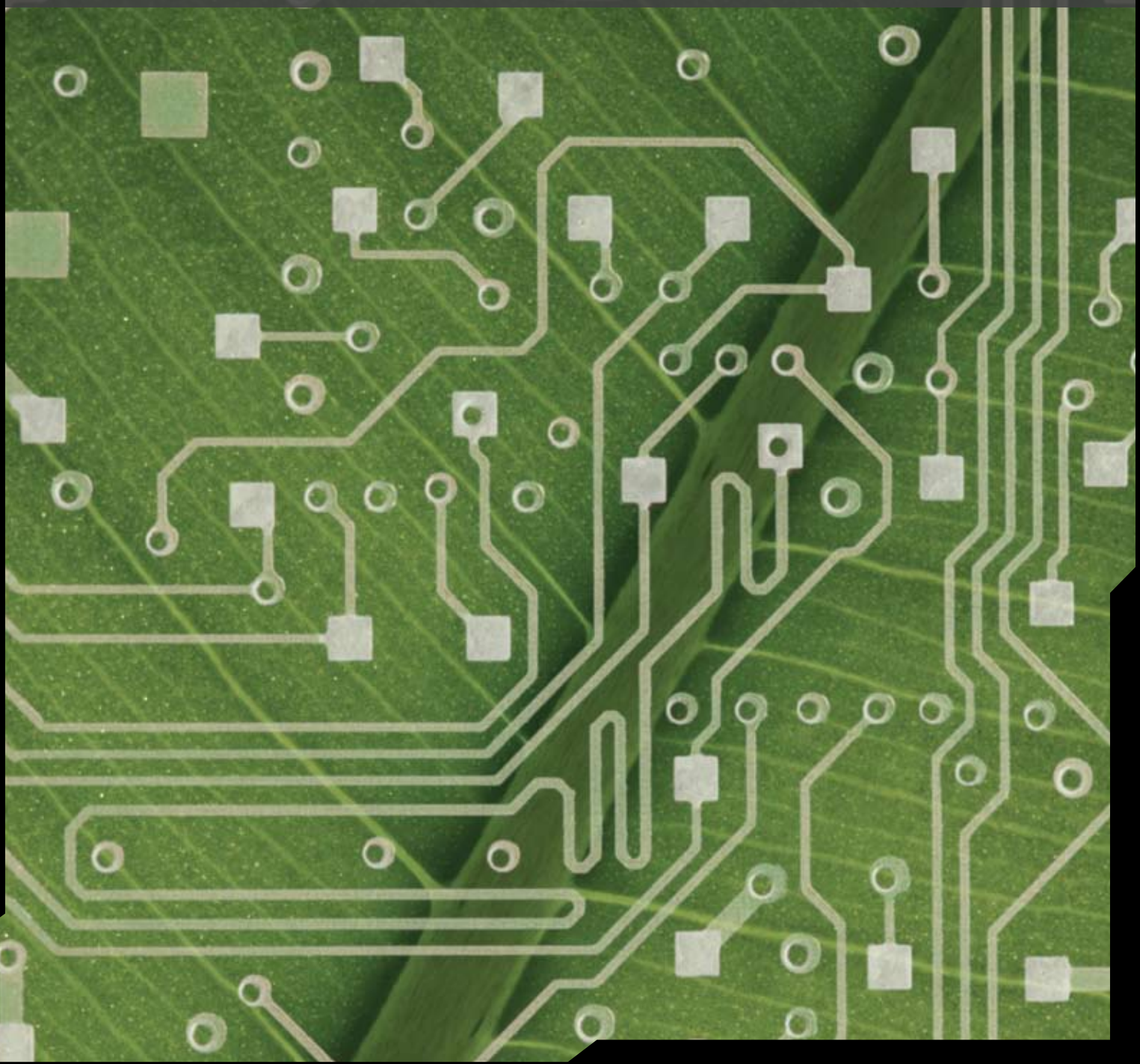
Global Biodiversity Informatics Outlook

Delivering Biodiversity Knowledge in the Information Age



Global Biodiversity Informatics Outlook

Contributing authors

Donald Hobern (coordinating author)
Alberto Apostolico
Elizabeth Arnaud
Juan Carlos Bello
Dora Canhos
Gregoire Dubois
Dawn Field
Enrique Alonso García
Alex Hardisty
Jerry Harrison
Bryan Heidorn
Leonard Krishtalka
Erick Mata
Roderic Page
Cynthia Parr
Jeff Price
Selwyn Willoughby

For affiliations of authors,
see page 36.

Editorial assistance

Tim Hirsch
Sally Hinchcliffe
Samy Gaiji

Acknowledgements

The authors extend their gratitude to all organizers, workshop leads, sponsors, participants and facilitators at the Global Biodiversity Informatics Conference (see Annex on page 36), whose creative discussions and inputs gave rise to the framework outlined in this document. Special thanks are due to the following for additional helpful comments during the drafting period: Donat Agosti, Ana Casino, Walter Berendsohn, Lee Belbin, Don Doering, Gregor Hagedorn, Keping Ma, Michelle Price, Hugo von Linstow and Zheping Xu.

Funding for the Global Biodiversity Informatics Conference and Outlook was contributed by:

GLOBAL BIODIVERSITY INFORMATION FACILITY

CREATIVE-B

AAGE V. JENSEN CHARITY FOUNDATION

JRS Biodiversity Foundation

UNIVERSITY OF COPENHAGEN

Delivering Biodiversity Knowledge in the Information Age

The Global Biodiversity Informatics Outlook

The Global Biodiversity Informatics Outlook helps to focus effort and investment towards better understanding of life on Earth and our impacts upon it. It proposes a framework that will help harness the immense power of information technology and an open data culture, to gather unprecedented evidence about biodiversity and to inform better decisions.

This document is accompanied by a website, www.biodiversityinformatics.org, that will report progress towards each part of the framework and provide a forum for ideas, projects and funding sources supporting the goals of the Outlook.

Foreword

Our knowledge of the natural world and its complexity continues to grow at a staggering rate. We continue to understand more and more of the mind-bending intricacy of DNA-based life and how the various products of evolution interact. In many ways we are still engaged in the same endeavour as the naturalists of earlier centuries. We are trying to develop an understanding of this complex reality and how it works — only now we see even more levels to that complexity than early naturalists could have imagined.

We are fortunate to have in our hands an increasing number of tools to assist us with observing, recording and measuring this complex system. We have rapid sequencing technologies, a wealth of imaging systems, remote-sensing systems, physical and chemical sensors of all kinds, global-positioning tools, the information backbone and processing power of the web and modern high-performance computing, a global workforce of biologists with greater understanding of evolutionary processes than ever before, and an army of amateur observers contributing their skills and efforts. We also have political recognition of the importance of understanding this system and applying that understanding to support a sustainable future for mankind, the planet and all the other species around us.

This outlook proposes a framework for making better use of all these opportunities, to benefit us all. We hope you will join us in building and developing it.

Donald Hobern

Executive Secretary, Global Biodiversity Information Facility
Coordinating author, Global Biodiversity Informatics Outlook

Contents

The origins of the Global Biodiversity Informatics Outlook

This document has been developed in consultation with the community, in a process initiated by the Global Biodiversity Informatics Conference (GBIC).[1] The conference gathered together around 100 experts from a wide variety of disciplines to meet in Copenhagen in July 2012.[2] Scientists, informatics experts, policy makers, and others were invited to identify how best to harness the power of information technology, biodiversity science and social networks to improve our understanding of life on Earth. Through a series of workshops they identified the highest priority questions and the tools that would be needed to answer them and outlined the steps that would need to be taken to create those tools and deploy them effectively.

The ideas and priorities identified during the discussions at GBIC were subsequently distilled and structured by the workshop leads and a core writing team. This document aims to present a consensus view and framework that will be widely adopted by the people and institutions that will be key actors in its implementation.

Executive Summary

In order to preserve the variety of life on Earth, we must understand it better. The world's governments missed their target to reduce significantly the rate of biodiversity loss by 2010. One of the main reasons for this was the shortage of available information. To create appropriate policies to protect habitats we must understand what they contain, how the species within them interact, and how they might respond to changes and pressures, natural and manmade. With the adoption of the Strategic Plan for Biodiversity 2011-20, including the Aichi Biodiversity Targets, governments have re-affirmed the importance of preserving and restoring biodiversity and maintaining the planet's ecosystem services. **Biodiversity informatics does not merely contribute towards meeting these goals: it is fundamental to achieving them.**

The last 250 years of biodiversity research have produced a wealth of information, but too much of it is still locked away and inaccessible. At the same time, new technologies and scientific approaches are today unleashing a flood of new data that could help us towards this fundamental understanding, but only if we are able to harness it effectively. **Mobilizing all biodiversity data, old and new, in a structured and standardized form would enable a vast range of uses, creating new opportunities for research and putting biodiversity-related policy making on a sounder footing.**

Much progress has been made in the past ten years to fulfil the potential of biodiversity informatics. However, it is dwarfed by the scale of what is still required.

The Global Biodiversity Informatics Outlook (GBIO) offers a framework for reaching a much deeper understanding of the world's biodiversity, and through that understanding the means to conserve it better and to use it more sustainably.

The GBIO identifies four major focal areas, each with a number of core components, to help coordinate efforts and funding. The co-authors, from a wide range of disciplines, agree these are the essential elements of a global strategy to harness biodiversity data for the common good.

In summary, the GBIO proposes actions in the following key areas:

- Creating a **culture** of shared expertise, robust common data standards, policies and incentives for data sharing and a system of persistent storage and archiving of data.

- Mobilizing biodiversity **data** from all available sources, to make them promptly and routinely available. Data should be gathered only once, but used many times. This includes data in all forms from historic literature and collections to the observations made by citizen scientists; from the readings of automated sensors to the analysis of the genetic signatures of microbe communities.

- Providing the tools to convert data into **evidence** by enabling those data to be discovered, organizing them into views that give them context and meaning. This includes major collaborative efforts to improve the accuracy of data and their fitness to be used in research and policy; to provide a taxonomic framework; and to organize information about the traits of species and the interactions between them.

- Generating **understanding** of biodiversity and our impacts upon it, by applying the evidence in models, tools for visualization and identifying gaps to prioritize future data gathering.

We invite funders, policymakers, researchers, information technology specialists, educators and the general public to unite around the framework detailed in the following pages. The rewards of coordinated action will be as exciting and significant as the great scientific collaborations to advance our understanding of space, the human genome and the fundamental particles of matter.

Introduction

"I am convinced that the lack of adequate biodiversity monitoring is at the heart of our difficulties to make convincing arguments. A Government that sees what its policies do to biodiversity because it has access to reliable data will be less likely to risk biodiversity loss and more likely to find solutions that embrace biodiversity as a part of such solutions." – Braulio Dias, Executive Secretary, Convention on Biological Diversity (CBD), message to Global Biodiversity Informatics Conference.

"Not only will biodiversity informatics projects need to deal with an explosion in the amount of biodiversity-relevant data, they may well need to accommodate data that are of a conceptually different form." – Bob Robbins, Global Biodiversity Informatics Conference.

A fundamental problem confronts us as we seek to preserve the diversity of life on Earth: we need to know much more about biodiversity if we are to understand how best to protect it. We missed a target to reduce significantly the rate of biodiversity loss by 2010 partly because of a lack of information. As the 2010 deadline arrived, we had few reliable indicators that could provide us with a clear picture of the status of global biodiversity, and we lacked the information and tools to foresee the impact of human activities.

And yet, we have a wealth of data — it is just that too often it is locked up in museum drawers or printed publications, in isolated desktop computers or in incompatible digital formats, and in multiple human languages. Fortunately we live in an increasingly networked world where more and more people understand that information gains most value when it is shared. The new norms of social media and the development of the semantic web, as well as advances in data capture and handling 'big data', have transformed our approach to information. There are improvements in modelling and information processing, and increasingly collaborative developments within the biodiversity community itself. Combined, these developments give us the opportunity to close the current gap in our understanding, by mobilizing all available biodiversity data – past, present and future – and making them useful for science and society.

