Accelerating Research through Creation of an International Structure for Biodiversity Information

30 November 2017

Lawrence M. Page, iDigBio Director; and Curator, Florida Museum of Natural History, University of Florida, Gainesville, 32605. Email: lpage1@ufl.edu

Donald Hobern, GBIF Executive Secretary, GBIF Secretariat, Universitesparken 15, DK-2100, Copenhagen Ø, Denmark. Email: dhobern@gbif.org

John La Salle, Director Atlas of Living Australia, CSIRO National Collections and Marine Infrastructure, Canberra, ACT, 2601, Australia. Email: john.lasalle@csiro.au

Volker Mosbrugger, Director General, Senckenberg Gesellschaft für Naturforschung, Frankfurt, Germany. E-mail: volker.mosbrugger@senckenberg.de

Rosemary Gillespie, Professor & Schlinger Chair in Systematic Entomology, University of California, Berkeley, 94720-3114. Email: gillespie@berkeley.edu

James Hanken, Director, Museum of Comparative Zoology & Alexander Agassiz Professor of Zoology, Harvard University, Cambridge, Massachusetts 02138. Email: hanken@oeb.harvard.edu

Dave Tazik, Senior Research Scientist, National Ecological Observatory Network, Boulder, Colorado, 80301. Email: dtazik@battelleecology.org

Key terms: Biodiversity, collections, global change

The programmatic need exists for support of a global collaboration to leverage expertise, data, facilities, and other resources in existing networks of biodiversity information to advance critical research on biodiversity. Elements of the collaboration include networking across four major focus areas identified in the Global Biodiversity Informatics Outlook (GBIO) document—culture, data, evidence, and understanding—which together form an interconnected map of information management needs. Each area forms an excellent basis for in-depth review and development of a corresponding implementation roadmap. The **Culture** area provides the open data and open science context for all other aspects. The **Data** area ensures the web accessibility of well-formed streams of data from all relevant sources. The **Evidence** area organizes these streams into accessible integrated information resources. The **Understanding** area builds on these resources to provide modeled representations of best available understanding of actual biodiversity patterns and properties, which enable strategic research on biodiversity.

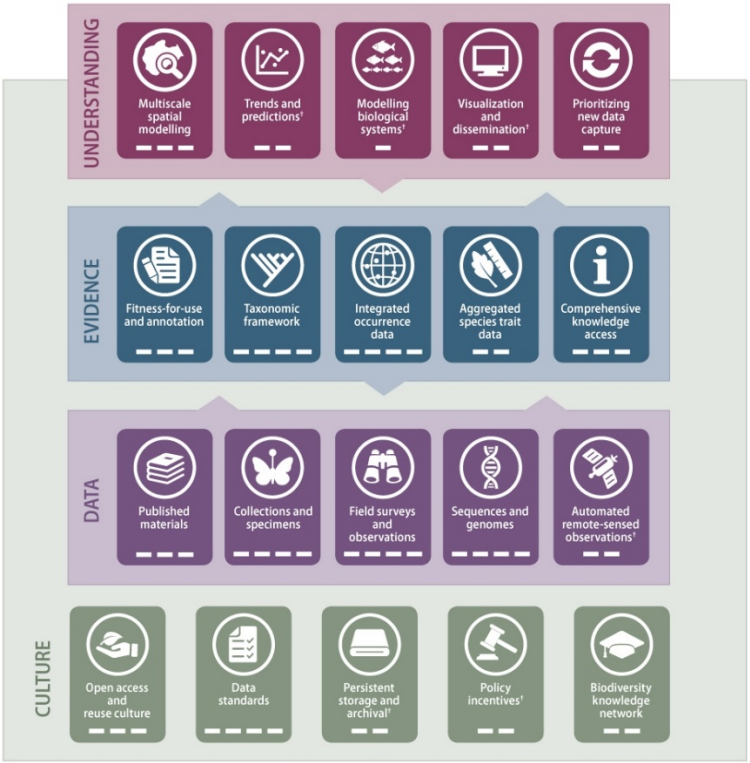

Figure 1. Elements of the Global Biodiversity Information Outlook (GBIO).

