

The Global Biodiversity Information Facility

by Meredith A. Lane

Meredith A. Lane is a public relations and communications officer with the Global Biodiversity Information Facility (GBIF) Secretariat. She can be reached at Universitetsparken 15, 2100 København Ø, Denmark; mobile: +45 28 75 14 84; phone: +45 35 32 14 84; fax: +45 35 32 14 80; e-mail: mlane@gbif.org.

Biodiversity or biological diversity means the variability among living organisms from all sources including, *inter alia*, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems” (Convention on Biological Diversity, Art. 2, para. 1).

The Global Biodiversity Information Facility (GBIF) (www.gbif.org) mission is to make the world’s biodiversity data freely and universally available via the Internet. GBIF was established and is being maintained by an ever-growing consortium of forward-thinking countries, economies and international organizations. Its charge is to carry out specific tasks that are essential to a world-wide infrastructure that can overcome current barriers to the universal availability of species-level biodiversity information.

The GBIF Vision

Science. Significant parts of the tripartite (gene, species, ecosystem) biodiversity information resource are already online, such as the DNA data served by GenBank, EMBL and DDBJ and other sequence (RNA, protein, etc.) data served by various sources such as RNABase, SCOR and ExPASy. These community research resources have already contributed substantially to medical, pharmaceutical and agricultural industries and through these to society, which has footed the bills for the establishment and maintenance of the digital information resources.

Ecological, ecosystem and planet-wide data are

provided online by, for example, MABnet and LTERnet, the U.S. National Aeronautics and Space Administration (NASA) through its Mission to Planet Earth (MPE) and other national and bi- or trilateral consortia. These online ecological resources are just beginning to be thoroughly analyzed and the results synthesized into larger understandings of the functioning of the natural systems of our planet. Nonetheless, they hold great promise for predictive modeling of global climate change and other large-scale ecological phenomena.

Though there are areas of the world that still do not have either full access to the Internet or the on-site infrastructure to fully utilize data that might be received (one hopes that this situation will be rectified in the near future), the molecular and ecological data discussed above are available to anyone and everyone with an Internet connection. One of the reasons that this is so is that the available data have all been collected in the past few decades, during the age of computers.

The piece, in fact the anchor stone, of biodiversity information that to-date has been missing from digital availability is data about individuals, populations and species of whole organisms. These data are not online because most pre-date the computer age. They have been collected over nearly three centuries and recorded in the only medium available: paper and ink. These data are on the labels of natural history specimens, in libraries and in handwritten notebooks or typewritten card files. Species-occurrence data are essential to many kinds of analyses, and one of GBIF’s four major areas of

