT) ecosystems requires interoperable, reusable, and scalable integration of data, services, and workflows across distributed environmental infrastructures. This work proposes an assessment-driven framework for biodiversity and bioflow DT ecosystems based on service mapping, reusable analytical components, and workflow integration to support operational deployment across biodiversity and marine domains. The approach, tested through HE BioDT and continued in HE BioFlow, systematically identifies, evaluates, and connects datasets, models, computational services, and stakeholder requirements into interoperable workflows adaptable across applications and regions.

The assessment framework reduces fragmentation between platforms and solutions while enhancing transparency, reproducibility, and findability of biodiversity DT implementations. We propose an approach integrating environmental resources through standardised interfaces and modular tools, where assessment-driven matching enables identification of suitable execution and operational environments.

In this context, Data Spaces can benefit from biodiversity DT ecosystems by providing operational environments where FAIR data, services, and analytical workflows can be shared, reused, and connected across domains.

The proposed framework contributes to the objectives of the European Mission “Restore our Ocean and Waters” by supporting assessment of DT ecosystems for improved environmental intelligence and evidence-based decision support for marine and coastal resilience.

Speaker



Syrine Ben Aziza

Scientist Integrator | TNO

Discussion

10:13 AM – 10:30 AM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

Moderator



Claus Weiland

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main

SYM20 - Meeting Biodiversity Data Standards Using Specify Software

8:30 AM – 10:30 AM | Location: SAL C

COLLECTIONS DIGITISATION & MANAGEMENT

As the biodiversity community moves toward pursuing FAIR principles and the concept of the Extended Specimen, the need to produce high quality, biodiversity standard compliant data has never been greater. However, implementing these standards can present a challenge for collections managers and institutions.

This session aims to showcase the role Specify software has played in empowering collections staff to digitize, manage, and export their collections' data according to Biodiversity Data Standards. The session will feature presentations from Specify users at institutions across the globe who will highlight data management workflows using Specify tools and techniques.

In addition to user presentations, the Specify Collections Consortium (SCC) will also deliver an overview of Specify's core and newer features that allow users to meet evolving Biodiversity Data Standards and ensure data quality, standardization, and interoperability.

Moderator



Theresa Miller

Membership and Fulfillment Officer | Specify Collections Consortium

Introduction

8:30 AM – 8:33 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Speaker



Theresa Miller

Membership and Fulfillment Officer | Specify Collections Consortium

SpecifyaaS: Managing Ephemeral Repositories in Collections-Based Research

8:33 AM – 8:46 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Specify 7 has substantially expanded the capacity of institutions to standardize efficient biodiversity data capture, management, and exchange during collections-based research. At the Swedish Museum of Natural History (NRM), data are maintained across multiple Specify instances, which includes databases for research programmes and projects. They support collaborative initiatives characterized by operational constraints that often differ greatly from those of the public repositories where research specimens will ultimately be deposited.

Here I describe the use of Specify 7 instances as "ephemeral repositories" within the biodiversity data network at the NRM. Although they are often broader in scope with respect to the data compiled, and less likely to persist as independent data aggregations, ephemeral repositories are critical staging arenas for data standardization, mobilization, and exchange. As an low-investment outcome from using Specify 7, that reduces reliance on bespoke or isolated data management solutions, Specify-as-a-Service is intended to support efficient transfer of information to downstream systems after phases of active research.

Investment in early standardization improves data quality and reduces duplication of effort during rounds of data capture. Extending IT infrastructure into the broader suite of institutional tools supporting collections-based research promotes the implementation of FAIR practices and consistent workflows. Providing SpecifyaaS to staff and research collaborators will strengthen data integration across research and collections systems, and increase the long-term data usability and applicability across a wider range of use cases.

Speaker



Kevin Holston

Applying Data Standards at the Royal Botanic Garden Edinburgh

8:46 AM – 8:59 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Data standards play a crucial role in enabling the dissemination, interoperability, and reuse of biodiversity knowledge. At the Royal Botanic Garden Edinburgh (RBGE), efforts over more than four decades have focused on aligning herbarium collection data with a wide range of cultural, historical, technological, and biodiversity data standards. This ongoing work reflects both the evolving nature of standards and the need to integrate legacy data with emerging frameworks.

Since the transition to the Specify Collections Management System (CMS) in 2022, RBGE has further refined its use of data standards to enhance data quality, interoperability, and connectivity. These developments support improved sharing and linking of RBGE's herbarium data with collections and research initiatives worldwide, contributing to a more integrated global biodiversity knowledge infrastructure.

Speaker



Robyn Drinkwater

Digitisation Informatics Coordinator | Royal Botanic Garden Edinburgh

From Legacy Data to FAIR Workflows: Standardising Biodiversity Collections Using Specify

8:59 AM – 9:12 AM

COLLECTIONS DIGITISATION & MANAGEMENT

The KwaZulu-Natal Museum manages eleven natural history collections supporting biodiversity research, education, and long-term preservation of African heritage. Like many institutions, historical collection data were distributed across spreadsheets, paper records, and inconsistent legacy systems, creating challenges for data quality, interoperability, specimen tracking, and data mobilisation.

To address these challenges, the museum implemented Specify as its primary collections management system. This enabled the development of standardised workflows for specimen cataloguing, accessioning, georeferencing, taxonomic management, loan tracking, image attachment, and data cleaning across eleven collections.

A major focus of implementation has been aligning institutional workflows with biodiversity informatics standards, FAIR data principles, and machine-readable data practices. Centralised data management has improved metadata consistency, reduced duplication, and strengthened quality control. Specify's querying and export functions have also improved readiness for data sharing and publication to global biodiversity infrastructures such as Global Biodiversity Information Facility.

The museum has observed improved efficiency in data capture, transaction management, researcher support, and standardised staff training. This presentation highlights practical lessons from implementing standardised, interoperability-focused workflows within a South African museum context and demonstrates how legacy collection data can be transformed into FAIR, mobilisation-ready biodiversity datasets.

Speaker



Siyabonga Zamisa

Research Data Technician | KwaZulu-Natal Museum

Digitize, Standardize, Mobilize: Wildlife Biodiversity Biobanking with Specify

9:12 AM – 9:25 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Biodiversity biobanks are increasingly recognized as critical research infrastructures supporting conservation, genomics, species management, and One Health initiatives through the long-term preservation of samples and associated metadata. While the physical preservation and curation of samples remains central to biobanking, the scientific value of these collections depends on the generation of high-quality, standardized, and interoperable data. As biodiversity research advances toward implementation of FAIR (Findable, Accessible, Interoperable, and Reusable) principles and the Extended Specimen concept, these biobanks face challenges in managing multidimensional datasets that extend beyond individual samples to include taxonomic, spatial, genetic, regulatory, and derivative information. Within the SANBI Wildlife Biobank, collections include diverse taxa and sample types such as blood, tissues, DNA extracts, and cell cultures linked to permits, storage records, images, molecular data, and metadata. Historically, these data were distributed across disconnected spreadsheets, paper records, folders, and external drives, creating challenges in data consistency, accessibility, and long-term sustainability. The implementation of Specify has transformed collection workflows from fragmented data management practices into an integrated relational framework capable of linking specimens, tissue collections, storage information, compliance documentation, and taxonomic data within a single system. The use of hierarchical taxonomy structures, controlled vocabularies, unique identifiers, and customizable workflows has improved data consistency, sample traceability, quality control processes, and transformed biodiversity biobanks from static repositories into dynamic digital resources.

Speaker



Kim Labuschagne

Wildlife Biobank Curator | South African National Biodiversity Institute

Integrating Specify in broader Science infrastructure

9:25 AM – 9:38 AM

COLLECTIONS DIGITISATION & MANAGEMENT

At the Royal Botanic Gardens Victoria, Specify is used to deliver high-quality data to the Australasian Virtual Herbarium (AVH), as well as managing the herbarium's holdings and transactions like loan and exchange. Specify is embedded in our broader science infrastructure and we have built various extensions and integrations.

We make use of the flexibility of Specify to handle complex workflows. We use a separate discipline, rather than the Borrows table, to manage loans from other institutions. This way we can link up images and DNA sequences we have taken to specimens data, while keeping the data separate from the National Herbarium of Victoria (MEL) collection. We also use "staging" collections for exchange data, so we can make sure we have the specimens and validate the data before it is incorporated into the main collection. There is also a dedicated collection for the Victorian Conservation Seed Bank, which shares most of its data with the main collection.

We have a "bystander" application that provides functionality that Specify does not have or where the Specify functionality does not quite fit our workflows, e.g., handling ZPL-based label-printing, storage assignment and loan returns.

The Specify taxon tree functions as the taxonomic point of truth for our Digital Asset Management System (DAMS). At the same time, we have applications to sync the Specify taxon tree with national and global taxonomies, including VicFlora, APC, AusMoss, AusFungi, Plants of the World Online (PoWo), Bryonames and Tropicos.

The AVH is our collection's portal in to the world. We use the Global Biodiversity Information Facility (GBIF) Integrated Publishing Toolkit (IPT) to publish collections data to AVH.

Speaker



Niels Klazenga

Biodiversity Informatics Developer | Royal Botanic Gardens Victoria

The Good, the Bad, and the Standardised: A Specify Journey Through 200 Years of the MfN Berlin's Biological Collections

9:38 AM – 9:51 AM

COLLECTIONS DIGITISATION & MANAGEMENT

The foundations of the Museum für Naturkunde Berlin (MfN)'s biological collections laid as early as 1810, houses today around 30 million specimens from the fields of palaeontology, zoology, botany and geology, as well as archives and modern media such as a sound library, reflecting 200 years of scientific, technical, colonial and political upheavals. Colonial collections (such as the Tendaguru excavation which returned 225 tons of dinosaur fossils), wars (the 1945 bombings destroyed a considerable number of specimens) and the Cold War isolation have left behind fragmented, multilingual and ethically complex data. Following reunification, resistance to digital transformation and the heterogeneity of legacy systems (paper labels, Excel, local databases), which persist to this day, hinder the standardisation of this complex yet essential biodiversity data.

Today, we are in the process of transitioning to Specify as our sole collection management system (CMS): this powerful tool enables us to standardise metadata (Darwin Core), manage the MfN's 22 collections, fill scientific gaps (missing labels, obsolete taxonomy) and manage permissions for over 100 users, with specific needs, whilst meeting modern requirements (GBIF exports, 3D scanning). We are also exploring ways to fill historical gaps by linking colonial archives to specimens (for example, the roles of workers at Tendaguru via Wikidata in the DTE project), using specific tables to reintegrate overlooked elements.

This presentation will share the lessons learnt from 200 years of upheaval: how Specify helps to preserve biodiversity data and reconstruct history, as well as the work that remains to be done to meet all these needs.

Speaker



Marion Depraetere

Data steward | Museum für Naturkunde Berlin

Meeting Biodiversity Standards with Specify 7

9:51 AM – 10:04 AM

COLLECTIONS DIGITISATION & MANAGEMENT

The Specify Collections Consortium (SCC) supports research collections around the world with Specify Software, an intuitive, robust, and highly-customizable software platform for digitizing holdings and managing collections. For over 25 years, Specify has supported biological research collections, biorepositories, and now geological research collections, with its open-source software for managing and publishing collections data.

Data sharing and data publishing is a fundamental requirement for collection management and research. Ensuring your collections data meets Biodiversity Standards means it will be more accessible to the scientific community and research. Over the last year, the SCC has focused our efforts on enabling members to more easily meet biodiversity standards during data entry and making shared data more accessible and interoperable.

In this presentation, SCC staff will discuss the soon to be integrated Darwin Core standards in the data model, making publishing to data aggregators such as the Global Biodiversity Information Facility (GBIF), the Atlas of Living Australia (ALA), the Global Genome Biodiversity Network (GGBN), and Integrated Digitized Biocollections (iDigBio) easier while maintaining the customizability and flexibility upon which Specify was built.

The SCC is a collaborative collections community governed, funded, and sustained by its member institutions. Our software and technical support priorities address member and system infrastructure needs. Specify software enables natural history museums to curate, integrate, and publish data from a single, robustly-engineered platform for biological and geological collections.

Speaker



Grant Fitzsimmons

Assistant Director | Specify Collections Consortium

From Data Entry to Distribution: Streamlining Biodiversity Workflows with Specify at CSIRO

10:04 AM – 10:17 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Since 2022, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) has migrated five of its collections into separate instances of Specify. This undertaking and subsequent support of the curation staff represents a challenge to the Data Management team in terms of ensuring data standards are met consistently for different specimen types across the National Research Collections of Australia (NRCA).

This talk will explain how we configured the Specify schema to facilitate clean data entry, including building standardised queries and 'workbench' bulk data entry templates across all collections. Our promotion and adherence to biodiversity standards such as Darwin Core enabled more efficient internal data discovery and the establishment of clear data ingest principles. This in turn allows for a more streamlined process for sharing data and the use of ETL tools to export the data into external aggregators such as the Atlas of Living Australia (ALA). It also facilitates the use of API-based integration to connect to extended specimen metadata repositories, increasing the reach and interoperability of our data.

We will also cover our strategy of maintaining close relationships with the curation staff and users of Specify to further improve processes, identify data quality issues, and employ the batch edit features to help provide the most accurate specimen data to researchers and the general public alike.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Dan Baker

Data Manager | CSIRO

Authors

Zoe Warner, Corinna Paeper

Updates on Data Mobilization in the KU Entomology Collection

10:17 AM – 10:30 AM

COLLECTIONS DIGITISATION & MANAGEMENT

The [KU Entomology Collection](#) (KU SEMC), in Kansas, USA, houses around 5 million insect specimens collected worldwide. Since the early 2000s, KU SEMC has been actively managed and digitized using various versions of the Specify Software, becoming one of the largest entomological datasets available today in the [Global Biodiversity Information Facility](#) (GBIF). Recently, KU SEMC management was migrated to Specify 7, enabling new tools for online data management and mobilization. During the migration process, data compliance with international standards (e.g. Darwin Core, Multimedia Core) was evaluated and updated to facilitate data and media mobilization to external biodiversity platforms. In the last year, enhanced mobilization protocols have been implemented, achieving a more effective workflow to share FAIR (e.g. Findable, Accessible, Interoperable, Reusable) specimen data and images to [regional](#) and [international](#) aggregators. The ability to mobilize these enriched records more efficiently represents a valuable resource to advance research projects at KU SEMC, while also increasing opportunities to continue contributing with biodiversity science at a global scale.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Samanta Orellana

Postdoctoral Researcher | University of Kansas Biodiversity Institute

Authors

Victor Gonzalez, Rachel Osborn, Zachary Falin, Nico Franz

SYM38 - Sustaining the Geo- and Biodiversity Data Ecosystem: Standards-Driven Approaches to Long-Term Resilience

8:30 AM – 10:30 AM | Location: FORUM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Over the past several decades, geo- and biodiversity sciences have been reshaped by digitization, data mobilization, and data interlinking efforts from which arose the community-developed Digital Extended Specimen concept, a rallying framework akin to transformative large-scale scientific innovations like the Human Genome Project. These initiatives have delivered substantial physical and digital infrastructure, new tools and workflows, and an increasingly connected global community.

As funding cycles end, challenges emerge: sustaining infrastructure, maintaining software and data quality, retaining expertise, coordinating governance across fragmented systems, and supporting workforce development. Machine-actionable, FAIR-compliant data require sustaining the entire ecosystem—data providers, aggregators, platforms, repositories, users, and underlying standards.

This session frames geo- and biodiversity data infrastructure as a distributed socio-technical system where standards enable coordination, reduce duplication, and support research while shaping infrastructure needs. The presentations and discussion in this symposium will explore how coordinated approaches combined with research- and robot-ready data standards can align social, physical, and digital infrastructures across domains, leverage economies of scale, strengthen governance and collaboration, and build relevance beyond geo- and biodiversity sciences.

Moderators



Libby Ellwood

Director | iDigBio



James A. Macklin

Agriculture and Agri-Food Canada

Session organizers



José Fortes

Professor | University of Florida



Ely Wallis

Engagement Team Lead | Atlas of Living Australia, CSIRO

Authors

Kerstin Lehnert, Jutta Buschbom, Andrew Bentley, Pamela Soltis

Sustainability and collaboration in a time of uncertainty

8:30 AM – 8:35 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Over the past several decades, geo- and biodiversity sciences have been reshaped by digitization, data mobilization, and data integration efforts, giving rise to the community-developed Digital Extended Specimen as a unifying framework. These initiatives have produced substantial infrastructure, new tools and workflows, and a more connected global community. However, as funding cycles conclude, urgent challenges remain, including sustaining social, physical, and digital infrastructure; maintaining and aligning software and data quality work across networks; retaining expertise; coordinating governance; and supporting workforce development.

This presentation frames geo- and biodiversity data infrastructure as a distributed socio-technical system in which semantically structured and linked data, standards development, and data harmonization enable coordination, reduce duplication, and support research. It emphasizes that achieving machine-actionable, FAIR- and CARE-compliant data depends on sustaining a standards-based ecosystem of data providers, aggregators, platforms, repositories, and users.

Drawing on current efforts within the networks of the International Partners for the Digital Extended Specimen and the Biodiversity Collections Network, along with the broader community, this talk explores how coordinated approaches and research- and machine-ready standards can better align social, physical, and digital infrastructures needed to support community goals. It highlights opportunities to leverage economies of scale, strengthen governance and collaboration, and expand the relevance of biodiversity data across domains, offering a forward-looking perspective on building a more sustainable and integrated data ecosystem.

Speaker



Libby Ellwood

Director | iDigBio

Authors

Andrea Thomer, Kerstin Lehnert, Jutta Buschbom, Andrew Bentley, José Fortes, James A. Macklin, Brooke Long-Fox, Deborah Paul, Dori Contreras, Dimitris Koureas

GBIF after 25 years

8:35 AM – 8:47 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

This year GBIF celebrates its first 25 years. GBIF has grown from a big idea to a highly productive global network and biodiversity infrastructure. GBIF originated from the natural history museum community as an infrastructure to share data from specimens. GBIF now mediates data from over 285 million specimens and GRSciColl contains information from nearly 11,000 collections around the world. This success is due to the global GBIF network of over 2,600 data publishers and national nodes. The integration of data from different sources allows connection to related data with images, DNA sequences and even phylogenies. This data sharing allows reuse for many scientific applications beyond taxonomy and systematics such as conservation, invasive species research and climate change. At GBIF we track data usage and confirm the use of collection data in strategic documents that inform policy such as the IPBES assessment on invasive species. GBIF users link their GBIF-enabled research to achieving UN Sustainable Development Goals. Perhaps most importantly the easy access to shared digitized data is immediately available to national reporting and the CBD Global Biodiversity Framework. GBIF has succeeded by providing needed support for the nations that require. GBIF sustainability has benefited by its unique governance structure and community engagement. This collaboration will be even more important as move forward together for the next 25 years.

Speaker



Joe Miller

Executive Secretary | GBIF

Unifying Natural Heritage for a Resilient Future

8:47 AM – 8:59 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

In an era defined by rapid environmental shifts and global instability, the mobilization of biodiversity data is no longer a luxury—it is a necessity for informed survival. DiSSCo (Distributed System of Scientific Collections) represents a transformative response to this uncertainty, transitioning fragmented physical collections with specimens into a synchronized, FAIR-compliant digital ecosystem built on the Digital Extended Specimen concept.

This presentation explores how DiSSCo fosters collaboration by unifying over 1.5 billion biological and geological objects across hundreds of institutions. At the heart of this infrastructure is the open Digital Specimen, using a FAIR Digital Object compliant architecture that enables researchers and collection curators to link, analyze, and enrich collection data with the help of machines at an unprecedented scale.

By harmonizing institutional efforts, DiSSCo ensures the long-term sustainability of scientific assets, providing the evidentiary base required to tackle climate change, biodiversity loss, and emerging health risks. We will discuss the technical and organizational frameworks that allow DiSSCo to bridge the gap between isolated silos and a shared knowledge base. Ultimately, DiSSCo demonstrates that through integrated digital infrastructure and collaborative governance, natural heritage can pivot from a state of fragmented, fragile silos to a future of data-driven resilience.

Speaker



Wouter Addink

Manager Software Development | Naturalis Biodiversity Center

Requirements of Sustainability Models for Geo-Biodiversity Data Cyberinfrastructure

8:59 AM – 9:11 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

The sustainability of cyberinfrastructure (CI) for Geo-Biodiversity (GB) research depends on the existence of a sustainable GB community of entities (aka ecosystem) which includes data providers, data aggregators, users, professional organizations and others. Conversely, a healthy progressive GB ecosystem also needs a sustainable CI for its own sustainability. This interdependence originates significant CI requirements which fall into three categories, namely, functional requirements, operational requirements and economic-model requirements. Functional requirements relate to the research activities that the CI must support such as biodiversity data management, data analytics, FAIR practices, educational tools and workforce training. Operational requirements follow from the need for efficiency, scalability, security and resiliency of CI services. Economic-model requirements relate to tracking usage of resources (including data) and associated costs, generation and allocation of revenue, and accounting to funding sources and stakeholders.

This presentation reviews the possible funding models for scientific CI in the context of the GB ecosystem and discusses the CI features that they require, elaborating on the nature of these requirements and possible solutions. These requirements need to be addressed in the context of existing software, design patterns and data standards for economic activities in cyberspace. While paid commercial solutions are provided, open-source solutions that can be adopted and integrated with relevant standards are also possible. This presentation will also provide an example of how a prototype agentic system for biodiversity data exploration, integration and analytics, called iChatBio, could be extended to address some of these requirements.

Speaker



José Fortes

Professor | University of Florida

From preservation to perseverance: bridging disciplines for data resilience

9:11 AM – 9:23 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Since 2025, major federal funding cuts in the US have put multiple scientific data infrastructures at risk, including many in the Earth and environmental sciences. For example, multiple data resources managed by NOAA, USGS and other federal agencies have been taken offline, typically without plans for replacement or restoration. Many NASA science programs have been cut, and multiple NSF programs for research and infrastructure in the Earth sciences are canceled or rumored to be canceled. The funding programs that remain are increasingly skewed toward AI and partnerships with industry. Earth science data infrastructures are having to quickly pivot to alternative funding models, and are discussing worst case scenarios in case of closure.

The Earth and biodiversity sciences have much to gain from coordination and partnership in this increasingly uncertain funding landscape. In this talk, I'll share sustainability efforts taken by data infrastructures in the Earth and environmental sciences, including the System for Earth and Extraterrestrial Sample Registration (SESAR), which include diversifying our funding streams and refactoring our codebase and vocabularies. I'll also discuss broader resilience efforts in the US-based Earth science community, including nascent community-building initiatives from the American Geophysical Union. I'll conclude by reflecting on existing digital preservation insights from scholars in the library and information sciences, and how these present us with important guidance as we seek to persevere through times of uncertainty and resource scarcity.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Andrea Thomer

University of Arizona, College of Information Science

Authors

Kerstin Lehnert, Natalie Raia

NEON, Community Standards, and the Hyper-Extended Specimen

9:23 AM – 9:35 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Each year, the National Ecological Observatory Network (NEON) collects over 400 terabytes of ecological data and more than 100,000 spatially and temporally co-located geochemical samples and biological specimens across the United States. NEON monitors whole ecosystems through repeated field sampling, continuous sensor measurements, remote sensing, and downstream analyses, creating a unique opportunity to support "hyper-extended specimens" while minimizing common data challenges related to standardization.

NEON primarily maintains and publishes its data using internal data models, vocabularies and identifiers. However, many of NEON data conventions align with community standards, such as GBIF vocabularies, Darwin Core, IGSN IDs, Veg-X, MlxS, and USA-NPN phenophase definitions. Selected data are formally translated and republished through standards-based infrastructures such as GBIF, BOLD, AmeriFlux, and Google Earth Engine. In addition, we have explored how ecological monitoring datasets can be represented using emerging standards such as the Humboldt Extension and Darwin Core Data Package. We are also working with broader community efforts to examine and improve the AI readiness of our datasets to increase interoperability and relevance for next-generation analytical tools.

We outline the current state of the Digital Extended Specimen concept within the NEON ecosystem, including ongoing efforts and challenges to connect our physical collections, digital infrastructure, and diverse downstream research communities, while improving provenance tracking and discoverability.

Speaker



Chandra Earl

National Ecological Observatory Network (NEON)

Authors

Edward Gilbert, Kelsey Yule

Canada's need for synthesis for Adaptation and Resilience to Climate Change

9:35 AM – 9:47 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

The exponential growth of genomic data presents an unprecedented opportunity to mitigate the ecological impacts of climate change. Genomic data is estimated at 30 petabytes available publicly, with likely at least half as much unpublished. The Genomic Research and Development Initiative's (GRDI) Genomic Adaptation and Resilience to Climate Change (GenARCC) will amass nearly a petabyte of data. Consequently, we may already have the raw genomic data required to address some of the most pressing ecological problems that can be solved using genomics.

To fully realize this potential, a paradigm shift towards systematic reuse of data and synthesis is essential. Synthesis centres have previously demonstrated the value of reanalyzing existing assets to accelerate discovery. This requires realigning incentives to prioritize data reuse over data generation such as by recognizing algorithms, associated pipelines and applications, and the data itself like long term scientific infrastructure. This involves standardization of data, processes, and products, which can only be realized by the community's will and/or incentivized through funding-related obligations. With improved access and curation of shared scientific resources, researchers can develop novel data management architectures and leverage artificial intelligence, to integrate diverse datasets and reveal cross-system patterns of resilience that isolated studies overlook. To achieve this goal, future efforts should work towards FAIR data, and validation, maintenance, and automation of code and underlying algorithms. Together, these efforts can enhance how scientific data is shared and utilized in conservation and ecology. Prioritizing synthesis will ensure we get the highest return on our collective investment.

Speaker



James Macklin

Research Scientist, Botany and Biodiversity Informatics | Agriculture and Agri-Food Canada

Authors

Ian Bradbury, Tony Kess, Etienne Low-Decarie

Architecting the National Biodiversity Information System: Standards-Driven Workflows for South Africa's Botanical and Zoological Data

9:47 AM – 9:59 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

To sustain a robust Geo-Biodiversity Data Ecosystem, SANBI has implemented standardised data management strategies for South Africa's national botanical and zoological collections. The project focuses on transitioning legacy data into FAIR (Findable, Accessible, Interoperable, and Reusable) datasets using globally recognised informatics standards. Central to this architecture is the National Biodiversity Information System (NBIS), which integrates data from institutional partners using standardised platforms.

A key challenge in national data integration is metadata harmonisation, particularly taxonomic inconsistencies. To address this, SANBI uses the TaxaMatch algorithm, which matches misspelt or variant species names to an authoritative taxonomic backbone. This ensures high-quality, interoperable data regardless of differing institutional interpretations. For botanical data, SANBI distributes authoritative dictionaries, including taxa with authors and gazetteer data, to improve quality at capture.

Institutions push new or updated data to staging instances where automated quality checks are applied before records move to production. ETL packages synchronise data regularly. For zoological collections, Specify's taxonomic ranking system accommodates diverse research traditions while maintaining referential integrity.

The system supports spatial searches, enabling users to draw polygons on maps to retrieve species coordinates. SANBI also uses a separate NBIS indexing database with persistent identifiers linked across repositories. By providing shared infrastructure and standards, SANBI enables a scalable web portal that promotes data cleaning, reduces duplication, and supports environmental monitoring and research.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Brenda Daly

Deputy Director: Biodiversity Information Management | South African National Biodiversity Institute

Authors

Fhatani Ranwashe

Discussion

10:00 AM – 10:30 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

WKS14 - New approaches and tools to maximise the value of the Biodiversity Heritage Library

8:30 AM – 10:30 AM | Location: BALDER

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

This workshop aims to be a first step in rethinking how we can interact with the Biodiversity Heritage Library (BHL), and how we extract information to build interfaces and datasets appropriate for a user landscape dominated by mobile devices and new AI tools. It will be

run as a series of live demos of ideas for interfaces and information extraction tools. Participants will be invited to explore the demos and provide feedback on them. We will also encourage participants to share their views on what BHL could do to increase its value to their research.