Imagine a world in which every field observation, bird-ringing record, specimen image or species description was permanently stored in a way that it was accessible, searchable and usable, through standards that were commonly agreed and well understood. Imagine that the efforts of scientists and experts to improve, validate and synthesize these records were also automatically captured and stored in a transparent and accessible form, along with the information held in the biodiversity literature from Linnaeus to the present day. Nothing is lost or wasted, no effort needs to be repeated, freeing researchers to concentrate on the areas where we know the least. It would be possible to develop large-scale ecological models, based on a constantly improving and growing body of data. Indeed, it would be possible to do quite unexpected things with the data, outcomes we have not anticipated. Not only would this contribute to our efforts to understand and track the rate of biodiversity loss — it would enable better policy choices to be made to slow and even halt this loss.

After the failure to meet the 2010 biodiversity target, governments agreed the ambitious 2011-20 Strategic Plan for Biodiversity, including the new and more detailed Aichi Biodiversity Targets. The plan aims to halt the loss of biodiversity in order to ensure that by 2020, ecosystems are resilient and continue to provide essential services (see page 7).[3] As a result, policy makers now urgently require the means to monitor the status and trends of life on Earth, to model the impact of changes, and to support the right policies to slow and ultimately to end the depletion of the planet's biological diversity. Aichi Target 19 explicitly sets the goal of improving, sharing and applying knowledge about biodiversity (see page 7); but in fact biodiversity information will be fundamental to the achievement of all of the Aichi Targets.

Improved access to information will also be critical for the new Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)[4], which aims to strengthen the science-policy interface for biodiversity and ecosystem services for the conservation and sustainable use of biodiversity, long-term human well-being and sustainable development.[5] Improved access to information is not only vital for the assessments that IPBES is expected to deliver, but it has also been identified as one of the key capacity building needs the platform is required to meet as part of its functions.[6]

The GBIO and the framework it proposes will serve as essential support to these and many other policy needs, by extending our understanding of ecosystems and the services they provide, making conservation and biodiversity management policies more effective.

We can provide the information tools that researchers and policy makers need, but only if we work together to mobilize the data we already have and ensure that the data we collect in future are fit for purpose. We will also need to collaborate to put in place the core infrastructure and data-related policies that will provide a solid and sustainable foundation for future research and future decisions affecting biodiversity.

Actions must be both local and global. Each country should be committed to increasing the knowledge base and to strengthening local, national, regional and global infrastructures by making its data and information openly available. When a country is committed to producing data and knowledge, the chance that this data and knowledge will be used in policies and decision support systems is greater. Such buy-in will be essential if the tools enabled through the following framework are to achieve their potential in addressing the biodiversity crisis.

Biodiversity informatics

Biodiversity research seeks to understand the variation within and between species, and their relation to geographical, ecological, temporal and anthropogenic factors. It explores the interactions among organisms, including with humans; and between organisms and the environment. Further research can then explore trends, analyse drivers of change and make predictions about the future. All of these activities depend on access to the best available data on recorded observations for each species, supported by the best possible understanding of the biases and uncertainties associated with each dataset.

Biodiversity informatics relates to the use of information technology (IT) to support these needs, by organizing knowledge about individual biological organisms and the ecological systems they form. Over time, biodiversity informatics will deliver an increasingly interconnected digital resource supporting scientific research of the natural world.

The first decade of the millennium witnessed growing interest in biodiversity informatics and the emergence of a community of scientists and IT professionals who have collaborated to develop new capabilities in the field.

A number of global initiatives have been established to further the goals of biodiversity research through informatics, including the Species 2000 Catalogue of Life (CoL),[7] Biodiversity Information Standards (TDWG),[8] the Global Biodiversity Information Facility (GBIF),[9] Encyclopedia of Life (EOL),[10] the Consortium for the Barcode of Life (CBOL),[11] the Biodiversity Heritage Library (BHL)[12] and the Group on Earth Observations Biodiversity Observation Network (GEO BON).[13]

Despite these initiatives, and many more at global, regional and national scales, the biodiversity informatics landscape remains very fragmented. The challenge for the GBIO is to help coordinate not just these efforts but the contributions of all biodiversity research.

Contribution of the GBIO to Aichi Targets

Focus area A: Culture

The whole of the GBIO framework will contribute to Aichi Target 19, improving the world's biodiversity knowledge base, and this is especially true of the components outlined in focus area A. This foundational layer will underpin contributions to other targets from the remaining focus areas as detailed below.

Focus area B: Data

Increasing the use of crowd-sourcing and volunteers, especially with field observations, will support **Target 1**, making people aware of the value of biodiversity, in a very direct and immediate way. Indigenous and local communities will need to be engaged in the effort, so this will also support **Target 18**, respecting traditional knowledge and practices. Making field and remote sensed data immediately available will be key to tackling **Target 9**, identifying and targeting invasive species, while better integration of genetic data will also support efforts to meet **Target 13**, safeguarding genetic diversity.

Focus area C: Evidence

Having access to increasingly comprehensive data about species, their occurrences, traits and interactions will be essential to achieve **Target 9**, identifying and targeting invasive species and **Target 12**, preventing extinction of threatened species, while integrated occurrence data will be important for **Target 4**, implementing plans for sustainable consumption, **Target 5**, halving the rate of loss of natural habitats, **Target 6**, managing aquatic stocks sustainably, and particularly **Target 11**, the creation and expansion of protected areas.

Focus area D: Understanding

The real impact this focus area will have is on improving the effectiveness of decision making and policies, vital to **Target 3**, phasing out harmful subsidies, **Target 4**, implementing plans for sustainable consumption, **Target 5**, halving the rate of loss of natural habitats, **Target 6**, managing aquatic stocks sustainably, **Target 9**, identifying and targeting invasive species, **Target 11**, creation and expansion of protected areas, and **Target 12**, preventing the extinction of threatened species. Arguably as important, improved visualization and communication of the information will transform people's understanding of biodiversity, underpinning **Target 1**, making people aware of the value of biodiversity, and hence providing the political will to make the other targets a reality.

Aichi Targets: an overview

The Aichi vision is that "*By 2050, biodiversity is valued, conserved, restored and wisely used, maintaining ecosystem services, sustaining a healthy planet and delivering benefits essential for all people.*"

This is to be achieved through five strategic goals, including 20 individual targets. The goals are:

Strategic Goal A: Address the underlying causes of biodiversity loss by mainstreaming biodiversity across government and society
Strategic Goal B: Reduce the direct pressures on biodiversity and promote sustainable use
Strategic Goal C: Improve the status of biodiversity by safeguarding ecosystems, species and genetic diversity
Strategic Goal D: Enhance the benefits to all from biodiversity and ecosystem services
Strategic Goal E: Enhance implementation through participatory planning, knowledge management and capacity building

Of the individual targets, the GBIO will form a cornerstone of efforts to meet Target 19: "*By 2020, knowledge, the science base and technologies relating to biodiversity, its values, functioning, status and trends, and the consequences of its loss, are improved, widely shared and transferred, and applied.*"

The GBIO will also provide important support for meeting several other targets, across all of the five strategic goals.

Source: Aichi Biodiversity Targets. https://www.cbd.int/sp/targets/

What you can do

The overall aim of this document is to outline the proposed GBIO framework and its interdependent components, and to make the case for its adoption. It also offers a snapshot of what progress has already been made and priorities for the short, medium and long-term future, as well as the steps needed to take it forward. The GBIO is intended to be a dynamic and interactive process, and the website at www.biodiversityinformatics.org will be updated with projects, ideas and funding sources proposed by the community, providing much more detail about the individual components as they evolve over time. Please visit this website and share your ideas and expertise.

If you're a policy maker

We invite you to look at the relevant legislation and make the changes necessary to underpin a culture of data sharing, to provide the right policy incentives, and above all to fund the data mobilization effort. We also invite you to work with researchers to develop the decision-making tools you need to support your conservation policies and to make use of them.

The priority steps will be to communicate data and analysis needs in relation to biodiversity, to fund digitization efforts, to develop long-term national or regional data repositories, and to put in place open access legislation to ensure data are made freely and persistently available in appropriate and usable forms.

If you're a funder

We invite you to align your funding criteria with any or all of the components listed here, and make collaboration with the GBIO effort a condition for funding relevant projects. We also invite you to consider how long-term data mobilization and storage can be supported.

The priority steps will be to develop policies ensuring project data are made freely and persistently available in appropriate and usable forms and to fund projects working within the GBIO framework outlined here.

If you run a regional, national or international biodiversity organization

We invite you to build the GBIO framework into your mission and align your ongoing and project work to take into account the components outlined here. We also invite you to mobilize any relevant collections you hold, in whatever form, and share them freely and openly.

The priority steps will be to identify which components you can best support and how you can align your planned and existing work with the GBIO framework. If you have run successful pilots or projects in any of the component areas you should share these and scale them up, while piloting any promising new approaches and adopting best practice from elsewhere. Any disincentives to sharing or annotating data should be dismantled.

If you are an owner or custodian of biodiversity data

We invite you to make this information permanently, freely and openly available for reuse, so that it can form part of the wider data resource. If you have information that is not yet digitized, we invite you to take advantage of the incentives and resources available to mobilize these resources as quickly as possible.

The priority steps will be to agree common systems with other data custodians for serving data, managing annotations and recognizing onward use of data; and to adopt common standards, tools and licences where possible. It will also be important to collaborate with other data providers to recognize, develop and share best practice in priority areas such as digitization, error detection and correction, annotation systems, and crowd sourcing.

If you're a biodiversity researcher

We invite you to consider how your research and project work could contribute to any or all of the individual components described here and to build their goals into any future proposals you make. We also invite you to consider contributing to any follow-up activities to this framework in whatever capacity as well as contributing corrections and annotations to data where possible.

The priority steps will be to help us understand in more detail where the knowledge and information gaps are that biodiversity informatics can address, and to identify how you can help to bridge them. You may wish to identify the most promising modelling and visualization approaches and work with others to help establish the requirements (data, standards, workflows and techniques) needed to make them a reality.

If you're an IT professional or biodiversity informatics specialist

We invite you to consider what standards, technologies, protocols and tools you could deliver in support of the components described here, and to contribute your technical expertise to any follow-up activities to this framework. We also invite you to consider what opportunities the mobilization of biodiversity information offers to create exciting new tools.

The priority steps will be to work with researchers to identify gaps in existing standards and identify commonalities and map overlaps in standards from other disciplines. Developing and refining modelling and visualization tools will also be a high priority area.

If you're a member of the public

We invite you to investigate the opportunities to share your knowledge and passion for biodiversity through participation in biodiversity-related science projects and to lobby your government and other organizations to increase support for biodiversity research and for policies to encourage free and open access to data.

The priority steps will be to encourage development of resources making it easier for the public both to contribute biodiversity records, and to discover biodiversity data and information in accessible formats.

What happens next?

The GBIO framework outlines the priority steps we are inviting individuals, organizations and nations to take. Each component also provides a more detailed set of actions that, over the next five to ten years, would make the vision outlined in the framework a reality. A **GBIO working group** will plan follow-ups to this document, helping to track projects, funding and ideas around the framework, and develop mechanisms for monitoring progress in each of the focus areas and components. The website (www.biodiversityinformatics.org) will enable people and organizations to register projects and offer support for particular activities. The website will also act as a clearing house for ideas and tools, and showcase some of the results as they are developed.

Members of the GBIO working group will also collaborate with key networks and processes such as the Convention on Biological Diversity and the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) to ensure that the framework and its components help to meet key biodiversity information needs in coming years.

A second Global Biodiversity Informatics Conference is planned for 2014, and this will be an opportunity to follow up on the framework presented here. Please watch the website for details.

The GBIO framework

The framework described here is organized into four focus areas, each of which is broken down into several core components. All four interconnect and strengthen each other; all four are needed if biodiversity informatics is to achieve its full potential. Each focus area – and the individual components within them – can be progressed independently but as they develop they should start to feed into and reinforce each other, making them together far greater than the sum of their parts.

At the root lies the **culture** focus area which puts in place the necessary elements to turn biodiversity information into a common and connected resource – stable and persistent storage, pooled expertise, the culture and policies to support sharing, and common data standards. Building on those foundations, the **data** focus area aims to accelerate the mobilization of data from all sources, unlocking the knowledge held in our collections and literature, improving data quality and filling in gaps, and bringing observations and data from all sources from satellites to genomes online. The **evidence** focus area deals with refining, structuring and evaluating the data, to improve quality and place it within a taxonomic framework that organizes all known information about any species. Finally, the **understanding** focus area enables a broader synthesis, providing the modelling tools to enable us to look at whole ecosystems, make better policy decisions and react to any changes.