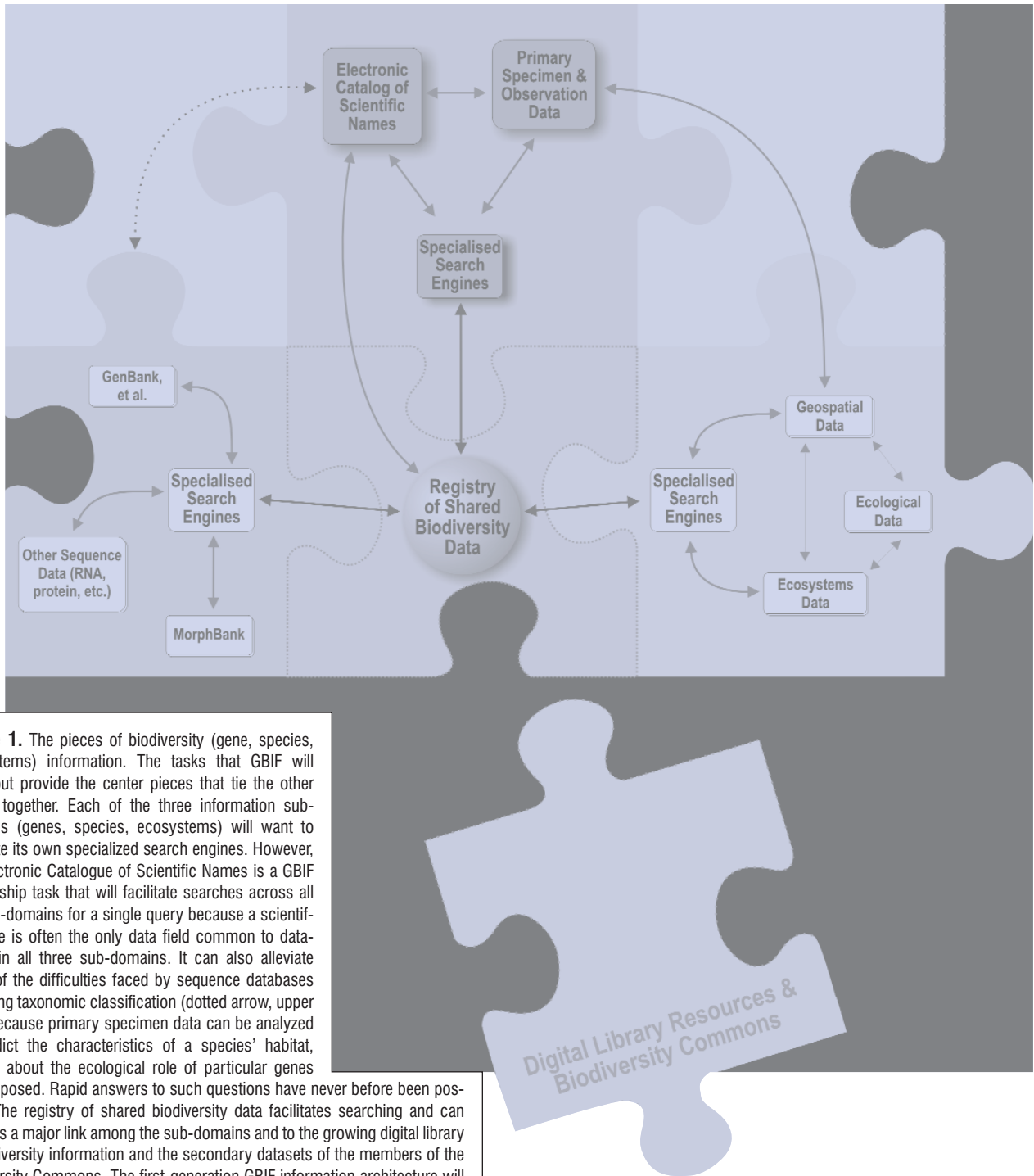


Figure 1. The pieces of biodiversity (gene, species, ecosystems) information. The tasks that GBIF will carry out provide the center pieces that tie the other pieces together. Each of the three information sub-domains (genes, species, ecosystems) will want to generate its own specialized search engines. However, the Electronic Catalogue of Scientific Names is a GBIF partnership task that will facilitate searches across all the sub-domains for a single query because a scientific name is often the only data field common to databases in all three sub-domains. It can also alleviate some of the difficulties faced by sequence databases regarding taxonomic classification (dotted arrow, upper left). Because primary specimen data can be analyzed to predict the characteristics of a species' habitat, queries about the ecological role of particular genes can be posed. Rapid answers to such questions have never before been possible. The registry of shared biodiversity data facilitates searching and can serve as a major link among the sub-domains and to the growing digital library of biodiversity information and the secondary datasets of the members of the Biodiversity Commons. The first-generation GBIF information architecture will be online by the end of 2003; the growth of the electronic catalogue is underway, but is expected to take approximately 10 years to achieve 90% of all names. Digitization of natural history specimens has begun, but needs substantially more investment to speed the work. (Source: GBIF)

work is aimed at making these data digital and therefore more useful and available.