Topics would include:

- Taxonomic names
 - tree-based taxonomic browsing using classifications such as the Catalogue of Life
 - Google n-gram like tools tracking changes in taxonomic names over time
- Geography
 - map-based interfaces to explore BHL content geographically
 - finding images of maps in BHL content
- Images
 - image extraction and classification tools using LLMs
 - image search using embeddings (find images that look like this)
- Entities
 - extracting entities from text, such as taxonomic names, people, places, and museum specimens
- AI-friendly interfaces
 - model content protocol (MCP) interfaces for AI tools to query BHL using natural language

We think that the chance for people to explore new tools, and discuss those with active members of BHL will be beneficial for BHL as it restructures and rethinks its platform, and to the wider biodiversity data community for whom BHL is a valuable, if underutilised, resource.

Workshop leaders



Roderic Page



Nicole Kearney

Communications Director | Biodiversity Heritage Library

BREAK

10:30 AM – 11:00 AM | Location: 1st Floor

SYM29B - From Cabinets to Clouds – Bridging the gap between collection- and database management

11:00 AM – 12:30 PM | Location: SAL A

COLLECTIONS DIGITISATION & MANAGEMENT

Physical collections and digital databases are often managed in parallel by collection managers and data managers. As dedicated development teams become more common in institutions, opportunities for misalignment between those working in the collection cabinets and those in the data clouds arise.

Across institutions, this relationship is handled in different ways. A wide range of approaches for coordinating collection digitisation pipelines exist, reflecting differences in institutional structures and ongoing integration of new development roles into long-established collection management practices and workflows. In some cases, workflows are closely aligned across teams, while others are more loosely defined. These differences offer an opportunity to learn how institutions balance collection management expertise with emerging digital practices.

We aim to bring together a range of examples to learn from each other's institutional workflows to implement biodiversity data standards and contribute to global biodiversity knowledge. Topics could include experiences and best practices in facilitating functional relationships between collection and data managers; decision-making processes for data standards; having one expert in two domains vs. two experts; enric

Moderators



Caitlin Thorn

Data Manager | Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science



Nora Lentge-Maaß

Head of Coordination Office for Digitization on Demand, Deputy Head of Department Services for Collection Future | Museum für...

Session organizers



Frederik Berger

Museum für Naturkunde Berlin



Mareike Petersen

Lead Science Programme Collection Future | Museum für Naturkunde Berlin



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

Introduction

11:00 AM – 11:10 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Speakers



Caitlin Thorn

Data Manager | Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

Drawers in the dark harboring data that could spark - finding the sweet spot in collection digitization and mobilizing metadata from museum's collections

11:10 AM – 11:25 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Natural history museum's collections have a long history in collecting, storing and giving access to an – usually – science-based community. But in today's technologically connected society, there is expectation for access to data instantaneously. Traditional museums are taking up the journey to open up and make their collections publicly available to all. With this in mind, the Museum für Naturkunde Berlin started an action Plan in 2020 with the goal of digitizing all of the approximately 30 million collection objects in 10 years' time. Since then, collection-, digitization-, media- and database managers work together and explore different formats and scales of project workflows. Here we examine how museum collections can move from slow and fragmented efforts to a scalable and integrated strategy. We argue why aiming for perfection in every digital record slows down the digital transformation. This tradeoff between slowly acquired high quality, "gold standard" digital assets and a straight forward interoperable dataset with FAIR data needs to be managed. In the end, museums must decide when perfection slows the trail, and when "good enough" leads to collections that spark and knowledge that's shines at scale.

Speaker



Nora Lentge-Maaß

Head of Coordination Office for Digitization on Demand, Deputy Head of Department Services for Collection Future | Museum fü...

Accelerating specimen identification through Virtual Collections while promoting collaboration

11:25 AM – 11:40 AM

COLLECTIONS DIGITISATION & MANAGEMENT

Natural science collections are essential benchmarks for biodiversity research. Accelerating digitization highlights the need for taxonomic identification at scale, which Virtual Reference Collections (VRCs) provide by gathering expert-identified digital specimens. Funded by the TETTRIs project, we implemented VRCs into the DiSSCo infrastructure. By utilizing Persistent Identifiers (PIDs), VRCs ensure a stable reference to these digital objects, enabling reliable data citation and global collaboration. Consequently, users can seamlessly consult data and taxonomic experts can identify specimens remotely.

However, implementing effective VRCs requires addressing critical data challenges. Foremost, specimens must be expertly curated to ensure accuracy. Data must not remain siloed across separate portals; a VRC requires a single, user-friendly entry point. High-quality imagery is mandatory since standard digitization often misses diagnostic traits. Furthermore, these collections must represent complete taxonomic groups and meet machine-readable FAIR standards, including complete event dates and unique identifiers.

Supported by TETTRIs, DiSSCo addresses these hurdles by leveraging Digital Specimen architecture. Virtual Collections harmonize federated data into unified reference sets. Using PIDs and standardized schemas, experts aggregate specimens into citable, stable sets. This facilitates "curation-at-distance," allowing taxonomists to remotely validate and enrich metadata, bridging the gap between physical repositories and global expertise. In this session, we present DiSSCover's new capabilities for creating and maintaining VRCs to solve these identified issues.

Speaker



Melanie de Leeuw

Software developer | Naturalis Biodiversity Center

Managing high-volume specimen movements in the entomological collection of Naturalis with a new type of record

11:40 AM – 11:55 AM

COLLECTIONS DIGITISATION & MANAGEMENT

The entomological, pinned collection at Naturalis contains approximately 1.7M specimens, of which about 10% are digitized. Collection managers mostly use a higher level of data (the storage units) to retrieve and manage the collection. The physical storage units contain multiple taxa, which is reflected in the storage unit record as multiple identifications. New acquisitions and taxonomic reshuffling would cause collection managers to have to move the specimens of several taxa to adjacent storage units, and to remove an identification and reenter the data in the destination storage unit record. Given the high curation activity in entomology, this created a critical workload bottleneck with high risk for errors. At first, investigations were done to make identifications a separate subrecord, to allow moving between two storage units more easily. This was deemed too disruptive to the database architecture. Instead a new type of record was developed: the lot, a conceptual bridging entity. The lot record reflects an aggregation of specimens belonging to the same lowest-determined taxon stored within a contiguous physical location. It can be connected to a specific storage unit, optimizing digital workload following physical actions. Also, in the future it will be easier to publish this data at short intervals (as opposed to the static dataset of storage units which is now 8 years old). Ultimately, this implementation satisfies the most frequent inquiries from both external and internal users: confirming whether a specific taxon is present in the Naturalis collection, and pinpointing exactly where it is stored.

Speaker



Myriam van Walsum

3D scanning specialist for heritage objects | University of South Florida - CVAST

Authors

Oscar Vorst, Ester Judith Slaa

Everything is coded: How to prevent a game of telephone in a system full of interfaces

11:55 AM – 12:10 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Building a basic data infrastructure for a digital collection is not a particularly cutting-edge project, although, of course, there are many design decisions to be made that have a major impact on the system's performance, interoperability, and resilience. However, such projects typically do not start from scratch but evolve within socio-technical environments whose subsystems are characterized by huge differences, particularly with regard to digital literacy as well as in terms of understanding scientific methods and the significance of research data. In our case, this is particularly relevant for the involvement of natural and social scientists, collection managers, project coordinators, technical developers, and data users in exhibition and education.

Designing cross-disciplinary processes for the generation, transformation, and publication of specimen-related data faces various challenges under these circumstances. For example, interfaces are needed throughout the entire data pipeline where translations must take place: from unstructured to structured forms, from human-readable to machine-readable, from one professional group's specialized terminology to another's, and from strategic thinking to actual agile development work.

In our presentation, we'll use specific examples to demonstrate how successfully – or unsuccessfully – we manage to transfer data from one expert to another, while on the one hand preserving the essential integrity and significance of the information across all transformation steps, and on the other hand building trust among stakeholders for whom parts of the infrastructure are inevitably a black box.

Speaker



Franziska Schuster

Co-head of dept. Infrastructure and Development | Museum für Naturkunde Berlin

Authors

Jasper Funk-Smit

Discussion

12:10 PM – 12:30 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Moderators



Caitlin Thorn

Data Manager | Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science



Sabine von Mering

Data scientist | Museum für Naturkunde Berlin

SYM46 - AI-Readiness Metrics and Metadata for Biodiversity Data

11:00 AM – 12:30 PM | Location: SAL B

AI AND ROBOT READY

Artificial intelligence is transforming biodiversity research, yet many datasets remain hard for automated systems to interpret, validate, and integrate. "Robot-ready" (AI-ready) data requires more than FAIR: structured, machine-actionable metadata, clear semantics, provenance, quality metrics, and reproducible workflows.

This session will examine emerging approaches for defining, measuring, and operationalizing AI-readiness in biodiversity data infrastructures. We invite contributions addressing: (1) quantitative AI-readiness metrics and validation frameworks; (2) extensions or refinements to Darwin Core and related TDWG standards to support machine reasoning; (3) automated metadata generation and enrichment; (4) integration of image-, trait-, or sensor-derived data into standards-compliant records; (5) provenance capture and workflow reproducibility; (6) semantic alignment and ontology-driven interoperability; and (7) benchmarking and evaluation of datasets for AI applications.

We especially welcome case studies from repositories, aggregators, and infrastructures implementing automated quality checks, semantic annotation, or containerized workflows, including cross-domain links to climate or geoscience data.

The goal of this session is to identify concrete standards gaps, promote interoperable solutions, and advance a shared community framework for AI-ready biodiversity data aligned with TDWG principles.

Moderator



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute

Introduction

11:00 AM – 11:03 AM

AI AND ROBOT READY

Speaker



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute

Toward Quantitative AI-Readiness Metrics for Biodiversity Data Infrastructures

11:03 AM – 11:15 AM

AI AND ROBOT READY

As AI becomes increasingly integrated into biodiversity research, the need for biodiversity data that are not only FAIR but also machine-actionable, interoperable, and reproducible has become increasingly apparent. While biodiversity repositories continue to expand access to specimen, image, and trait data, there remains no widely adopted framework for evaluating how effectively these datasets support AI-driven workflows. Current biodiversity standards primarily emphasize data sharing and interoperability for human use, but often lack mechanisms for assessing suitability for machine reasoning, computer vision, automated metadata generation, or large language model integration.

This presentation examines the emerging concept of "AI-readiness" in biodiversity informatics and proposes a framework for developing quantitative AI-readiness metrics for biodiversity data infrastructures. Potential dimensions include metadata completeness, semantic consistency, provenance transparency, image processability, workflow reproducibility, interoperability of derived data products, and machine-actionable linkages among specimens, media, and traits. The talk will explore how AI-readiness evaluation may support automated quality assessment, benchmarking across repositories, and reproducible AI workflows.

Using examples from biodiversity imaging and AI-assisted collection workflows, this presentation discusses how quantitative evaluation frameworks may help identify standards gaps and support the development of interoperable and provenance-aware biodiversity infrastructures aligned with emerging AI applications. Particular attention will be given to the role of TDWG standards and community coordination in defining shared metrics and best practices for AI-ready biodiversity ecosystems.

Speaker



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute

Traits Thesaurus: toward semantic harmonization of FAIR trait-based data on aquatic organisms

11:15 AM – 11:27 AM

AI AND ROBOT READY

Trait-based approaches are increasingly important in biodiversity and ecological research, particularly for assessing community and ecosystem responses to biodiversity loss and environmental change. Functional traits provide essential information linking species and community characteristics to ecosystem functioning across spatial and temporal scales. However, the growing use of trait-based methods across ecological domains has highlighted the lack of harmonised trait definitions, terminologies, and measurement units. This semantic heterogeneity limits interoperability, data integration, and reuse of trait information across studies, taxa, and ecosystems, representing a major barrier to large-scale biodiversity synthesis.

To address these limitations, we developed the Traits Thesaurus, a controlled vocabulary containing more than 1,500 standardised trait concepts for aquatic organisms. The thesaurus unifies and significantly expands five existing vocabularies dedicated to Phytoplankton, Macroalgae, Zooplankton, Macrozoobenthos, and Fish traits within a single interoperable framework. Its development followed a collaborative, community-endorsed development and validation process involving more than 40 experts from different ecological and biodiversity research domains. The resource provides stable URIs, clear definitions, preferred and alternative labels, and semantic mappings to external vocabularies and ontologies. Designed according to FAIR and Linked Open Data principles, the Traits Thesaurus supports semantic annotation, interoperability, and integration across biodiversity information systems and ecological disciplines, facilitating large-scale data mobilisation and advancing trait-based biodiversity and ecosystem research.

Speaker



Jessica Titocci

Research Technologist | National Research Council of Italy (CNR), Research Institute on Terrestrial Ecosystems (IRET)

Authors

Martina Pulieri, Alberto Basset, Ilaria Rosati, Senem Onen Tarantini, Angela Carluccio

AI Ready Standards with Croissant for Type Specimens Catalog Datasets

11:27 AM – 11:39 AM

AI AND ROBOT READY

The FAIR principles are fundamental to Open Science. Publishing datasets used in Artificial Intelligence (AI) experiments in accordance with these principles is essential for ensuring reproducibility. Additionally, these FAIR datasets enhance the performance of Machine Learning (ML) applications, e.g., species labeling using named entity recognition models. Recently, [ML Commons](#) introduced Croissant, an ML-ready dataset format designed for seamless integration with ML models and with ML-ready dataset-sharing platforms such as Hugging Face.

Biodiversity data mobilization necessitates adherence to FAIR principles that go beyond merely sharing data in structured formats. We present an experimental approach to converting unstructured text from entomological type specimen catalogs into [Darwin Core](#) (DwC) metadata using Large Language Models, including GPT and Qwen. Subsequently, we map Croissant metadata fields to DwC terms—such as *dwc:genus*, *dwc:scientificName*, and *dwc:individualCount*—because these terms can be used to make type specimen catalogs machine-readable.

To evaluate whether our datasets align with FAIR principles for ML, we generate and conduct experiments on subsets of them. These metadata are then integrated into the [NFDI4DataScience metadata portal](#), where FAIR metrics are visualized using the [Piveau](#) metric service. This portal provides metadata on datasets for AI use and metrics on FAIR principles. Complementing this quantitative assessment, we validate our metadata format using the [FAIR Checker](#) service.

Our proposed metadata framework for type specimen catalogs serves as a reference model for broader applications within biodiversity research domains. In addition, we provide a standardized protocol for the open mobilization of entomological type specimen data.

Speaker



Sefika Efeoglu

Research Scientist | Senckeneberg Nature Research

Authors

Jana Hoffmann

Unifying Biodiversity Images: Content-Based Identification with the ISCC standard (ISO 24138:2024)

11:39 AM – 11:51 AM

AI AND ROBOT READY

Digital images are vital to modern biodiversity science for e.g. studying specimens, identifying species, and extracting data by AI. However, sharing imagery faces critical challenges: image duplication across repositories, broken data links during reformatting, and AI-generated fakes undermining trust. Traditional PIDs (like DOIs) manage location but ignore content, while standard hashes (SHA-256) break with any pixel modification.

This presentation introduces the International Standard Content Code (ISCC, ISO 24138:2024) as a content-aware verification layer. The ISCC-CODE functions as a digital fingerprint derived from visual media, enabling similarity-based matching so that cropped, compressed, or converted images retain highly similar codes.

Because the code only describes the media itself, the ISCC Discovery Protocol (IDP) is proposed by T. Pan & M. Etzrodt (ISCC Foundation) to enable content-based discovery and provenance. The core of IDP is the Declaration Event - a FAIR Digital Object (FDO) identifying who declared what content at what time via a persistent ISCC-ID. Reference implementation: <https://github.com/iscc/iscc-hub>.

The IDP solves three critical problems:

- Automated Deduplication: Content based clustering of identical or near-identical images across infrastructures without relying on inconsistent metadata.
- Annotation Resilience: Keeping downstream annotations anchored to an image even when it is manipulated across workflows.
- AI Fake Detection & Provenance: Using cryptographically signed Declaration Events to verify the origin and authenticity of images, isolating synthetic fakes.

A content-derived fingerprint implementation like IDP offers a scalable, decentralized path to securing trust and interoperability in biodiversity imagery.

Speaker



Wouter Addink

Manager Software Development | Naturalis Biodiversity Center

Provenance, Lineage, and Auditability in AI-Driven Biodiversity Image Workflows

11:51 AM – 12:03 PM

AI AND ROBOT READY

As the role of artificial intelligence in biodiversity research becomes increasingly significant, the traceability of AI-processed image data has become critical to ensuring research reproducibility, data reusability, and workflow auditability. However, in biodiversity image collections, persistent identifiers are still primarily limited to original specimens. Derived images generated through AI pipelines lack systematic provenance tracking, and their chain of lineage is broken once the data leaves the repository. This talk presents FishAIR's provenance model as a case study of how a production biodiversity image system extends persistent identifiers beyond original specimens to AI-derived images, maintaining an unbroken chain from digitized collection images through bounding box, segmentation, and landmark processing. At the core of this model is the treatment of processing activities as citable first-class objects: each "batch" representing a set of image processing runs is treated as a traceable workflow event. It is assigned a unique ARK identifier recording the responsible institution, pipeline type, software configuration, and parameters. This design ensures that provenance information is not merely optional metadata but an inherent property of every derived image, without relying on external workflow management infrastructure. By maintaining this chain, FishAIR enables data users to trace the origin and transformation history of any training image and identify which processing stage or source collections those images originated from when model performance varies across different data subsets. The framework further supports transparent and auditable AI biodiversity workflows by preserving lineage across all stages of image derivation and reuse.

Speaker



Xiaojun Wang

Tulane University, Biodiversity Research Institute

Authors

Yasin Bakış, Bahadır Altıntaş, Brian Sidlauskas, Henry L. Bart

Some approaches to robot-ready biodiversity data work backwards: collect first, then retrofit standards-compliant metadata before submission. The Ecological Monitoring System Australia (EMSA) takes the opposite approach. Its Monitor platform captures field data across 38 ecological survey protocols, and the export pipeline emits standards-aligned RDF at source — so submissions arrive at the national Biodiversity Data Repository (BDR) robot-ready by construction, not by remediation.

The pipeline resolves every attribute, property, and observation-procedure URI live against EMSA vocabularies in a GraphDB SPARQL endpoint, with caching and vocabulary harmonisation mapping field terms to controlled vocabularies. Data quality is enforced at both boundaries: JSON Schema validation governs what enters the system from the field, and per-protocol SHACL shapes gate what leaves — submissions that fail validation are not exported. Provenance is tracked end-to-end from field collection through to BDR ingest, and a companion XLSX gives curators a human-readable view of what the machine submitted. The full pipeline is config-driven, containerised, and reproducible.

We discuss what works well — SPARQL resolution keeps the pipeline aligned as vocabularies evolve, and SHACL gating catches structural errors before they reach BDR curators — and the practical lessons of maintaining a config-driven exporter across 38 protocols as both the monitoring program and its target vocabularies continue to grow.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Andrew Tokmakoff

TERN Ecosystem Surveillance

Authors

Walid al Naim, Junrong Yu, Siddeswara Guru, Luke Derby, Ben Sparrow

Artificial Intelligence Amplifies Paleontology's Reproducibility Problem

12:15 PM – 12:27 PM

AI AND ROBOT READY

Paleontology has become a data-intensive science. Along with the specimen, large amounts of associated data are produced. These data are distributed across supplementary files, institutional databases, and discipline-specific repositories, and reproducibility problems are common: missing datasets, character matrices that fail to execute, incomplete documentation, unverifiable specimen vouchers, and missing methodological detail. Many of these problems predate AI, but AI methods are now being applied to paleontological data faster than existing infrastructure was designed to support, particularly LLMs that extract structured data from legacy literature and vision models that classify specimens from images. AI offers reproducibility benefits, including standardization at scale and recovery of data from inaccessible literature. However, it also introduces a specific risk: it can scale and conceal existing data quality problems by converting incomplete, ambiguous, or unverifiable source material into structured outputs that meet FAIR criteria and appear authoritative. A dataset can pass automated validation, receive a persistent identifier, and be reused in computational workflows while lacking the provenance, expert validation, and traceable interpretation required to assess its reliability. AI-readiness in paleontology must therefore be defined by verifiability, including provenance to source evidence, executability of analyses, and traceability of expert judgment. Journals should require disclosure of AI involvement in dataset construction and extend peer review to AI-generated data; repositories should treat AI provenance, validation status, and source linkage as first-class metadata.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Brooke Long-Fox

MorphoBank Data Curation Scientist | Phoenix Bioinformatics

Authors

Mark Young

Closing remarks

12:27 PM – 12:30 PM

AI AND ROBOT READY

Speaker



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute

SYM25A - From Mobilizing Data to AI-Ready Knowledge: Infrastructure for Multimodal Biodiversity Data

11:00 AM – 12:30 PM | Location: SAL C

AI AND ROBOT READY

More ecological and biodiversity data are online than ever before, in part thanks to the enormous decades-long efforts by the biodiversity community in mobilizing observational data, supporting shared standards, and building aggregation networks. Publicly accessible data are also more varied than ever before, including images of many modalities, videos, and streaming sensor data. However, using this sea of data for increasingly ML-heavy and AI-driven research pipelines remains fraught. Significant ecological context, locality detail, trait information, and interaction data are buried in unstructured media. Records are often incomplete, unevenly georeferenced, and documented inconsistently across platforms. Metadata regarding terms of use, provider expectations, and desired round-tripping of data amendments are unclear or missing. Without downloading massive amounts, assessing the quality and fitness-for-use of data is often difficult or impossible. As AI-enabled and agent-driven methods mature, developing shared, community-aligned infrastructure and practices that transform multimodal biodiversity data into interoperable, provenance-aware, and research-ready knowledge becomes a central challenge. Contributions to this session will explore how standards, human–AI collaboration, uncertainty representation, ontology integration, and data governance frameworks can enable AI-orchestrated pipelines to understand fitness-for-use, respect provider expectations, and track provenance.

Moderators



Hilmar Lapp

Neuromatch, Inc.



Robert Guralnick

Professor and Curator | University of Florida

Session organizers



Eric Sokol

Quantitative Ecologist | National Ecological Observatory Network (NEON), Battelle



David Bloom

University of North Carolina at Greensboro

Authors

Kelsey Huelsman, Eric Sokol, David Bloom

Introduction

11:00 AM – 11:05 AM

AI AND ROBOT READY

Speaker



Hilmar Lapp
Neuromatch, Inc.

A metadata-first FAIR4AI checklist and automated evaluation agent for biodiversity datasets

11:05 AM – 11:20 AM

AI AND ROBOT READY

As AI methods become increasingly integral to ecology and biodiversity science, the data that are used must meet standards beyond traditional FAIR principles. In a cross-community workshop with data providers, infrastructure operators, and AI/ML practitioners, we refined definitions for three tiers of data readiness: FAIR data meet established open-science norms; FAIR4AI criteria extend those norms with metadata sufficient for AI systems to discover, assess, and stage datasets to enable their use at scale without manual intervention; and AI-ready data have been fully prepared for specific AI-enabled applications. These distinctions are important because the gap between FAIR and FAIR4AI appears to be primarily a metadata gap, not a lack of technical infrastructure, and closing it is necessary to successfully use emerging AI tools in ecological synthesis.

To make these distinctions actionable, we developed a FAIR4AI checklist adapted from the Earth Science Information Partners (ESIP) AI-Readiness Checklist and the Imageomics Institute's Dataset Card and Metadata Checklists, with additions for biodiversity-specific concerns including taxonomic resolution, spatial and temporal extent, annotation provenance, and instrumentation metadata. Evaluation questions span Provenance, General Information, Data Structure, Data Origin, Data Processing, Data Quality, Guidance and Recommendations, and Data Access. We then built an open-source, LLM-powered evaluation tool to ingest a dataset's metadata and return evidence-based responses to each evaluation question along with per-category FAIR scores.

Advancing FAIR4AI standards in biodiversity will require community input to ensure they support AI applications in ways that are consistent with community ethics and goals.

Speaker



Tanya Berger-Wolf
Professor, Computer Science and Engineering | The Ohio State University

Corresponding author



Eric Sokol
Quantitative Ecologist | National Ecological Observatory Network (NEON), Battelle

Authors

Jenna Kline, Elizabeth Campolongo, Tanya Berger-Wolf, Marty Downs, Christine Laney, Hilmar Lapp, Greg Maurer, Sydne Record, Max Tarjan, Matt Thompson

Phenobase: a harmonized, AI-ready knowledge base of global plant phenology

11:20 AM – 11:35 AM

AI AND ROBOT READY

Plant phenology shapes ecosystem dynamics, species interactions and serves as a sensitive indicator of environmental change. Sources of global phenology data have exponentially increased but remain siloed and fragmented. In situ monitoring networks produce rigorous repeated observations, but are typically regional and disconnected. Digitized herbarium specimens and community-science platforms such as iNaturalist offer vast spatial, temporal, and taxonomic breadth but phenology information is latent in images and unstructured metadata. Phenobase addresses these gaps by assembling these sources into a harmonized knowledge base. We first applied computer-vision models (PhenoVision and PhenoVision-Leaf) to score presence of open flowers, fruits, breaking leaf buds, green leaves, and colored leaves across tens of millions of iNaturalist records. We used an different, ensemble-based ML model approach trained via expert annotations for labeled flowers on herbarium sheets to score millions of available imaged herbarium sheets. Finally, we assembled in-situ monitoring data from USA-NPN, NEON, PhenObs, Budburst, and SeasonWatch. Phenobase harmonizes these sources using the Plant Phenology Ontology which maps source-specific phenophase terms to shared trait concepts while preserving contextual metadata. Model certainty thresholds, validation accuracies, ML model version, and family-level performance metrics travel with each record, so downstream users can filter on quality. Phenobase data are openly available and show how machine-learning extraction, ontology-based integration, and standardized provenance metadata turn heterogeneous biodiversity collections into AI-ready data products. We close by discussing next-step challenges with maintaining and growing such resources.

Speaker



Robert Guralnick
Professor and Curator | University of Florida

Authors

Robert Guralnick

From 1.5 Billion Raw Queries to AI-Ready Biodiversity Data: Human–AI Collaborative Curation Pipelines in PI@ntNet

11:35 AM – 11:50 AM

AI AND ROBOT READY

Large biodiversity platforms collect massive volumes of opportunistic observations that are valuable for ecological research and AI model training. However, transforming these heterogeneous user contributions into trustworthy and reusable biodiversity data remains a major challenge. PI@ntNet currently processes more than 1.5 billion user queries consisting of plant images and associated metadata submitted by a large international community. Only a small fraction of these data ultimately become AI-ready training data (~25M images) or research-grade occurrences shared with GBIF (~2.5M images).

We present the human–AI collaborative curation pipeline used in PI@ntNet to progressively transform raw multimodal observations into fitness-for-use biodiversity datasets. The pipeline combines automated filtering, community moderation, collaborative taxonomic revision, image-quality assessment, and governance constraints. Automated filters remove non-plant content, malformed observations, duplicate or low-quality images, and sensitive media. Taxonomic quality is refined through a collaborative validation process combining AI predictions, community interactions, and expert feedback across species galleries and observation streams. Community moderation signals help identify problematic observations, including off-topic or unreliable content. Dataset construction also incorporates licensing and provenance constraints, excluding observations that cannot be redistributed or reused.

We further show how downstream objectives require different curation levels. While AI training datasets prioritize taxonomic consistency, visual relevance, and diversity, GBIF publication requires stricter filters on geolocation, cultivation status, metadata completeness, and interoperability standards.

Speaker



Alexis Joly

research scientist | INRIA

Authors

Pierre Bonnet

From Sensor to Species: AI-Assisted Segmentation, Deep Learning Workflows, and Persistent Biodiversity Records from Earth Observation Systems

11:50 AM – 12:05 PM

AI AND ROBOT READY

Advances in airborne and spaceborne Earth observation systems are enabling increasingly detailed ecological analyses across global spatial and temporal scales. RGB, multispectral, and hyperspectral imaging platforms (ie: Landsat, Sentinel-2, NASA AVIRIS-NG, EMIT, and the National Ecological Observatory Network (NEON) Imaging Spectrometer) provide opportunities to identify vegetation structure, phenology, and potentially species-level ecological patterns from remotely sensed imagery. However, airborne and spaceborne remotely sensed biodiversity observations remain comparatively underrepresented within biodiversity informatics infrastructures such as the Global Biodiversity Information Facility (GBIF).