Several major initiatives to increase online access to biodiversity information have been launched in the past 20 years. These initiatives vary in their objectives, but all converge on the belief that natural biological diversity is fundamental to a healthy, sustainable planet and that enhancing access to integrated data is necessary to improve our understanding of biodiversity and provide an informed response to human impacts.

Outstanding among these initiatives because of their degree of community engagement, networks of participants, ability to integrate a range of data types, sizes of their databases and contributions to advancing research are NSF's Advancing Digitization of Biodiversity Collections (ADBC) program (idigbio.org), which promotes digitization of U.S. natural history collections data; the Global Biodiversity Information Facility (GBIF), which provides access to species occurrence data; the Atlas of Living Australia (ALA), which provides occurrence records and mapping tools to explore and analyze information; the Encyclopedia of Life (EOL), including the Biodiversity Heritage Library, which aggregates ecological and other species data from global resources; the Barcode of Life (iBOL), a public reference library of DNA sequences that can be used to assign unidentified specimens to known species; and the National Ecological Observatory Network (NEON), a continental-scale facility sponsored by NSF that gathers and synthesizes data on the impacts of climate change, land use change and invasive species on natural resources.

These initiatives were designed specifically to aggregate and integrate sources of information from multiple domains related to biodiversity to enrich data analysis and visualization. They have successfully connected myriad small silos of data

(institutional collections, individual researchers' data, etc.) as well as provided or improved computing and research tools that previously were difficult to access. Each initiative provides expertise, data, facilities, tools and other resources aimed at aggregating and advancing free and open access to biodiversity data to enable scientists to more effectively address the most significant questions about the diverse life of our planet.

Although each of these initiatives has made substantial progress in providing open access to biodiversity data, the overarching vision of developing a single, unified global resource for biodiversity conservation, natural resource management and sustainable development is far from realized. Isolated silos of information have been replaced by larger silos that are difficult to cross-search in ways required for large-scale, integrative research. Relevant data remain largely opaque to much of the data science community and are rarely included in societally motivated analyses that seek to resolve connections between biodiversity metrics and urban planning, agriculture, education, human health, etc.

Realizing this shortcoming and knowing that all biodiversity information systems face similar challenges—in particular, managing the huge amount of available data, insuring long-term sustainability and keeping pace with rapid technological innovations that require investments in personnel and maintenance—a three-day workshop, "Exploring Synergies and Sustainability for Biodiversity Information Systems," was held in Frankfurt, Germany, in March 2017. The workshop brought together representatives from eight countries and several major biodiversity data and research initiatives to compare infrastructures and goals, identify potential opportunities for enhancing collaboration and interoperability, and develop strategies for resource sharing and sustainability. Workshop participants agreed that development of an integrated global structure has the potential to dramatically improve access to and utility of biodiversity information. Realization of such a structure, however, will require existing systems to improve engagement with one another, identify their unique roles, eliminate redundancy, identify gaps in data and services, share technology and other resources, and develop strategies for sustainability based on collaboration instead of competition.

The ultimate goal of the workshop was development a single sustainable global network in which biodiversity data centers and their networks of participants can most effectively mobilize and deliver biodiversity data. Initiation of such a structure could be launched through development of a network-to-network collaboration among, at a minimum, iDigBio, GBIF, ALA and NEON. Such a collaboration would immediately begin to address:

1) Technical problems and issues that current limit collaboration:
   - Lack of scalable and interoperable data standards
   - Lack of a robust taxonomic and phylogenetic framework
   - Difficulties in assessing and improving data quality
   - Antiquated software infrastructure
   - Limited access to advanced computational resources
   - Diverse requirements of a large number of data providers and users

- The need to link phenotypic, genomic and environmental data
- Insufficient resources for outreach and capacity-building activities

2) Social and cultural issues that prevent collaboration:
- Lack of sufficient information about other biodiversity information systems
- Overlapping mandates
- Reluctance to forfeit independence
- Competing timelines and funding sources