All of GBIF's tasks are aimed at making digitally available the bridging pieces (data about individuals, populations and species) that tie the whole biodiversity information pattern together (see Figure 1). Without the Taxonomic Name Service (a function that uses both the Registry of Shared Biodiversity Information and the Electronic Catalogue of Scientific Names) that will be provided by GBIF, there is no way to obtain a seamless response to a query that requires a call upon both ecological and molecular data, such as "*What other organisms live in the same kinds of habitats as this one from which I have extracted a gene that enables it to tolerate high levels of lead acetate, and are able to do so because of either homologous or analogous gene function?*"

Society and Sustainability. The beauty of digital data is that they can be used over and over again and in many ways. The same information can be put to many purposes by the whole variety of users in the world. Investments made in digitizing scientific data and bringing them online via well-planned information architecture are paid back many times over. Not least of these repayments is that these data are now usable by people other than scientists.

The ultimate source of global and national wealth is natural resources. Biodiversity, the living portion of natural resources, provides an ever-increasing portion of that wealth. Biodiversity has provided the basis of human survival (clean water and air, food, fuel and fiber), not to mention prosperity, since *Homo sapiens* first set foot on Earth. If we humans are to continue to prosper and to leave future generations a healthy place in which to live, we must learn to use living resources in a sustainable way and make it possible for all peoples to share in the benefits of the sustainable use of biodiversity.

If the economic and survival benefits provided by biodiversity are to be equitably shared globally, as required by the Convention on Biological Diversity, a number of factors must be overcome. Primary among these is the need for access to scientific data and information about biodiversity to be as easy and complete in Mongolia or Madras as it is in Madrid or Munich. If scientific advances regarding biodiversity are to be made by any and all of the talents around the world that can contribute to them, access to the data and results that have already been generated must be as easily and fully available, wherever the researcher.

The GBIF Strategy

The GBIF biodiversity informatics infrastructure has two equally important components:

- 1) Computational (standards, interoperability and search engines)
- 2) Content (particularly primary scientific data that are currently "imprisoned" by paper and ink or other, non-digital media)

GBIF is also aware of the digital divide faced by many of the most biodiverse countries and, therefore, undertakes capacity-building activities to help overcome these challenges.

At present GBIF is focusing on making primary, species-level biodiversity data available via the Internet. Primary data are those derived from the direct observation of nature, such as the labels on natural history specimens or culture tubes. Eventually, GBIF will include in its purview secondary data which have been derived through some manipulation of primary data, such as an analysis of pattern or process. This task will likely be done in close collaboration with digital library efforts, as well as the Biodiversity Commons. However, the immediate need is to provide, via the Internet, the label data from more than two billion natural history specimens and uncounted culture collections.

GBIF is an international organization in its own right. Importantly, GBIF is open to participation by *all* countries, economic entities and organizations that can benefit from the open sharing of biodiversity information on a global scale.

Unlike other megascience facilities that are built of *bricks and mortar*, GBIF operates as a virtual facility. The bricks of this facility are the databases, other information resources and informatics tools made available by GBIF participants. The mortar that holds the bricks together is the informatics infrastructure (software tools and the Internet). It is staffed by a small secretariat (14 positions) that works internationally to coordinate national, regional and local biodiversity informatics efforts and bring focus to the activities of the organization as it develops GBIF.

GBIF is distributed and encourages cooperation and coherence; it is global in scale, though implemented regionally, nationally and locally. GBIF, through relationships built by its secretariat staff, works closely with many producers and providers of biodiversity information as well as with its users. In some ways GBIF has the characteristics of a large, distributed public-domain database with a number of interlinked and interoperable modules such as data stores, software and networking tools, search engines and analytical algorithms that enable users to navigate and use the data. It differs from such a unitary database, however, in that it is more comprehensive in content and much more complex in its interconnections.

In summary, the GBIF strategy includes:

- Focus on its mission and specific goals, with intermediate milestones identified in each year's work program;
- Outreach to developing countries;
- Inclusiveness in the manner in which it seeks advice;
- Openness in data sharing and software developments;
- Cost-effectiveness in its partnerships with like-minded organizations; and
- Fund-raising efforts to enhance its product and speed up its activities.