This presentation explores a biodiversity informatics workflow for transforming sensor-derived ecological observations into persistent, interoperable biodiversity records. The workflow integrates foundation-model segmentation approaches, including the Segment Anything Model (SAM), object-based image analysis [8], and DeepForest tree crown delineation workflows. These approaches are combined with AI-assisted species classification and Darwin Core standards to generate candidate biodiversity occurrence records from remotely sensed vegetation objects.

A central focus of the project is the introduction of persistent identifiers for segmented vegetation objects, enabling longitudinal monitoring of ecological entities across repeated temporal observations. The presentation will discuss interoperability challenges, provenance and uncertainty representation, and opportunities for integrating AI-generated ecological observations into biodiversity informatics infrastructures and biodiversity monitoring workflows.

Speaker



Kit Lewers

PhD Candidate, FINESST Fellow | CU Boulder, NASA

On-Device AI for Data Cleaning, Standardisation, and Exploration in Collections Management

12:05 PM – 12:20 PM

AI AND ROBOT READY

Museums operate across a diverse landscape of legacy data systems and formats, reflecting the long history of collections management and digitisation efforts. These include heterogeneous database structures, varied metadata standards, and unstructured or semi-structured records derived from historical catalogues and specimen labels. A key component of ongoing work is the standardisation and transformation of these legacy data into consistent, interoperable formats aligned with modern biodiversity data standards. This presentation focuses on the use of local/on-device AI systems to support the cleaning, standardisation and exploration of collections data spread across legacy systems, including collections management records, storage location data, specimen locality information, geographic coordinates and bibliographic references. We demonstrate how AI can be applied to transform unstructured and semi-structured records into consistent, interoperable data aligned with modern biodiversity standards. Using dedicated hardware such as Dell GB10 systems and high-performance development devices, the platform provides AI-assisted suggestions for improving data quality, identifying inconsistencies and highlighting missing or ambiguous information across both collections management systems and exported datasets. It also enables interactive exploration, allowing users to query, interrogate and summarise collections data in real time. This allows the platform to generate summarised insights, identify patterns and relationships, and highlight areas requiring further curation or standardisation. A live demo will show how local AI systems can support large-scale data transformation and contribute to the development of standardised, high-quality and AI-ready collections data.

Speaker



Jack Hollister

Senior Robotics Technology Engineer | The natural History Museum, London

Authors

Steen Dupont

FAIR, CLEAR, and Causal: A semantic framework for ecological knowledge synthesis

12:20 PM – 12:30 PM

AI AND ROBOT READY

Ecological synthesis produces knowledge that is human-interpretable but rarely machine-actionable. Attempts to address this through RDF/OWL-based semantic representations introduce a new problem: structures become opaque even to domain experts. We present the Semantic Units Framework, a modular approach to structuring ecological data and knowledge as logic-aware, independently addressable units aligned with the FAIR (Findable, Accessible, Interoperable, Reusable) and CLEAR (Cognitively interoperable, semantically Linked, contextually Explorable, easily Accessible, human-Readable and -interpretable) Principles.

The central contribution is embedding causal knowledge within this architecture. Universal causal statement units encode causal hypotheses as addressable graph entities forming a navigable semantic grid, like a spatial map, to which evidence, publications, and contextual qualifiers can be anchored. Individual observations can be linked to causal hypotheses they support or contradict, enabling queries such as “Which data support this causal claim?” This supports semi-automated causal reasoning, confounder detection, and AI-ready evidence-gap maps.

The framework decomposes knowledge graphs into (a) statement units as single propositions with a persistent identifier, provenance, and method links and (b) compound units grouping related statement units into reusable knowledge objects, both serializable as nanopublication FAIR Digital Objects. Machine-interpretable structures are decoupled from human-facing representations via dynamic labels and graphs, keeping OWL/RDF accessible without specialist knowledge. Currently conceptual with partial prototypes, we invite discussion toward its community-driven realisation.

Authors

Tim Alamenciak, Robert Fröhstückl, Carlos Alberto Arnillas, Tina Heger, Birgitta Koenig-Ries

CO4 FAIR data in practice - eDNA, modelling and biodiversity concepts

11:00 AM – 12:30 PM | Location: FORUM

CONTRIBUTED ORALS

Moderator



Andre Zenger

Director | Atlas of Living Australia

FAIR ancient environmental DNA data in practice: developing and implementing interoperable metadata standards in AEGIS

11:00 AM – 11:15 AM

CONTRIBUTED ORALS

Ancient Environmental Genomics Initiative for Sustainability (AEGIS) is a seven-year programme investigating how records of adaptive responses stored in ancient environmental DNA (eDNA) coupled with other biomolecular approaches can support understanding of ecosystem resilience and sustainable food production, while preserving biodiversity, under climate change scenarios.

The data engine, a core component of AEGIS, includes a strong commitment to FAIR and open data, with enriched metadata to promote findability, access, interoperability and re-use, with eDNA data shared via the European Nucleotide Archive (ENA) and other EMBL-EBI resources. To ensure FAIR, the team engaged with the Genomics Standards Consortium (GSC), which maintain metadata standards for eDNA under MxS (Minimum Information for any Sequence), the community-led FAIRe (Making eDNA FAIR) project, and with the ancient DNA community working on MInAS (Minimum Information about any Ancient Sequence). The metadata collected under AEGIS was mapped to these existing and developing standards to construct metadata checklists and accurately describe experiment data in AEGIS. To facilitate compliance, AEGIS uses standardised metadata collection templates.

This presentation will describe the implementation process, including the development of an ENA ancient DNA metadata checklist, approaches for capturing provenance and contextual metadata, and challenges in harmonising standards across biodiversity, eDNA, and ancient genomics domains. The work demonstrates how machine-readable and interoperable metadata standards can support scalable reuse of ancient eDNA data, improve reproducibility, and strengthen long-term integration of palaeogenomic resources into wider biodiversity and environmental data ecosystems.

Speaker



Joana Pauperio

European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

Authors

Peter Woollard, Carl Baden, James A Fellows Yates, Alisha Ahamed, Colman O'Cathail, Antonio Fernandez Guerra, Magnus Johannsen, Line Olsen, Thorfinn Sand Korneliusen, Eske Willerslev

Rethinking “absence” in biodiversity data collection and modeling

11:15 AM – 11:30 AM

CONTRIBUTED ORALS

The rapid growth of biodiversity data through [GBIF](#) and related sites has accelerated the development of species distribution models (SDMs) and occupancy models. Simultaneously, the literature has accumulated many partially overlapping terms related to “absence”, including absence record, confirmed absence, negative detection, zero count, and pseudo-absence. A working group with experience in biodiversity data collection, curation, and modeling met monthly for approximately 18 months to sort out this confusing terminology.

Most biodiversity sampling compendia did not explicitly define absence-related terms. When they did, usage varied: some sources used presence-absence language, whereas others used detection-non-detection language.

The occupancy-modeling literature provided a stronger conceptual framework for data collection (MacKenzie (2005). Here detection and non-detection characterize observations, whereas presence and absence are inferred ecological states that can be estimated probabilistically with repeated survey data. This perspective motivated a [petition](#), introduced by members of our working group and others, to change the values of occurrenceStatus in the [Humboldt Extension](#) from “present/absent” to “detected/nonDetected.”

SDMs use another set of absence-related terms, such as pseudo-absence, presence-pseudoabsence data, and true absence. For modeling, ordinary language is insufficient. These terms have technical meanings that depend on model class, data type, sampling assumptions, and historical context. During this innovative phase of model development, it seems difficult to create standardized terminology,

MacKenzie, D. I. 2005. What are the issues with presence-absence data for wildlife managers? *Journal of Wildlife Management* 69: 849-860.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Rob Stevenson

University of Massachusetts Boston

Authors

Debora Arlt, Abigail Benson, Wesley Hochachka, Yi-Ming Gan, Rob Stevenson, Rubén Pérez-Pérez

Challenges and Opportunities for FAIR Marine eDNA Data Publication Workflows

11:30 AM – 11:45 AM

CONTRIBUTED ORALS

Marine environmental DNA (eDNA) dataset volumes are increasing rapidly as eDNA becomes integrated into routine sampling workflows on research vessels and in monitoring programs. However, challenges related to interoperability, metadata standardisation, and repository integration remain and require further work to support future interoperable biodiversity data infrastructure and FAIR data publication workflows.

In this project, we aim to assess current challenges and community needs associated with marine eDNA data publication and interoperability.

We are designing evaluation instruments to assess the current landscape of marine eDNA data publication workflows. We are reviewing existing workflows, tools, metadata standards, and FAIR requirements associated with biodiversity publication pathways. Through surveys and consultations with key stakeholders, we are mapping workflows from sample collection through to data mobilisation and publication. Stakeholder engagement has included collaboration and discussions with Australian and international biodiversity organisations, data infrastructure initiatives, and sequencing and biodiversity data communities.

Preliminary findings indicate that there is a strong interest within the eDNA community in publishing and reusing biodiversity data. However, results suggest inconsistent adoption of metadata standards, fragmented publication workflows, time-consuming publication processes, and incomplete metadata collection practices that may reduce FAIR compliance and data reusability.

This work will inform approaches to interoperable infrastructure for marine eDNA data publication and support FAIR adoption, thereby increasing reusability of marine biodiversity datasets.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Berenice Talamantes Becerra

Senior Technical Officer | CSIRO NCMI, Information and Data Centre (IDC), Black Mountain, ACT, Australia

Authors

Sachit Rajbhandari, Miwa Takahashi, Katherine Tattersall

Visualising specimens as event chains on the Australian Reference Genome Atlas

11:45 AM – 12:00 PM

CONTRIBUTED ORALS

The Australian Reference Genome Atlas (ARGA) was built in partnership by the Atlas of Living Australia, Australian BioCommons, and Bioplatforms Australia to index whole genome sequence data from Australian biodiversity species and other eukaryote species relevant to Australia's needs for the bioinformatics community. As part of the project, we created a new model for tracking the provenance of genomic data, tracing specimens from their wild state through to laboratory processing, and appending data objects to their relevant source samples.

Here, we outline the data requirements, the selection of fields, and the backend database structure used to facilitate the provenance chain from specimen to data object. We developed a dynamic and extensible demonstration UI, supported by these fields and architecture, to guide users through a discovery journey within the specimen area of the ARGA interface. Using an example, we will demonstrate how users can navigate from a specimen or tissue to a genomic data packet to reveal collection information, protocols, and agents for a rich provenance history display. Underpinning a card-based interface was the aim of ARGA to facilitate deeper ways of exploring and interacting with data for the genomics user community.

The ARGA project has now concluded, and we aim to extract lessons and components that we can reuse within the Atlas of Living Australia (<https://www.ala.org.au>) and which may be repurposed by other platforms that share DNA-derived biodiversity data. We will explore the applicability of this component within other collections-based online interfaces, such as for museums collections and biobanks.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Kathryn Hall

Scientific Product Manager | Atlas of Living Australia, CSIRO

Authors

Jack Brinkman, Keeva Connolly, Christopher Mangion, Sarah Richmond, Goran Sterjov, Andrew Turley, Nigel Ward

Bridging Biodiversity Data Standards and Predictive Ecology: Operationalizing Interoperable Workflows for Species Distribution Modeling in West Africa

12:00 PM – 12:15 PM

CONTRIBUTED ORALS

Biodiversity data standards have enabled large-scale aggregation of species occurrence records, yet their translation into reproducible analytical workflows remains uneven, particularly in data-limited regions. This study operationalizes standardized biodiversity data for machine learning-based species distribution modeling, using threatened tropical timber species in West Africa as a case study. We develop an interoperable workflow integrating species occurrence records structured with Darwin Core and high-resolution bioclimatic variables. The approach emphasizes data harmonization, metadata completeness, and spatial validation to address challenges such as sampling bias, coordinate uncertainty, and inconsistent taxonomy. By aligning heterogeneous datasets within a standards-compliant framework, the workflow supports reproducible and reusable analyses. Preliminary results suggest that adherence to biodiversity data standards improves model robustness and transparency, particularly in underrepresented regions. Predicted distributions under current and future climate scenarios highlight both suitable habitats and areas where data limitations affect model reliability. This work extends the role of biodiversity data standards from data mobilization to fully integrated analytical pipelines. By linking standards-based data management with predictive ecology, it demonstrates how interoperable workflows can enhance biodiversity data use for conservation decision-making and global assessments, especially in data-sparse regions such as West Africa.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Peter Ugege

DS40 - National and policy-relevant species checklists: why official, updated and validated data matter

11:00 AM – 12:30 PM | Location: ODIN

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

National and policy-relevant species checklists, like protected or alien species lists, and national inventories, are essential inputs for biodiversity policy and decision making, serving as regulatory tools and foundations for governmental strategies. National checklists are typically scientifically driven instruments supporting natural resources management, while policy-relevant such as IUCN, CITES, or GRISS build on scientific evidence and incorporate additional economic, regulatory, or societal considerations.

Variation in origin, scope, structure, licensing, and quality has limited their interoperability and use in global biodiversity infrastructures, despite ongoing progress toward improved standardization.

Through ChecklistBank, the Catalogue of Life (COL) already hosts over 1,000 national and 400 policy-relevant lists, positioning it as a key platform for improving access, visibility, and reuse.

This session will combine short presentations with an open, moderated discussion to explore the needs around generating and using these lists, and how COL can support processes from production through publication. Discussion topics include translating heterogeneous lists into shared data standards, improving metadata and validation for policy use, and documenting taxonomic differences transparently. By bringing together taxonomists, data managers, and policymakers, we seek pathways to build trust, improve integration, and translate taxonomic work into sustained policy impact.

Discussion leaders



Diana Hernandez

Editor | Catalogue of Life



Camila Plata

Catalogue of Life



Olaf Banki

Catalogue of Life Foundation

WKS43 - How Literature Services can Support & Benefit from Biodiversity Publication and Data Standards ?

11:00 AM – 12:30 PM | Location: BALDER

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Rapid growth of biodiversity literature makes it difficult to retrieve precise, relevant, and interoperable knowledge. Traditional bibliographic catalogues mainly rely on metadata and abstracts, limiting fine-grained discovery and large-scale reuse.

This hands-on workshop explores how literature services can support biodiversity publication and data standards. Using BiodiversityPMC, developed within the SIB Literature Services (SIBiLS), we demonstrate how structured, machine-readable publications enable advanced indexing, semantic enrichment and concept-based retrieval across heterogeneous sources.

The SIBiLS platform integrates taxonomic treatments from Plazi, biodiversity-related documents from Zenodo and full-text articles from biodiversity journals (e.g., EJT, Pensoft). By leveraging structured formats, persistent identifiers and domain ontologies, it transforms publications into AI-ready contents that can be cross-linked and queried at scale. We will also demonstrate how pre-computed literature selection through Triage supports Biodiversa+ communities.

Through guided exercises, attendees will perform advanced searches via the web interface, explore semantic annotations and test selected API calls for programmatic access, including MCP calls from AI bots.

The session also fosters discussion on feedback loops between publishers and literature platforms, showing how literature services can strengthen biodiversity data standards toward interoperable ecosystems.

Workshop leader



Emilie Pasche

Research associate | HES-SO & SIB Swiss Institute of Bioinformatics

LUNCH

SYM37A - A consensus taxonomic reference for improved freshwater biodiversity data and knowledge

2:00 PM – 3:30 PM | Location: SAL B

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Freshwaters are among the least comprehensively documented habitats, despite showing some of the fastest rates of biodiversity decline. Fragmented taxonomic knowledge, heterogeneous data standards and limited interoperability between infrastructures continue to constrain global assessments and policy uptake. This session will bring together initiatives within the freshwater community to relaunch and align key components of the freshwater data landscape, including the Freshwater Animal Diversity Assessment (FADA), the Freshwater Information Platform (FIP) and freshwater data mobilisation efforts in FWBON, GBIF and Catalogue of Life. We will explore how curated, realm-specific taxonomic checklists and biodiversity data can be mobilised as interoperable machine-actionable resources adhering to global Biodiversity Information Standards feeding into emerging freshwater specific infrastructures and reproducible workflows.

The session will focus on operationalising freshwater taxonomy through standards, sustainable infrastructures, and linking taxonomic backbones with trait, occurrence, and monitoring databases and conservation status information from the IUCN Red List of Threatened Species. We also invite discussion on community-driven approaches, including regional networks, e.g. the Freshwater Biodiversity Information System (FBIS), governance models, and long-term sustainability, aiming to advance a coherent, FAIR, and globally integrated freshwater biodiversity data ecosystem.

Moderators



Astrid Schmidt-Kloiber

Senior Scientist | BOKU University



Koen Martens

Royal Belgian Institute of Natural Sciences, Brussels

Session organizers



Vanessa Bremerich

Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB)



Helen Dallas

Executive Director | Freshwater Research Centre (FRC)



Olaf Banki

Catalogue of Life Foundation

Not just another database: building the future of freshwater taxonomy with FADAtims

2:00 PM – 2:20 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Freshwater biodiversity is declining rapidly, yet freshwater taxonomy remains fragmented across databases, standards and expert communities. Maintaining reliable species lists is challenging because names change continuously, new taxa are described, and taxonomic concepts differ between regions and organism groups. These issues limit interoperability between biodiversity infrastructures and constrain large-scale ecological assessments and policy applications.

To address this gap, we are developing FADAtims, an online taxonomic information management system for collaborative curation of freshwater biodiversity checklists, which is developed within the Belspo-funded infraFADA project. The platform enables registered experts to curate and continuously update species names, synonyms, classification, distribution and status information through a transparent and traceable workflow.

FADAtims is designed as a FAIR, interoperable and machine-actionable infrastructure component for freshwater biodiversity data. Built with open-source technologies and aligned with Biodiversity Information Standards, it supports data exchange with major biodiversity infrastructures including GBIF, Catalogue of Life and the Freshwater Information Platform. Interoperability with systems such as WoRMS and TaxonWorks further supports harmonisation and long-term sustainability.

The platform aims to provide an authoritative freshwater taxonomic backbone that can be linked to occurrence, trait and monitoring databases as well as conservation assessments. We present the current status of FADAtims, its role within the emerging freshwater biodiversity data ecosystem and discuss community-driven approaches for sustaining curated freshwater taxonomy infrastructures.

Speaker



Astrid Schmidt-Kloiber

Senior Scientist | BOKU University

Authors

Vanessa Bremerich, Helen Dallas, Tim Sutton, Dimas Tri Ciputra, Géraldine Mertens, Koen Martens

Catalogue of Life and Freshwater species data

2:20 PM – 2:35 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

30 years ago the activities that led to the Catalogue of Life started. For 25 years the catalogue of described organisms of all life, has been freely available online. GBIF, also celebrating a 25 year anniversary, is now making use of the Catalogue of Life as the taxonomic reference for the 3.7 billion species occurrences it mediates. Although a lot of progress has been made, there are still obvious taxonomic gaps. Some groups are notoriously under-represented and incomplete (like insects). In other cases, taxonomic data from specific ecosystem realms like freshwater is scarce. The Catalogue of Life infrastructure, ChecklistBank, now promotes the open data publishing of taxonomic and nomenclatural resources, and different types of species lists (such as policy-relevant lists). The infrastructure also provides the means to share species interaction data. We will provide an overview of the available information in Catalogue of Life and in its infrastructure. We also discuss how this data could be improved, for example through collaboration with initiatives like the Freshwater Animal Diversity Assessment.

Speaker



Olaf Banki

Catalogue of Life Foundation

A species-level taxonomic catalog of Neotropical freshwater insects: building a consensus reference for occurrence data integration, and conservation planning.

2:35 PM – 2:50 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Freshwaters are among the most biodiverse yet data-deficient habitats globally, and the Neotropical region exemplifies this paradox — exceptional aquatic insect diversity combined with critical taxonomic and distributional gaps rooted in historical undersampling. To our knowledge, no equivalent species-level, multi-order catalog of freshwater insects exists for the Neotropical region. Here we present an ongoing effort to compile a DarwinCore-compliant catalog of freshwater aquatic insects for the Neotropical lowlands, covering seven orders: Trichoptera, Coleoptera, Odonata, Ephemeroptera, Plecoptera, Hemiptera, and Megaloptera. All compiled species names are verified against the GBIF Backbone Taxonomy and the Catalogue of Life, documenting nomenclatural discrepancies including synonyms, combinationes novae, and typographic errors. Country-level distribution data are linked to georeferenced occurrence records from GBIF and regional sources. For each verified species, occurrence records are cleaned and linked to stream network segments, with environmental and topographic variables extracted at the sub-catchment scale. This dataset will underpin distribution modelling across the Magdalena–Cauca Basin in Colombia, assessing how river network structure shapes richness, rarity, and endemism, and prioritising conservation areas accounting for cumulative hydropower impacts on freshwater connectivity.

Speaker



Alejandra Correa-Bedoya

Doctoral researcher | IGB Berlin

Biodiversity data at your fingertips: open-access freshwater biodiversity information systems for South Africa and Africa

2:50 PM – 3:05 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Freshwater biodiversity data are essential for monitoring ecosystem change, informing conservation, and supporting evidence-based management. However, such data often remain fragmented, inaccessible, or poorly integrated across scales, institutions, and regions.

The Freshwater Biodiversity Information System (FBIS) and FBIS-Africa platforms, developed for South Africa and the broader African region respectively, respond directly to this need. FBIS is a powerful, visual, data-rich system designed to host, share, and explore freshwater biodiversity data. Its design and functionality have been shaped by key user groups, including water resource managers, conservation practitioners, planners, researchers, and environmental consultants.

Both platforms interoperate with global biodiversity infrastructures, including GBIF and the IUCN Red List of Threatened Species, contributing to a more connected and accessible freshwater biodiversity data ecosystem. By combining regional expertise, open infrastructure, and interoperable data workflows, FBIS and FBIS-Africa demonstrate how community-driven information systems can support FAIR biodiversity data mobilisation and translate knowledge into actionable conservation and management outcomes.

The South African platform already supports improved freshwater science, conservation, and decision-making at a national scale, with approximately 14 million occurrence records across eight biodiversity modules. The recently developed FBIS-Africa platform is rapidly expanding in both data coverage and user uptake.

Together, these platforms play a critical role in translating freshwater biodiversity data into informed monitoring, management, and conservation action across local, regional, continental, and global scales.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Helen Dallas

Executive Director | Freshwater Research Centre (FRC)

Authors

Tim Sutton, Dimas Tri Ciputra, Jeremy Shelton, Nazley Liddle

The FIPbio data portal: towards an integrated freshwater biodiversity data ecosystem

3:05 PM – 3:20 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Freshwater biodiversity research, monitoring and policy increasingly rely on interoperable data infrastructures linking taxonomy, biodiversity observations, abiotic sampling data and remote sensing products. With the renewal of the freshwater biodiversity data portal FIPbio within the Freshwater Information Platform (FIP), we present a framework integrating freshwater occurrence records, eDNA observations, sampling data and satellite-derived environmental time series, including chlorophyll-a dynamics and land-use change, across lakes, wetlands and river catchments into a unified freshwater data ecosystem.

The renewed portal builds on the Biodiversity Information Management System (BIMS), originally developed for the South African Freshwater Biodiversity Information System (FBIS). Central to FIPbio is the global Freshwater Animal Diversity Assessment (FADA) taxonomic backbone, continuously curated through the FADAtims platform. This freshwater-specific backbone enables a comprehensive overview of occurrence data mobilised through infrastructures such as GBIF.

Sampling sites are spatially linked to global freshwater environmental context layers including river networks, basins, lakes, wetlands and land-water interface classes. Occurrence records are automatically annotated with contextual dataset IDs and snapped to river network segments, enabling reproducible freshwater-specific analytical workflows using resources such as Hydrography90m.

FIPbio enhances the discoverability, interoperability and long-term reuse of freshwater biodiversity and monitoring data from local to global scales. The platform supports freshwater data mobilisation efforts, including those within the FWBON network and contributes to an integrated global freshwater biodiversity information landscape.

Speaker



Vanessa Bremerich

Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB)

Authors

Hellen Dallas, Tim Sutton, Dimas Tri Ciputra, Sonja C. Jähnig, Astrid Schmidt-Kloiber

Discussion

3:20 PM – 3:30 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Moderator



Astrid Schmidt-Kloiber

Senior Scientist | BOKU University

COLLECTIONS DIGITISATION & MANAGEMENT

Collection Management Systems (CMS) face evolving challenges in managing, curating, publishing, and integrating increasingly complex research data from diverse sources. This session seeks contributions that address practical and conceptual solutions for modern, research-oriented CMS, with a focus on enabling richer, interoperable, and sustainable data flows across the global biodiversity data landscape. We welcome talks on data mobilization for research infrastructures, e.g., GBIF, OBIS, and DiSSCo. We especially seek submissions focused on emerging standards such as Darwin Core Data Package (DwC-DP), and alignment with TDWG and related domain standards to support semantic consistency and FAIR principles. Presentations exploring implications of the Darwin Core Conceptual Model - including clearer treatment of relationships among material entities, organisms, and occurrences - are encouraged. We also invite contributions on the maintenance and use of persistent identifiers and the relationship between material samples and primary literature in the context of material sample management systems and their data consumers. Further topics include machine-assisted curation workflows, community-driven development, capacity building, long-term maintenance, CMS-as-a-Service, and governance models. By bringing together CMS developers, data architects, curators, and aggregators, the session aims to advance practical solutions to current and emerging biodiversity data exchange challenges.

Moderators



Jonas Grieb

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research



James A. Macklin

Agriculture and Agri-Food Canada

Session organizers



Christian Bölling

Museum für Naturkunde Berlin - Leibniz Institute for Evolution and Biodiversity Science



Sam Leeflang

Lead Developer Collection Data Infrastructure and Access team | Naturalis Biodiversity Center



Wouter Addink

Manager Software Development | Naturalis Biodiversity Center



Claus Weiland

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main

Authors

Michaela Grein, Volker Lohrmann, Etta Grottrian

Introduction

2:00 PM – 2:06 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Speaker



James A. Macklin

Agriculture and Agri-Food Canada

From Specimens to Networks: Darwin Core Data Packages in Practice

2:06 PM – 2:18 PM

COLLECTIONS DIGITISATION & MANAGEMENT

This year the Darwin Core standard gained a Conceptual Model and a Data Package guide detailing its practical application. These additions represent significant semantic evolution and the capacity to integrate a greater variety of primary biodiversity data within and between datasets of biological collections. By moving beyond the Darwin Core Archive, with its inherent structural limitations, the model enables institutions to capture the depth and richness of specimen and associated data through material object hierarchies and relationships to Media, Nucleotide Analyses and Sequences, Surveys, and more. We will explore the implications for collection management systems and the practical challenges of mapping existing institutional data models into the relational structures enabled by the Darwin Core Data Package framework.

Speaker



John Wieczorek

Information Architect | Rauthiflor LLC, GBIF

Authors

Tim Robertson

Aligning An Evolving Global Data Model To Software With Global Ambitions: GBIF x TaxonWorks

2:18 PM – 2:30 PM

COLLECTIONS DIGITISATION & MANAGEMENT

[TaxonWorks](#) (TW) has an implementation schema of over [70 models](#) that it uses to facilitate the description of Earth's biodiversity. While it exports DarwinCore Occurrences and Checklists, Catalogue of Life Data Packages, and other standards its development was not constrained to those standards. As a proxy to the needs of a growing community of users it therefore acts as a lens to explore potential emerging standards. The Global Biodiversity Information Facility (GBIF) is in the process of finalizing changes with respect to how they represent data under their umbrella via the [DarwinCore Data Package framework](#). The new semantics introduced within DwC-DP create a clean space within which to compare TW concepts. Both models are strongly similar in how they seek to "shape" data, they strongly overlap conceptually. To identify the infrastructure changes required to efficiently export from TW to the DwC-DP standard, specifically to minimize data loss (all data-mappings are lossy to some degree) we map core concepts in both. We focus on 1) the interrelationship between DwC-DP's "Assertion" class and TW Observation and DataAttribute classes; 2) the nature of DwC Occurrences and related entities in TW (specifically the nature of "material" entities); and 3) abstractions related to the handling of annotations (e.g. Identifiers, Media). Some of our exploration has been facilitated by the [Species File Group](#)'s "grand schema" and "SFGBorg" projects, internal efforts focused on data exchange and interoperability, we contextualize those efforts in this comparison. Finally, as potential fodder for future standards, we re-introduce the conceptual domains of Process and Rendering, and describe how TW is exploring development in these realms.