3) Efforts that are likely to significantly ameliorate problems:
- Enhancing communication among biodiversity information systems to facilitate identification of shared goals and problems, reduce redundancy and, where appropriate, increase specialization
- Establishing an active global network that can share experiences and agree on techniques, standards and division of labor
- Mapping all data pipelines and processes according to a common metamodel
- Linking data architectures to enable more rapid exchange of information
- Developing robust APIs built for specific research needs
- Developing a common data portal and data format for various types of data
- Building a robust platform for linking occurrence, phenotypic, genetic and environmental data
- Sharing community-maintained vocabularies and semantic technologies, including a single taxonomic backbone
- Promoting a shared approach to collection registration
- Standardizing formats for integrating data
- Forming working groups to address particular challenges
- Launching joint projects that leverage technical expertise; e.g., linking technical expertise of ALA in GIS with bioinformatics expertise in iDigBio
- Sharing advisory staff, software experts, etc., to improve strategic alignment
- Improved reporting of metrics/usage

4) Activities that could promote long-term sustainability:
- Developing high-profile ‘demonstration’ research projects
- Effective delivery of solid, high-value data streams for research, conservation and governmental processes
- Demonstrating value through reports on data use, publications, importance for conservation, etc., as well as more formal impact evaluation assessments.
- Building products that attain an integral and near-essential role in broadly recognized applications and uses
- Aligning with broader-scale initiatives around open-research data infrastructures
- Reducing what needs to be sustained
- Identifying mechanisms for joint funding and sustaining collaborations

Frankfurt workshop attendees agreed on the need for a shared mechanism to develop a truly interconnected approach for improving the global biodiversity data infrastructure. The goal is to enable parties jointly and transparently to develop and agree on priorities on a decadal or longer timescale, and then work in a coordinated fashion to secure funding and sustainable delivery of all agreed components. The

attendees proposed that GBIF should explore inclusion of this role as part of its mission.

In response, the GBIF Governing Board recently approved funding to plan a workshop in mid-2018 in Copenhagen, which will bring together the widest possible array of relevant stakeholders to develop a model for how such a coordination mechanism might work. The workshop will seek to deliver the following products:

1) A whitepaper proposing how such a collaborative planning mechanism might be established, including governance and funding models.
2) Two examples of the type of roadmap documents that the mechanism would be expected to deliver. An initial suggestion is to focus on:
   a) Integrated occurrence data **(Evidence).**—Aligning activities globally to aggregate spatial evidence of species occurrence, removing duplication of effort and ensuring delivery of the most comprehensive integrated data resource possible. This is a pragmatic choice since several key infrastructures are already starting discussions on how to achieve this.
   b) Biodiversity knowledge network **(Culture)**.—Developing models, social networks, reputation and recognition systems to empower and reward professionals and knowledgeable amateurs to contribute to the curation and improvement of digital biodiversity information. This component involves sociological integration to bring together skills from a different set of contributors.

Application to Research

An integrated global network of biodiversity data, expertise and tools is required to enable emerging research on many big-data questions related to human health, climate, agriculture, species discovery, species extinctions, rates of evolutionary change and ecosystem services. Research topics that require digitized data on a global basis include:

- What are the likely impacts of climate change?
- What are the effects of invasive species?
- What are the effects of landscape modification?
- What is the history of life on Earth?
- How are species distributed in geographical, ecological and temporal space?
- What factors lead to speciation, dispersal and extinction?
- What information is needed to develop effective conservation strategies?
- How are specific genotypes distributed in geographical and ecological space?
- How do species (and genotypes) covary across the landscape, and are those covariances likely to persist into the future?
- How are ecologically significant traits distributed geographically?
- What is the extent of phylogenetic constraints on ecological niche evolution?
- What connections between biodiversity and ecosystem services contribute to human welfare?

Answers to these questions are needed to understand and appreciate the diversity of life on Earth as well as the potentially profound societal and economic consequences of global change. Research systems that will be linked with the biodiversity data networks include the Berkeley Initiative for Global Change Biology (BiGCB), which uses

state-of-the-art tools and technologies to mobilize historic and modern biological data to understand how organisms and ecological systems biological systems will respond to future global change; Edaphobase, the database for distribution and ecology of soil organisms at Senckenberg Research Institute; and NEON, investigating impacts of climate change, land use change, and invasive species on natural resources. Although the global network would initially include only North American and European programs, once it is in place linkages will be established with other networks, such as the African Open Science Platform (AOSP), and the Pacific Rim Applications and Grid Middleware Assembly (PRAGMA).