The diagram below shows how the focus areas interconnect, and breaks them down into their individual components.

†Considered to be of high urgency, but have made limited progress to date

Focus area A: Culture

Putting the foundations in place to make biodiversity data an openly shared, freely available, connected resource.

Without stable foundations, all of the other components outlined here risk proving both fragile and fragmented. We cannot fill the gaps in our understanding of biodiversity without pooling the past, present and future efforts of many projects and individuals — and making sure that the resulting data are neither lost nor wasted. In an ideal world, every fragment of biodiversity knowledge generated would be held as part of a common global resource — able to be preserved, searched, found, reused and linked in ways its creator never imagined.

With the development of information and communication technologies in the last decades, knowledge production increasingly involves actors from different disciplines, specialisms, institutions, countries and cultures.[14] This is particularly true in areas such as biodiversity and the promotion of sustainable development. Communication is crucial throughout the process, and the traditional paradigm of sharing scientific data and results only through publications in books and specialized journals is not sufficient.[15] Nor is it enough simply to make data available on the Internet. The data infrastructure must be capable of offering long-term preservation and curation of data, searchable both by humans and by machines, and of serving the data in useful and usable formats which are interoperable with other systems.

The culture focus area addresses this by directing efforts on the changes in community attitudes and processes needed to make biodiversity data open, shared and reusable; and the social, legal and technical underpinnings needed to maintain those data resources in a stable and permanent form. It also covers incentives to encourage researchers to make full use of the opportunities offered by this infrastructure. The goal is to engage the whole community in developing and managing the world's biodiversity data and ensuring they remain freely available to all. This will build on some of the same tools and cultural changes as those used by other fields of research and data management.

Components:

- **A1. Open access and reuse culture**: make open sharing of data standard practice through public funding and other incentives and through proper attribution and recognition of primary data resources, data creators and curators, including individuals as well as institutions.
- **A2. Data standards:** deliver a flexible set of data standards that support the reuse and interoperability of all biodiversity data.
- **A3. Persistent storage and archives:** provide a distributed network of data repositories for all types of biodiversity data, along with consistent handling of metadata, identifiers, licences, tools and services.
- **A4. Policy incentives:** ensure that public policies, legislation and funding initiatives at all scales combine to reinforce this strategy and support its individual components.
- **A5. Biodiversity knowledge network:** create the technical infrastructure to support curation and annotation of data using the best-available community expertise, in a way that makes such curation immediately visible to future users as well as providing feedback to data holders.

Progress: With the exception of data standards, where there has been a long-standing community effort to develop common vocabularies and structures, progress at the global scale has generally been limited in this area although individual countries have made significant advances towards making scientific data openly accessible.

Priorities and dependencies: This focus area acts as a foundation, enabling all of the other areas rather than directly depending on them. However, we anticipate that as the other elements develop, they in turn will feed back into this focus area, setting up a cycle of positive reinforcement. For instance, as more systems are built combining multiple data sources, the **underlying data standards (A2)** and **persistent storage and archives (A3)** will need to be refined to support them, in turn enabling more sophisticated systems to be built. Within the focus area, the first priority has to be the development of policies that promote an **open access and reuse culture (A1)** as without a willingness to allow data not just to be used but also to be annotated and combined in unanticipated ways, the whole enterprise will fail. In practice, this will require

having the right **policy incentives (A4)** in place – or at least to remove disincentives to providing access to data. Both of these components will also benefit from a positive feedback loop: data sources will gain prestige from being involved in high profile projects, encouraging other data providers to join in. Similarly, as users and data sources alike see benefits from the **biodiversity knowledge network (A5)**, through **fitness-for-use and annotation (C1)** it will be more widely adopted and used, generating network effects.

Encyclopedia of Life
http://eol.org/

An initiative to bring together information on species into a single database, accessed via an online portal. Data contributions to EOL come from individuals and organizations, and are reviewed by a community of voluntary curators who help improve the quality of content on the website. EOL has created three levels of curators, based on each person's expertise and experience. Each curator's work is displayed on their profile page, and community members can assess each other's contributions.

PANGAEA
www.pangaea.de

An open access library for georeferenced data from earth system research. Each dataset is assigned a Digital Object Identifier (DOI), with which it can be identified. The system ensures long-term availability of its content through a commitment of the operating institutions.

GenBank
http://www.ncbi.nlm.nih.gov/genbank/

An open access, annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which also comprises the DNA DataBank of Japan and the European Molecular Biology Laboratory. Data sharing between these three organizations occurs daily. In addition, most journals require DNA sequences cited in articles be submitted to a public sequence repository, such as GenBank.

Darwin Core Archive
http://rs.tdwg.org/dwc

A biodiversity informatics data standard which uses terms from the Darwin Core body of standards to produce datasets for species occurrence or checklist data, as well as accompanying metadata. It helps provide a stable, standard reference for sharing information on biological diversity.

Canadensys
www.canadensys.net/

A Canada-wide network which aims to unlock the information held in biological collections. Data are published using the Darwin Core standard, and a central web portal provides access to the network's specimen and geospatial data, as well as names from the Database of Canadian Vascular Plants (VASCAN) and the Catalogue of Life. As of August 2013, Canadensys hosted more than 1.2 million records from 20 collections, also available through the Global Biodiversity Information Facility (GBIF).

Office of Science and Technology Policy
www.whitehouse.gov/administration/eop/ostp/about

Established by the United States Congress to advise the President on the effects of science and technology on domestic and international affairs. In 2013, the OSTP announced new requirements for publicly funded research projects to adopt policies on open data access. An Executive Order made open and machine readable data the norm for government information.

The projects and initiatives highlighted here are for illustration only – many more contribute to the objectives of GBIO. Each will often cut across several focus areas, as indicated by the icons (see page 11 for the icon key).

A1. Open access and reuse culture

Making data sharing the norm.

Progress: some progress (issues understood, needs operational implementation)

At the national and institutional level, individual researchers and projects need to build data sharing into their daily work. Researchers are not necessarily rewarded for providing and improving raw data, but are judged largely on their publication record. This gives them an incentive to control access to their datasets until after they have published their results. As a consequence, datasets are either built for a particular project or publication and access is restricted; or sharing of data becomes a neglected effort, operated on an inadequate budget despite its great value to the community at large. The whole of this framework relies on freely available, reusable data, yet even now some new data are being published in restricted forms, so it is important to tackle this as a matter of urgency.

In some fields, such as genomic research, publication relies on the underlying data being deposited in a common data store such as GenBank.[16] The next step will be to broaden this to other fields, so that funding or publication is compromised if the underlying data are not made permanently or openly available.

In the short term, the priority will be to implement mechanisms for citing data, including use of DOIs and **data standards (A2)**, and promoting recognition of data owners and data managers, in concert with changes to **policy incentives (A4)** to use data sharing as a criterion for awarding funding. In the medium term, providing data and making improvements to data quality should become valued as a service to science, giving institutions and individuals an incentive to make data available. In the long term, data sharing through permanent archives or national repositories will become part of the language of science, just as citing publications or type specimens is now.

A2. Data standards

Ensuring data can be understood and used across systems and across disciplines.

Progress: significant (significant progress made, further investment needed to complete)

Simple and clear but rigorous data standards ensure that both machines and humans can interpret and use data arising from thousands of different sources. As community-wide efforts accelerate the digitization and collection of data, including from unstructured and non-traditional sources (**focus area B**), having the right standards in place from the start will be crucial. Aggregating, integrating and simply discovering data (**focus area C**) all depend on common structures and vocabularies to work well. Closing gaps in the existing set of standards and driving the uptake of robust and well-supported standards are therefore urgent tasks.

The importance of common data structures has long been understood in the biodiversity informatics community, with Biodiversity Information Standards (TDWG)[17] leading efforts to develop common standards. The next steps will be to work with the other component areas to identify where the most important gaps are in existing standards and whether there are commonalities and overlaps with standards from other disciplines.

In the short term, the priority will be to develop interoperable standards and common vocabularies to support planned data use and reuse in all components of this framework. In the medium term, as the use of the underlying data sources becomes more sophisticated these standards will evolve from simple data structures and vocabularies. In the long term, they will develop into structures capable of supporting full semantic reasoning, fully integrated with the relevant standards from other disciplines, from geo-sensing to socio-economics.

*A3. Persistent storage and archives

Creating a stable data archiving infrastructure to ensure no data are lost or mislaid.

Progress: limited (needs further development)

The costs of digitizing and organizing data are significant and we cannot afford either to lose data or to digitize a second time. Many of the records already created are held in legacy systems or systems created for short-term projects whose funding is ended, making them vulnerable to being lost or taken offline. Internet technologies are fast moving and constantly evolving, meaning that even when records still exist, links to them may be broken or identifiers changed. Data sources may disappear, go offline or change protocols, making any systems built on top of them unreliable and increasingly costly to maintain. Without persistent, long-term digital archives being available to house the data gathered in **focus area B**, or to form the foundation for the tools and aggregators in **focus area C** and **focus area D**, we risk wasting time, money and effort. As a key underpinning for the whole of the GBIO framework – and one where very limited progress has been made – providing stable data storage is a matter of very high urgency.

Biodiversity Information Standards (TDWG) and the Global Biodiversity Information Facility (GBIF) have already worked to standardize some common protocols and to build a consensus around persistent universal identifiers for biodiversity data.[18] Such identifiers are essential to allow all users and applications to refer reliably to particular datasets or records. However, these offer limited benefits unless the associated data also remain reliably accessible in stable locations on the web and in usable formats. Significant planning and investment is required to deliver persistent repositories that guarantee long-term access to data, and that can be developed to offer additional services to support community peer-review and annotation and to ensure that the data remain accessible and usable as data access technologies change. Most datasets are today held in private or institutional databases which cannot guarantee this stability.

In the short term, clear recommendations are needed for researchers and projects on how best to organize their data to simplify future archiving and curation including providing stable identifiers, and for institutions and research infrastructures to plan storage facilities that will guarantee long-term access and interoperability. In the medium term, national publicly funded data repositories should be established which follow these recommendations and provide a free, or low-cost, persistent home for research data and key citizen science data products. In the long term, global collaboration should ensure that all data sets are maintained through replicated copies.

* Considered to be of high urgency, but have made limited progress to date.

* A4. Policy incentives

Creating a policy framework that actively encourages the sharing and reuse of biodiversity data, however the data have arisen.

Progress: limited (needs further development)

Science is built on shared knowledge, yet the incentives and funding mechanisms that support research sometimes act against an open data culture. Governments are understandably keen to protect national interests, including data gathered at taxpayers' expense, while institutions sometimes see their data stores as a possible source of income or the basis of future research funding. Even where policies support open access, short-term project funding does not support the ongoing maintenance and serving of data, and funding bodies are reluctant to support the sort of day-to-day running costs that are needed to provide long-term data access. As this component enables the vital **open access and reuse culture (A1)** it should be tackled as a matter of urgency.

Many governments now have open access legislation at least for government-funded data. For example, the United States Office of Science and Technology Policy (OSTP)[19] recently mandated all federal research agencies to have clear policies to increase open access,[20] and the South African National Biodiversity Institute (SANBI)[21] has legislative responsibility to organize national biodiversity information.[22] The next step will be to encourage all governments to follow suit.

In the short term, the priority should be to concentrate funding on projects that make their data freely and openly available to all users, targeting the components in **focus area B**. In the medium term, governments and funding bodies should ensure that their guidelines support open access data, both *de jure* and *de facto*, by ensuring the resulting data archives are on a sustainable footing both technically and financially, in concert with the efforts to create persistent **storage and archives (A3)**. Governments and legislators should ensure that laws favour open access and enhancing biodiversity knowledge. Sensitive data must be dealt with as an exception. In the long term, all projects should build permanent data access into their plans, with sub-national, national and international structures in place to support it.

A5. Biodiversity knowledge network

Benefitting from the expertise of the whole global community.

Progress: limited (needs further development)

Researchers in biodiversity have long had a culture of curating and annotating data — from identifying specimens to correcting and cleaning up entire downloaded datasets. These efforts are a key part of the data validation process: even with the best automated tools, identifying and correcting most errors still requires an expert, human eye. Yet these annotations are not always made available to the original data owners, and even when they are, there may be neither the resources nor the mechanisms in place to incorporate them. As a result, mistakes get replicated or have to be repeatedly corrected, duplicating effort, while there is little incentive for researchers to continue to correct and annotate records more widely.