Speaker



Matthew Yoder

University of Illinois, PRI, Illinois Natural History Survey, Species File Group

Authors

Geoff Ower, Deborah Paul, Dmitry Mozzherin

Material Samples and Occurrences in Darwin Core Data Packages: Practical Experiences from DINA

2:30 PM – 2:42 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Research collection management systems (CMS) are the principal tools used to host and mobilize material sample data to biodiversity data infrastructures. The mobilization process traditionally depends on the export of the collection data in the Darwin Core Archive (DwC-A) format. This data export often involves the flattening of information from complex relational structures and the synthetic generation of occurrences. The new Darwin Core Data Package (DwC-DP) format provides an opportunity to overcome these limitations by (1) following a relational data model and (2) relaxing requirements so that valid Darwin Core data can also be transported without instantiating occurrences. However, new challenges arise from the alignment of local, material-focused CMS data models and the conceptual model powering DwC-DP. A prominent example is the treatment of organism attributes such as sex or life stage, which are grouped under the Occurrence class in DwC-DP as they possibly change over time. In a CMS, however, usually data about preserved specimens or material parts derived from them are managed, where these attributes do not change and are directly associated with the material entity. As a result, exporting this type of information from a CMS still requires the generation of occurrences, which raises the question if and how occurrence IDs should be generated.

In this talk, we will discuss practical examples of complex material entities stored in DINA, a collaboratively developed specimen and sample-based CMS, and how they can be serialized to DwC-DP. We also demonstrate how DINA's data model with managed attributes and field extension features provides a native way to store standards-conformant data and dynamically adapt to evolving and emerging biodiversity data standards.

Speaker



Jonas Grieb

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research

Authors

James A. Macklin, Christian Bölling, Michaela Grein, Volker Lohrmann, Etta Grotian, David P. Shorthouse

Sharing Extended Specimen Data from Symbiota Portals using Registered Extensions

2:42 PM – 2:54 PM

COLLECTIONS DIGITISATION & MANAGEMENT

As both a collection management system (CMS) and aggregation software, Symbiota (<https://symbiota.org/>) faces compounding challenges of making extended specimen data easy to enter, manage, and re-use for research. Symbiota portals are designed to be user friendly for people of varying levels of database experience—from new students to professional data managers—underscoring the need to keep interfaces and data structures as simple and intuitive as possible. To address these challenges, Symbiota code designers have chosen the seemingly straightforward approach of aligning extended specimen data with registered extensions, e.g., Extended Measurement or Facts, Resource Relationship, and GGBN Material Sample. Here we will describe the timeline and current use of registered extensions in Symbiota portals, including the newest tools, challenges, and community implementations of these extensions.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Katie Pearson

Symbiota Support Hub, University of Kansas

Authors

Lindsay Walker, Nico Franz, Edward Gilbert

Mapping the Relational Data Schema of the Diversity Workbench (DWB) to the Darwin Core Data Package Data Model to fit better to GBIF's New Data Model

2:54 PM – 3:02 PM

COLLECTIONS DIGITISATION & MANAGEMENT

The publication workflow from the German Federation for Biological Data (GFBio) Data Centers at Natural History Collections to GBIF is relying on the BioCASE Provider Software and the ABCD standard. This approach emerged because the classical Darwin Core Archive (DwC-A) model was often insufficient for representing complex relationships between collection objects, identifications, multimedia, agents, and related entities. However, the DWB–ABCD publication pathway also revealed some limitations within the GBIF ingestion workflow, as important dataset metadata, including multiple global identifiers, associated persons, institutions, and their roles, could not always be fully represented or displayed on the GBIF portal.

The Darwin Core Data Package (DwC-DP), currently evolving within the GBIF and TDWG communities, aims to combine the strengths of both approaches. Aligned with the Darwin Core Conceptual Model and new ABCD versions, DwC-DP introduces a more flexible and semantically explicit framework for biodiversity data publishing.

This contribution presents a practical, AI-assisted mapping approach from DWB-based collection management software model to the emerging DwC-DP model. During the mapping process, information distributed across different DWB database modules is consolidated into DwC-DP entities while respecting existing relations, withhold settings and publication restrictions. Using real-world datasets from GFBio Data Center SNSB IT Center, we analyze how relational structures, identifiers, artefacts, material storage, agents and general metadata can be transformed into DwC-DP entities and properties. The results demonstrate the potential of DwC-DP for richer biodiversity data integration and improved metadata visibility within GBIF publishing workflows.

Speaker



Tanja Melanie Weibulat

Bavarian State Collections of Natural History, SNSB IT Center

Authors

Stefan Seifert, Ariane Grunz, Dagmar Triebel, Markus Weiss

Diversity Workbench (DWB) as a Standards-Based Collection Management System for Heterogeneous Research Data and Biodiversity Data Mobilization

3:02 PM – 3:10 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Research collection management systems must support heterogeneous data, distributed workflows, and standards-based publication to enable interoperable and sustainable biodiversity data flows. At the Leibniz Institute for the Analysis of Biodiversity Change, Diversity Workbench (DWB) is the central platform for managing, standardizing, and publishing collection and project data across different sites. As a modular relational database system with multi-tenant capabilities and granular rights management, it supports the differing requirements of zoological, geological, palaeontological and mineralogical collections as well as diverse research projects.

DWB integrates data from internal and external sources through standardized and customizable import workflows, including its Import Wizard and API-based ingestion pipelines. These support collection digitization, project-based data capture, and laboratory-derived workflows such as metabarcoding. Its modular architecture enables collection-specific curation while maintaining a common institutional data framework.

Close collaboration among all stakeholders ensures data quality, semantic consistency, and domain-appropriate mappings. DWB bridges internal workflows and external biodiversity infrastructures. Data are prepared and published according to community standards, including ABCD via BioCASE and Darwin Core via the GBIF IPT, enabling dissemination to GBIF, GGBN, GFBio, and NFDI4Biodiversity. Stable CETAF identifiers and DiSSCo DOIs support FAIR and sustainable publishing.

We present DWB as a practical example of a research-oriented CMS that combines modularity, standards compliance, and sustainable publication workflows across the full data lifecycle.

Speaker



Birgit Rach

Digital collection manager | LIB - Leibniz institute for the analysis of biodiversity change

Authors

Tanja Melanie Weibulat, Peter Grobe, Björn Quast

From Client to App: Extending Diversity Workbench with REST and graphical web interfaces

3:10 PM – 3:18 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Diversity Workbench is a collection management system that supports the diverse requirements of managing biological and geological collections. The Leibniz Institute for the Analysis of Biodiversity Change uses the Diversity Workbench as its central collection database. The system is organized into modules for specific tasks, including *DiversityCollection* for the collection digitization, *DiversityTaxonNames* for taxonomy management, and *DiversityProjects* for project metadata administration. Each module uses its own relational SQL Server database and dedicated client software.

While this modular architecture enables highly flexible data management, it also results in complex desktop applications with limited transparency in user guidance and a dependency on Windows-based systems.

To address these limitations, we are developing two complementary components that use the existing relational databases as backend infrastructure: 1) a REST-compatible API that provides access to data in JSON, RDF, and HTML formats for machine interoperability, and 2) a configurable web application that allows users to focus on their specific data management tasks through graphical browser-based interfaces.

The API and web application are already being used in the upcoming version of our online collection catalog. In addition, the integration of the API into our ASV-Registry demonstrates its applicability beyond collection management, for example in specialized research domains such as metabarcoding data management. Together, these developments represent an important step toward improved usability, platform independence, and machine interoperability for Diversity Workbench.

Speaker



Björn Quast

Leibniz Institute for the Analysis of Biodiversity Change (LIB)

Authors

Anthony Gray, Sheikh Mastura Farzana

The More Extended, The More Open: Semantic Integration of 3D Natural Heritage Data into Biodiversity CMS

3:18 PM – 3:30 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Biodiversity platforms are evolving into 'Open Knowledge Spaces' connecting primary data, research, and exploration. While the Digital Extended Specimen (DES) framework integrates derivative data, current Collection Management Systems (CMS) using models like openDS limit the ability to accommodate the 3D complexity and broader utility of physical specimens.

Mobilizing biodiversity data beyond taxonomy—into exhibitions, citizen science, and AI—requires high-fidelity 3D resources. For these to act as reliable knowledge assets, an expanded schema must track 3D modeling paradata and 'utilization contexts' of how specimens are preserved and exhibited. Through the temporal dynamics of human curation, biological specimens transform from mere evidence into 'Natural Heritage' with socio-historical value.

This study proposes a 'Natural Historical Specimen (NHS)' data model based on 3D data from 320 natural heritage animal species in South Korea. To maximize interoperability, this schema integrates Darwin Core (DwC) for occurrences, Prov-O and MorphoSource for 3D digitization provenance, and CIDOC-CRM for institutional curation and exhibition contexts.

Mapping biological resources, digital objects, and utilization contexts ensures traceability from field collection to public engagement. We implemented an interactive dashboard using this model, enabling users to explore morphological attributes and historical trajectories via a Knowledge Graph. This study offers practical solutions for next-generation CMS, demonstrating that integrating DwC and CIDOC-CRM standards elevates 3D resources into hubs within global biodiversity networks.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Yeeun Lee

Master Student | Korea National University Of Heritage

Authors

Jongwook Lee

SYM25B - From Mobilizing Data to AI-Ready Knowledge: Infrastructure for Multimodal Biodiversity Data

2:00 PM – 3:30 PM | Location: SAL C

AI AND ROBOT READY

More ecological and biodiversity data are online than ever before, in part thanks to the enormous decades-long efforts by the biodiversity community in mobilizing observational data, supporting shared standards, and building aggregation networks. Publicly accessible data are also more varied than ever before, including images of many modalities, videos, and streaming sensor data. However, using this sea of data for increasingly ML-heavy and AI-driven research pipelines remains fraught. Significant ecological context, locality detail, trait information, and interaction data are buried in unstructured media. Records are often incomplete, unevenly georeferenced, and documented inconsistently across platforms. Metadata regarding terms of use, provider expectations, and desired round-tripping of data amendments are unclear or missing. Without downloading massive amounts, assessing the quality and fitness-for-use of data is often difficult or impossible. As AI-enabled and agent-driven methods mature, developing shared, community-aligned infrastructure and practices that transform multimodal biodiversity data into interoperable, provenance-aware, and research-ready knowledge becomes a central challenge. Contributions to this session will explore how standards, human-AI collaboration, uncertainty representation, ontology integration, and data governance frameworks can enable AI-orchestrated pipelines to understand fitness-for-use, respect provider expectations, and track provenance.

Moderators



Hilmar Lapp

Neuromatch, Inc.



Robert Guralnick

Professor and Curator | University of Florida

Session organizers



Eric Sokol

Quantitative Ecologist | National Ecological Observatory Network (NEON), Battelle



David Bloom

University of North Carolina at Greensboro

Authors

Kelsey Huelsman

Mobilising and integrating extracted information from the literature into a biodiversity data ecosystem - the BIOfid approach

2:00 PM – 2:15 PM

AI AND ROBOT READY

Insufficient data availability is a critical bottleneck for biodiversity research. The fact that countless species are going extinct before they can even be collected and described is a particularly alarming aspect of our lack of knowledge and data. Another aspect is that vast quantities of data already exist but remain largely inaccessible—namely, data locked in unstructured, unindexed literature. These biodiversity data must be urgently mobilized to support research. The term 'mobilization' can take on different meanings. The most basic approach is to digitize printed publications and make them freely available online, allowing them to 'circulate globally' and allowing their content be discovered through keyword searches. However, since the vocabulary, organism names, and geographic terms used in publications vary considerably over time, normalization is essential to enable both efficient searches and integration with other data sources.

This normalization is achieved through annotation based on generic and domain-specific standards, with TDWG playing a central role in defining the latter. Two examples illustrate this: Only when species names are linked to a taxonomic backbone like GBIF's, and when terms for morphological traits are connected to TraitBank, can researchers efficiently explore the wealth of data on a species' characteristics hidden in literature—despite its heterogeneity in language, publication year, or writing style. Against this background, this presentation demonstrates how the Discipline-Specific Information Service Biodiversity Research (BIOfid) addresses these challenges and how biodiversity data from the BIOfid corpus can be transformed into a graph-compatible format and made robot-ready for AI applications.

Speaker



Gerwin Kasperek

Curator of the Biology Collection | University Library Johann Christian Senckenberg, Goethe University Frankfurt

Authors

Katrin Peikert, Mevlüt Bağcı, Jana Hoffmann, Marius Böyng, Sefika Efeoglu, Alexander Mehler

Camtrap DP: extending data standard designed for camera trapping research to support AI workflows and new data sources

2:15 PM – 2:30 PM

AI AND ROBOT READY

Passive sensor monitoring is generating multimodal biodiversity data at unprecedented scale. As AI becomes central to processing these datasets, standards must evolve to capture not only what was observed, but how, preserving the full provenance of human and machine annotations.

Camtrap DP is an open data exchange standard for camera trap data, developed under TDWG. Built on the Frictionless Data framework, it natively integrates AI-relevant data (bounding boxes, classification probabilities, method attribution and model name). With multiple datasets published via GBIF, it has demonstrated clear value for interoperable, research-ready wildlife monitoring data.

Expanding adoption has now exposed important gaps. The Safe and Sound bioacoustics project and the Insect AI community — both evaluating Camtrap DP as a candidate standard for their data — highlight shared needs alongside feedback from traditional camera trap users: including more expressive hardware description, flexible recording schedule specification, and richer support for AI-generated data.

While the framework's flexibility permits ad-hoc extensions, community-wide standardization is essential to ensure data remains FAIR and comparable across projects. We are therefore developing a new version of Camtrap DP that formalizes these needs through new standardized tables covering AI model provenance and the complete record of detections and annotations, rather than only final consensus results.

We present the community consultation process driving these developments and the proposed changes to the data model and controlled vocabularies, situating this work within the broader challenge of evolving domain-specific standards to meet a rapidly changing, AI-driven data landscape.

Speaker



Karolina Kuczkowska

Software developer, open data specialist | Open Science Conservation Fund

Authors

Jakub Witold Bubnicki, Peter Desmet, Sanne Govært, Jamie Alison, Wouter Koch, Julia Wiel, Benjamin Cretois

FAIR² Drones: An AI-Ready Standard for Cross-Domain Wildlife Drone Dataset

2:30 PM – 2:45 PM

AI AND ROBOT READY

Animal ecology data collection using drones requires a substantial investment of time, expertise, and financial resources. Yet most existing datasets serve only a single research community, limiting interdisciplinary reuse. We propose a unified drone dataset standard, FAIR² Drones, that bridges ecology, robotics, and computer vision by building on existing FAIR and AI-ready data frameworks while adding essential platform metadata and annotation specifications. Our standard enables datasets to simultaneously support ecological analysis, robotics algorithm development, and computer vision model training and benchmarking. We provide open-source validation tools, reference implementations, and multimodal extensions linking drone imagery with complementary sensors, such as camera traps, GPS, and acoustics. By standardizing metadata across disciplines, this framework maximizes the scientific return on investment for costly field deployments and accelerates cross-domain collaboration in environmental monitoring.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Jenna Kline

Postdoctoral Associate | Massachusetts Institute of Technology

Authors

Jenna Kline, Tanya Berger-Wolf

Causal Mosaic Schema: Encoding Causal Claims in Ecological Literature as Machine-Actionable Knowledge Graphs

2:45 PM – 3:00 PM

AI AND ROBOT READY

Ecological evidence synthesis increasingly relies on large language models to extract knowledge. Yet most extraction pipelines discard the very information that makes causal claims scientifically defensible: what type of evidence supports the claim, how strong is it, and what kind of causal relationship is established. Without a schema that encodes these properties, extracted claims are superficially machine-readable but epistemically opaque.

Causal Mosaic is a LinkML schema that addresses this gap. It merges the Ecolink Model (ELM) entity structure with an annotation framework based on Illari and Russo's "causal mosaic" approach. We represent causation as a relation between state changes. Each causal node decomposes into an entity, a measurable attribute, and a direction of change. Each edge carries four annotation layers: claim strength, philosophical account of causation (mechanistic, interventionist, probabilistic, and others), fifteen features that further characterize the causal relationship, and an evidential basis grounded in the view that mechanistic and probabilistic evidence is jointly required to establish causation.

The schema produces a single canonical labeled property graph renderable as natural language (cf. Rosetta Statements), Fuzzy Cognitive Maps, Evidence Gap Maps, and RAG retrieval indices. Edges are grounded in source text spans and carry full provenance, supporting round-tripping from graph back to literature.

We present the schema design, an LLM-assisted extraction pipeline validated against author-reviewed annotations, and a construct validity study for a DFG-funded expert elicitation workshop. Causal Mosaic enables biodiversity knowledge graphs that represent not just what causes what, but how we know it and how confident we should be.

Speaker



Tim Alamenciak

Authors

Robert Fröhstückl, Phyllis Illari, Lars Vogt

LifeGate: Extending a Global Visual PhyloMap into a Structured Biodiversity Knowledge System

3:00 PM – 3:15 PM

AI AND ROBOT READY

LifeGate (<https://lifegate.idiv.de/>) is a biodiversity information platform centered on the “PhyloMap”, a visual representation of biodiversity that enables exploration of taxa across multiple hierarchical levels of the Tree of Life. Unlike conventional biodiversity portals based mainly on text search, the PhyloMap provides a navigable overview of phylogenetic relationships from kingdoms to species. The platform integrates taxonomy, images, phylogenetic information, and links to biodiversity literature.

A new project aims to transform LifeGate into a structured and interoperable biodiversity knowledge system by converting descriptive taxonomic knowledge into machine-interpretable biodiversity data. We combine ontology-based modeling, biodiversity knowledge graphs, and AI-assisted information extraction workflows. Domain-specific templates and OWL ontologies are being developed to represent morphological, ecological, and taxonomic characteristics and to support semantic integration with infrastructures such as GBIF.

The workflow integrates expert curation with semi-automated extraction from taxonomic literature using natural language processing and large language models. Extracted entities and relationships are validated through ontology constraints before integration into the LifeGate knowledge graph. A major long-term objective is the development of structured, visually supported identification and faceted search workflows derived from taxonomic descriptions and trait information.

This contribution presents the existing PhyloMap and its extension toward structured biodiversity knowledge integrating visualization, semantic interoperability, and AI-assisted curation.

Speaker



Birgitta König-Ries

Friedrich Schiller University Jena/German Centre for Integrative Biodiversity Research (iDiv)

Authors

Elke Freiberg (Systematic Botany and Functional Diversity Lab, Institute of Biology, Leipzig University), Cornelia Fürstenau (Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena), Erik Kleinstaub (Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena), David Schellenberger Costa (Institute of Biology, Leipzig University/ German Centre for Integrative Biodiversity Research iDiv), Sebastian Schuch (Systematic Botany and Functional Diversity Lab, Institute of Biology, Leipzig University), Franziska Zander (Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena), Martin Freiberg (Institute of Biology, Leipzig University/ German Centre for Integrative Biodiversity Research iDiv)

Discussion

3:15 PM – 3:30 PM

AI AND ROBOT READY

Moderator



Hilmar Lapp

Neuromatch, Inc.

SYM45 - Planning the Libroscope: creating research ready biodiversity from scientific publications

2:00 PM – 3:30 PM | Location: FORUM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Scientific publications systematically present original or synthesised research and incorporate it into the formal scholarly record through specialist journals and books. They use standardised vocabularies, cite prior work, and—particularly in biodiversity science—reference studied specimens and DNA sequences, thereby linking the literature to natural science collections (NSC) and DNA sequence databases. However, although they contain critical factual data, publications are written for human consumption. They are rarely semantically enhanced with machine-actionable links that would make them Artificial Intelligence-ready.

The [Disentis Roadmap](#) is a decadal initiative to liberate global biodiversity knowledge from scientific literature, endorsed by 110 scientists and organizations. Its primary goal is to deliver open machine-actionable data by building the Biodiversity Libroscope instrument. Envisioned as a shared framework, the instrument aims to systematically discover, structure, publish, liberate, and connect biodiversity data from published sources in FAIR-compliant formats. It builds on existing projects, infrastructures, and use cases to define the standards needed for global collaboration, interoperability, and scalability.

This symposium will present exemplary contributions demonstrating key elements of the Libroscope

conversion workflow, followed by a critical discussion of the approach, adopted standards, and strategies for meaningful community engagement.

Moderators



Donat Agosti
Director | Plazi GmbH



Laurence Benichou

Introduction

2:00 PM – 2:02 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Speaker



Donat Agosti
Director | Plazi GmbH

The future is the future - thought about access to data in publications

2:02 PM – 2:14 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Scientific publications contain a large share of humanity's knowledge on biodiversity and are becoming increasingly important as training and reference sources for Artificial Intelligence applications. Yet most biodiversity literature remains difficult for machines to access and interpret. Estimates suggest a corpus of hundreds of millions of printed pages, with 50,000–200,000 biodiversity-related articles and around 2,000 taxonomy-relevant journals published annually. Resources such as the Biodiversity Heritage Library and Plazi demonstrate both the scale of the challenge and the feasibility of extracting structured biodiversity knowledge from publications.

Biodiversity publications not only describe taxa and synthesize research, but also cite specimens, observations and datasets, thereby linking natural science collections, occurrence records and scientific interpretations. However, most publications are still produced primarily for human reading and lack machine-actionable semantic structure. Unlocking this knowledge at scale therefore requires strategic prioritisation.

Within the framework of the Disentis Roadmap, which aims to liberate biodiversity knowledge from the literature by 2034, key questions arise: should limited resources focus on continued digitisation of historical print literature, semantic enhancement of legacy publications, or the promotion of future-ready publishing workflows that produce structured and FAIR-compliant data from the outset, as demonstrated by initiatives such as Pensoft and CETAF?

This contribution argues that a practical compromise is the development of a Biodiversity Libroscope: a framework to systematically digitise, extract and publish semantically standardised biodiversity knowledge from well-defined corpora of literature.

Speaker



Donat Agosti
Director | Plazi GmbH

Authors

Laurence Bénichou

Biodiversity literature as FAIR and AI-ready data by design (and how to get there)

2:14 PM – 2:26 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Since the earliest observations of the living world, scientific literature has been the primary medium for documenting and disseminating biodiversity knowledge. Over centuries, this process has generated a vast corpus of taxonomic, ecological, biogeographic, and recently also molecular information. Yet, much of these data remain embedded in PDF-based publications, limiting their findability, accessibility, interoperability, and computational reuse.

Although data extraction and annotation tools have been produced to address this issue and unlock biodiversity data from PDFs, it is crucial to stop extending the legacy backlog. According to the Disentis vision, this presentation proposes to shift toward semantically enhanced publishing workflows designed to readily produce FAIR, interlinked, and AI-ready data, a process that started more than a decade ago as part of the ViBRANT FP7 project and was subsequently carried out successfully through the pro-iBiosphere, EU BON, and BiCIKL projects.

XML-first workflows for prospective literature (MetoTaxa, ARPHA) will be presented from semantic markup to enrichment and publication through interconnected infrastructures. Particular emphasis will be placed on the standards (e.g., JATS-TaxPub), vocabularies (e.g., Darwin Core), and resolvable persistent identifiers.

By supporting large-scale knowledge integration and machine-actionable biodiversity data, such workflows create new opportunities for emerging AI applications and also improve support for many other machine driven forms of reuse.

Speaker



Chris Le Coquet

Muséum national d'Histoire naturelle

Authors

Guido Sautter, Lyubomir Penev

Realtime Access to Data in new Taxonomic Publications

2:26 PM – 2:38 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Taxonomic knowledge is published at an accelerating rate across hundreds of journals, creating major challenges for aggregators and biodiversity databases to keep up. Many articles are available only as PDFs, lack machine-actionable structure, and are often hidden behind paywalls, making automated extraction of taxonomic names, type specimens, and diagnostic data increasingly difficult.

This work presents Plazi's semantics-aware, modular workflow that converts in realtime taxonomic publications published as PDFs or XML into open, FAIR, and machine-actionable biodiversity data and semantically enhanced articles. Structured PDFs are decoded to infer document structure, while XML publications (e.g. JATS-TaxPub) are processed directly through embedded semantic markup. Publication streams are harvested through RSS feeds or direct ingestion as part of editorial workflows.

Extracted taxonomic treatments, figures, tables, materials citations, and bibliographic data are semantically enhanced and linked to external repositories and taxonomic reference systems.

Automated quality control is complemented by targeted manual correction where required. Outputs are disseminated in near real time through standards-based formats including DwCA, ColDP, and JATS-TaxPub XML. TreatmentBank currently contributes more than 58,000 datasets to the Global Biodiversity Information Facility, over 80,000 new species records to GBIF and the Catalogue of Life, more than one million treatments to BiodiversityPMC, and 800,000 figures as FAIR digital objects in the Biodiversity Literature Repository.

By transforming legacy and newly published literature into interoperable machine-actionable knowledge, this workflow accelerates biodiversity data access for aggregators, infrastructures, and AI applications.

Speaker



Guido Sautter

Developer | Plazi

Building digital-native species: Transitioning from PDF liberation to semantic-by-design publishing

2:38 PM – 2:50 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Scientific publications remain the primary vehicle for communicating biodiversity knowledge. The Bouchout Declaration marked an important step toward opening biodiversity literature and mobilising the taxonomic knowledge embedded within publications. Since then, major efforts have focused on liberating biodiversity data from legacy PDFs through text mining, semantic annotation and automated extraction workflows, enabling large-scale mobilisation of taxonomic names, treatments and material citations.

However, PDF-centric approaches remain limited by the publication format itself. Extracting structured knowledge from PDFs is retrospective, heterogeneous and costly, often requiring complex workflows to recover semantics that were never encoded during publication.

In the framework of the Disentis Roadmap and the Biodiversity Libroscope initiative, this contribution argues for a transition from retrospective data liberation to digital-first and XML-first publishing workflows that generate machine-actionable biodiversity knowledge by design. Central to this vision is the concept of “digital-native species”, in which species descriptions and supporting evidence are created, published and maintained in interoperable semantic environments from the outset.

Publications remain the formal scholarly record but are generated from, and linked to, structured specimen-level evidence including images, sequences, measurements, treatments and collection records. Semantic XML publishing enables fine-grained markup, interoperability across infrastructures and direct integration with natural science collections and global biodiversity platforms.

This contribution discusses how semantic-by-design publishing can create interconnected, computable and AI-ready biodiversity knowledge ecosystems.

Speaker



Donat Agosti
Director | Plazi GmbH

Authors

Rudolf Meier, Amrita Srivathsan

Literature triage to support Island Biodiversity Monitoring

2:50 PM – 3:02 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Biodiversity research increasingly relies on vast quantities of scientific literature, therefore developing screening tools to better scan the literature is essential. Island biodiversity was selected as a test bed thanks to the support of Biodiversa+. An automatic triage system was developed and applied to Biodiversity Island monitoring.

We started by querying OpenAlex with broad coverage Boolean equations. Further, a binary classifier for island ecosystem publications (BioMoQA project) was developed using pre-trained bidirectional language models. After manual selection, the articles are then indexed into SIBiLS/BiodiversityPMC (<https://biodiversitypmc.sibils.org/>). Using 5-fold cross-validation on 1,434 abstracts, we compared domain-specific biomedical transformers (e.g., BioBERT, BiomedBERT) for life sciences against general-purpose models (e.g., BERT, RoBERTa), which are combined with ensemble learning. Binary cross-entropy ensembles achieved best performance with ROC-AUC of 0.926. The reported model performance is competitive with curator inter-rater agreement (expert kappa ~0.78). The classifier is now integrated in the SIBiLS/BiodiversityPMC platform, and through weekly screening ensures consistent up to date coverage of island biodiversity literature, thus supporting the harmonization of global island biodiversity monitoring efforts. The service platform is currently being applied to other biodiversity assessment projects and could likely serve a large set of biodiversity communities.