On May 20, 2003, the National Biological Information Infrastructure (NBII) and CSA (Cambridge Scientific Abstracts) announced the launch of the Biocomplexity Thesaurus, a major new resource for the bioinformatics community. The thesaurus is now online and available for use by the NBII nodes, their partners and the public-at-large. The Biocomplexity Thesaurus can be accessed at <http://thesaurus.nbio.gov>.

The Biocomplexity Thesaurus will be integrated into NBII products and services to facilitate more relevant retrieval of and intellectual access to these resources by NBII users. It will be the required thesaurus for all keyword and subject metadata created by NBII nodes. The thesaurus will be used for the cataloging of Web resources, the creation of HTML metadata for Web pages and the creation of new Federal Geographic Data Committee (FGDC)-compliant metadata records for the NBII Metadata Clearinghouse. Accordingly, researchers, scientists, librarians and the general public will be able to use the thesaurus as an aid in creating strategies for searching NBII databases.

This thesaurus represents a major step forward for the biological sciences and the NBII network. It is one of the most comprehensive and freely accessible biological thesauri available online.

The Biocomplexity Thesaurus is a living resource that will be updated quarterly. To this end, the NBII Thesaurus Working Group (TWG) has been established to review recommended additions and modifications to the thesaurus. The goal is to have representation from every node on the TWG. Two NBII regional nodes, Pacific Basin Information Node and Southern Appalachian Information Node; CSA, the thesaurus developer; and the NBII Program Office already have representation on this working group. Representatives from the thematic nodes and the other regional nodes are encouraged to participate as well.

Additionally, a Biocomplexity Thesaurus community has been set up in the NBII portal. All nodes may submit requests for term additions or modifications through a gadget located in this online community. NBII node partners may also submit comments and requests through this gadget. Contributors need not be members of the TWG to access this Community.

Jessica L. Milstead is the primary architect of the Biocomplexity Thesaurus. She is one of the foremost experts in the world on thesaurus development for the biological and natural sciences. She holds a doctorate in library science and

Biocomplexity Thesaurus Launched

Reprinted from U.S. Geological Survey. (2003). *NBII Access*, newsletter of the National Biological Information Infrastructure, 6 (3).



is the founder of the index and thesaurus development company, JELEM (www.jelem.com). Dr. Milstead has extensive experience in both industry and academe. Under contract to CSA, she brought the massive Biocomplexity Thesaurus project to fruition in just 18 months instead of the 24 months originally allocated.

To create the Biocomplexity Thesaurus, Milstead merged, vetted and reconciled the terminology in five large existing thesauri plus one smaller specialty thesaurus collectively covering the biological, environmental, aquatic, ecotourism and sociological sciences. These thesauri include

- CERES/NBII Thesaurus (California Environmental Resources Evaluation System)
- CSA Life Sciences Thesaurus
- CSA Pollution Thesaurus
- CSA Aquatic Sciences and Fisheries Thesaurus
- CSA Sociological Thesaurus
- CSA Ecotourism Sciences Thesaurus.

Milstead used the high-powered MultiTes 8.0 thesaurus development

software package to create the Biocomplexity Thesaurus. Among many other features, MultiTes provides for an unlimited number of hierarchies per thesaurus, an unlimited number of relationships for each individual term, the validation of conflicting relationships and the automatic generation of reciprocal relationships.

The online Biocomplexity Thesaurus is very user-friendly. Tips are offered for easily navigating and searching the thesaurus. Explanations are provided for best understanding search results, such as concepts related hierarchically or associatively to the search entry term.

The NBII (www.nbio.gov) is a broad, collaborative program to provide increased access to data and information on the nation's biological resources. The NBII links diverse, high quality biological databases, information products and analytical tools maintained by NBII partners and other contributors in government agencies, academic institutions, non-government organizations and private industry.

CSA (www.csa.com/csa/about/Biocomplexity.shtml) is a leading producer of bibliographic citation databases and Web resources databases. CSA recently extended its government/private industry partnership with the NBII from 20 to 56 months. CSA's original designation as the "Biocomplexity Information Node" has now been broadened to "NBII Infrastructure and Knowledge Integration Node."