Data aggregators generally encourage users to report mistakes; several GBIF national nodes have developed systems of data curation, including amateur networks to curate citizens' observations[23] while the EU OpenUp! project[24] includes a data quality toolkit for GBIF data. Some projects are already using expert curation for aggregated data, for example the Encyclopedia of Life[25]

* Considered to be of high urgency, but have made limited progress to date.

and the Fish Barcode of Life Initiative (FISH-BOL).[26] However, too often these use *ad hoc* systems and require an extra effort on the part of the contributors, especially if they want to make corrections in many different sites, while data providers or publishers may not feel confident in trusting changes submitted over the Internet. The next step will be to agree with individual institutions and projects how data cleanup efforts can be recognized and valued, putting the incentives in place to ensure that annotations are made and fed back into the system. In combination with the **fitness-for-use and annotation (C1)** component – which considers the systems needed to enable annotations to be integrated into the data – this will be the first step towards making distributed data curation the norm.

In the short term, the priority should be developing a shared identity management system for contributors, whether professionals or citizen scientists, so that they can have a common identity and contribution history across platforms — particularly the key data networks and publishers. In the medium term, key data networks will be able to trace back any changes to the original contributor and over time it will be possible to use metrics to value contributions automatically, based on the contributor's past history. In the long term, annotating data will become the norm and the curation of data will come to be considered a shared responsibility among the biodiversity community.

Focus area B: Data

Mobilizing biodiversity data from all sources and organizing it in forms that can support large-scale analysis and modelling.

Over the centuries, our knowledge of biodiversity has been built from many millions of observations and measurements, as well as countless publications which are now held in a variety of digital and non-digital forms. These resources may have been originally collected for a single purpose, but could further our understanding in many other fields. While some of these are structured and standardized so that they are automatically accessible, the majority are not, meaning that they can only be reused with difficulty. As a result, the efforts of both professional biologists and, increasingly, citizen scientists are not achieving their full potential, leading to wasted time and duplicated effort.

The data focus area addresses this by putting in place the tools and standards to ensure that such information is gathered once but used many times. It offers ways to accelerate and coordinate current digitization and data gathering efforts and to ensure that the resulting information is as useful and freely accessible as possible.

Components:

- **B1. Published materials:** developing mechanisms to extract the biodiversity data currently embedded in publications and other, multimedia formats and to provide them as freely available, standardized and structured information.
- **B2. Collections and specimens:** developing and sharing more efficient techniques to accelerate the efforts to digitize and capture historic data from collections.
- **B3. Field surveys and observations:** capturing all biodiversity observations, including sounds and images, and making them available as soon as they are made, or within a defined period.
- **B4. Sequences and genomes:** capturing all relevant data from genomic activity, including vouchered reference sequences, environmental metagenomics, genetic variation and full genomes.
- **B5. Automated and remote-sensed observations:** exploiting opportunities for automated and semi-automated recording and identification of species and populations from sources ranging from satellite images down to automated gene sequencing.

Progress: This focus area is probably the most advanced of the four, with the community already undertaking significant digitization projects and most new information now automatically held in digital form. However, such is the scale of the task that future projects will need to use more automation and algorithmic techniques. Moreover, digitized information of all kinds needs to be standardized and given a structure that enables it to be automatically processed: the importance of this task should not be underestimated.

Priorities and dependencies: Completing this focus area depends primarily on the right **policy incentives (A4)** and political will: governments and funding bodies need to provide the substantial resources required. An **open access and reuse culture (A1)** will encourage institutions and copyright holders to share the resulting information while **persistent storage and archives (A3)** will ensure the digitized material has a permanent home. In order to scale up efforts, the work can no longer be confined to the professionals: projects will need to harness the skills and enthusiasm of the amateur community of 'citizen scientists' both to gather and to annotate data, and the infrastructure will need to be in place to handle these contributions (**A5. biodiversity knowledge network; C1. fitness-for-use and annotation**). With more distributed data gathering efforts – and more disparate sources of data becoming available – clear and interoperable **data standards (A2)** will be essential and the parallel development of a comprehensive **taxonomic framework (C2)**, adequate spatial and temporal tags through **integrated occurrence data (C3)** and online identification tools based on **aggregated species trait data (C4)** will do much to improve data quality. The components of this focus area will to a certain extent depend on each other: accelerated **collection and specimen (B2)** digitization efforts will benefit from data extraction techniques developed for **published materials (B1)**, while **remote and automated sensing tools (B5)** will undoubtedly benefit from refinements developed for **field surveys and observations (B3)**.

Biodiversity Heritage Library
www.biodiversitylibrary.org/

A consortium of natural history and botanical libraries that cooperate to digitize the public domain books and journals held within their collections. BHL has also obtained permission from rights holders to make available content that is under copyright. As of August 2013, the BHL portal provided access to more than 41 million pages from over 60,000 separate titles.

Atlas of Living Australia
www.ala.org.au

The ALA aims to create a national database of all of Australia's flora and fauna, accessed through a single, easy-to-use website. It engages the public through a number of innovations such as the Volunteer Portal, encouraging users to help digitize information from specimen labels, field notes and survey sheets from various Australian museums. ALA also produces software tools to help capture field data.

iNaturalist
www.inaturalist.org/

An online community of naturalists and citizen scientists built on sharing species observations via the iNaturalist website or from a mobile application. Data records may include images and geographical coordinates, and can be annotated by the community. Data with confirmed identifications are published through the Global Biodiversity Information Facility (GBIF).

Moorea Biocode project
http://mooreabiocode.org/

A DNA barcoding survey to build a genetic library of all non-microbial life on Moorea, an island in French Polynesia. Specimens are systematically collected in the field and genetic sequences recorded to build this open access library. The Moorea Project is part of a network of Genomic Observatories aimed at taking the 'biological pulse' of the planet.

Movebank
https://www.movebank.org/

An online database which allows the sharing, managing and archiving of animal tracking data. The information helps us understand how individuals and populations move within local areas, migrate across oceans and continents and evolve through millennia. Researchers who contribute data retain full ownership and control over the level of access to their data. Movebank also provides tools for making basic edits to the data.

The projects and initiatives highlighted here are for illustration only – many more contribute to the objectives of GBIO. Each will often cut across several focus areas, as indicated by the icons (see page 11 for the icon key).

B1. Published materials

Using data mining and semantic tools to turn unstructured and inaccessible data into information.

Progress: some progress (issues understood, needs operational implementation)

Published materials – primarily printed literature, but also images, videos and other multimedia forms such as sound recordings – have long served as the primary means for disseminating biodiversity knowledge. Along with **collections and specimens (B2)** they also form the primary source of species-level trait and descriptive data, vital for identifications and taxonomic research. Much progress has been made by research institutions and by the Biodiversity Heritage Library (BHL),[27] scanning historical materials into digital formats, while new materials are almost exclusively developed as digital objects. Nevertheless the information in these resources remains largely inaccessible to automated processing due to a lack of internal structure and mark-up, and for older literature errors introduced during the scanning process. Multimedia objects need consistent indexing to make them properly discoverable. Consistent standards, including text recognition standards, keywording and indexing, data mining techniques and crowd sourcing will enable the community to step up the rate at which such data are made fully accessible. The resulting information, and the tools to generate it, also need to be made freely available to all. Despite progress in recent years, the scale of the task and its importance to the framework means this component requires continued and long-term investment and it will be urgent to find ways to accelerate and streamline the process.

There are already a number of initiatives working on the relevant tools and techniques with research projects investigating automated image extraction and crowd-sourced tagging (BHL),[28] data mining (EOL)[29], semi-automatic (GoldenGATE[30]) and automatic markup of taxonomic descriptions (MARTT [MARkupper forTaxonomic Treatments],[31] TaxonFinder,[32] TaxonGrab[33], FAT [Find All Taxa][34]), and handling multiple languages (SciELO [Scientific Electronic Library Online][35]) as well as using structured data in taxonomic publications. The next steps will be to catalogue and define the types of unstructured biodiversity data available (TaxPub, TaxonX, taXMLit)[36] and understand their particular challenges, and to agree standards for future publication in a form that makes the data immediately available not just to experts but to searches and automated processing.

In the short term, the priority will be to build on some of the existing pilots and implement some full-scale crowd sourcing and automated data mining projects. In the medium term, the software behind such projects should be made available as open source tools. Countries will start to establish their own bibliographies of national biodiversity. New publications will increasingly come in an enhanced, semantically structured form.[37] In the long term, such enhanced publication will be the norm and automated and semi-automated data mining tools will be freely available for unstructured biodiversity data. As a result, complete bodies of thematic or geographical information will be progressively made available as linked datasets.

B2. Collections and specimens

Accelerating the rate at which historic specimen-based data are made discoverable and accessible.

Progress: significant (significant progress made, further investment needed to complete)

The past 250 years of biodiversity research have resulted in a treasure trove of preserved specimens held in the world's natural history collections, and they are still being added to today. These collections form the irreplaceable foundation of our knowledge of biodiversity, as well as a source of DNA samples for future analysis. Digitizing the data embedded within these specimens

dramatically improves our understanding of species distributions, morphology and population variation, including changes over time. As with **published materials (B1)**, the scale of the task and its importance to the framework means this component requires continued and long-term investment. Widespread development and adoption of the most efficient techniques could dramatically accelerate current digitization efforts, making continued improvement of methodologies an urgent task.

In recent years, museums and herbaria have increasingly begun to capture the data contained in these specimens and their labels, and in accompanying field notes, making them available through aggregators such as GBIF, its network of national nodes and data publishers, and through thematic networks such as the Ocean Biogeographic Information System (OBIS).[38] But the work is labour intensive and their efforts are dwarfed by the scale of the task. Some have begun to accelerate digitization efforts through exploring automation, adopting highly-efficient workflows, and the use of volunteers. Others have pioneered crowd sourcing through making specimen images available online, for example the Biodiversity Volunteer portal of the Atlas of Living Australia (ALA),[39] and the 'Herbonautes' initiative of the Muséum National d'Histoire Naturelle in Paris.[40] The next step will be to document and share current best practices to help institutions choose the optimum approach to accelerate digitization. Institutions will need to prioritize digitization, while funding bodies and governments will have to make the resources available to develop the skills and deploy enough people to accelerate the task. Data quality improvements are also fundamental as digitization rates accelerate, whether through automated tools, or feedback mechanisms via the **biodiversity knowledge network (A5)** and **fitness-for-use and annotation (C1)**.

In the short term, natural history collections should continue to develop and document accelerated digitization techniques which organizations like GBIF and its nodes can use to develop training materials and programmes for smaller institutions. In the medium term, GBIF and digitization projects should develop global infrastructure to support accelerated workflows, including the generation of identifiers and crowd-sourcing clearing houses. In the long term, fully automated mass digitization should eliminate most bottlenecks, clearing the way to complete the digitization effort.

B3. Field surveys and observations

Making field data immediately accessible and interoperable from the moment it is collected and engaging the public in its collection.

Progress: significant (significant progress made, further investment needed to complete)

Much of the biodiversity data collected in the field stays trapped for years in notebooks or stand-alone databases before being transcribed into more accessible locations, if at all. Many citizen science projects engage the wider community in data gathering projects that could add significantly to our knowledge if they were integrated with other sources of biodiversity data. Increasing numbers of projects are using georeferenced images and sound recordings to support and enhance field observations. Improvements in mobile and handheld technology now make it possible to enter data directly from the field into institutional databases and to make them immediately accessible.

Some small-scale projects have piloted mobile field capture systems, such as the Moorea Field Information Management System[41] and the FieldData software developed for Atlas of Living Australia,[42] while hundreds of citizen science projects have developed tools for distributed data capture and online identification. The next step will be to review existing tools and approaches and document the best, providing guidance for projects and users on methodologies suitable for citizen science groups, consultants or professional researchers. The United Kingdom's Natural Environment Research Council (NERC) Centre for Ecology and Hydrology (CEH)[43] has made a valuable contribution to this effort through a detailed review and analysis of existing citizen science projects and a guide to future activities.[44]

In the short term, the priorities will be to build on the existing pilot data capture tools, ensuring that the tools integrate with resolution services using globally unique identifiers (GUIDs)[45] and that they use common standards for data exchange, integration, identity management and contribution tracking. In the medium term, fully integrated open source mobile applications should be widely available to capture field data and support species identification, while networked communities will develop around common interests (taxonomic, regional or thematic). In the long term, some funding and permits might be dependent on projects using real-time mobile data capture, and it will be possible to target gaps dynamically based on data gathered in this way.