Speaker



Patrick Ruch
HES-SO & SIB Swiss Institute of Bioinformatics

Literature usage in IPBES - Past, Present and Future with Libroscope

3:02 PM – 3:14 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

The Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) produces global assessments that rely heavily on scientific literature as a primary source of evidence. Over time, the literature search practices within IPBES have evolved considerably — from reliance on closed, proprietary databases such as Web of Science and Scopus, toward open infrastructure like OpenAlex combined with scripted, reproducible, and transparent search strategies.

This talk traces that evolution and reflects on what has been learned along the way. It examines the trade-offs between different database choices, the shift toward programmatic and auditable search workflows, and the persistent challenge that scientific publications — while rich in factual biodiversity data — remain written for human consumption and are rarely machine-actionable.

A further complication is the role of grey literature — reports, technical documents, government publications, and other non-peer-reviewed sources that contain an abundance of important information and knowledge, yet remain largely invisible to standard bibliographic infrastructure. Grey literature is at best inconsistently indexed, poorly structured, and often inaccessible at scale, making systematic discovery and extraction particularly difficult.

Looking forward, I discuss how the Libroscope vision addresses the machine-actionability gap: by systematically liberating, structuring, and connecting biodiversity knowledge from the published literature in FAIR-compliant, AI-ready formats. The IPBES experience provides a concrete use case illustrating both the current limitations and the transformative potential that a shared Libroscope instrument could deliver for evidence-based biodiversity policy.

Speaker



Rainer M Krug

Senior Data Specialist | SIB Switzerland

Species trait literature mining to support integrative biodiversity distribution modelling

3:14 PM – 3:26 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Combining biodiversity genomics and AI-driven text mining has the potential to transform how biodiversity data are integrated for species and habitat protection and restoration. The project's central focus is the large-scale extraction of species traits from biodiversity and taxonomic literature to improve integrative species distribution and taxonomic richness modelling. It aims to deliver an open-access knowledge graph and modelling portal linking text-mined traits, genomic indicators, and environmental layers for transparent, reusable analyses.

Current species distribution models increasingly incorporate traits and genomic data, but progress is limited by difficulties accessing and harmonising heterogeneous data sources. We will address this through AI-assisted literature digitisation, named-entity recognition and normalisation, relationship extraction, and semi-automated expert curation to convert biodiversity literature into machine-actionable trait datasets. These text-mined traits will be integrated with genomic analyses and benchmarked against occurrence-only models to quantify gains in predictive power and ecological realism.

The project is also addressing uncertainties from incomplete taxonomy and undescribed “dark taxa” through taxonomic reconciliation and uncertainty propagation within modelling workflows. Research focuses on birds, bats, and fish in Switzerland, and butterflies, bumblebees, and amphipods across Europe, providing case studies for evaluating literature-mining approaches. Resulting workflows and decision-ready maps will support biodiversity research and conservation planning in Switzerland and across Europe.

Speaker



Robert Waterhouse

SIB Swiss Institute of Bioinformatics

Closing remarks

3:26 PM – 3:30 PM

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Speaker



Donat Agosti

Director | Plazi GmbH

DS31 - Panel Discussion: Critically Assessing the Role of AI in Biodiversity Research

2:00 PM – 3:30 PM | Location: ODIN

AI AND ROBOT READY

AI promises much to scientists working with biodiversity data. There has been a recent surge in projects based on LLMs and related technologies: systems that go beyond species identification from images and also identify species traits, agents answering complex ecological questions in natural language based on real data, machine transcription that not only reads handwritten labels but understands them. Such projects, and many more, hope to harness AI in answering some of the most pressing questions in biodiversity research, multiplying the efforts of scientists through more efficient

research, larger studies, and novel insights into data.

But while much is promised, critics might say that little has been delivered. Studies focus on the exciting potential of new models, but often lack performance metrics to prove their abilities. Are the data centres in which our models run destroying the same ecosystems we study? Does public backlash to AI risk losing the trust of citizen scientists whose data is used in training models without attribution or consent? Is AI amplifying our abilities, or are we outsourcing our thinking and neglecting hard-won expert knowledge? How do we reconcile a reliance on closed models and secret training data with our obligations to FAIR and open science?

This discussion session invites a panel of experts to discuss these topics with our audience, with the aim of co-creating guidelines for effective and ethical use of AI in biodiversity research.

Session organizers



Wouter Koch

Senior adviser biodiversity informatics | Norwegian Biodiversity Information Centre



Arianna Salili-James

Natural History Museum UK



Rukaya Johaadien

GBIF Norway



Michal Torma

Head engeneer | University of Oslo



Maarten Trekels

Biodiversity Data Scientist | Meise Botanic Garden



Kit Lewers

PhD Candidate, FINESST Fellow | CU Boulder, NASA

Discussion leader



David Williamson

Postdoctoral Researcher | NTNU University Museum

WKS10 - Geological Collections with Darwin Core: Compound Specimen Model and Mineralogy Extension, Accelerated with AI

2:00 PM – 3:30 PM | Location: BALDER

EARTH SCIENCES & GEOLOGICAL COLLECTIONS

Earth sciences collections present digitization challenges that traditional biodiversity standards were not designed to address. Geologic material – naturally occurring matter generated by geologic processes, encompassing minerals, rocks, meteorites, and fossils – requires vocabularies and data structures that go beyond those developed for biological specimens. This workshop introduces a standards-based workflow that leverages the Compound Specimen model to represent part-whole and associated-object relationships using Darwin Core and the Mineralogy Extension, a community-developed vocabulary now entering public review.

Participants will map real earth sciences collection datasets to Darwin Core and Mineralogy Extension terms, working through examples drawn from mineral, rock, and fossil collections. Facilitators will demonstrate how the Compound Specimen model handles objects comprised of multiple identifiable components – such as a thin section linked to its parent rock sample, or a mineral specimen associated with chemical analyses. The session closes with a structured feedback exercise on the Mineralogy Extension and an introduction to the next phase of integrating earth sciences collections into the broader TDWG community.

No prior experience with the extension is required. Familiarity with Darwin Core is helpful but not essential.

Workshop leader



Ben Norton

Yale University

BREAK

3:30 PM – 4:00 PM | Location: 1st Floor

SYM37B - A consensus taxonomic reference for improved freshwater biodiversity data and knowledge

4:00 PM – 5:00 PM | Location: SAL B

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Freshwaters are among the least comprehensively documented habitats, despite showing some of the fastest rates of biodiversity decline. Fragmented taxonomic knowledge, heterogeneous data standards and limited interoperability between infrastructures continue to constrain global assessments and policy uptake. This session will bring together initiatives within the freshwater community to relaunch and align key components of the freshwater data landscape, including the Freshwater Animal Diversity Assessment (FADA), the Freshwater Information Platform (FIP) and freshwater data mobilisation efforts in FWBON, GBIF and Catalogue of Life. We will explore how curated, realm-specific taxonomic checklists and biodiversity data can be mobilised as interoperable machine-actionable resources adhering to global Biodiversity Information Standards feeding into emerging freshwater specific infrastructures and reproducible workflows.

The session will focus on operationalising freshwater taxonomy through standards, sustainable infrastructures, and linking taxonomic backbones with trait, occurrence, and monitoring databases and conservation status information from the IUCN Red List of Threatened Species. We also invite discussion on community-driven approaches, including regional networks, e.g. the Freshwater Biodiversity Information System (FBIS), governance models, and long-term sustainability, aiming to advance a coherent, FAIR, and globally integrated freshwater biodiversity data ecosystem.

Moderators



Astrid Schmidt-Kloiber

Senior Scientist | BOKU University



Koen Martens

Royal Belgian Institute of Natural Sciences, Brussels

Session organizers



Vanessa Bremerich

Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB)



Helen Dallas

Executive Director | Freshwater Research Centre (FRC)



Olaf Banki

Catalogue of Life Foundation

A collaborative perspective on a globally standardised river basin name database

4:00 PM – 4:15 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

When categorising and aggregating data across a given landscape, Earth scientists tend to use grids, broad geographic delineations such as continents, or administrative units such as countries, counties or states. However, when the environmental characteristics and energy flows in a given region come into focus, then the drainage basin becomes an appropriate spatial unit. This is because of the topographic structure of the landscape along rivers, streams, lakes and wetlands that are laterally and longitudinally connected. Especially in light of global change, spatially pinpointing and being able to search a certain drainage basin by name - e.g. to analyse the observed and projected changes - represents a task that is important for the study of biodiversity as well as for neighbouring areas such as transnational water management. To date, though, a comprehensive and globally standardised resource of basin names is missing.

In this contribution, we will outline how such a resource could be developed, maintained and refined collaboratively on the basis of the Wikimedia and OpenStreetMap platforms and their respective communities. The Wikimedia ecosystem comprises of a suite of interconnected open collaborative platforms that already cover multiple facets of both river basins and biodiversity and that are multilingual. Wikidata in particular is a repository of semantic web-compatible general reference information whose entries include river basins, taxa, territorial entities, ecological concepts and scientific publications about any of these. Wikidata and its sister sites are also closely integrated with OpenStreetMap and citizen science efforts, and Wikidata identifiers are widely used in scholarly contexts and frequently mapped to entries in other relevant databases.

Speaker



Daniel Mietchen

Wikimedia Germany

Authors

Vanessa Bremerich, Sami Domisch

From “Lost in Translation” to Fit-for-Purpose Data: Why Ecologists and Conservation Practice Still Need groupings and cF

4:15 PM – 4:30 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

This talk highlights mismatches between biodiversity data standards and practical ecological and conservation work in freshwater ecosystems, with a botanical focus. Drawing on NIVA's experience, we demonstrate why ecologists still need taxonomic groupings such as *agg.*, *col.*, *vel.*, */*, *sensu lato*, and *sensu stricto*. The talk gives aims to show that honest and effective biodiversity reporting requires standards that allow uncertainty to be expressed—rather than erased.

We will present graphs illustrating how substantial amounts of monitoring data can effectively be cut out from time series, as cF and grouping annotations are not supported in GBIF. At the same time, we emphasize that management categories should be applied as downstream filters, not embedded in global biodiversity infrastructures.

Further, as hybrids and unresolved taxa are common across many organism groups. Mappers are indirectly pushed towards making uncertain species records that are not flagged for uncertainties, due to requirements linked to red listing, Water Framework Directive monitoring, and invasive species mapping. Combined with a lack of verifiability, this creates serious challenges for biologists—both in carrying out their commissioned work and in submitting truthful, high-quality data to GBIF.

Speaker



Bernhard Kløw Askedalen

Research assistant | NIVA

Authors

Nathalie Marquesin Risbakk

How does human impact change species diversity and composition in 50 types of inland water habitats?

4:30 PM – 4:45 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

The revision of inland water habitats for the European Nature Information System (EUNIS) is based on the common types used for the EU Water Framework Directive and related habitat types of the Habitats Directive, as well as on more specific habitats of the Habitats Directive. The presentation describes natural biodiversity for major biological groups in 50 inland water habitat types and how these groups change with human impact. Altogether 23 standing water habitat types and 27 running water habitat types are presented in terms of their natural abiotic conditions and their truly aquatic biological communities, including species richness and characteristic, common and dominant taxa of phytoplankton, aquatic vegetation and fish in standing waters, as well as benthic algae, aquatic vegetation, benthic invertebrates and fish in running waters. The data are based on harmonised taxa lists for each major taxonomic group. While habitats in reference condition have distinct biological communities shaped mainly by geology, altitude, and size, impacted habitats converge towards degraded states with fewer characteristic species due to major pressures, such as nutrient pollution, hydromorphological change and water abstraction. With the EU Biodiversity Strategy 2030 and Nature Restoration Regulation, Europe now has a shared scientific framework to support the monitoring of whether restoration actions effectively improve inland water ecosystems. The work is published as a report from the European Topic Centre for Biodiversity and Ecosystems funded by the European Environment Agency (EEA).

Speaker



Anne Lyche Solheim

Chief scientist | NIVA

Authors

Petra Mutinova, Tor Erik Eriksen, Ute Jandt

Discussion

4:45 PM – 5:00 PM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Moderator



Astrid Schmidt-Kloiber

Senior Scientist | BOKU University

SYM42B - Solutions for Research Collection Management Systems Challenges

4:00 PM – 5:00 PM | Location: SAL A

COLLECTIONS DIGITISATION & MANAGEMENT

Collection Management Systems (CMS) face evolving challenges in managing, curating, publishing, and integrating increasingly complex research data from diverse sources. This session seeks contributions that address practical and conceptual solutions for modern, research-oriented CMS, with a focus on enabling richer, interoperable, and sustainable data flows across the global biodiversity data landscape. We welcome talks on data mobilization for research infrastructures, e.g., GBIF, OBIS, and DiSSCo. We especially seek submissions focused on emerging standards such as Darwin Core Data Package (DwC-DP), and alignment with TDWG and related domain standards to support semantic consistency and FAIR principles. Presentations exploring implications of the Darwin Core Conceptual Model - including clearer treatment of relationships among material entities, organisms, and occurrences - are encouraged. We also invite contributions on the maintenance and use of persistent identifiers and the relationship between material samples and primary literature in the context of material sample management systems and their data consumers. Further topics include machine-assisted curation workflows, community-driven development, capacity building, long-term maintenance, CMS-as-a-Service, and governance models. By bringing together CMS developers, data architects, curators, and aggregators, the session aims to advance practical solutions to current and emerging biodiversity data exchange challenges.

Moderators



Jonas Grieb

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research



James A. Macklin

Agriculture and Agri-Food Canada

Session organizers



Christian Bölling

Museum für Naturkunde Berlin - Leibniz Institute for Evolution and Biodiversity Science



Sam Leeflang

Lead Developer Collection Data Infrastructure and Access team | Naturalis Biodiversity Center



Wouter Addink

Manager Software Development | Naturalis Biodiversity Center



Claus Weiland

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main

Authors

Michaela Grein, Volker Lohrmann, Etta Grottrian

Old habits and new visions: Exploring the RECODE data model, future integration and data enhancement

4:00 PM – 4:12 PM

COLLECTIONS DIGITISATION & MANAGEMENT

The NHM London RECODE programme tasked with developing and implementing a new collections management system has been actively implementing since late 2023. With a vision of a sustainable and community focused system. RECODE has tackled a conscious shift to a more normalised data model and a focus on the use of bidirectional links to better represent and manage the NHM collections data and museum procedures.

In this talk we will share key insights into the new Dosis system, considerations around the developed data model and key functionality that we will leverage. As the system stabilises and data migration tests bring real data into the system the programme is also able to reflect on the impact of the data model and how it performs against the complexity, inconsistency and scale of NHM's data.

Presentation points will include:

A complex object model that supports derived objects, groups, parts and preparations as a significant step in disambiguating objects from specimens and from collections management procedures.

In system access to the data model has provided better visibility of system databases, classes, fields, field types, relationships and configuration dependencies. The DM-viewer supports understanding the Dosis system and is expected to provide significant core functionality in terms of visualising data standards, search pathways, permissions and user insight into data structures.

Internal system data validation to support data integrity, quality assurance (QA) and quality control (QC) to support heterogeneous data and data standard constraints.

The ambition to surface validated data from authority data repositories (e.g. GBIF, Bionomia) as well as agent data cleaning to the user within the system to help continuous enhancement and enable data round tripping

Speaker



Steen Dupont

Programme Manager- RECODE | Natural History Museum London

Authors

Recode Team

The Unruliness of Higher Geography: What is it doing for us?

4:12 PM – 4:24 PM

COLLECTIONS DIGITISATION & MANAGEMENT

In natural science collections, the concept of a collecting locality intermixes human and physical geography. This breadth is necessary to facilitate sharing location information, not only for research but also to ensure legal compliance and to encourage ethical practices. Geography is a primary facet for discovery and analysis of collections-based data, and we increasingly need ways to share geographic data accurately regardless of how precise it is. For example, a mollusk specimen collected from a given watershed in 2020 likely has location data that includes precise coordinates recorded from a GPS unit, and a mollusk specimen collected in 1920 may *only* have the watershed recorded, but either way, “watershed” may be the most useful concept. Geography is fluid, no matter what flavor of geography is being considered. Maps get redrawn and tectonic plates move, which means that geographic concepts are most useful when associated with time.

In this talk, we review existing conceptual models for capturing and sharing geographic information from common Collections Management Systems, data standards, and data aggregators. We then highlight pain points within these models, and illustrate their impact on discoverability. For example, although the “Atlantic Ocean” is a fundamental geographic concept, finding specimens that were collected in the Atlantic Ocean can be challenging. These pain points are connected by a throughline of forcing physical and administrative geographies together. We close by suggesting the need for a new TDWG task group under the Darwin Core Maintenance Group to assess the current terms and recommend modifications or additions.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Erica Krimmel

Biodiversity Information Scientist | independent

Authors

Teresa Mayfield-Meyer

Mobilizing rich, interoperable, and secure fossil data from underserved collections using Symbiota

4:24 PM – 4:36 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Symbiota software underlies over 70 data portals that are used as collections management systems and data aggregators. While U.S. federal funding catalyzed the use of Symbiota for the digitization of biological collections, Symbiota’s uptake by the paleontological community was previously limited by a lack of functionality for paleontological context data. To address these needs, new software enhancements for managing and publishing fossil specimen data are now available in Symbiota 3.4. Notably, these enhancements implement full functionality for Darwin Core GeologicalContext terms in Symbiota’s public interfaces, at the database level, and for data harvesting. Substantial development efforts have been leveraged to reconfigure the handling of chronostratigraphic information to align output data with GBIF’s GeoTime vocabulary and improve downstream interoperability. These features are complemented by revised documentation available in the Symbiota Docs user information resource and via a new Paleo Data Knowledge Hub. In addition to making these enhancements available to all Symbiota portals, a new Symbiota portal, the “Paleo Data Portal” (PDP), was launched to provide secure infrastructure to fossil collections that have not yet received support through U.S. federal digitization funding, allowing these collections to engage in the global paleo data ecosystem. The PDP currently serves as a testing ground to better understand gaps in existing data standards and digitally available taxonomy to inform a forthcoming “Paleo Roadmap” intended to guide future paleo-related cyberinfrastructure improvements. All Symbiota software enhancements are synergistic with greater community-wide efforts to improve the utility of existing cyberinfrastructure for paleontological research.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Erica Krimmel

Biodiversity Information Scientist | independent

Authors

Holly Little, Talia Karim, Erica Krimmel, Nico Franz, Edward Gilbert, Gregory Post, Nikita Salikov, Logan Wilt

Discussion

4:36 PM – 5:00 PM

COLLECTIONS DIGITISATION & MANAGEMENT

Moderators



James A. Macklin

Agriculture and Agri-Food Canada



Jonas Grieb

Senckenberg – Leibniz Institution for Biodiversity and Earth System Research

CO5 Digital Tools for Data Discovery, Resolution and Exchange

4:00 PM – 5:00 PM | Location: SAL C

Moderators



Ben Norton

Yale University



Andre Zerger

Director | Atlas of Living Australia

How Literature Services can Support & Benefit from Biodiversity Publication and Data Standards ?

4:00 PM – 4:15 PM

AI AND ROBOT READY

Rapid growth of biodiversity literature makes it difficult to retrieve precise, relevant, and interoperable knowledge. Traditional bibliographic catalogues mainly rely on metadata and abstracts, limiting fine-grained discovery and large-scale reuse.

This presentation explores how literature services can support biodiversity publication and data standards. Using BiodiversityPMC, developed within the SIB Literature Services (SIBiLS), we demonstrate how structured, machine-readable publications enable advanced indexing, semantic enrichment and concept-based retrieval across heterogeneous sources.

The SIBiLS platform integrates taxonomic treatments from Plazi, biodiversity-related documents from Zenodo and full-text articles from biodiversity journals (e.g., EJT, Pensoft). By leveraging structured formats, persistent identifiers and domain ontologies, it transforms publications into AI-ready contents that can be cross-linked and queried at scale. We will also demonstrate how pre-computed literature selection through Triage supports Biodiversa+ communities.

Through simple examples, attendees will perform advanced searches via the web interface, explore semantic annotations and test selected API calls for programmatic access, including MCP calls from AI bots.

Speaker



Patrick Ruch

HES-SO & SIB Swiss Institute of Bioinformatics

Authors

Emilie Pasche

From NetCDF to Web Maps: Building an Open-Source Web GIS for Biodiversity Data Cubes

4:15 PM – 4:30 PM

DIGITAL TWINS & EMERGING INFRASTRUCTURES

Essential Biodiversity Variables (EBVs) provide a standardized framework for monitoring global biodiversity change and supporting international biodiversity assessments. However, there is still a lack of accessible web-based infrastructures that allow multidimensional biodiversity datasets to be explored consistently across spatial and temporal scales. This limits the usability of biodiversity information for research, policy support, and decision-making.

In this presentation, we introduce the EBV Visualizer, a lightweight open-source Web GIS for the interactive exploration of biodiversity data cubes. The application complements the EBV Data Portal by enabling users to visualize and analyze multidimensional raster datasets directly in the browser. The visualizer is built entirely with open-source technologies using vanilla JavaScript, Leaflet.js for interactive web mapping, and Chart.js for temporal analytics and visualization. The backend infrastructure relies on THREDDS Data Server and ncWMS to serve NetCDF datasets efficiently through dynamic tile rendering and caching of multidimensional raster data.

The platform enables users to explore biodiversity variables across space, time, and thematic dimensions, while supporting interactive analyses such as temporal trends for countries, protected areas, or other regions of interest. By combining OGC data standards (netCDF, WMS) with lightweight web technologies, the EBV Visualizer lowers technical barriers to biodiversity data access and contributes to open, interoperable, and FAIR biodiversity infrastructures.

Speaker



Christian Langer

Web-GIS Programmer | German Centre for Integrative Biodiversity Research – iDiv / University of Halle-Wittenberg

Authors

Emmanuel Ocegüera-Conchas, Néstor Fernández, Lina M. Estupinan-Suarez, Andrés Mármol-Guijarro, Henrique M. Pereira

Role of Large Language Models for Biodiversity Data Discovery, Integration, and Curation: A Case Study of Dehradun Herbarium

4:30 PM – 4:45 PM

COLLECTIONS DIGITISATION & MANAGEMENT

The Dehradun Herbarium (DD), housed at the Forest Research Institute (ICFRE), Dehradun, is India's second-largest herbarium, comprising over 330,000 specimens including 1,300 invaluable Type specimens collected since 1890. Despite ongoing digitization efforts since the late 1990s and the launch of the Digital Herbarium Specimen Database (ddherbarium.icfre.gov.in), significant challenges persist in metadata completeness, specimen label transcription, taxonomic data integration, and global discoverability of these biodiversity records. This study presents a case study exploring how Large Language Models (LLMs) can enhance biodiversity data discovery, integration, and curation workflows at the DD Herbarium. We demonstrate LLM applications for: (i) automated extraction and structuring of specimen label data into Darwin Core-compliant fields; (ii) natural language querying of herbarium records to improve researcher accessibility; (iii) AI assisted metadata gap detection and enrichment; and (iv) semantic integration of DD specimen data with global repositories such as GBIF. Results indicate that LLM assisted pipelines substantially improve metadata completeness and reduce manual curation effort, particularly for legacy and handwritten label transcription. We also examine challenges including hallucination, taxonomic bias, provenance tracking, and ensuring reliability of AI generated metadata. Our findings suggest that LLMs, when integrated responsibly with structured biodiversity data standards, offer transformative potential for mobilising underutilised natural history collections and making forestry biodiversity data more discoverable, reusable, and actionable for conservation science.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Neelesh Yadav

Forest Research Institute Dehradun India

Local scientific name resolution using custom data sources

4:45 PM – 5:00 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Scientific name verification is crucial for biodiversity informatics workflows, enabling researchers to reconcile name-strings against authoritative taxonomic databases, detect misspellings, resolve synonyms, and retrieve accepted names and vernacular names. GNverifier is a tool that provides this capability with high quality and a throughput exceeding 2,000 name-strings per second. For several years, it has operated exclusively as a centralized web service.

Here, we present GNverifier as an easily installed local application that supports custom datasets, including private ones, not available in the centralized service. Data can be seamlessly converted from Darwin Core Archives, CSV, TSV, pipe-delimited files, Catalogue of Life Data Packages, and plain scientific name-string text files.

This functionality is driven by two new applications: `sf` and `gndb`. The `sf` tool converts the aforementioned data formats into an intermediate Species File Group Archive (SFGA) format, which `gndb` then imports into the GNverifier database. While standard GNverifier datasets are downloaded automatically by `gndb`, custom and private datasets are imported directly from the local file system. Once the database is built, the local GNverifier instance exposes the same REST API and web interface as the centralized service, but provides the specific datasets required for a given research project. This enables workflows that demand strict data sovereignty or reproducibility.

GNdb is open-source under the MIT license at <https://github.com/gnames/gndb>. Pre-built binaries for Linux, macOS, and Windows are available on the releases page, and the tool can also be installed via 'go install' or compiled from source.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Dmitry Mozzerin

University of Illinois Urbana-Champaign

SYM26 - From WDSRPD to GeoSchemes: Recording Biodiversity Distributions for Science and Policy

4:00 PM – 5:00 PM | Location: FORUM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

The TDWG GeoSchemes task groups are finalising an updated World Geographic Scheme for Recording Species Distributions (WGSRPD). The new version expands the scope of the original standard to cover all taxonomic groups, not only plants, and to include the marine domain. GeoSchemes will provide a versioned, and citable framework for representing biogeographic areas following political boundaries.

By bringing together standard developers and implementers / users, the session will foster dialogue on how GeoSchemes can support interoperable, policy-relevant biodiversity data across terrestrial and marine realms. It aims to present recent developments in the TDWG GeoSchemes from WGSRPD, as well as their use in recording, exchanging, and interpreting biodiversity distribution information for scientific research, policy, and data management.

The session explicitly invites users of biodiversity distribution schemas—across data infrastructures, collections, publishing workflows, and policy-facing applications—to share practical implementation

experiences. We expect that contributions will highlight current uses, integration challenges, and limitations encountered in real-world contexts. The discussion aims to identify emerging needs, inform maintenance and governance priorities, and help delineate future directions for WDSRPD and GeoSchemes as community-driven TDWG standards.

Moderators



Francisco Pando

Researcher | CSIC / GBIF.ES



Serge Gofas

Honorary associate | Universidad de Málaga

Session organizers



Pierre Bonnet

Ecologist | CIRAD



Britt Lonneville

Ghent University



Visotheary Ung

Research Engineer | Institut de Systématique, Evolution, Biodiversité (ISYEB) CNRS-MNHN-SU-EHE-PSL-UA

Authors

Alexandre Monro

Introduction

4:00 PM – 4:02 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Speaker



Francisco Pando

Researcher | CSIC / GBIF.ES

Mapping geographical distributions in the World Register of Marine Species

4:02 PM – 4:14 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

The World Register of Marine Species (WoRMS) is an authoritative database hosted at the Flanders Marine Institute (VLIZ), aiming to provide a comprehensive list of all marine organisms. Associated information on the species includes “Documented distribution” backed by published sources. Since the beginnings of WoRMS this representation has been hampered by the lack of an agreed set of geounits, of homogeneous extension. For this reason the Geoschemes task group of TDWG was set up to undertake a proposal, which will parallel the current “World Geographic Scheme for Recording Plant Distribution” used on land.

Purposes of this set of geounits is twofold: (1) to provide a sensible visual representation of a species’ range, also able to handle imprecise country-level records from the old literature and (2) to provide bounding boxes for the retrieval of distribution data and species lists, particularly at the country level.

The countries’ Exclusive Economic Zones (EEZs) were either taken as recording units or subdivided where needed, and the High Seas were subdivided according to physiographical features in the bathyal and abyssal stages respectively. A hierarchical set of five Level-1 units, 26 Level-2 units, 232 Level-3 units and 536 Level-4 units is currently under review. Subdivision of the large or complex EEZs followed Marine Ecoregions of the World or the IHO Limits of Oceans and Seas when appropriate, and existing local schemes were adopted for Europe and Australia. Attempts to distinguish habitats were disregarded in view of the advantages of having a standard representation for all the species: therefore, representing a polygon as a distribution record assumes that the species exists “somewhere” inside, not everywhere.

Speaker



Serge Gofas

Honorary associate | Universidad de Málaga

Authors

Britt Lonneville, Nicolas Bailly

Marine Regions as a backbone for the TDWG Marine GeoScheme

4:14 PM – 4:26 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Marine Regions provides a standard of marine georeferenced locations, boundaries and regions for scientific and educational purposes. At its core is a hierarchical gazetteer that serves as a geographic backbone for a wide range of applications, from biodiversity infrastructures such as the World Register of Marine Species (WoRMS), to global fisheries initiatives such as Global Fishing Watch.

Beyond place names and boundaries, Marine Regions brings together diverse geospatial ecological classifications relevant to the recording and interpretation of marine biodiversity distributions. These include global and regional frameworks, such as the Marine Ecoregions of the World (MEOW), Longhurst provinces, DFO Marine Bioregions and IMCRA Provincial Bioregions. In addition, it develops and maintains data products that are important for distribution recording, but are not always readily available or consistently managed elsewhere, including maritime boundaries such as Exclusive Economic Zones (EEZ). Marine Regions has also integrated the TDWG Terrestrial Scheme since 2015.

This experience positions Marine Regions as an active contributor to the development of the proposed TDWG Marine GeoScheme within the broader GeoSchemes framework. Its contribution is twofold: providing relevant source data, including Exclusive Economic Zones and IHO Sea Areas, and contributing technical expertise on how to build, version and maintain such a standard in a consistent and sustainable way. Once finalised, the TDWG Marine GeoScheme will be integrated into Marine Regions, where its areas will receive persistent unique identifiers (PIDs), Marine Regions Geographic Identifiers (MRGIDs), enabling stable, unambiguous citation, machine to machine linkage, and use in (biodiversity) data publication.

Speaker



Britt Lonneville
Ghent University

Authors

Lawrence Whatley, Michelle Guevara, Bart Vanhooorne

The new World Geographical Scheme for Recording (not only) Plant Distributions: process, progress and challenges

4:26 PM – 4:38 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

The *World Geographical Scheme for Recording Plant Distributions* (WGSRPD) has been one of the most widely used TDWG standards, supporting biodiversity data management, floristic synthesis, conservation assessments and scientific publishing for more than two decades. Its combination of administrative boundaries and biogeographical concepts within a hierarchical geographical framework proved both practical and broadly interoperable.

However, the current version of WGSRPD, published in 2001, requires substantial revision to remain relevant and useful. Changes in political boundaries, evolving community practices, the emergence of global biodiversity infrastructures, and new expectations regarding interoperability, transparency and machine-actionability all demand an updated approach.

The TDWG GeoSchemes Task Group is developing a new framework intended to extend and modernize WGSRPD beyond its original plant-focused scope. The work aims to preserve compatibility with the conceptual strengths of WGSRPD while introducing clearer governance, persistent identifiers, explicit versioning, improved documentation, and alignment with contemporary biodiversity data standards and policy needs.

This presentation will provide an overview of the principles guiding the revision process, the structure and intended outputs of the new GeoSchemes framework, current progress of the terrestrial domain work, and some of the main conceptual and technical challenges being addressed. It will also outline remaining work and opportunities for community participation and review.

Speaker



Francisco Pando
Researcher | CSIC / GBIF.ES

Digitising Plant and Fungal Collections at Royal Botanic Gardens, Kew: Applying International Standards to Support Data Sharing and Geospatial Analysis

4:38 PM – 4:50 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

RBG, Kew holds over 8 M plant and fungal materials across 12 diverse collections and recently completed a major collection-based initiative to create a comprehensive digital catalogue, implement an Integrated Collections Management System (ICMS), and launch the Kew Data Portal. ICMS enhances data standardisation, curation, accessibility, and interoperability, enabling a strategic opportunity for integrated analysis across collections within a single digital platform. We present our experience in application of WGSRPD standards. Data derive from migrated legacy databases with varying standards, as well as entry of new records. A semi-automated process assigns TDWG levels based on ISO 3166 country codes for original provenance. ICMS currently holds over 7.8 M records linked to physical materials. While some materials are of historical significance dating back to BCE, collection events from original localities range from the late 17th century, with the most from the 19th century onward. Over 1.44 M records could not be assigned to a country or territory without further investigation, while a high proportion met international data standards: ISO codes - 81.3%; TDWG L1 - 80.4%, L2 - 71.6%, L3 - 53.1%, and L4 - 52.9%. Provenance spans all 249 countries and territories, mainly from Africa, followed by Asia-Tropical and Southern America. Key challenges include missing hierarchical locality data, historical references, geopolitical boundary changes, transcontinental nations, multi-jurisdictional archipelagos, absent geocoordinates, language and translation inconsistencies, and variable user familiarity with the TDWG system. Our results highlight both the value and the limitations of applying contemporary standards to heterogeneous, historically complex biodiversity data.

Speaker



Udayangani Liu

Science Data Resources Manager | Royal Botanic Gardens, Kew

Authors

Robert Turner, David Quinn, Evgeniy Meyke

Discussion

4:50 PM – 5:00 PM

DARWIN CORE, STANDARDS DEVELOPMENT & IMPLEMENTATION

Moderator



Francisco Pando

Researcher | CSIC / GBIF.ES

DS49 - What Makes Biodiversity Data “Robot-Ready”? Standards, Infrastructure, and AI

4:00 PM – 5:00 PM | Location: ODIN

AI AND ROBOT READY

This discussion session will examine what technical and infrastructural conditions are required for biodiversity data to become “robot-ready” usable by automated systems and AI pipelines without manual intervention. Despite the scale of global biodiversity repositories, many datasets remain difficult for automated systems to interpret because key metadata fields are incomplete, data quality indicators are rarely available, and existing standards lack terminology designed for AI workflows. As a result, researchers and data scientists often spend significant effort preparing data before it can be used in computational pipelines. The session will begin with short invited perspectives from biodiversity data infrastructure providers, standards developers, and AI researchers. These perspectives will highlight gaps in metadata completeness, missing data quality metrics, and the lack of machine-actionable descriptors needed for automated analysis. The discussion will then focus on identifying a community framework for robot-ready biodiversity data, including: 1) essential metadata fields, 2) standardized data quality metrics, 3) terminology for AI and machine-generated annotations, and 4) priorities for extending existing standards. The goal of the session is to identify community priorities and outline a roadmap toward robot-ready biodiversity data standards. Outcomes will inform a collaborative community paper and future standards work within the TDWG community.

Discussion leaders



Henry L. Bart

Professor Emeritus of Ecology and Evolutionary Biology & Curator Emeritus of the Royal D Sutkus Fish Collection Collection | Tu...



Yasin Bakır

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute

Friday, September 25, 2026

CHECK-IN

8:00 AM – 8:30 AM | Location: 1st Floor

SYM47A - Large Language Models for Biodiversity Data Discovery, Integration, and Curation. Part A: AI-enabled knowledge creation and extraction

8:30 AM – 10:00 AM | Location: SAL A

AI AND ROBOT READY

Large language models (LLMs) and other generative AI systems are changing how researchers interact with data and digital infrastructure. In biodiversity informatics, these technologies create new opportunities for improving data discovery, automating metadata generation, assisting with data curation, and enabling natural language interfaces for biodiversity databases.

This session will explore emerging applications of large language models in biodiversity data infrastructure and research workflows. Topics may include natural language querying of biodiversity repositories, extraction of biodiversity information from literature and specimen labels, AI-assisted metadata generation and validation, and integration of LLM-based tools with biodiversity data standards such as Darwin Core.

We encourage case studies demonstrating how LLM-based tools can support biodiversity data discovery, improve metadata completeness, and assist with data publishing and curation workflows. Contributions addressing challenges such as reliability, provenance, bias, and integration of generative AI with structured biodiversity data are also welcome.

The goal of the session is to examine how large language models may reshape biodiversity data access and curation while identifying opportunities and limitations for their responsible use.

Moderators



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute



Bahadır Altıntaş

Abant İzzet Baysal University

Introduction to session: AI-enabled knowledge creation and extraction

8:30 AM – 8:36 AM

AI AND ROBOT READY

Speaker



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute

The Planetary Knowledge Base: an infinite solutions engine converting data into action for nature

8:36 AM – 8:48 AM

AI AND ROBOT READY

The accelerating nature crisis demands more than fragmented evidence and retrospective reporting. It requires trusted systems that can turn biodiversity knowledge into action. The Planetary Knowledge Base (PKB), led by the Natural History Museum, is being developed as an AI-enabled global infrastructure to meet that need. Its success will depend on wider community participation, bringing together collections institutions, researchers, source communities, data infrastructures and users to ensure biodiversity evidence is connected, governed and applied in reciprocal ways. The PKB combines a biodiversity foundation model, trained across literature, specimen records, observations, images and genetic and environmental data, with a provenance-aware knowledge graph linking species, specimens, traits, interactions and change.

This emerging neuro-symbolic architecture is designed to extract information at planetary scale while grounding every inference in traceable evidence. Through retrieval-augmented generation, APIs and natural-language interfaces, the PKB aims to make biodiversity knowledge accessible to researchers, policymakers, communities and businesses, while reducing hallucination and opaque AI decision-making.

The PKB is conceived as an infinite solutions engine. This living infrastructure that can absorb data, test hypotheses, identify relationships and support scenario exploration for nature-positive decisions. Coupled with ecological modelling, it is intended to move from evidence synthesis to intervention design, evaluating restoration, migration and resilience planning. By linking centuries of natural history evidence with AI, digitisation and global data networks, the PKB seeks to transform collections into actionable intelligence for planetary stewardship.

Speaker



Vincent Smith

Natural History Museum

Authors

Ben Hartley

SCRIBE: Structured Collection Record Interpretation and Bio-entity Extraction

8:48 AM – 9:00 AM

AI AND ROBOT READY

There are over 120 million objects housed in natural history collections in the UK. With DiSSCo UK, 90+ institutions will be coming together, in order to make over 50% of UK natural history collections digitally available and accessible.

As well as specimen imaging, a crucial goal is the building of datasets extracted from the millions of textual records. These records can include labels and index cards associated with specimens, collection registers and more. Our aim is to provide a service to aid the automatic extraction of structured data from collection records.

SCRIBE (Structured Collection Record Interpretation and Bio-entity Extraction) is a new platform, being developed by a team at the NHM UK for automatic data extraction. The proposed software has the potential to revolutionise data extraction, for curation, as well as ecological analysis, by utilising Artificial Intelligence (AI), and providing an easy-to-use interface, to help unlock a plethora of data. While traditionally this task would involve manual transcription, or traditional OCR methods, taking up days and weeks, the proposed tool will for some collections, reduce this to minutes. The tool follows a computer vision and LLM based workflow that utilises customised prompts depending on visual differences within images uploaded by users, enabling users to upload datasets from varied collections concurrently.

SCRIBE is a DCMS (UK Department for Culture, Media, Sport) funded and externally-funded program that will sit within the DiSSCo UK workflows. The software will be shared with institutions and evaluations will be jointly collected to further improve the online tool. Finally, the SCRIBE online tool will be supported in-house, helping to scale AI-assisted digitisation across museum and heritage sectors.

Speaker



Arianna Salili-James

Natural History Museum UK

Authors

Ben Hartley, Ginger Burns, Rachel Wiles, Al Turner

From Historical Oology Cards to Structured Biodiversity Data: Multimodal LLM Transcription, Uncertainty Metrics, and Human Review

9:00 AM – 9:12 AM

AI AND ROBOT READY

Historical biodiversity collections contain vast quantities of ecological information embedded in handwritten and typed archival records that remain difficult to access at scale. This project explored the use of multimodal large language models (LLMs) to transcribe and structure oological information found in historical specimen cards from the Cornell University Museum of Vertebrates. These cards record collectors' data associated with bird egg sets, breeding phenology, nest materials, and habitat descriptions collected between the late 19th and mid-20th centuries.

Using GPT-4o and Claude Sonnet 4.6-based workflows, I evaluated methods for extracting both literal transcriptions and semantically structured biodiversity data from heterogeneous card layouts containing handwriting, typewriting, annotations, and mixed narrative content. Beyond transcription accuracy, my project investigates methods for estimating review effort and transcription reliability using token-level uncertainty signals and a human-in-the-loop validation workflow. I developed a "hesitation" score based on token-level probabilities and found that this score correlates with human correction effort and may support scalable triage of records requiring expert review.

This project demonstrates how historical narrative records can become machine-actionable biodiversity data while preserving provenance, interpretability, and curatorial oversight. I discuss implications for biodiversity data mobilization and the integration of semantically enriched collection records into standards-based biodiversity infrastructures.

Speaker



Grete Pasch

CUMV

Using Large Language Models to enhance biodiversity knowledge extraction: From taxonomic treatments to graph-based knowledge systems

9:12 AM – 9:24 AM

AI AND ROBOT READY

Taxonomic treatments contained in checklists, floras, monographs and ecological literature compile centuries of detailed documentation on the world's biodiversity. They typically appear as structured, peer-reviewed sections within scientific publications describing the taxonomy, morphology, distribution and ecology of a taxon, representing valuable and authoritative resources on past and present biodiversity.

Although a large collection of treatments is now freely available as Findable, Accessible, Interoperable and Reusable (FAIR) digital objects, ecological information contained within these data, such as habitat notes, remains largely underutilised, leaving untapped a vital source of knowledge for ecological modelling and conservation assessment.

Here, we present a practical framework for structuring, extracting and mapping habitat information stored in taxonomic treatments using an Artificial Intelligence (AI)-assisted, ontology-aware knowledge graph system (a formal, machine-readable representation of domain entities and their relationships). This work forms part of a broader initiative aimed at improving the mapping of data-poor species, with a focus on vascular plants.

We propose to (a) create a knowledge graph to store evidence related to plant species habitats, (b) link this graph with semantic environmental and geospatial layers, and (c) connect this system to a large language model (LLM) to automate, query and reason over this complex knowledge base.

By reducing hallucinations and linking extracted habitat information to spatial layers, we demonstrate a viable pathway towards more reliable and transparent geospatial AI applications, supporting a more responsible use of generative AI for species distribution and habitat mapping.

Speaker



Ian Ondo

Programme Officer | United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC), Cambridg...

Authors

Rosa Dunkley, Ben Leo Hollis, Neil Burgess

Automated data transcription and Darwin Core field itemization from insect specimen labels using machine learning

9:24 AM – 9:36 AM

AI AND ROBOT READY

A mass digitization project of pinned insects is currently undertaken at the Smithsonian National Museum of Natural History (NMNH). With a conveyor system, up to 1,500 specimens and all associated data labels (two to several) are digitized each day. To mobilize the primary biodiversity data from these specimens/labels, adding them to the institutional database, and sharing them through GBIF, the label data need to be transcribed and itemized into Darwin Core fields. The two currently available options are (1) manual transcription and itemization by a human or (2) automated transcription by a tool and manual data itemization by a human. Here, we present a tool, *Labellum*, that goes further and is developed to transcribe all written information found on labels (whether handwritten or printed) from a specimen and automatically itemize all data into Darwin Core fields. The tool has been developed and trained on a dataset of labels from prior digitization projects at the NMNH. It uses Yolo (You Only Look Once) (v8.3.222) and Llama (v4) libraries and the Gemma 3 (12B) model for transcription and itemization. Initial, promising results using these libraries were obtained in the fall of 2025, and with further improvements, especially training on Darwin Core fields and institution-specific label data, *Labellum* has an accuracy of >97% transcription efficacy and a similar accuracy on Darwin Core field assignment. *Labellum* will be made available for use on a local machine in a self-contained environment without the need for installation of specific software. The tool has been trained on insect labels, but can be utilized on labels of other organism groups. This will allow other natural history collections to utilize *Labellum* in transcribing and itemizing digitized labels.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Torsten Dikow

Research Entomologist | Smithsonian National Museum of Natural History, Department of Entomology

Authors

Jessica Bird, Yaal Dryer, Aidan Caruso, Samuel Cleary, Trevor Moore, Anuraag Karunakaran, Maiwand Raheem, Alexander White, Evan Economo, James Purtill

AI AND ROBOT READY

The International Plant Names Index (IPNI) is a critical global resource for plant nomenclature, supporting biodiversity research, conservation and taxonomic data [IPNI]. However, nomenclatural acts published in taxonomic journals are still largely dependent on manual extraction to populate the resource. This is a labour-intensive process requiring IPNI staff to screen publications and manually transcribe nomenclatural details into the database, creating bottlenecks that reduce the pace of biodiversity data collection.

We developed an automated extraction pipeline that processes taxonomic PDFs and outputs structured nomenclatural data using Large Language Models (LLMs). The pipeline combines PDF-to-markdown conversion, section filtering and a two-pass extraction architecture. The first pass identifies relevant taxonomic treatments and nomenclatural acts. The second pass extracts comprehensive nomenclature information from each treatment using schemas. These schemas are designed to capture various key features including name components, authorship, publication details, type specimen information, geographic details, basionyms and other metadata such as distribution.

The pipeline supports configurable schemas, LLMs, and extraction targets. This enables adaptation to different biodiversity informatics workflows and document types.

By transforming unstructured literature data into structured database-ready formats, this project demonstrates how LLM based processes can reduce manual extraction effort and accelerate taxonomic data mobilisation.

IPNI (2026). International Plant Names Index. Published on the Internet <http://www.ipni.org>, The Royal Botanic Gardens, Kew, Harvard University Herbaria & Libraries and Australian National Herbarium. [Retrieved 21 May 2026].

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Ishaipiriyam Karunakularatnam

Research Associate | Royal Botanic Garden, Kew

Authors

Helen Hartley, William Fullwood, Nicky Nicolson

Extracting Structured Biodiversity Knowledge from Literature with AI-Assisted Workflows

9:48 AM – 10:00 AM

AI AND ROBOT READY

While AI and large language models (LLMs) are increasingly used to accelerate scientific discovery and support knowledge generation, their effectiveness strongly depends on the availability of structured, interoperable, and machine-readable data. In biodiversity and ocean sciences, however, a significant portion of valuable information remains embedded in scientific articles, reports, and grey literature, making it difficult to access, reuse, and integrate into digital infrastructures.

To address this challenge, the St Lawrence Global Observatory (SLGO) developed an AI-assisted workflow capable of extracting not only structured metadata, but also tabular datasets and sensitive information from unstructured scientific documents. OCR-based table extraction automates the conversion of tabular information embedded in scientific PDFs into structured, machine-readable datasets, including the reconstruction of fragmented multi-page tables into unified data structures. Unlike narrative text, extracted tables can be readily integrated into ETL pipelines and downstream analytical workflows, reducing manual preprocessing while improving interoperability and reuse within biodiversity and ocean data systems, such as the Ocean Biodiversity Information System (OBIS) and Global Biodiversity Information Facility (GBIF).

In this session, we will present the extraction pipeline, from document ingestion and preprocessing to metadata generation, table extraction, validation, and packaging. We will also discuss current challenges and opportunities related to AI-assisted data extraction, including reproducibility, provenance and model bias. This work highlights AI's potential to support FAIR principles while ensuring transparency and quality control.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Yagmur GULEC

Programmer analyst | St-Lawrence Global Observatory (SLGO)

Authors

Maria-Emilia Rodriguez-Cuicas, Yagmur Gulec, Benjamin Veyrac, Omnain Kutos, Simon Beauvilliers

SYM32 - May the Data Be Structured: Linking Descriptions, Identification and AI

8:30 AM – 10:00 AM | Location: SAL B

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Taxonomic identification relies on explicit, character-based knowledge derived from published descriptions. Beyond assigning names, taxonomy depends on structured representations of morphological, ecological, and other diagnostic information that make comparisons possible and reproducible. Identification keys and structured knowledge bases play a central role in organizing this information and making it accessible to both specialists and non-specialists. Standards such as Structured Descriptive Data (SDD) provide a formal framework to encode taxonomic descriptions in a computable and interoperable way. They support the generation of identification keys, ensure traceability to the scientific literature, and allow long-term reuse of curated data. In this sense, structured knowledge bases are not only tools for identification, but infrastructures for preserving and mobilizing taxonomic knowledge. Recent developments in Artificial Intelligence create new opportunities, but also new challenges. AI can assist in accessing or navigating taxonomic information, yet it still depends on well-structured, high-quality data. This session will focus primarily on the role of SDD and structured knowledge bases in identification workflows, and will explore how they can interact with emerging AI approaches while maintaining transparency, scientific rigor, and interoperability.

Moderators



Adeline Kerner

National Centre for Scientific Research (CNRS)



Wouter Koch

Senior adviser biodiversity informatics | Norwegian Biodiversity Information Centre

Ctrl+Alt+SDD: A Historical Review Toward a Standard Reboot

8:30 AM – 8:50 AM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Structured Descriptive Data (SDD) is a TDWG standard, ratified in 2005, that structures taxonomic descriptions in a machine-readable format. SDD aims to standardize descriptive biodiversity data while preserving the richness and complexity of taxonomic knowledge. Built around descriptive concepts, characters, states, and semantic relationships, SDD allowed adaptation to different taxonomic contexts.

This flexibility became both a strength and a limitation. SDD allows users to freely create additional concepts and vocabularies, enabling rich and highly adaptable descriptive practices. However, this freedom also limited interoperability: similar concepts could be created under different names depending on the application, complicating the reuse of an SDD file in another application.

SDD supports complex descriptions, multilingual versions, and additional resources, making it particularly complete. However, its XML-based structure is often perceived as verbose, technically heavy, difficult to implement, and less accessible compared to newer web technologies. As a result, SDD adoption progressively declined, and today, only a handful of projects still actively use it.

Interest in SDD recently re-emerged following discussions during the TDWG 2023 Unconference on the future of descriptive data standards. The objective is not simply to revive the original standard, but to rethink how structured descriptive data could better support interoperability, FAIR principles, and modern workflows.

Current discussions around a possible “SDD 2.0” emphasize greater plasticity, modularity, and easier integration with modern biodiversity data infrastructures while preserving the conceptual completeness and adaptability that made the original SDD valuable.

Speaker



Adeline Kerner

National Centre for Scientific Research (CNRS)

Authors

Elie Saliba, Regine Vignes Lebbe

Structured identification keys complementing AI: closing the corpus bottleneck with AI-assisted digitisation

8:50 AM – 9:04 AM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

AI image recognition and structured identification keys are complements, not competitors. Image models narrow the candidate set from thousands of species to a handful; a well-formed key then guides the user the rest of the way. But the structured keys themselves remain the bottleneck - building a new key by hand is a large, complex task even for experts, and most curated taxonomic knowledge sits in printed books in non-machine-readable form.

We present a working pipeline that addresses this bottleneck. AI-driven extraction and audit produce Clavis JSON, a deliberately minimal capturing format with built-in correctness guarantees. Each character state captures a single trait, sibling states together cover every value the source asserts without overlap, shared traits are stored on the parent taxon, rare variants stay reachable through frequency values, and an automated checker confirms references resolve and every species ends up with a distinguishing combination of traits. Clavis is polythetic by design (no fixed paths) and uses the same taxon/character/state primitives as SDD; it was built for verifier simplicity and a low barrier for authors, and is mappable to SDD where downstream consumers need it.

The resulting key runs in an open web key player, which has been integrated into a working demo version of Norway's national species-recognition app, where image-based AI provides candidates that the key then refines. The engineering of the AI-assisted digitisation pipeline behind these files - detailed methods and costs - is treated in a companion talk in "AI for Biodiversity Data" (SYM30).

Speaker



Wouter Koch

Senior adviser biodiversity informatics | Norwegian Biodiversity Information Centre

Beyond traditional identification keys: an example of hybrid identification method using deep learning and Xper3 keys for insect identification

9:04 AM – 9:18 AM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Taxonomic identification keys are essential tools for species identification, but remain prone to user error when morphological characters are ambiguous or numerous. Deep learning offers powerful alternatives through automatic extraction of morphological descriptors from images, yet fully automatic approaches lack the interpretability and expert validation that interactive keys provide. Here, we develop a hybrid identification framework that combines metric learning, which builds morphological representation spaces enabling nearest-neighbour identification, with interactive identification keys. We apply this approach to French ladybirds (Coccinellidae) and show that deep learning pre-filtering complements the Xper3 interactive key, reaching 90% accuracy in suggesting nearest taxa and reducing the mean number of identification steps by more than half while maintaining user control over the final identification.

This hybrid approach leverages the power of automatic image recognition as well as the interpretability of morphological keys. Its implementation uses the Xper3 knowledge bases in the SDD exchange format and the associated interactive key service (Kerner et al., 2024). For the image AI component, only the training phase remains time-consuming, as the application subsequently relies solely on nearest neighbours retrieval, enabling a low environmental cost of the hybrid method. This is generalisable to other taxa, thereby offering a scalable and accessible framework for AI-assisted identification of specimens.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Agathe Puissant

Postdoctoral researcher | Sorbonne-Université

Authors

Regine Vignes Lebbe, Romain Nattier

Specimen identification has traditionally relied on printed single-access (polytomous) keys using morphological and anatomical characters. These approaches later evolved into more flexible and interactive free-access identification keys. DNA barcoding also emerged as a method for identification based on molecular data.

Over the last two decades, automatic identification software using images or sounds (e.g., [Merlin Bird ID](#), [Pl@ntnet](#)) has become increasingly popular. During TDWG 2025, [Vignes Lebbe et. al](#) proposed several hybrid scenarios combining connectionist AI and human interpretation through free-access keys. Two of these scenarios have been implemented in 2026 by integrating AI with the [Xper3](#) identification software. Such hybrid approaches provide an explanatory dimension in addition to the automation enabled by AI. However, they lead to an open question: how can these hybrid identification processes be represented and stored for later analysis, and how can these be used to improve future identifications?

In 2015, the Xper3 development team implemented a system to record identification pathways within the SPIPOLL citizen science program called Xperience. This enabled analyses of the users' choices, interpretation errors, and collected feedback for improving the key ([Delaunay, 2019](#)). However, these records remained confidential and relied on a format created for that purpose.

We will discuss the value of a standardized format for storing and exchanging hybrid identification workflows and associated keys, as well as addressing data representation and the potential of a system for the automatic improvement of keys.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Regine Vignes Lebbe

Sorbonne Université

Authors

Elie Saliba, Regine Vignes Lebbe, Romain Nattier, Sylvain Bouquin

The Iterative Signal-Based System (ISBS): Structuring Bioacoustic Data for Interpretable Identification

9:32 AM – 9:46 AM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Passive acoustic monitoring produces large volumes of heterogeneous data, yet their integration into biodiversity information systems remains limited by the lack of structured and interpretable representations. Machine learning approaches often rely on large annotated datasets and provide limited interpretability, making it difficult to explicitly represent the knowledge supporting identification decisions.

We propose the Iterative Signal-Based System (ISBS), a descriptor-based framework designed to structure bioacoustic signals into explicit knowledge representations. The approach relies on a three-phase process (detection, description, classification) in which acoustic events are characterized using acoustics descriptors combined with contextual metadata. These elements are organized within a graph-based knowledge structure linking descriptors, semantic entities, and target classes.

This representation supports interpretable identification through symbolic comparison with previously validated cases and descriptor configurations. The approach aligns with principles of structured descriptive data by organizing observations into reusable and semantically coherent entities.

The current implementation was developed as a proof of concept on a limited dataset of 518 cetacean vocalizations. Classification was evaluated on broad acoustic categories (e.g. harmonic, impulsive, and stationary), validating the representation framework while maintaining interpretability of the decision process. The framework highlights how structured representations can capture domain knowledge and support expert-driven validation.

ISBS is designed as a preparatory layer for machine learning, enabling the structuring and enrichment of bioacoustic data prior to large-scale automated analysis.

Speaker



Cyprien Pankowski

Phd Student | Université de la Réunion

Discussion

9:46 AM – 10:00 AM

TAXONOMY, CHECKLISTS & STRUCTURED DESCRIPTIONS

Moderator



Adeline Kerner

National Centre for Scientific Research (CNRS)

SYM07 - Publishing and Communications Promoting Biodiversity, Natural History Collections, People, Data and Data Standards

8:30 AM – 10:00 AM | Location: SAL C

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Publish or perish, the catchphrase used commonly in the academic world, is increasingly relevant for biodiversity data, collections and standards. The pressure to publish applies to peer-reviewed journals, databasing and sharing information, publishing and promoting data standards, documenting best practices, highlighting expertise in collections (e.g. Bionomia), creating harvestable and linked information, and producing success metrics and new collaborations. Social media, blogs, websites and formal publications are all ways to share and amplify the needs, values and potential of our work. Can we better promote the exchange of knowledge, ideas and best practices between our leaders, biodiversity data professionals, researchers, allies and enthusiasts above and beyond our local audiences and conferences for lasting impact? Are there tools and software available to make publishing in all formats less challenging, time consuming or stressful? We invite submissions from across the globe, from any discipline, data or physical, software and standards developers, publishers, education or policy that demonstrate or explore examples of publishing impact and success within our domain.