B4. Sequences and genomes

Incorporating data arising from genomic and genetic exploration of the living world.

Progress: significant (significant progress made, further investment needed to complete)

Molecular research – including the analysis of genomic information, or biodiversity genomics – contributes to our understanding of the biology and evolution of species, and increasingly offers a way of detecting and monitoring organisms in the environment. Eco-genomic sampling and sequencing can provide data on the composition of biological communities; for some communities, especially in the microbiome, it may be the only feasible way to determine the species they contain. Mapping genes could enable the discovery of individual species traits, while measuring genetic variation across a population or populations provides information about their overall health, their habitat and the extent to which they are isolated or interconnected.

Databases of reference sequences such as DNA barcodes can map genetic data into the wider world of biodiversity information by tying them to species and vouchered individuals. Collaboration through initiatives such as the Genomic Standards Consortium (GSC)[46] has started the process of mapping genomic-level data standards to those recording biodiversity observations of named species.[47] Place-based genomic studies offer an unrivalled opportunity to combine best practice, standardized data capture, uniform methods of analysis and rigorous and systematic characterization of DNA, the foundational layer of biodiversity, to advance our knowledge of biodiversity on earth.

Of note in this respect is progress towards a network of Genomic Observatories (GOs),[48] a combined initiative of the GSC and the Group on Earth Observations Biodiversity Observation Network (GEO BON).[49] This is an international effort to bring together premier sites engaged in long-term research programmes to help contextualize ongoing and new genomic observations at the DNA level. Leading sites in this consortium include the island of Moorea (see also B3 and project profile); and the 'L4' site in the Western English Channel, which has been studied for more than a century and is now one of the best studied sites in the world in terms of metagenomic information on microbial communities.

In the short term, the priority will be to develop the Genomic Observatories network of sites into a range of case studies on the integration of biodiversity genomic work in the content of long-term ecological, evolutionary and environmental studies, with the production of training materials and best-practice guidelines for gathering genomic-level data. In the medium term, research sites will work to generate well-contextualized genomic observations in line with global data standards. In the longer-term, genomic observations will enable more systems-based approaches to the study of biodiversity and ecosystem services, providing data on interactions within an ecosystem and contributing ecosystem-wide biodiversity models.

*B5. Automated and remote-sensed observations

Harnessing automated monitoring technology to provide planet-wide surveys, filling in the gaps left by traditional field research.

Progress: limited (needs further development)

Field research is labour intensive and it will never be possible for the whole planet to be sampled and monitored in detail, leaving large gaps in our models. Remote sensing technologies allow the repeated capture of information from any location and/or over large areas. While camera traps, acoustic sensors and buoys can provide regular observations from the most remote areas, satellites and drones can cover very large areas in a short space of time, generating huge numbers of high-resolution images. Tracking devices further allow us to follow the movements of species, large and small, in real time. Remote-sensed data can be accompanied by abiotic observations (such as temperature or humidity), adding further depth to our understanding of ecosystem interactions. This component is in its infancy in the area of biodiversity research, but its potential to provide otherwise unobtainable information makes it a priority area for investment.

The Icarus Initiative (International Cooperation in Animal Research using Space)[50] is working to establish a remote sensing platform for tracking small organisms while MoveBank[51] provides an online platform for sharing tracking data. Continuing improvements in the technology, especially in earth observation imagery, and the reduced costs of accessing the data, will dramatically increase the volume of data available. Currently and for some time to come, human expertise will still be needed in some forms of remote sensing such as acoustic monitoring and camera traps to extract and process the data if they are to be correctly interpreted, for example in identifying species, although there is scope for using crowdsourcing for these tasks.

In the short term, the priorities will be to put more standard format data into open repositories from existing land and sea-based automated sensors and remote sensing, and to develop the algorithms needed to extract the information automatically. In the medium term, the priorities will be to expand remote sensing networks to create a complete global picture of biodiversity at multiple resolutions. In the long term, automated change detection and data integration will help signal unexpected change and trigger more intensive conventional field research, helping to prioritize new data capture (D5).

* Considered to be of high urgency, but have made limited progress to date.

Focus area C: Evidence

Providing the tools to support consistent and comprehensive global discovery and use of data from all sources about the biodiversity of any defined area over time, covering all taxonomic groups.

Ultimately, our understanding of the natural world is based on the data sources addressed in **focus area B**. For most purposes, however, what users need are organized views and discovery tools that enable them to access efficiently the relevant information at the right level of detail, without being hampered by the need to find obscure data sources, allow for differences in taxonomic views or correct for data quality.

This focus area addresses these requirements by providing services to index content from all relevant resources, the infrastructure to organize species-level data (traits, interactions and occurrences) and providing the means for data quality to be improved, whether manually or automatically.

Components:

- **C1. Fitness-for-use and annotation:** an efficient mechanism to enable amateurs, experts and automated tools to correct and annotate individual data elements to improve quality and their fitness to be used for particular purposes, and to ensure that these annotations have to be made only once.
- **C2. Taxonomic framework:** a comprehensive, expert-curated catalogue of species, including data on names, classification and phylogeny (evolutionary relatedness) and incorporating taxa lacking formal names.
- **C3. Integrated occurrence data:** bringing together data from all sources to document the known occurrences of all species in time and space.
- **C4. Aggregated species trait data:** providing the tools to bring together all available data on species traits and interactions and ensure it is held in forms suitable for use in digital analysis and modelling.
- **C5. Comprehensive knowledge access:** making all published biodiversity knowledge linked and accessible through the rich indexing of biodiversity literature, data, multimedia and other resources, including presentation of the information as species pages and via web services.

Progress: Some of this work is already on course and significant progress has been made, although further investment will be needed to complete the process. International collaboration has delivered key components for building the **taxonomic framework (C2)** and work on **integrated occurrence data (C3)** is already mature and operating at a global scale. Mechanisms for recording **fitness-for-use and annotations (C1)** are starting to be developed in pilot while many of the individual pieces needed for **comprehensive knowledge access (C5)** have been trialled in different projects, but not brought together in one coherent framework. Standards for species trait data and interactions are starting to be adopted but there is as yet no common infrastructure in place to capture and query such data in a consistent way.

Priorities and dependencies: Completing the **taxonomic framework (C2)** will underpin almost all of the other work in this focus area as names provide the key to most biodiversity data. Enabling a community-wide effort to improve data quality and fitness for use will need widely accepted **data standards (A2)** and depend heavily on the **biodiversity knowledge network (A5)** and on the **policy incentives (A4)** and **open access and reuse culture (A1)** being in place to encourage it. The more data can be mobilized into digital and structured forms (**focus area B**), the richer and more accurate the information in this layer will be.

Catalogue of Life
www.catalogueoflife.org/

A global index of species with information on their names, relationships and distributions. The Catalogue compiles data from 115 peer-reviewed taxonomic databases maintained by specialist institutions, and helps provide a taxonomic backbone for other data portals on biodiversity. The list for 2013 included more than 1.4 million species.

Global Biodiversity Information Facility
www.gbif.org

An intergovernmental scientific infrastructure aimed at providing free and open access to biodiversity data, via the Internet. GBIF offers a single online access point to over 400 million biodiversity records from over 10,000 datasets published by nearly 500 institutions, ranging from museum specimens collected from the earliest days of natural history exploration, to current observations by 'citizen scientists' and monitoring from research expeditions.

SpeciesLink
http://splink.cria.org.br/

An information system developed by the Centro de Referência em Informação Ambiental (CRIA) in Brazil, to provide access to biodiversity data records. The system offers a number of data quality tools including an annotations system enabling users to register comments, for example on the identification of species, that are reported to the data curator and remain available for future users consulting those records.

Morphobank
www.morphobank.org/

An online database which allows researchers to upload morphological images and data about organisms, and use these for the study of evolutionary relationships. The Morphobank web application provides a virtual platform for scientists to collaborate and build phylogenetic matrices with image data.

Ocean Biogeographic Information System
http://www.iobis.org/

An online open access database of marine species distributions, now an activity under the International Oceanographic Data and Information Exchange programme of UNESCO's Intergovernmental Oceanographic Commission (IOD). Integrating data from institutes around the world, OBIS allows users to identify biodiversity hotspots and large-scale ecological patterns, analyse species dispersions over time and space, and plot species' locations with temperature, salinity and depth.

The projects and initiatives highlighted here are for illustration only – many more contribute to the objectives of GBIO. Each will often cut across several focus areas, as indicated by the icons (see page 11 for the icon key).

C1. Fitness-for-use and annotation

Creating a network of expertise to manage and curate biodiversity data and permanently capture data cleanup efforts.

Progress: some progress (issues understood, needs operational implementation)

Over the years, millions of biodiversity data records have been created from a variety of sources, offering varying degrees of accuracy and quality, with even the best-quality records affected over time by changes in taxonomy, vocabulary and geospatial precision. While the **biodiversity knowledge network (A5)** component will encourage people to contribute their expertise in improving these records, currently there are few systems in place to make the resulting annotations immediately available or to link them securely to the original record. This can result in the creation of a new and contradictory version of the original data, or – more often – the changes are completely lost.

A number of projects have begun to look at how annotations can be incorporated into or combined with the original records. AnnoSys and the Filtered Push project have been developing networked annotation system for biodiversity data,[52] while the SpeciesLink network from Centro de Referência em Informação Ambiental (CRIA) allows users to exclude records from searches if they have been flagged with geographical inconsistencies.[53] The next step will be to agree common standards and mechanisms for incorporating annotations for data curators to adopt.

In the short term, the priority will be to pilot easy-to-use tools and systems that can capture annotations and tie them to back the original data record. In the medium term, integrated or online tools for data cleanup will be widespread and available to both professionals and amateurs while data records will increasingly have a curation status indicating what checks and corrections they have received. In the long term, it should be possible to filter most data by curation status and robust systems will be in place to provide real-time delivery of community annotations, including mechanisms to resolve any conflicts or contention.

C2. Taxonomic framework

Providing a stable and comprehensive catalogue of all species.

Progress: significant progress (significant progress made, further investment needed to complete)

The classification of species has been developed over the last few centuries, and will continue to change in the future as our understanding of evolutionary history develops, and new species are discovered and described. This means that the actual names applied to specimens, observations or populations are subject to change over time – or according to the person doing the naming – while specimens or observations may be of currently undescribed species without a formal scientific name at all. Yet names – vernacular or scientific – are one of the primary means for retrieving and grouping information. It is crucial to be able to draw correlations between names used (currently and in the past) and the taxa they relate to, according to the major classification schemes and phylogenies in use, and to map between different classifications. Although progress has been made in this area, it is one of the key underpinning components to make data fully available, and it should be completed as a matter of urgency.

Scientists have been working to create a comprehensive formal taxonomic classification since work began on Species 2000/ Catalogue of Life.[54] Some projects such as Centro de Referência em Informação Ambiental (CRIA) already identify taxonomic and

geographical gaps in the record to help identify priorities for research.[55] Collaborations are creating architectures that can handle multiple taxonomies as well as informal and vernacular names through the Global Names Architecture and most recently the i4Life project, contributing to resources such as the GBIF backbone or nub taxonomy.[56]

In the short term, a clear road map is required to consolidate all these existing activities and deliver a suite of reliable, robust and open tools for accessing basic information on species names and classifications. In the medium term, global taxonomic expertise must be organized to fill remaining gaps in the underlying datasets and to address linkages with key species lists such as the IUCN Red List and CITES (the Convention on International Trade in Endangered Species),[57] and with approved national species lists. In the long term, all new species names and descriptions should automatically be integrated into this framework.

C3. Integrated occurrence data

Making accessible all data about when and where any given named organism has been recorded.

Progress: significant progress (significant progress made, further investment needed to complete)

Documenting the distribution of species in time and space may be viewed as the 'weather observations' of biodiversity: the fundamental underpinning for any accurate model of existing patterns and trends. All of the data sources outlined in **focus area B** will contribute species occurrence records in some form or another, but the key elements of every record need to be brought together in a more readily accessible and usable form to enable efficient discovery and use.

Many national, regional and thematic efforts such as the Ocean Biogeographic Information System (OBIS)[58] and VertNet[59] are already mobilizing significant quantities of data on specimens and biodiversity observations. Mature data standards such as Darwin Core and the Access to Biological Collection Data (ABCD) schema[60] have enabled GBIF to index and organize data from thousands of such sources. The GBIF data portal already offers access to more than 400 million species occurrence records.[61]

In the short term, national, regional and thematic networks that already handle occurrence data should work to link their data to global networks. Global activities such as GBIF should enhance their processes to improve understanding and fitness-for-use of all mobilized data, to provide support for data on species abundance and sampling events, and to ensure that all contributors receive appropriate acknowledgement and feedback for their work. In the medium term, additional sources of data on species occurrence, including **published materials (B1), sequences and genomes (B4)** and **automated and remote-sensed observations (B5)**, should also be integrated. In the long term, globally-connected networks should continuously and automatically process all new observations and samples of biodiversity.

C4. Aggregated species trait data

Providing the framework to capture all available trait information for any species, and interactions between species.

Progress: limited (limited existing standardization; needs further development)

Scientists need access to analysable data about the characteristics of organisms and the interactions between them. Being able to capture, index, query and curate such data will revolutionize how we understand large-scale biological systems. This will enable us to build more accurate models of how they will behave over time and under changing conditions, and to create a new generation of identification tools.

The community has already developed some major data repositories. For traits these include: TraitNet, TRY, LEDA, MorphoBank, Animal Diversity Web, the Phenoscape KnowledgeBase, species traits in the Encyclopedia of Life and a growing number of model organism and agricultural resource databases.[62] Interactions are recorded in the Interaction Web Database (IWDB), Semantic Web Informatics for Species in Space and Time (SWISST) and the Encyclopedia of Life.[63] Species trait data standards include Structured Descriptive Data (SDD), the Plinian Core, the Phenotypic Quality Ontology, and the Animal Natural History ontology, which also includes species interactions.[64] The next tasks are to standardize across these vocabularies; ensure that the data recorded reflect the complex relationships among traits, their genetic base and the environment; and report traits and interactions at multiple scales.

In the short term, the priority will be to produce summaries of which organisms interact with each other in defined ways such as predator/prey or parasite/host relationships or in more general terms (pollinators, invasive species) and to continue to develop and improve identification tools. In the medium term, the priorities will be systems to serve rich trait data for the best-understood organisms, and identifying techniques for large-scale trait data capture building on **focus area B**. In the long term, the priorities will be to capture high-quality trait and interaction information at increasingly large taxonomic and geographic scales using high-throughput descriptions equivalent to NextGen sequencing.

C5. Comprehensive knowledge access

Delivering access to all published biodiversity knowledge.

Progress: some progress (issues understood, needs operational implementation)

The other elements in this focus area have related to the provision of central indexes of particular types of structured information which can feed directly into modelling and analysis tools. However the biodiversity data that have been gathered and published are much richer than these core subsets – consisting of anything from video recording to identification tools to conservation assessments. Locating all information about a species – or the biodiversity of a defined area– requires seamless access to the underlying information resources, structured and unstructured. Users may also want to retrieve information in other ways, such as by usage or threat level, or in other languages.

Many national and thematic activities – and other web communities such as Wikipedia and Wikispecies[65] – organize and deliver web content for particular species, either as online databases or as species pages. The Encyclopedia of Life (EOL), now represents

an international partnership of institutions and agencies committed to providing open access to authoritative species information, including multimedia,[66] while EUBrazilOpenBio aims to combine open access resources including data, tools and services in a single 'e-infrastructure'.[67] Other organizations manage key expert-curated data sets such as the IUCN Red List and the United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC) World Database on Protected Areas.[68] Users need efficient mechanisms to discover, access and integrate all of these resources.

In the short term, **relevant data standards (A2)** and links into the **taxonomic framework (C2)** need to be widely adopted to support the sharing of different types of species information along with mechanisms enabling users to locate materials in a language that they can understand. In the medium term, comprehensive catalogues should be developed to simplify access and reuse of authoritative species descriptions, images, identification keys, etc. In the long-term, all species information should come to be managed and curated as an inter-connected digital knowledge base.

Focus area D: Understanding

Using the combined biodiversity data from multiple sources to generate new information, inform policy and decision makers, and help educate wider society to improve the way we manage the Earth's resources.

Society needs systems that deliver the best possible estimate of the abundance and distribution of all species in all areas and at all scales, now and at any time in the recent past, and projections into the future depending on human actions and decisions. We need the best possible understanding of how different species interact and how communities and ecosystems function. For all our efforts in the other focus areas, our data resources can only provide a partial and fragmented picture of the world's biodiversity today, let alone in the past. As importantly, all the data in the world will not help achieve the Aichi Biodiversity Targets unless they are presented to policy makers and the public they serve in a compelling and easily understood way.

This focus area addresses these challenges by integrating biodiversity data sources with other disciplines – from geology to economics – to bring to bear all of our expert knowledge to create the best possible models of existing biodiversity patterns and how they might respond to human activities. It addresses how to make these immediately and vividly understandable through visualizations and interactive tools, and how they can be used to support policy makers and monitor the effectiveness of the policies already in place, as well as alert us to potentially dangerous changes before it is too late.

It is envisaged that by integrating information from multiple sources – from local to global, and from different fields of knowledge – by acquiring new expertise, and by developing new tools, new questions will be asked which will require new types of analysis. Progress, therefore, depends on continuous and dynamic planning and evaluation, interaction, cooperation, suitable governance and adequate long-term funding.

Components:

- **D1. Multiscale spatial modelling:** Integrate data collected across disciplines and combine them with remote-sensed and geographic information systems (GIS) datasets to create the fullest-possible picture of geographical species distributions.
- **D2. Trends and predictions:** Integrate historical data and changes over time to create predictive modelling tools to support decision making, make biodiversity estimates and predict the potential impact of changing conditions anywhere on Earth.
- **D3. Modelling biological systems:** Build virtual models – from the molecular level to whole ecosystems – to improve understanding of biological systems and integrate that knowledge into other models.
- **D4. Visualization and dissemination:** Provide the tools to make biodiversity information accessible and understandable by diverse audiences, increasing biodiversity literacy among the public and policy makers.
- **D5. Prioritizing new data capture:** Use accumulating data to identify and prioritize new opportunities for data capture and to provide a timely response to changes in biodiversity patterns.

Progress: This is the least developed of the four focus areas, and one, **modelling biological systems (D3)** has not progressed much beyond the conceptual stage. The biodiversity informatics community will have to draw heavily on the knowledge of other disciplines — from climate modelling and socio-economics, to games development and web design.

Priorities and dependencies: Completing this focus area depends on having as much high-quality data mobilized as possible **(focus area B)** — both current observations of existing biodiversity, and historical records from the literature and natural history collections. Synthesizing new knowledge requires mobilizing data in sufficient volume from across multiple disciplines, making the right **data standards (A2)** crucial. Displaying and using data in novel ways, and even in real time, relies on an **open access and reuse culture (A1)** as well as robust **persistent storage and archives (A3)** to serve the underlying data stores. Vertical integration of **species traits** and **occurrences (C4, C3)** around a strong **taxonomic**

framework (C2), including vernacular names, will be essential when developing tools to make the data relevant to policy makers and the general public. This focus area is also the most closely interdependent, with the three modelling components (**D1, D2 & D3**) needing to work in parallel to provide a framework of coupled models based on interlinked data and tools. Finally, both **visualization and dissemination (D4)** and **prioritizing new data capture (D5)** will depend on having increasingly robust models to draw on.

The projects and initiatives highlighted here are for illustration only – many more contribute to the objectives of GBIO. Each will often cut across several focus areas, as indicated by the icons (see page 11 for the icon key).

D1. Multiscale spatial modelling

Estimating biodiversity patterns from available evidence.

Progress: some progress (some progress, issues understood, needs operational implementation)

Many issues in biodiversity research, conservation and management depend on understanding the spatial occurrence, distribution and abundance of species and the communities of which they form a part. These questions operate at all scales from the planetary down to small research plots. Underlying them is a general requirement – estimating the range of species that are to be found in any area, their relative abundance and their functions and inter-relationships. **Focus area B** mobilizes the data needed to meet this requirement while the **integrated occurrence data (C3)** component organizes these data into evidence for what has actually been recorded of biodiversity patterns. However, these data will never be comprehensive enough to document which species occur in every part of the planet. Some of the data may only provide very coarse indication of species occurrence, for example just the fact that a particular species has been recorded within a particular country. We need to build models that make appropriate use of all available evidence to provide the best possible estimate of the actual set of species in each area and of the relative abundance of each. Such models can simultaneously help to compensate for errors in primary data and to address gaps in our knowledge, for example providing estimates of species and species numbers where records are incomplete.

Progress has already been made in building models for individual species using approaches such as Maxent[69] and OpenModeller,[70] an online environment providing access to a range of published modelling algorithms, to assess and map the suitability of different habitats and environments for the species based on similarity to known localities for the species. The Map of Life[71] aims to integrate models with a wide variety of spatial biodiversity data sources to build a knowledge base and platform for species distribution map development. Other approaches such as generalized dissimilarity modelling (GDM)[72] support exploration of patterns involving whole communities of species. The next step will be to refine these foundations over time to use all available sources of evidence, biotic or abiotic (e.g. environmental variables like temperature and precipitation), to build improved spatial models to support research and decision making. These models themselves should be made available in a standard format for further research and shared using the infrastructure and culture addressed by **focus area A**.

In the short term, the priority will be to develop best practices for combining and using data with varying degrees of precision, from high-precision coordinates for samples or observations through to national species lists, and to organize consistent access to the abiotic datasets required for spatial modelling. In the medium term, repositories are required to support efficient archival and reuse of models and modelled data. In the long term, insights and models derived from the **modelling biological systems (D3)** component must be accommodated to maximize the biological validity of spatial models.

*D2. Trends and predictions

Using predictive modelling to assess status, trends and the impacts of any potential changes.

Progress: limited (limited existing efforts, needs further development)

Being able to model future trends in biodiversity under different conditions (such as the outcomes of particular policies or climate change) are key to sound decision making. The wealth of historical data that **focus area B** will unlock from the literature and museum collections, and the existing long-term ecological monitoring sites, provide a temporal dimension to the available

evidence for our spatial models. Mobilizing all of these data will enable not only **multiscale spatial modelling (D1)** of current biodiversity patterns, but also to construct a series of historical models, at least for some taxonomic groups and for some regions and scales of detail. These models will in turn help us to identify trends, to explore the key drivers behind these trends and to predict responses to potential future changes. Having good predictive tools are a key element for the better management of biodiversity, so developing this component should be a matter of high priority.

A number of projects, such as the Wallace Initiative, ClimaScope, eHabitat and the Ermitage project, already use biodiversity data to explore the impacts of environmental changes on biodiversity,[73] while the Digital Observatory for Protected Areas (DOPA) is looking at building interoperable modelling services for biodiversity.[74] The next step will be to create a group equivalent to the Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA) of the Intergovernmental Panel on Climate Change (IPCC)[75] to establish the baseline data and methodologies, workflows and repositories needed.

In the short term, mechanisms are required to assess the suitability of integrated occurrence data for different species to support development of a series of historical distribution models, and to establish guidelines for the work in **focus area A** and **focus area B** to maximize the suitability of the data for temporal modelling. In the medium term, work should focus on visualization of such models and on algorithms to explore the drivers behind apparent changes in biodiversity. In the long term, these insights should support development of tools to model predicted changes in biodiversity in response to human-induced pressures such as climate change, land use change and the expected impacts of invasive species on ecosystems.

*D3. Modelling biological systems

Building virtual models from molecules to ecosystems.

Progress: minimal (needs further investigation)

Biodiversity is a complex and massively interconnected system with interactions that range from the molecular to the entire biosphere. Current solutions in the **multiscale spatial modelling (D1)** and **trends and predictions (D2)** components primarily depend on knowledge of recorded species occurrence and environmental spatial data to build models that estimate actual distributions. Clearly such models do not reflect the complex realities of diverse ecosystems. Actual species distributions are driven by specific aspects of the physiology and behaviour of each species, where an organism is in the food chain, and how it competes and interacts with other species. Incorporating information from **aggregated species trait data (C4)** will make many refinements and enhancements to spatial and temporal models possible. Understanding which taxa exist in marine, freshwater and terrestrial environments is fundamental to predicting their distributions. Knowledge of reproductive strategies, metapopulation characteristics, dispersal and migration strategies, food requirements and many other features may support much more fine-grained modelling. Some aspects of species and community biology may themselves be handled through models that could be integrated into sophisticated models that better simulate the behaviour of complex systems in space and time. Predictive models based on phylogeny may also allow inference of likely traits even for species for which these data have not been recorded. As this is a key component for good decision making, and the least-developed area of the whole framework, it should be pursued as a matter of urgency.