All TDWG 2026 speakers will submit an extended abstract to the Pensoft journal*B/SS*; we encourage you to gain experience

as a reviewer for the session, and explore publishing a full paper in this or other relevant journals, like *Natural History Collections and Museomics*, *Biodiversity Data Journal* or *RIO*.

Moderators



Shelley James

Collections Manager | Western Australian Herbarium - DBCA



Ely Wallis

Engagement Team Lead | Atlas of Living Australia, CSIRO

Session organizers



Julia Percy-Bower

Database Management Officer | Western Australian Herbarium - DBCA



Deborah Paul

Species File Group (INHS), University of Illinois

Authors

Elycia Wallis, Deborah Paul

Approaches to the Automated Tracking of Taxonomic Research Activities

8:30 AM – 8:45 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Modern taxonomy is a dynamic field of research, with activities from individual scientists working on their field of special expertise to large-scale multi-national research projects. However, no dedicated and extensive overview exists that tracks the current state of and trends in taxonomic research. The systematic aggregation of data about the present taxonomic research landscape could be used to identify gaps and help policy makers, funders or research leaders to strategically address them.

As part of the European [TETTRIs project](#) (*Transforming European Taxonomy through Training, Research and Innovations*) we previously assessed the taxonomic expertise of researchers in Europe based on taxonomic publications ([De Nolf et al. 2023](#)) and specimen data ([Fichtmueller et al. 2025](#)). In the follow-up project TETTRIX (*TETTRIs neXt level for the taxonomy of tomorrow*) we will build on this experience by automatically tracking taxonomic research activities from multiple sources and aggregating them into an open platform. This will enable the compilation of landscape reports that provide a consolidated overview of active research areas as well as highlight gaps, e.g. in taxonomic or geographic coverage. The resulting insights can inform prioritization of research agendas and strategic decision-making in taxonomy, both on national and international levels.

This presentation differs from conventional talks in that the project is still in its early stages. Rather than presenting a final product or workflow, we will outline our current conceptual framework, planned data sources, and methodologies. We want to devote the second part of our time slot to engaging with the community: we welcome early feedback on our approach and encourage suggestions for additional data sources or methods.

Speaker



David Fichtmueller

Computer Scientist | Botanic Garden and Botanical Museum (BGBM) Berlin

Authors

Jana Hoffmann, Anton Güntsch

Empowering and Expanding Communities from Knowledge to Best Practices in Data Mobilization: The TaiBIF Collaborator Training System

8:45 AM – 9:00 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Taiwan Biodiversity Information Facility (TaiBIF), like many GBIF nodes, has long promoted biodiversity data mobilization through training, technical support, and community engagement. However, the growing demand for data publishing and support from local communities increasingly exceeds the capacity of small node teams. To address this challenge, TaiBIF established a training system ("TaiBIFer") that can offer paid collaboration opportunities inspired by the GBIF volunteer community model.

The TaiBIFer system cultivates translators, ambassadors, data mobilizers, and species name checkers through structured training courses connected to TaiBIF capacity-building activities. The system aims not only to certify collaborators who can support communication and technical assistance between TaiBIF and local communities, but also to strengthen a more sustainable national biodiversity information collaboration network and enhance the attractiveness of biodiversity informatics as a career pathway.

Since late 2022, TaiBIF has trained 11 translators, 21 ambassadors, 11 data mobilizers, and 19 species name checkers. These collaborators have contributed to translation projects, workshops, biodiversity data publishing, and taxonomic validation work for the Catalogue of Life in Taiwan (TaiCOL). Through this expanding network, TaiBIF has been able to extend its outreach and support capacity beyond the limitations of a small core team.

As biodiversity data publishing needs continue to grow in Asia, the TaiBIFer training system demonstrates how community-based collaboration and capacity building can help create a more resilient and sustainable biodiversity informatics network.

Authors

Chun-I Chang

Bionomia, Scholia, Wikidata and Zenodo : broadcasting biodiversity data contributions

9:00 AM – 9:15 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Promoting biodiversity data, the people and institutions that provide that data, as well as the data standards that empower the sharing and reuse of that data, is a challenge. This presentation will discuss several tools, techniques and workflows that can be used by researchers, natural history collections, and institutions to increase the visibility and impact of their contributions to biodiversity knowledge.

The tools covered will include the use of Wikidata, and its associated scholarly profile tool Scholia, to highlight researcher and institutional impact, Bionomia to assist with the showcasing of expertise in collecting, determining and attributing natural history specimens, and the platforms Wikidata, Zenodo and Wikimedia Commons as alternative repositories for research data, workflows, and best practice documentation empowering both individuals and communities to access and reuse biodiversity focused contributions.

Speaker



Siobhan Leachman

Wikimedian | Wikimedia Aotearoa New Zealand

Handling evidence: a cross-disciplinary perspective

9:15 AM – 9:30 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

In documenting biodiversity, empirical evidence dominates over intellectual evidence, and direct evidence co-exists with circumstantial evidence: for example, a wasp sitting on a tree is an example of direct evidence, while a wasp's nest on a tree is circumstantial evidence. In the same example, however, these two situations bear different signals for biological interpretation: a standalone observation of an adult wasp on a tree supports a broad range of inferences from an occasional migrant to an established resident, but an observation of a wasp nest infers presence of a breeding family and of a population.

While historically cross-disciplinary roots of TDWG are shared with at least library science, the present times show symptoms of professional encapsulation with little interaction with sister communities such as [GSC](#) or [ISO](#). In my presentation I will invite the audience to combine the in-depth field specific approaches to standards perfection with a broader outlook and seeking inspiration from other disciplines with long tradition of figuring out evidence handling – from the resource science and history as well as from criminal evidence and forensic science.

Moreover, we witness inflation of biodiversity evidence, and a number of reasons can be blamed for the trend – one is the technology-driven acceleration of evidence production, another is the Popperian scientific paradigm, where scientific evidence is needed merely to support or counter hypotheses or theories and is disrespected for its extrinsic value. The increasing popularity of AI may lead to a more balanced co-existence of analytical research with syntheses, and we may expect higher demand and also new pressures to handling biodiversity evidence through standards.

Speaker



Dmitry Schigel

Lead scientific officer | Global Biodiversity Information Facility (GBIF), Secretariat

A Network That Is Social: A Community-Based Model and Plan for Sustaining MorphoSource, a Scholarly 3D Data Repository

9:30 AM – 9:45 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

MorphoSource (<https://www.morphosource.org>) is a public web repository specializing in 3D images representing physical objects of scholarly importance, containing almost a quarter of a million datasets contributed by over 2,800 scholars and with >250 museum organizational units (from 90+ institutions) actively engaged in long-term management of contributed data. Data in the repository facilitates over 268 unique NSF awards and other funded grants, and the repository has been cited for data access or deposit in ~3,500 scholarly publications. As a data resource, MorphoSource is unique in that stakeholders from research, education, publishing, collections, and imaging communities interactively contribute, manage, and access repositied 3D data representing natural history and cultural heritage objects. The design model for this infrastructure is explicitly community-based, with an emphasis on empowering groups and resources from different domains to maximize efficiency (e.g., linking to specimen databases and aggregators for object records, improving researcher data contribution workflows, providing publications with transparent persistent identifiers, enabling museum collections staff to exercise long-term data management where appropriate, etc.). Similarly, long term support for this resource must also involve the communities that find it most useful. MorphoSource is currently transitioning to a community support model engaging different stakeholders in different ways. After considering many approaches, this model was created specifically to help sustain MorphoSource without compromising its core mission of democratizing access to data. We will discuss this model and our progress in implementing it, focusing on how we have involved our user communities in the process.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Julie Winchester

Technical Director for MorphoSource 3D Data Repository / Biodiversity Informatics Architect | Duke University / University of Kan...

Authors

Doug Boyer

Hit or miss? Creative promotion of biodiversity data and standards for success and recognition

9:45 AM – 10:00 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Let's think outside the box as to how biodiversity data, standards, collections, and their human creators can be better promoted and recognized through forms of publication. What are the journals, data aggregators, or other formats in which you publish information, and what audiences do they reach (or not)? Are they local, national or international in scope? If you contribute as a reviewer or editor, tell us about your experiences. Do they effectively link information or use standards in their publishing? Did you know there are article types that could be more effectively used, including standards, guidelines, case studies, collection descriptions, data papers, and methods? When considering authorship, who gets included or excluded? Does your leadership value non-formal publication, and if it does, what models have been successful? It's also important to share our disappointments in order to strategize how to move forward. . Please share what skills, knowledge or tools do you feel you are lacking to pitch your work. Might collaboration be the answer? While time is often a limiting resource, what else is preventing you from getting creative, gaining more experience and promoting your work for success? What tools, for example, AI and APIs, could help? What might improve our ways to track success more effectively? Let's ensure the works we publish, or the materials we develop, make a strategic impact on people, departments, organizations and funders. This introductory and interactive presentation encourages everyone to bring experiences and ideas to share.

Speaker



Shelley James

Collections Manager | Western Australian Herbarium - DBCA

Authors

Elycia Wallis, Deborah Paul, Julia Percy-Bower

CO6 AI Driven Monitoring and Species Identification

8:30 AM – 10:00 AM | Location: FORUM

CONTRIBUTED ORALS

Moderator



Javier Molina

Project Manager / TDWG 2026 Communications Manager | Atlas of Living Australia / CSIRO

Standardizing Multimodal Insect Monitoring Data for AI-Ready Pipelines: Lessons from an InsectAI Community Datathon

8:30 AM – 8:45 AM

CONTRIBUTED ORALS

The growing availability of automated insect monitoring systems is capturing huge volumes of media containing insects. Translating these heterogeneous datasets into interoperable, AI-ready resources remains a major bottleneck for biodiversity research. Within InsectAI, an international network advancing AI for insect monitoring, we organized a community "datathon" — a hackathon for implementing data exchange standards. Teams mapped eleven datasets from platforms including Mothbox, Diopsis, Lepmon, and AML traps to the Camtrap DP standard, producing (1) curated before/after examples, (2) concrete proposals to refine Camtrap DP, and (3) conversion scripts, templates, and documentation.

We summarize key data management challenges exacerbated by entomological sensors as "the five Cs": Certainty (confidently separating taxa from morphology or sound alone); Consensus (reconciling determinations from multiple experts and/or AI models); Counting (hundreds of insects entering and leaving the frame); Cropping (storing regions of parent media while retaining provenance); and Coverage (spatial, temporal, and taxonomic scope, crucial for quantifying sampling effort and true negatives).

We proposed targeted adaptations to Camtrap DP addressing each challenge, and reflect on how this community-driven process offered insights unlikely to emerge from a top-down approach. All datasets, scripts, templates, and documentation are openly available at <https://github.com/InsectAI-COST-Action/example-datasets>.

Speaker



Jamie Alison

Tenure-track Researcher | Department of Ecoscience, Aarhus University

Authors

Sanne Govært, Wouter Koch, Chirag Padubidri, Karolina Kuczkowska, Asger Svenning, Lorenzo Girotti, Valentin Stefan, Issy Russ, Hubert Szczygieł, Anna Viklund

From Standardized Field Methods to AI-Ready Knowledge: The LEPMON Project with the LAUP Infrastructure Stack for Automated Moth Monitoring

8:45 AM – 9:00 AM

CONTRIBUTED ORALS

Automated biodiversity monitoring can deliver repeatable, high-throughput observations across sites and years, but only if standardization extends beyond field methods into the data infrastructure itself. In practice, images, locality information, detections, taxonomic predictions, human annotations, expert validations, and quality-control decisions are often fragmented across files, spreadsheets, and project-specific workflows. This fragmentation increases harmonization costs, weakens comparability, and limits the creation of trustworthy AI-ready biodiversity data.

The LAUP software stack, developed for the German LEPMON project, addresses this by linking deployments, localities, media records, detections, AI-assisted classifications, human annotations, expert review, and publication outputs in one traceable workflow. Rather than treating standards as a final export step, LAUP makes structured curation, provenance tracking, and quality control part of routine operations. This preserves context, supports corrections, and documents how records change over time.

A key extension is the integration of Camtrap DP as a core standard for data exchange and reuse. The aim is not merely to generate a Camtrap DP export at the end of the workflow, but to reflect its key concepts in the data infrastructure itself: by aligning deployments, media, and organism records with Camtrap DP concepts LAUP produces machine-actionable outputs that are easy to integrate into downstream analytical pipelines.

We present LAUP as a practical bridge between standardized field monitoring, non-standardized data realities, and AI-ready biodiversity knowledge infrastructure, showing how domain-specific systems can turn raw image streams into curated, comparable, and reusable biodiversity evidence.

Speaker



Peter Grobe

Leibniz Institute for the Analysis of Biodiversity Change

A Provenance-Aware Geospatial AI Pipeline for Auto-Labeling Multimodal Biodiversity Data: Application to Invasive *Prosopis juliflora*

9:00 AM – 9:15 AM

CONTRIBUTED ORALS

Occurrence data is only useful downstream if a researcher can tell where it came from, how certain it is, and whether they are allowed to use it. Most biodiversity pipelines stop at the label. Ours does not. *Prosopis juliflora*, an invasive spreading across arid India, is hard to tell apart from native *Vachellia* and *Senegalia*, which carry the same pinnate leaves and thorned branches. Getting the species call right is half the problem. Making that call trustworthy at scale is the other half.

Three data streams feed the pipeline. Hyperspectral imagery from PRISMA, EMIT, and UAV push-broom sensors picks up *P. juliflora* in the red-edge and SWIR bands, where its evergreen canopy and leaf chemistry diverge from native species that shed seasonally. Sentinel-2 and Planet time series add phenology. Field photos, geotagged through a mobile app, ground the signal and feed retraining. A two-stage model returns per-polygon calls with uncertainty vectors, tagged with the sensor, model version, training cohort, and reviewer who signed off.

Data moves through three zones in a cloud lake. Raw API pulls land in a Drop Zone with acquisition metadata held intact. Orchestrated ELT pushes it into a Refined Zone for cleaning and georeferencing, then into a Curated Zone where records sit in Darwin Core schemas extended for model-derived occurrences. A governance layer tracks lineage, license terms, and access rules across all three; IAM, logging, and monitoring services keep the audit trail.

The same curated layer drives spread modelling, risk mapping, and habitat-loss accounting. Multimodal biodiversity labels need to carry enough context — uncertainty, provenance, license terms — for downstream AI pipelines to judge fitness-for-use before consuming them.

Speaker



Pranav Jha

Authors

Dr.Khyati Thacker, Shaurya Sharma, Aryaman Mihir Seth, Shubhankar Mihir Seth

Integrating advanced monitoring technologies for spatially explicit, ecosystem-based biodiversity conservation across diverse habitats in Saudi Arabia

9:15 AM – 9:30 AM

CONTRIBUTED ORALS

This study evaluates a spatially explicit, ecosystem-based framework across 35 Biodiversity Protection Areas (BPAs) in Saudi Arabia. The network spans diverse ecosystems, from the hyper-arid Rub' al Khali, supporting reintroduced populations of Arabian Oryx, Sand Gazelles, and Red-necked Ostrich, to the high-altitude Abha BPA, a global biodiversity hotspot currently undergoing Nubian Ibex reintroduction. Coastal and marine sites further protect Red Sea mangrove forests, sea turtle breeding grounds, and Arabian Gulf marine megafauna.

Ecological assessments have recorded 906 species, including 44 native and 189 avian species, with 26 taxa listed on the IUCN Red List. Key fauna include the Arabian Striped Hyena, Asir Magpie, Arabian Wolf, and various shark and fish species. Integrated monitoring utilizes a technological stack comprising satellite transmitters, camera traps, Remote Sensing (RS), GIS, and Unmanned Aerial Vehicles (UAVs) to facilitate high-resolution population assessments. Strategically located within the African-Eurasian flyways, these BPAs also serve as vital stopovers for billions of migratory birds, including the threatened Socotra Cormorant and Steppe Eagle. This model demonstrates how advanced technology and spatial prioritization can effectively safeguard global biodiversity across complex, high-value ecosystems.

Speaker



Mohammad Zafarul Islam

Senior Environmental Scientist | Saudi Aramco

Evaluating Four Machine Learning Using Google Earth Engine for Analyzing LULC Dynamics of Sandwip Island, Bangladesh

9:30 AM – 9:45 AM

CONTRIBUTED ORALS

Sandwip Island is a highly dynamic coastal environment experiencing continuous geomorphological changes driven by erosion, accretion, and sediment deposition. This study analyses shoreline changes from 1995 to 2026 using the Digital Shoreline Analysis System (DSAS) in ArcGIS and evaluates Land Use Land Cover (LULC) changes from 2000 to 2024 using Google Earth Engine and machine learning algorithms, including Classification and Regression Tree, Gradient Boosting Trees, Random Forest, and Support Vector Machines. A total of 140 ground-truth points representing seven LULC classes were used to assess model accuracy. Among the models, RF demonstrated the highest overall accuracy (93.57%), producing clearer boundaries and better separation of major land-cover classes. GBT achieved an accuracy of 92.14%, showing reliable classification across most LULC types. SVM reached a moderate accuracy of 90%, although some overgeneralization was observed in spectrally similar classes. CART showed the lowest accuracy (87.14%), with increased mixed-pixel occurrence in dense and heterogeneous regions. The findings reveal a gradual increase in the total area of Sandwip Island due to sediment deposition, contributing to the expansion of sandy areas (61.628 km²), mangroves (27.412 km²), vegetation (58.618 km²), cropland (31.587 km²), and built-up areas (60.031 km²). These results highlight the significant influence of natural coastal processes on land-cover dynamics and demonstrate the effectiveness of remote sensing and machine learning techniques for monitoring dynamic coastal environments. The study further suggests that integrating sediment transport modeling and tidal-current analysis with machine learning approaches could improve the accuracy of future coastal change assessments.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Md Abrar Al Foysol

Postgraduate Student | Department of Oceanography Shahjalal University of Science and Technology, Sylhet

Authors

Abu Bokkar Siddique

WKS33A - Liberating and providing AI-ready data from the Natural History Literature

8:30 AM – 10:00 AM | Location: ODIN

BIODIVERSITY LITERATURE & KNOWLEDGE EXTRACTION

Ever since the first attempts to observe, document and describe the natural world—followed by the establishment of the discipline as a scientific field—literature has served as the primary channel for the validation, dissemination and archiving of natural history knowledge, resulting in an ever-growing corpus of 500 million pages. However, this huge amount of information is often inaccessible—lost in the corner of a library or locked in a PDF—and forms what we commonly refer to as "unknown known knowledge".

To unlock and transform it into FAIR, AI-ready, interlinked objects, the community has produced open source tools for the extraction and annotation of data from scanned or digital-born PDFs based on established standards and persistent identifiers. Meanwhile, the community has also produced XML tools (e.g., JATS, TEI) to address a similar challenge in prospective literature. Over the past decade, this unlocked knowledge has greatly contributed to a rich ecosystem of increasingly interconnected data infrastructures, creating new opportunities for research and bioinformatics communities.

This workshop will complement the proposed Libroscope symposium. As a starter, the participants will be introduced to the stakes and methods of literature FAIR-isation with a strong focus on data cleaning and linking. Then, the main course will have the participants explore new opportunities offered by the Libroscope using human interfaces as well as APIs to retrieve liberated AI-ready data.

Session organizers



Laurence Benichou



Donat Agosti

Director | Plazi GmbH



Patrick Ruch

HES-SO & SIB Swiss Institute of Bioinformatics



Rainer M Krug

Senior Data Specialist | SIB Switzerland

Workshop leader



Chris Le Coquet

Muséum national d'Histoire naturelle

WKS24 - Mobilizing Botanical Checklists and Red Lists to GBIF

8:30 AM – 10:00 AM | Location: BALDER

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

The MetaPlantCode (MPC) project develops harmonized best practices for collecting, analyzing, and publishing plant metabarcoding data, with a strong focus on making these data findable, accessible, interoperable, and reusable through GBIF.

A key challenge in plant metabarcoding is that many species cannot be distinguished solely by their DNA barcode sequences. Closely related taxa often share (near-)identical barcodes, making species-level assignment difficult. To overcome this limitation, researchers must integrate contextual information such as the geographic distributions of potential species. These data are typically found in national checklists, floras, and red lists, but many of these resources are not yet accessible or interoperable.

To address this gap, the MPC project has developed a practical manual that guides users through the complete workflow of preparing, structuring, and publishing botanical checklists and red lists via the GBIF Integrated Publishing Toolkit (IPT). The guidelines build on GBIF best practices and provide step-by-step instructions for data preparation, Darwin Core mapping, extension use, metadata completion, and quality control. Although developed in the context of plant metabarcoding, the manual is widely relevant for national GBIF node managers, biodiversity information specialists and conservation organizations. This workshop will demonstrate key steps in the workflow and support participants in applying the guidelines to their own resources.

Workshop leader



Niels Raes

Programme Manager Biodiversity Informatics | Naturalis Biodiversity Center

BREAK

10:00 AM – 10:30 AM | Location: 1st Floor

SYM47B - Large Language Models for Biodiversity Data Discovery, Integration, and Curation. AI applications, integration, and validation

10:30 AM – 12:00 PM | Location: SAL A

AI AND ROBOT READY

Large language models (LLMs) and other generative AI systems are changing how researchers interact with data and digital infrastructure. In biodiversity informatics, these technologies create new opportunities for improving data discovery, automating metadata generation, assisting with data curation, and enabling natural language interfaces for biodiversity databases.

This session will explore emerging applications of large language models in biodiversity data infrastructure and research workflows. Topics may include natural language querying of biodiversity repositories, extraction of biodiversity information from literature and specimen labels, AI-assisted metadata generation and validation, and integration of LLM-based tools with biodiversity data standards such as Darwin Core.

We encourage case studies demonstrating how LLM-based tools can support biodiversity data discovery, improve metadata completeness, and assist with data publishing and curation workflows. Contributions addressing challenges such as reliability, provenance, bias, and integration of generative AI with structured biodiversity data are also welcome.

The goal of the session is to examine how large language models may reshape biodiversity data access and curation while identifying opportunities and limitations for their responsible use.

Moderators



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute



Bahadır Altıntaş

Abant İzzet Baysal University

Authors

Cem Meydan

Automated Landmark Detection for Scalable Morphometric Data Extraction from Digitized Fish Collections

10:30 AM – 10:42 AM

AI AND ROBOT READY

Morphometric analysis is fundamental to studying biological shape variation, taxonomy, and biodiversity patterns, yet large-scale research remains constrained by the time and labor required for manual landmark annotation. As natural history collections increasingly digitize specimen images, the need for scalable and AI-ready approaches to morphological data extraction is rapidly increasing.

Here, we present a machine learning workflow for automated landmark detection in two-dimensional fish specimen images, built on a standardized anatomical annotation framework derived from established morphometric research. The model was trained and evaluated on manually annotated collection images to assess landmark detection accuracy across a diverse range of specimens.

Performance was evaluated using Euclidean distances between reference and predicted landmarks. Accuracy varied across anatomical regions, with external body outline landmarks achieving the highest precision. Species-level variation in performance was also observed, suggesting that morphology and imaging conditions both influence detection accuracy. Together, these results demonstrate that machine learning can reliably automate landmark detection in digitized collection images offering a scalable pathway toward AI-ready morphometric datasets for biodiversity research and natural history collections.

Speaker



Bahadir Altintas

Authors

Yasin Bakış, Brian Sidlauskas, Henry L. Bart, Xiaojun Wang

Contained Agentic Workflow for Literature Analysis and Data Extraction in Museum Collections

10:42 AM – 10:54 AM

AI AND ROBOT READY

As agentic AI systems become more capable, they offer new opportunities for interacting with complex digital infrastructures, specialist literature and collections data. These systems can move beyond single-purpose language models or fixed Retrieval-Augmented Generation workflows by supporting more flexible, context-aware interactions with knowledge. However, highly versatile agentic systems also raise important concerns around cyber security, governance, compute requirements and token costs, particularly when deployed at institutional scale.

This presentation explores the use of local LLMs and agentic workflows for the analysis of scientific literature and the extraction of structured data relevant to research and museum collections. Operating on local hardware, the system is designed to ingest, interpret and curate knowledge from research papers, reports and archival documents, linking extracted information to existing collections records and domain knowledge.

Using Dell GB10 systems and high-performance devices, we demonstrate how contained LLM workflows can identify taxonomic, morphological, geographic and ecological information within text, converting narrative scientific knowledge into structured, machine-readable data. This enables the enrichment of collections databases and supports the integration of external knowledge sources into digitisation workflows.

By operating within a local/on-device environment, the approach maintains control over data processing. Live demonstrations will show how local LLMs and agentic systems can augment research and curation workflows by helping researchers and curators unlock information contained within scientific literature and integrate it directly into research and collections-based work.

Speaker



Steen Dupont

Programme Manager- RECODE | Natural History Museum London

Authors

Jack Hollister

Curation through citation: using AI and a knowledge graph to curate DNA barcodes

10:54 AM – 11:06 AM

AI AND ROBOT READY

Recent work on visualising DNA barcode data in the BOLD database with [BOLD View](#) has revealed numerous records that are problematic or incomplete. Common problems are a lack of formal identification ("dark taxa" in the original sense), a lack of geographic coordinates, and missing information on type status.

Often the quality of a barcode record can be improved by consulting the paper that published the DNA barcode sequence. Papers frequently contain details on sampling localities, and an informally named taxon may subsequently have been described. For example, the barcode LEASZ362-22, published in the dataset "Lepidoptera of Cottian Alps" <https://doi.org/10.5883/DS-LEPICOTT>, is identified only to genus level as "*Denisia*". The DS-LEPICOTT dataset is cited by Huemer et al. (2023) <https://doi.org/10.3390/d15020214>, which in turn is cited by Huemer (2024) <https://doi.org/10.3897/alpento.8.127794>, which describes the new species, *Denisia cryptica*, for which the barcode LEASZ362-22 is from the holotype. Following this chain of connections resolves the identity of a dark taxon.

For this kind of curation to be feasible at scale we need to be able to discover these connections automatically. One way to do this is to create databases of the intermediate connections, especially citations between papers, and citations of data. Some of these links (such as papers citing papers) are available in existing databases, others have to be assembled by text mining or record matching.

This talk describes constructing this knowledge graph, and the role of AI tools both in constructing the graph and in enabling users to query the graph using natural language. Then talk will argue that a richly connected knowledge graph has the potential to improve the quality of biodiversity databases.

Speaker



Roderic Page

Integrating Many Data Sources in iChatBio Enables Emergent Scientific Use Cases

11:06 AM – 11:18 AM

AI AND ROBOT READY

iChatBio is an AI research platform for conversational exploration and integration of open biodiversity data, developed by the Advanced Computing and Information Systems (ACIS) Laboratory at the University of Florida. iChatBio operates by pulling data from online data sources into a unifying cloud-based workspace. Data interactions are performed by a team of autonomous large language model (LLM) agents, each specialized to interact with a single data source or data type. At its initial launch in February 2025, iChatBio only included three agents to access species occurrence data from Integrated Digitized Biocollections (iDigBio), the Global Biodiversity Information Facility (GBIF), and iNaturalist. To facilitate the development of new agents, we created an open source iChatBio Software Development Kit (SDK). The SDK has so far been used to develop 16 specialized agents, which have been incorporated into iChatBio to expand its scope to new data types and data sources, including Atlas of Living Australia (ALA), Biodiversity Heritage Library (BHL), Global Biotic Interactions (GloBI), GenBank, NASA Prediction of Worldwide Energy Resources (POWER), Ocean Biodiversity Information System (OBIS), World Register of Marine Species (WoRMS), and supporting agents for processing data, reading PDF documents, and generating maps from georeferenced data. iChatBio's design has also seen many improvements, including better intelligence, transparency, and configurability. This presentation will demonstrate how iChatBio's integration of diverse data sources and data processing tools supports emergent scientific use cases of online data beyond what is possible within existing Web-based data portals, while avoiding transparency and reproducibility issues that are common in LLM applications.