This is a complex area. Some use is already made of simple trait data to validate occurrence data, for example the indicators in the Interim Register of Marine and Non-marine Genera (IRMNG) of which genera are found in marine and terrestrial environments.[76] A wealth of models has been developed for model organisms and for ecological communities. The most relevant and useful of these models could be integrated using a coordinated coupled framework approach and associated usage guidelines.[77]

* Considered to be of high urgency, but have made limited progress to date.

In the short term, key species traits and interactions can be used to validate **integrated occurrence data (C3)** and to assess and refine **multiscale spatial modelling (D1)** results. In the medium term, standards will be required for sharing and reusing trait-based factors and biological models within spatial and temporal modelling systems. In the long term, models for different species and communities should be used to refine one another.

*D4. Visualization and dissemination

Giving people – from the public to scientists to policy makers – access to information in ways that will revolutionize how we understand biodiversity.

Progress: limited (limited existing efforts, needs further development)

The way we present information is often as important as the content itself. Recent advances in mobile technology have revolutionized the way we consume data, while information is being served in ever more inventive ways. The rapid development of processors and software is now enabling interactive visualization of data in ways that would have been impossible only a few years ago. We need to engage game developers, web designers, communications experts and policy makers in rethinking how biodiversity information is organized to support decision making and to communicate with the public. Building on the resources we have, and the vertical integration being developed in **focus area C**, the goal should be to give everyone everywhere the means to explore biodiversity information in context, increasing our understanding and thus increasing the value placed on biodiversity itself. As better understanding of biodiversity is key to generating the political will to protect it, this is another urgent task.

Many online projects are already using interactive maps, charts, taxonomic and phylogenetic trees, and other tools to enhance understanding and interpretation of biodiversity data. The next steps would be for projects to take a more multi-disciplinary approach and expand to encompass clear and informative presentation of biodiversity information in combination with other relevant knowledge, including environmental, climatic, and sociological data. Good interfaces should assist all users to understand the relevance and significance of available biodiversity information in relation to their interests and needs.

In the short term, more work is needed to bring experts in information science and interface design in conversation with biodiversity experts and users of biodiversity information. In the medium term, this work will drive refinements to the components in focus area C to deliver information in the most appropriate and effective forms. There will also be a need for developing new interoperability standards to maximize the usefulness of visualization tools and applications. In the long term, such tools and applications will lead to a new generation of code libraries and intuitive interactive platforms to support research, policy needs and public understanding.

* Considered to be of high urgency, but have made limited progress to date.

D5. Prioritizing new data capture

Making best use of limited resources by concentrating on the areas of greatest change and greatest uncertainty.

Progress: limited (limited existing efforts, needs further development)

If we are to respond effectively to changes in ecosystems we need to be alerted to them as soon as possible, and be able to reallocate monitoring resources to determine the causes, and to take appropriate action. Policy interventions also need to be monitored to ensure that they are having the desired effect and so modified where necessary. We can use our limited resources best by coordinating with other agencies and using all sources of information from remote-sensing satellites to the observations from local people on the ground.

A number of projects are involved in monitoring biodiversity and environmental changes, including GEOBON, Vital Signs Africa, the International Centre for Integrated Mountain Development (ICIMOD), Long Term Ecological Research Network (LTER Network), the Global Observation Research Initiative in Alpine Environments (GLORIA) and Eye on Global Network of Networks.[78] The next step will be to enable monitoring efforts to track and report changes promptly, feeding the information in to governmental and non-governmental action networks for rapid response. Existing monitoring agencies need to build in flexibility to share capacity with other organizations. Mechanisms developed under the new Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES) should help to identify priorities for filling gaps and to support efforts to fill them.

In the short term, the priority should be on understanding the coverage, the gaps and completeness of existing data and determining the suitability of these data for addressing questions at different scales for different species and taxonomic groups. In the medium term, gaps in this coverage can help to prioritize efforts to mobilize additional data, while understanding of which areas are already covered by good historical data can help to identify the most effective locations for establishing monitoring activities. In the longer term, research and citizen science activities can be focussed to maximize the expansion of our understanding of biodiversity and its patterns and processes. Regional, national and global intervention networks will be able to identify changes and threats rapidly, and support decision making regarding the best response.

Annex – The Global Biodiversity Informatics Conference

Organizing committee

Donald Hobern	Executive Secretary, Global Biodiversity Information Facility
Leonard Krishtalka	Chair, GBIF Science Committee
Wouter Los	CReATIVE-B (LifeWatch)
Norman MacLeod	Natural History Museum, London
Erick Mata	Encyclopedia of Life (EOL)
David Schindel	Consortium for the Barcode of Life (CBOL)

Workshop leads and the GBIO drafting team

Enrique Alonso García	Consejo de Estado, Spain
Alberto Apostolico	College of Computing, Georgia Institute of Technology
Elizabeth Arnaud	Bioversity International
Juan Carlos Bello	Alexander von Humboldt Institute for Research on Biological Resources, Colombia
Dora Canhos	Centro de Referência em Informação Ambiental (CRIA), Brazil
Gregoire Dubois	European Commission - Joint Research Centre
Dawn Field	Molecular Evolution and Bioinformatics Group, NERC Centre for Ecology and Hydrology, United Kingdom
Alex Hardisty	CReATIVE-B (Cardiff University)
Jerry Harrison	UNEP-WCMC
Bryan Heidorn	JRS Foundation & University of Arizona
Roderic Page	Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, United Kingdom
Cynthia Parr	Encyclopedia of Life (EOL)
Jeff Price	The Wallace Initiative, ClimaScope and University of East Anglia
Selwyn Willoughby	South African National Biodiversity Institute

Facilitators

Natasha Walker
Bart Slob
Niels Ferdinand

GBIF Secretariat support staff

Sampreethi Aipanjiguly
Olaf Bánki
Vishwas Chavan
Samy Gaiji
Alberto González-Talaván
Andrea Hahn
Tim Hirsch
Anne Mette Nielsen
Éamonn Ó Tuama
David Remsen
Tim Robertson
Susanne Lønstrup Sheldon
Ciprian Vizitiu
Hugo von Linstow
Katja Wolfhechel Christensen

Participants

Agosti, Donat	Plazi	Iran
Allen, Paul	Cornell University	United States
Alonso García, Enrique	Consejo de Estado	Spain
Ariño, Arturo H.	University of Navarra	Spain
Baadsvik, Karl	Norwegian Biodiversity Information Centre	Norway
Balslev, Henrik	Aarhus University	Denmark
Belbin, Lee	Atlas of Living Australia	Australia
Berendsohn, Walter	Botanischer Garten und Botanisches Museum Berlin-Dahlem	Germany
Blum, Stan	California Academy of Sciences	United States
Canhos, Vanderlei	Centro de Referência em Informação Ambiental (CRIA)	Brazil
Cao, Mingchang	Nanjing Inst. of Environmental Sciences, Ministry of Environmental Protection	China
Chenin, Eric	GBIF-France/Institut de Recherche pour le Développement	France
Chettri, Nakul	International Centre for Integrated Mountain Development	Nepal
Cochrane, Guy	European Bioinformatics Institute	United Kingdom
Costello, Mark	University of Auckland	New Zealand
Dalcin, Eduardo	Instituto de Pesquisas, Jardim Botânico do Rio de Janeiro	Brazil
Daly, Joanne	Commonwealth Scientific and Industrial Research Organisation (CSIRO)	Australia
Davies, Neil	Moorea Biocode Project	French Polynesia
Dulloo, Ehsan	Food and Agriculture Organization of the United Nations (FAO)	Italy
Faith, Daniel P.	Australian Museum	Australia
Ferrier, Simon	Commonwealth Scientific and Industrial Research Organisation	Australia
Flemons, Paul	The Australian Museum	Australia
Ganglo, Jean Cossi	Université d'Abomey-Calavi	Benin
Gärdenfors, Ulf	Swedish Species Information Centre	Sweden
Gray, William Alex	Cardiff University	United Kingdom
Guala, Gerald Stinger	U.S. Geological Survey	United States
Guralnick , Robert	University of Colorado	United States
Hammond, Thomas	UNEP-GEF Scientific and Technical Advisory Panel	United States
Hanner, Robert	International Barcode of Life (IBOL)	Canada
Hardisty, Alex	Cardiff University	United Kingdom
Häuser, Christoph	Museum für Naturkunde	Germany
Herrera Bachiller, Alfonso	University of Alcalá	Spain
Hilton-Taylor, Craig	IUCN UK Office	United Kingdom
Höft, Robert	CBD Secretariat	Canada
Hugo, Wim	South African Environmental Observation Network (SAEON)	South Africa
Jimenez Rosenberg, Raul	Comisión Nacional para el Conocimiento y Uso de la Biodiversidad	Mexico
Johansson, Anna Maria	European Commission - DG Research	Belgium
Kelly-Quinn, Mary	School of Biology and Environmental Science, University College Dublin	Ireland
Koleff Osorio, Patricia	Comisión Nacional para el Conocimiento y Uso de la Biodiversidad	Mexico
Kovaleva, Olga	N.I. Vavilov Research Institute of Plant Industry (VIR)	Russian Federation
La Salle, John	Atlas of Living Australia	Australia
Llosent de Nárdiz, Maria	University of Alcalá	Spain
Loarie, Scott R.	Carnegie Institution for Science	United States
Lund, Mette	European Environment Agency	Denmark
Ma, Juncai	Institute of Microbiology, Chinese Academy of Sciences	China
MacDevette, Monika	UNEP, Division of Early Warning & Assessment	Kenya

Annex – The Global Biodiversity Informatics Conference – continued

Participants

Mizrachi, Ilene	National Center for Biotechnology Information	United States
Moat, Justin	Royal Botanic Gardens, Kew	United Kingdom
Mora, Maria Auxiliadora	Instituto Nacional de Biodiversidad (INBio)	Costa Rica
Moritz, Tom		United States
Nicholls, Miles	Atlas of Living Australia	Australia
Owens, Ian	Natural History Museum-London	United Kingdom
Pagano, Pasquale	CNR-ISTI	Italy
Pando, Francisco	Real Jardín Botánico de Madrid (CSIC)	Spain
Roberts, Dave	Natural History Museum-London	United Kingdom
Saarenmaa, Hannu	Digitarium - Digitisation Centre of the Finnish Museum of Natural History and University of Eastern Finland	Finland
Saraiva, Antonio	University of São Paulo	Brazil
Schalk, Peter	ETI Bioinformatics	Netherlands
Scholes, Bob	Council for Scientific and Industrial Research	South Africa
Seberg, Ole	Botanic Garden & Museum	Denmark
Segers, Hendrik	Royal Belgian Institute of Natural Sciences	Belgium
Seltmann, Katja	American Museum of Natural History	United States
Simiyu, Stella		Kenya
Soberón Mainero, Jorge	University of Kansas	United States
Suárez-Mayorga, Ángela	Alexander von Humboldt Biological Research Institute	Colombia
Svenning, Jens-Christian	Aarhus University	Denmark
Thuveson, Maria	Swedish Research Council	Sweden
Turner, Woody	NASA	United States
Ulate, William	Biodiversity Heritage Library	United States
Valland, Nils	Norwegian Biodiversity Information Centre, Artsdatabanken	Norway
Vicario, Saverio		Italy
Vieglais, David A.	University of Kansas	United States
Vogel, Johannes	Museum für Naturkunde	Germany
Walters, Michelle	Group on Earth Observations Biodiversity Observation Network (GEO BON)	South Africa
Wang, Yu-Huang	Taiwan Forestry Research Institute	Chinese Taipei
Wieczorek, John	Museum of Vertebrate Zoology	United States
Wilson, Nathan	Encyclopedia of Life	United States
Yang, Xiangyun	Kunming Institute of Botany, Chinese Academy of Sciences	China