Speaker



Michael Elliott

PhD Student | University of Florida

Authors

José Fortes, Manuel Luciano

We describe a deployed LLM curation agent that surfaces data-quality issues beyond the reach of schema constraints across TERN's SWARM ecological-monitoring database (around 950 sites). This contribution reports the agent's design, operating envelope, and the choices we believe make LLM-assisted curation usable on research-grade biodiversity data.

The agent is a containerised Python service with a configurable backing model — Gemini Flash Lite by default — exposing seven tools: read-only SQL against SWARM, REST queries against the Paratoo Strapi API, and findings persistence. A ten-phase full scan covers coordinate validity, completeness, transect structure, lookup-table validity, ecological plausibility, orphans, soil characterisation, uniqueness, and cross-system value mismatch. A focused mode re-checks open findings only and runs daily at negligible cost. Findings persist to a curation table with severity, category, and where applicable a suggested SQL fix. The agent never writes to the database autonomously; fixes require explicit curator approval via a web dashboard.

The agent's key advantage over schema validation is ecological plausibility reasoning. Findings surfaced include growth-form heights orders of magnitude too large (decimal-place entry errors), shrubs recorded taller than the canopy, epiphytes on bare substrate, soil horizons with inverted depth ordering, visits spanning implausible durations, and near-monoculture transects at sites expected to be species-rich. These issues cannot be caught by schema constraints — they require domain knowledge about what is ecologically reasonable. We discuss the human-approval gate rationale and initial curator-led validation to characterise true-positive rate by category.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Andrew Tokmakoff

TERN Ecosystem Surveillance

Authors

Walid al Naim, Ben Sparrow, Carly Steen, Huy Vo, Donna Lewis

Enhancing Biodiversity Georeferencing Using Large Language Models and Knowledge Graphs

11:30 AM – 11:42 AM

AI AND ROBOT READY

Geographic location is one of the most important pieces of information in biodiversity occurrence records. However, many biodiversity databases document locations only as place names or textual descriptions without geographic coordinates, often introducing ambiguity due to historical names, inconsistent spellings, and overlapping administrative regions. Consequently, current georeferencing approaches either rely on gazetteer lookups, string matching, and supervised training that remain limited for low-resource biodiversity datasets where locality descriptions are sparse, ambiguous, and domain-specific.

In this study, we develop a georeferencing pipeline that combines Large Language Models (LLMs) with a knowledge graph, i.e., Wikidata, to disambiguate location names in biodiversity occurrence records. An LLM is used as a preprocessing component to normalise locality names, identify spelling errors, resolve alternate names, and determine whether a locality reference corresponds to a valid geographic entity. The cleaned locality names are then matched against Wikidata entities using the Wikidata API. Candidate entities are ranked using lexical similarity and locality context from biodiversity records. The most likely Wikidata entity is then used to retrieve and assign geographic coordinates to the record. The proposed approach is applied to herbarium data from the Manchester Museum comprising 2,442 records collected from African sites. We compared our approach with a gazetteer-only baseline using fuzzy string matching, simulating the manual process of entering a place name into a gazetteer and selecting the top-ranked result. Our study aims to improve the accuracy of biodiversity georeferencing, particularly for ambiguous and historically inconsistent locality records.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Roselyn Gabud

Associate Professor | Department of Computer Science, University of the Philippines Diliman

Authors

Riza Batista-Navarro, John Peel, Rachel Webster

Validating LLM-Extracted Biodiversity Data to Preserve Scientific Meaning

11:42 AM – 11:54 AM

AI AND ROBOT READY

Artificial intelligence (AI) tools now enable large-scale extraction of biodiversity data from published literature, offering transformative potential for repositories seeking to expand data mobilization. However, current approaches largely treat extraction as a technical problem, leaving a critical gap in preserving scientific meaning, ensuring that interpretations are reviewed and traceable, and enabling reliable reuse. As a result, biodiversity datasets may appear FAIR while subtly diverging from the scientific intent of their original sources. Here, we describe a curation and validation infrastructure for AI-assisted biodiversity data, drawing on implementation within the specialized data repository, MorphoBank. In this context, scientific meaning is preserved when extracted character and character state definitions accurately reflect the original author's intended distinctions, without altering those distinctions. We focus on the human validation required to achieve this: AI-generated draft data are reviewed through checklist-based verification to confirm that extracted data match the source, preserve meaning, and remain linked to their evidentiary context. Through case examples from this workflow, we show how human review identifies meaning-altering errors that are not detectable through automated validation alone. Together, these validation mechanisms enable scalable data extraction while preserving the scientific meaning of the original data. We present this as a generalizable model for biodiversity data systems that rely on AI-assisted extraction from literature. Automation accelerates data drafting, but reliable reuse depends on human validation.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Brooke Long-Fox

MorphoBank Data Curation Scientist | Phoenix Bioinformatics

Authors

Tanya Berardini

Concluding remarks

11:54 AM – 12:00 PM

AI AND ROBOT READY

Speaker



Yasin Bakış

Sr Manager of Biodiversity Informatics | Tulane University, Biodiversity Research Institute

SYM36 - Designing Institutional Knowledge Data Science Centres for Biodiversity

10:30 AM – 12:00 PM | Location: SAL B

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Natural history museums and biodiversity research institutes are evolving from collection holders into stewards of AI-ready biodiversity knowledge. As DiSSCo builds pan-European infrastructures for FAIR Digital Objects and extended specimens, a key question arises: **how can institutional biodiversity knowledge centres reinforce—rather than replicate—these global efforts?**

We envision an **Open Knowledge Space for Data Centers**: a semantic, web-based platform linking biodiversity knowledge with collection data, research outputs, and AI-generated insights. Designed for

machines and humans, it aims to serve as a **biodiversity hub where data and knowledge are FAIR, CLEAR, and AI-ready**—intuitive and meaningful for researchers, citizens, and policymakers.

This session brings together scientists and infrastructure architects to co-design **institutional knowledge spaces** that:

- Integrate semantic models and digital twins capturing data, workflows, and epistemic knowledge
- Align semantically and technically with DiSSCo
- Offer AI-ready services with operations for validation, prediction, and analysis
- Balance machine-actionability with human interpretability with intuitive interfaces

We welcome contributions on:

- Prototypes of semantic data pipelines from collections to knowledge graphs
- Interactive dashboards for exploring biodiversity knowledge
- Governance models connecting collections, data science, and infrastructure
- Lessons learned from building institutional knowledge spaces

Moderator



Peter Grobe

Leibniz Institute for the Analysis of Biodiversity Change

Session organizers



Carina Goretzky

Leibniz Institute for the Analysis of Biodiversity Change (LIB)



Björn Quast

Leibniz Institute for the Analysis of Biodiversity Change (LIB)

Authors

Sheikh Mastura Farzana

The LIB Open Knowledge Space: A semantic architecture for FAIR, CLEAR, and AI-ready biodiversity knowledge

10:30 AM – 10:45 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Natural history collections risk remaining isolated archives unless their data and knowledge are transformed into connected, semantically structured, and machine-actionable resources. The LIB Open Knowledge Space is a planned FAIR-, CARE-, and CLEAR-compliant knowledge infrastructure. It is central to the data and knowledge science strategy of the Leibniz Institute for the Analysis of Biodiversity Change (LIB) to integrate collection data, research outputs, and curated knowledge into a unified environment serving as LIB's living memory and an anchor of distributed semantic interoperability.

The architecture rests on two frameworks: (1) The Semantic Units Framework organises biodiversity data and knowledge into modular, identifiable units of meaning, each with a globally unique persistent identifier and explicit epistemic status, supporting assertional, existential, prototypical, and universal statements. (2) The Semantic Ladder defines a progressive formalisation pathway from natural language text snippets through Rosetta Statement anchors to Description Logics-based representations, enabling incremental FAIRness without full formalisation at data entry. Across all levels, vector embeddings support similarity-based search and AI-assisted exploration, enabling hybrid symbolic-vector access that we plan to complement with LIBTellMe, a RAG-based chatbot.

Together, these frameworks support semantic digital twins of specimens capturing not only associated data but also object and object-data journeys, going beyond the extended specimen concept.

We present the conceptual architecture and research programme, and invite discussion on alignment with DiSSCo and NFDI4Biodiversity.

Authors

Peter Grobe, Björn Quast, Sheikh Mastura Farzana, Carina Goretzky, Cristina Garilao, Birgit Rach

Strengthening Taxonomic capacity and Biodiversity Information in Education

10:45 AM – 11:00 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

"Name It To Save It" is the motto of the Taxonomy Recognition Day organised every 23rd of May, set up by the EU Project TETTRIs (Transforming European Taxonomy through Training, Research, and Innovations) (<https://tettris.eu/>). The underlying idea is that if people can identify species and better know their characteristics, habitats and needs, more awareness and means will be mobilised to protect them. To be effective, the naming and description of the species have to be done in a correct and user-friendly manner. It is thus essential that subjects such as nomenclature, taxonomy and biodiversity informatics and data standards are addressed properly both in academic and non-academic education. Within the TETTRIs project, 2 work packages were dedicated to training, resulting in a landscape analysis on taxonomic expertise needs as a result of a stakeholder process, a Memorandum of Understanding (MoU) for increased collaboration on taxonomic training and research and in the production of several pilot courses in taxonomy (<https://tettris.nhmc.uoc.gr/>). The long-term sustainability and availability of the courses is guaranteed by their inclusion into the CETAF Distributed European School of Taxonomy (DEST) (<https://cetaf.org/dest/>). Next to more classical taxonomic topics, the authors identified that transferable skills, including digital skills, are an essential part of the curricula. These include understanding what data standards are, how they work and are maintained, and how they can be used in various applications. This presentation will show examples of courses in the domain of Taxonomy and Biodiversity Information and discuss the opportunity to revive the TDWG Interest group on the Biodiversity Informatics Curriculum (<https://www.tdwg.org/community/curriculum/>).

Speaker



Patricia Mergen

Liaison officer | Meise Botanic Garden/Royal Museum for Central Africa/ Belgian Science Policy Office

Authors

Elie Saliba, Iasmi Stathi, Amelie Höcherl, Peter Warth, Louise Hendrickx, Sofie Meeus

Enabling machine-readable access to biodiversity data through federated semantic querying

11:00 AM – 11:15 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Biodiversity data are frequently published as distributed Darwin Core archives hosted across institutional repositories, creating major challenges for integrative ecological analyses. Existing approaches generally require either centralising datasets into large aggregators or querying heterogeneous provider-specific APIs. We present a framework that enables federated querying of distributed biodiversity datasets while leaving data in place. The approach combines Darwin Core semantics, ontology-based data access (OBDA), and virtual knowledge graphs to expose remote biodiversity datasets through a unified SPARQL endpoint.

The system performs spatiotemporal dataset discovery using dataset-level metadata annotations. Relevant datasets are dynamically assembled into a combined dataset by referencing remote Parquet assets stored in object storage. By combining semantic web standards with virtualised data access, the framework enables interoperable querying across independently managed biodiversity resources while preserving dataset autonomy and reducing data duplication.

This supports flexible cross-dataset discovery and analysis using shared biodiversity semantics. By integrating linked data principles with modern analytical database technologies, the framework demonstrates a scalable approach to federated biodiversity data access. The project, developed at the Quebec Centre for Biodiversity Science (QCBS), illustrates how semantic web technologies, virtual databases, and cloud-native analytical workflows can support reusable biodiversity research infrastructures aligned with FAIR data principles.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



El-Amine Mimouni

Research Professional | Quebec Centre for Biodiversity Science / McGill University

Authors

Guillaume Larocque, Fanie Pelletier, Andrew Gonzalez

Prototype: The Semantic Units Framework and Rosetta Statements for Knowledge Graph Construction Workflows

11:15 AM – 11:30 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

LIB's Open Knowledge Space proposes the implementation of the Semantic Units Framework and Rosetta statements as novel methods for FAIR (Findable, Accessible, Interoperable, Reusable), CLEAR (Cognitively interoperable, semantically Linked, contextually Explorable, easily Accessible, and human-Readable and -interpretable), and AI-ready research data. As both methods are formalized as technology-agnostic, we investigate their inclusion in end-to-end Knowledge Graph construction workflows and present our first prototype: A pipeline that leverages a novel user-facing input layer for knowledge graph mappings based on formalized natural language input according to Rosetta Statements and a series of modules to compile this input into executable knowledge graph mappings and SHACL shapes for validation. We propose a reusable and easily extensible template library. Templates define Rosetta statement input slots and both structural and semantic specifications to create and validate instances of Semantic Unit classes. This enables simplified mapping authoring and enables further graph and application extensions, such as natural language descriptions of Semantic Unit content that support cognitive interoperability for users and can be leveraged as features for retrieval systems. We present case-study results from the German Centre for Integrative Biodiversity Research (iDiv), in which the iAnswer and PhenObs projects collaborated to transform research data about phenological plant observations in botanical gardens into a Semantic Units Knowledge Graph using our pipeline. Therefore, this prototype is the first step towards mobilizing the wealth of iDiv research data into an institutional infrastructure providing machine actionable and human interpretable research data.

Speaker



Tarek Al Mustafa

German Centre for Integrative Biodiversity Research Halle-Jena-Leipzig – iDiv, University of Jena

Authors

Birgitta Koenig-Ries, Lars Vogt, Robert Rauschkolb

What is the Role of "AI" in the Centers for Biodiversity Informatics of the Future?

11:30 AM – 11:45 AM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

The Species File Group (SFG), now endowment funded for almost two decades, is an active center for biodiversity informatics. It therefore represents a use-case for exploring the impact of "AI" (generative or otherwise) within such organizations. AI has impacted our infrastructure (human and technical, costs), administration, software, development practices, services, teaching, and planning for the future. In the interests of collectively and collaboratively strengthening fellow centers we highlight examples spanning these areas, subjectively grouping them into positive, negative or neutral outcomes. We further place them along an axis of certainty, i.e. how strongly we feel about the longer-term reality of the issue. As an organization who manages most aspects of its infrastructure AI has impacted how we think about risk management. Empowering individuals, particularly with respect to enabling them to replicate core tooling and data locally is emerging as one priority. Architecting a balance between the management of the complex tools required to create rich networks of biodiversity data and creating a framework that distributes capacity, and risk, particularly one that persists over long periods of time, is a problem that is not unique to biodiversity informatics. AI is both pushing us to address this balance and introducing new considerations.

Speaker



Matthew Yoder

University of Illinois, PRI, Illinois Natural History Survey, Species File Group

Discussion

11:45 AM – 12:00 PM

RESILIENCE & SUSTAINABILITY OF DATA INFRASTRUCTURES

Moderator



Peter Grobe

Leibniz Institute for the Analysis of Biodiversity Change

SYM28 From Standards to Implementation: Connecting Observation Data in Asia to GBIF Infrastructure

10:30 AM – 12:00 PM | Location: SAL C

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Moderators



Kumiko Totsu

Data Specialist | National Institute for Environmental Studies



Christian Elloran

Director | Director - Biodiversity Informatics and Geospatial Analysis, ASEAN Centre for Biodiversity

Introduction

10:30 AM – 10:40 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Speaker



Kumiko Totsu

Data Specialist | National Institute for Environmental Studies

Building a Living Publication Workflow for Long-Term Social-Ecological Monitoring Data: The Taiwan LTSER Experience

10:40 AM – 10:55 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Long-term social-ecological monitoring data are growing, continuously updated resources that require living publication workflows rather than one-time releases. In long-term social-ecological research networks, biodiversity observations are often collected with environmental, social, and site-based monitoring data. This makes it challenging to extract biodiversity relevant records and routinely transform, validate, and publish them through GBIF compatible infrastructures.

Taiwan LTSER is developing a semi-automated workflow to help continuously updated monitoring datasets move toward Darwin Core alignment. The workflow uses depositor, a research data repository for managing and sharing datasets with metadata, to bridge local data management practices and downstream biodiversity data publication infrastructures.

From this repository-based entry point, scheduled checks detect dataset updates and evaluate whether records can be aligned with Darwin Core. In this approach, biodiversity data standards are implemented as workflow checkpoints rather than tasks left to individual researchers. Datasets that require revision can be returned to project contacts for correction, while datasets that meet the requirements are first presented on the LTSER portal and then prepared for further publication through TaiBIF IPT and GBIF.

The Taiwan LTSER experience shows how a living publication workflow can reduce manual work, lower technical barriers, and connect distributed long-term monitoring data to global biodiversity infrastructures. It demonstrates that the key implementation challenge is not only adopting data standards, but embedding them into practical, repeatable workflows that fit real research settings.

Speaker



Jhu-Jyun Jhang

Developer | Taiwan Biodiversity Information Facility, TaiBIF

Snapshot Japan: A Community-Driven Network for Large-Scale Wildlife Biodiversity Monitoring Using Camera Traps

10:55 AM – 11:10 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

To understand global biodiversity change, standardized data and common monitoring methods are increasingly important. This presentation introduces Snapshot Japan, a Japan-based network inspired by Snapshot Global. Since 2023, annual surveys have been conducted during September-October with participation from researchers, citizen groups, museums, schools, and private companies. Camera trap images are managed on Wildlife Insights and cross-checked prior to data paper publication.

Participation has been promoted through mailing lists, a dedicated website, camera trap lending, online orientation sessions, manuals, and email/chat support. Participant feedback is incorporated into annual operational improvements. Snapshot Japan has also developed a web application for visualizing Wildlife Insights data and calculating Relative Abundance Index (RAI).

The number of survey sites increased from 9 in 2023 to 65 in 2025. Participation from high schools and private companies notably increased in 2025, partly reflecting growing interest in Nature Positive activities. The network has facilitated collaboration among academic, citizen science, educational, and private-sector communities.

To promote international data reuse, datasets are published through GBIF in Camtrap DP format, an international standard for camera trap data. These data are also available through Wild Album. In addition, an R package for converting Wildlife Insights data into Camtrap DP has been developed, and a non-R application is under development. Although participation from Asian countries in Snapshot Global remains limited, Snapshot Japan could provide a practical model for expanding standardized camera trap monitoring networks across Asia.

Speaker



Kumiko Totsu

Data Specialist | National Institute for Environmental Studies

Corresponding author



Kana Terayama

National Institute for Environmental Studies

Authors

Keita Fukasawa, Kumiko Totsu, Yoshihiro Nakashima, Takahiro Morosawa

Reporting structure and sampling-bias correction in citizen-science bird data across Asian cities

11:10 AM – 11:25 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Citizen-science bird records are widely reused in urban biodiversity research through GBIF. Their value depends not only on record volume, but also on whether sampling effort, observer identity, checklist type, and observation locations are documented. Because these reporting features vary among cities, common bias-adjustment methods, including complete-checklist filtering, effort or observer covariates, occupancy models, spatial thinning, and accessibility-related corrections, may not be equally feasible.

We examine how reporting structure affects the diagnosis and correction of sampling bias in urban bird data. Using GBIF-mediated bird occurrence records from 2021 across 14 cities in Southeast, East, and South Asia, we compare record volume, checklist availability, observer participation, temporal continuity, and spatial concentration. We ask whether these reporting differences correspond to spatial bias patterns, and how biodiversity estimates and biodiversity–environment relationships change under different bias-adjustment choices.

For each city, we summarise reporting structure using record and checklist counts, unique-observer counts, observer dominance, temporal continuity, and spatial concentration. We explore spatial bias by modelling observation intensity against variables representing accessibility, repeated-site use, observer activity, and environmental-gradient coverage. We then compare biodiversity models fitted with effort covariates, observation-process proxies, and combined approaches.

The study identifies where standard correction strategies are feasible, where proxy-based approaches are needed, and where data should be limited to descriptive use. It links citizen-science data reuse to reporting standards and fitness-for-use discussions.

Speaker



Yeshan Qiu

PhD researcher | Singapore ETH center

Authors

Dengkai Chi, Mark McDonnell, Puay Yok Tan

From sampling-event datasets to the Humboldt Extension: a TaiBIF assessment and pilots for long-term monitoring in Taiwan

11:25 AM – 11:40 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Long-term monitoring data are key for tracking species trends and supporting policy, but global biodiversity infrastructures mostly hold single occurrence records. The Humboldt Extension offers a way to make monitoring data more interoperable. A practical way to scale adoption is to start with what is already published, and ask which datasets can be restructured into this format.

This talk reports an assessment by the Taiwan Biodiversity Information Facility (TaiBIF) of sampling-event datasets published from Taiwan to GBIF. Early checks suggest Taiwan holds a higher share of sampling-event datasets than most publishing countries. We describe these datasets by taxonomic scope, time coverage, sampling design, and repeat-visit structure, and ask which parts are ready to be mapped to the Humboldt Extension.

We then present two pilot cases chosen to cover different monitoring modes. The first is the Taiwan Breeding Bird Survey, a citizen-science program with structured repeat visits. The second is passive acoustic monitoring from ongoing research projects, which adds sensor-based and effort-rich data. For each pilot we show the current schema, the gap to Humboldt Extension terms, and a mapping workflow that provides restricted-access options for sensitive datasets.

We close by previewing a feedback service we are developing: site-level trend and data-quality plots returned to providers once they share data. We see restructuring and this feedback service together as a way to encourage future monitoring data to be published directly in the Humboldt format. We invite discussion on reuse across Asia and on how node-led work can shape community guidance for Humboldt Extension adoption.

VIRTUAL PRESENTATION

Speaker (VIRTUAL)



Jerome Chie-Jen Ko

Postdoctoral Researcher | Taiwan Biodiversity Information Facility, Biodiversity Research Center, Academia Sinica

Authors

Shih-Hung Wu, Tuanmu Mao-Ning

ASEAN Biodiversity Dashboard as a data sharing platform to monitor biodiversity trends and species occurrences in the ASEAN Region

11:40 AM – 11:55 AM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

The effective management of ASEAN Heritage Parks (AHPs) and protected areas across the ASEAN region is essential for conserving biodiversity, maintaining ecosystem integrity, and sustaining the ecosystem services that support livelihoods and human well-being. In response to the increasing need for accessible, integrated, and evidence-based biodiversity information, the ASEAN Biodiversity Dashboard was developed as a regional, open-access platform designed to monitor biodiversity trends, species occurrences, and conservation indicators across Southeast Asia.

The Dashboard demonstrates how digital technologies and data-driven approaches can strengthen protected area management by enabling biodiversity data integration, spatial analysis, and informed conservation planning. By consolidating information from ASEAN Member States alongside global biodiversity datasets and partners, the platform promotes harmonized, policy-relevant information that supports evidence-based decision-making and regional collaboration.

As part of its backend mechanisms, the Dashboard is supported by a GBIF-hosted portal that facilitates biodiversity data publishing, access, and interoperability with global biodiversity information systems. This infrastructure enables the integration of occurrence records, species observations, and other biodiversity datasets while ensuring alignment with international data standards and promoting open access to biodiversity information. The GBIF-hosted portal strengthens the technical foundation of the Dashboard by improving data discoverability, sustainability, and long-term accessibility.

Speaker



Christian Elloran

Director | Director - Biodiversity Informatics and Geospatial Analysis, ASEAN Centre for Biodiversity

Closing remarks

11:55 AM – 12:00 PM

COMMUNITY SCIENCE, ENGAGEMENT & REGIONAL NETWORKS

Speaker



Kumiko Totsu

Data Specialist | National Institute for Environmental Studies

WKS33B - Liberating and providing AI-ready data from the Natural History Literature

10:30 AM – 12:00 PM | Location: ODIN

Ever since the first attempts to observe, document and describe the natural world—followed by the establishment of the discipline as a scientific field—literature has served as the primary channel for the validation, dissemination and archiving of natural history knowledge, resulting in an ever-growing corpus of 500 million pages. However, this huge amount of information is often inaccessible—lost in the corner of a library or locked in a PDF—and forms what we commonly refer to as "unknown known knowledge".

To unlock and transform it into FAIR, AI-ready, interlinked objects, the community has produced open source tools for the extraction and annotation of data from scanned or digital-born PDFs based on established standards and persistent identifiers. Meanwhile, the community has also produced XML tools (e.g., JATS, TEI) to address a similar challenge in prospective literature. Over the past decade, this unlocked knowledge has greatly contributed to a rich ecosystem of increasingly interconnected data infrastructures, creating new opportunities for research and bioinformatics communities.

This workshop will complement the proposed Libroscope symposium. As a starter, the participants will be introduced to the stakes and methods of literature FAIR-isation with a strong focus on data cleaning and linking. Then, the main course will have the participants explore new opportunities offered by the Libroscope using human interfaces as well as APIs to retrieve liberated AI-ready data.

Session organizers



Laurence Benichou



Donat Agosti

Director | Plazi GmbH



Patrick Ruch

HES-SO & SIB Swiss Institute of Bioinformatics



Rainer M Krug

Senior Data Specialist | SIB Switzerland

Workshop leader



Chris Le Coquet

Muséum national d'Histoire naturelle

LUNCH

12:00 PM – 1:00 PM | Location: 1st Floor

UNC - TDWG 2026 Unconference

1:00 PM – 3:00 PM | Location: SAL A

UNCONFERENCE

The Unconference is a participant-driven session designed to explore ideas inspired by the main TDWG 2026 Conference event and this year's theme "Research and Robot-ready Biodiversity Data Standards". Attendees can propose topics, join breakout discussions, and collaboratively examine issues of shared interest. The format provides an opportunity for all participants, in person or online, to connect with both existing and new colleagues, explore subjects in greater depth or propose new areas that haven't been mentioned yet throughout the conference - sowing the seeds for new collaborations.

We invite participants to share their ideas on a topic board during the conference week, where others can also express their interest in the topic. With a little guidance from the organizers at the start of the session, we will work together to select and finalize discussion group topics and self-organize into breakout groups to learn, find synergy and gain new insights.

The session will end with breakout group leaders reporting insights or discussion summaries back to the whole group (e.g., what each group discovered, something learned). Participants are encouraged to carry discussions forward beyond the session through short writeups or position statements to invite broader engagement and continue collaboration.

Session organizers



Javier Molina

Project Manager / TDWG 2026 Communications Manager | Atlas of Living Australia / CSIRO



Deborah Paul

Species File Group (INHS), University of Illinois



Chandra Earl

National Ecological Observatory Network (NEON)



Shelley James

Collections Manager | Western Australian Herbarium - DBCA



Esteban Marentes Herrera

SiB Colombia



Nicky Nicolson

Senior Research Leader, Biodiversity Informatics | Royal Botanic Gardens, Kew

UNC - TDWG 2026 Unconference break out

1:00 PM – 3:00 PM | Location: SAL B

UNCONFERENCE

Speakers



Javier Molina

Project Manager / TDWG 2026 Communications Manager | Atlas of Living Australia / CSIRO



Deborah Paul

Species File Group (INHS), University of Illinois



Chandra Earl

National Ecological Observatory Network (NEON)



Shelley James

Collections Manager | Western Australian Herbarium - DBCA



Esteban Marentes Herrera

SiB Colombia



Nicky Nicolson

Senior Research Leader, Biodiversity Informatics | Royal Botanic Gardens, Kew

UNC - TDWG 2026 Unconference break out

1:00 PM – 3:00 PM | Location: SAL C

UNCONFERENCE

Speakers



Javier Molina

Project Manager / TDWG 2026 Communications Manager | Atlas of Living Australia / CSIRO



Deborah Paul

Species File Group (INHS), University of Illinois



Chandra Earl

National Ecological Observatory Network (NEON)



Shelley James

Collections Manager | Western Australian Herbarium - DBCA



Esteban Marentes Herrera

SiB Colombia



Nicky Nicolson

Senior Research Leader, Biodiversity Informatics | Royal Botanic Gardens, Kew

BREAK

3:00 PM – 3:30 PM | Location: 1st Floor

Closing Session and Business Meeting

3:30 PM – 4:30 PM

Speakers



David Bloom

University of North Carolina at Greensboro



Elie Mario Saliba

Post-doc/TDWG Deputy Chair | Bernice Pauahi Bishop Museum