Acronyms and abbreviations

ABCD	Access to Biological Collection Data
ALA	Atlas of Living Australia
BDRS	Biological Data Recording System
BHL	Biodiversity Heritage Library
CBD	Convention on Biological Diversity
CBOL	Consortium for the Barcode of Life
CEH	Centre for Ecology & Hydrology
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
CoL	Species 2000 Catalogue of Life
CRIA	Centro de Referência em Informação Ambiental
DOPA	Digital Observatory for Protected Areas
EOL	Encyclopedia of Life
FAT	Find All Taxa
FISH-BOL	Fish Barcode of Life Initiative
GDM	Generalized Dissimilarity Modelling
GEO BON	Group on Earth Observations Biodiversity Observation Network
GBIF	Global Biodiversity Information Facility
GBIO	Global Biodiversity Informatics Outlook
GLORIA	Global Observation Research Initiative in Alpine Environments
GSC	Genomic Standards Consortium
GUID	Globally Unique Identifiers
ICIMOD	International Centre for Integrated Mountain Development
IPBES	Intergovernmental Platform on Biodiversity and Ecosystem Services
IPCC	Intergovernmental Panel on Climate Change
IUCN	International Union for Conservation of Nature
IRMNG	Interim Register of Marine and Non-Marine Genera
IWDB	Interaction Web DataBase
LTER	Long Term Ecological Research Network
MARTT	MARkuper for Taxonomic Treatments
NERC	Natural Environment Research Council
OBIS	Ocean Biogeographic Information System
OSTP	United States Office of Science and Technology Policy
SANBI	South African National Biodiversity Institute
SciELO	Scientific Electronic Library Online
SDD	Structured Descriptive Data
SWISST	Semantic Web Informatics for Species in Space and Time
TDWG	Biodiversity Information Standards
TGICA	Task Group on Data and Scenario Support for Impacts and Climate Analysis
UNEP	United Nations Environment Programme
WCMC	World Conservation Monitoring Centre
WDPA	World Database on Protected Areas

Endnotes

1. For more information about the Global Biodiversity Informatics Conference, see http://www.gbic2012.org/
2. For a full list of participants, workshop leads and organizers of the conference, see Annex on page 36
3. Strategic Plan for Biodiversity 2011-2020, including Aichi Biodiversity Targets – https://www.cbd.int/sp/
4. http://www.ipbes.net/
5. Appendix I (Functions, operating principles and institutional arrangements of the Platform) of UNEP/IPBES.MI/2/9, accessible from http://ipbes.net/plenary/ipbes-1.html
6. Section B of IPBES/1/INF/10, accessible from http://ipbes.net/plenary/ipbes-1.html#three
7. Species 2000 – http://www.sp2000.org/; Catalogue of Life – http://www.catalogueoflife.org/
8. Biodiversity Information Standards (TDWG) – http://www.tdwg.org/
9. Global Biodiversity Information Facility – http://www.gbif.org/
10. Encyclopedia of Life – http://eol.org/
11. Consortium for the Barcode of Life – http://www.barcodeoflife.org/
12. Biodiversity Heritage Library – http://www.biodiversitylibrary.org/
13. Group on Earth Observations Biodiversity Observation Network – http://www.earthobservations.org/geobon.shtml
14. Gibbons *et al*. 1994
15. Schofield *et al*. 2010
16. GenBank – http://www.ncbi.nlm.nih.gov/genbank/
17. Biodiversity Information Standards (TDWG) – http://www.tdwg.org/
18. See for example the GUID and Life Sciences Identifiers Applicability Statement, accessible from http://www.tdwg.org/standards/150
19. United States Office of Science and Technology Policy – http://www.whitehouse.gov/administration/eop/ostp
20. See the Memorandum for the heads of executive departments and agencies, Feb 22, 2013. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
21. SANBI – http://www.sanbi.org/
22. See the National Environmental Management: Biodiversity Act 2004, Republic of South Africa Government Gazette, June 2004. Available at: http://www.sanbi.org/sites/default/files/documents/documents/biodiversityact2004pdf.pdf
23. Belbin, L. *et al*., 2013. A specialist's audit of aggregated occurrence records: An "aggregator's' perspective. *ZooKeys*, 305, pp.67–76. Available at: http://www.pensoft.net/journals/zookeys/article/5438/abstract/a-specialist
24. OpenUp! – http://open-up.eu/
25. See "EOL Curators" on the EOL website – http://eol.org/info/curators
26. FISH-BOL – http://www.fishbol.org/
27. Biodiversity Heritage Library – http://www.biodiversitylibrary.org/
28. See "Interested in improving access to millions of digital images?", Biodiversity Heritage Library website, 30 August 2012. Available at: http://blog.biodiversitylibrary.org/2012/08/interested-in-improving-access-to.html
29. See for example the "EOL Computable Data Challenge", EOL website, http://eol.org/info/323
30. GoldenGATE - see Agosti, D. and W. Egloff (2009), "Taxonomic Information Exchange and Copyright: The Plazi Approach", *BMC Research Notes* 2(53), http://www.biomedcentral.com/1756-0500/2/53, doi:10.1186/1756-0500-2-53
31. MARTT (MARkuper for Taxonomic Treatments). See Cui, Hong (2005) "MARTT: A General Approach to Automatic Markup of Taxonomic Descriptions with XML", http://www.cais-acsi.ca/proceedings/2005/cui_2005.pdf
32. TaxonFinder – http://taxonfinder.sourceforge.net/
33. Taxon Grab – http://taxongrab.sourceforge.net/
34. FAT (Find All Taxa). See Sautter, G., K. Böhm & D. Agosti (2006), "A Combining Approach to Find All Taxon Names (FAT) in Legacy Biosystematics Literature", *Biodiversity Informatics*, 3, 2006, https://journals.ku.edu/index.php/jbi/article/view/34/19
35. SciELO (Scientific Electronic Library Online) – http://www.scielo.org/
36. See Penev, L. *et al*. (2012), "Implementation of TaxPub, an NLM DTD Extension for Domain-Specific Markup in Taxonomy, from the Experience of a Biodiversity Publisher", *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012*, http://www.ncbi.nlm.nih.gov/books/NBK100351/, and, Penev, L. *et al.* (2011), "XML Schemas and Mark-up Practices of Taxonomic Literature", *Zookeys* 150, 89-116, doi: 10.3897/zookeys.150.2213
37. See for example Penev *et al.* (2010), "Taxonomy Shifts Up a Gear: New Publishing Tools to Accelerate Biodiversity Research", Zookeys 50, special issue: 1-4, doi: 10.3897/zookeys.150.2213
38. OBIS – http://www.iobis.org/
39. Atlas of Living Australia – http://www.ala.org.au/. For information about the Biodiversity Volunteer Portal see, http://volunteer.ala.org.au/
40. Herbonautes – http://lesherbonautes.mnhn.fr/
41. See the Moorea Biocode Project website for more information: http://mooreabiocode.org/pages/it-platform
42. FieldData (http://www.ala.org.au/getinvolved/citizen-science/fielddata-software/) is based on the Biological Data Recording System (BDRS) developed for the ALA by Gaia Resources (http:/www.gaiaresources.com.au)

43. Centre for Environment & Hydrology – http://www.ceh.ac.uk/index.html
44. See the *Guide to Citizen Science* available at http://www.ceh.ac.uk/news/news_archive/Citizen-Science-Review-Guide_2012_59.html
45. For more information on the relevance of GUIDs to the community, see the TDWG wiki http://wiki.tdwg.org/GUID
46. Genomic Standards Consortium – http://gensc.org/
47. See for example the outcomes of a GBIF-led workshop in Oxford in Tuama *et al*. (2012) "Meeting Report: Hackathon-workshop on Darwin Core and MIxS standards alignment (February 2012)" in *Standards in Genomic Sciences*, 7(1), http://standardsingenomics.org/index.php/sigen/article/view/sigs.3166513
48. http://genomicobservatories.blogspot.co.uk/
49. GEO BON – http://www.earthobservations.org/geobon.shtml
50. Icarus Initiative – http://icarusinitiative.org/
51. MoveBank – https://www.movebank.org/
52. The Filtered Push Project – http://wiki.filteredpush.org/; AnnoSys – http://wiki.bgbm.org/annosys/index.php/Main_Page; see Tschöpe, O. *et al*. (forthcoming), "Annotating Biodiversity Data", *Taxon*
53. Species Link Data Cleaning – http://splink.cria.org.br/dc
54. Species 2000 – http://www.sp2000.org/; Catalogue of Life – http://www.catalogueoflife.org/
55. See "Lacunas de conhecimento da flora e dos fungos do Brasil" – http://lacunas.inct.florabrasil.net/
56. Global Names Architecture – http://www.globalnames.org/; i4Life – http://www.i4life.eu/
57. Red Lists are the International Union for Conservation of Nature (IUCN) lists of threatened species (http://www.iucnredlist.org/). The Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) lists species which are protected against overexploitation through international trade (http://www.cites.org/)
58. OBIS – http://www.iobis.org/
59. VertNet – http://vertnet.org/
60. Darwin Core – http://rs.tdwg.org/dwc/; ABCD – http://www.tdwg.org/standards/115/
61. www.gbif.org
62. TraitNet – http://traitnet.ecoinformatics.org/; TRY Plant Trait database – http://www.try-db.org/; LEDA Traitbase – http://www.leda-traitbase.org/; MorphoBank – http://www.morphobank.org/; Animal Diversity Web – http://animaldiversity.ummz.umich.edu/; the Phenoscape KnowledgeBase – http://kb.phenoscape.org/
63. IWDB – http://www.nceas.ucsb.edu/interactionweb/; SWISST – http://peacelab.cloudapp.net/swisst.html
64. SDD – http://www.tdwg.org/standards/116/; the Plinian Core – http://www.pliniancore.org/; the Phenotypic Quality Ontology – http://obofoundry.org/wiki/index.php/PATO:Main_Page; the Animal Natural History ontology – http://animaldiversity.ummz.umich.edu/about/technology/
65. Wikipedia – http://en.wikipedia.org/; Wikispecies – http://species.wikimedia.org
66. See for instance the list of EOL global partners http://eol.org/info/global_partners and content partners – http://eol.org/content_partners
67. EUBrazilOpenBio – http://www.eubrazilopenbio.eu/
68. The World Database on Protected Areas (WDPA) can be searched at http://www.protectedplanet.net/
69. Maxent http://www.cs.princeton.edu/~schapire/maxent/; Phillips, S.J., R.P. Anderson and R.E. Schapire (2006), "Maximum Entropy Modeling of Species Geographic Distributions", *Ecological Modelling* 190, 231-259, http://www.cs.princeton.edu/~schapire/papers/ecolmod.pdf
70. OpenModeller http://openmodeller.sourceforge.net/; Muñoz, M.E.S. *et al*. (2009) "OpenModeller: A Generic Approach to Species' Potential Distribution Modelling", *Geoinformatica*, DOI: 10.1007/s10707-009-0090-7, http://link.springer.com/article/10.1007%2Fs10707-009-0090-7
71. Map of Life – www.mappinglife.org
72. Generalized dissimilarity modelling http://www.biomaps.net.au/gdm/ - see also Ferrier, S. *et al*. (2007), "Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment", *Diversity and Distributions* 13: 252-264, http://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2007.00341.x/abstract
73. Wallace Initiative – http://wallaceinitiative.org/; ClimaScope – http://climascope.tyndall.ac.uk; eHabitat – http://ehabitat.jrc.ec.europa.eu/; Ermitage – http://ermitage.cs.man.ac.uk/
74. DOPA – http://dopa.jrc.ec.europa.eu/
75. TGICA – http://www.ipcc.ch/activities/activities.shtml#tabs-4
76. IRMNG – http://www.obis.org.au/irmng/
77. For more information see Hardisty and Roberts (2013) "A decadal view of biodiversity informatics: challenges and priorities", *BMC Ecology,* 13(16), http://www.biomedcentral.com/1472-6785/13/16/abstract
78. GEO BON – http://www.earthobservations.org/geobon.shtml; Vital Signs Africa on the Conservation International website – http://www.conservation.org/about/centers_programs/funding/pages/vital_signs.aspx, ICIMOD – http://www.icimod.org/; LTER Network – http://www.lternet.edu/; GLORIA – http://www.gloria.ac.at/; and Eye on Global Network of Networks – http://www.eyeonearthsummit.org/special-initiative-global-networks

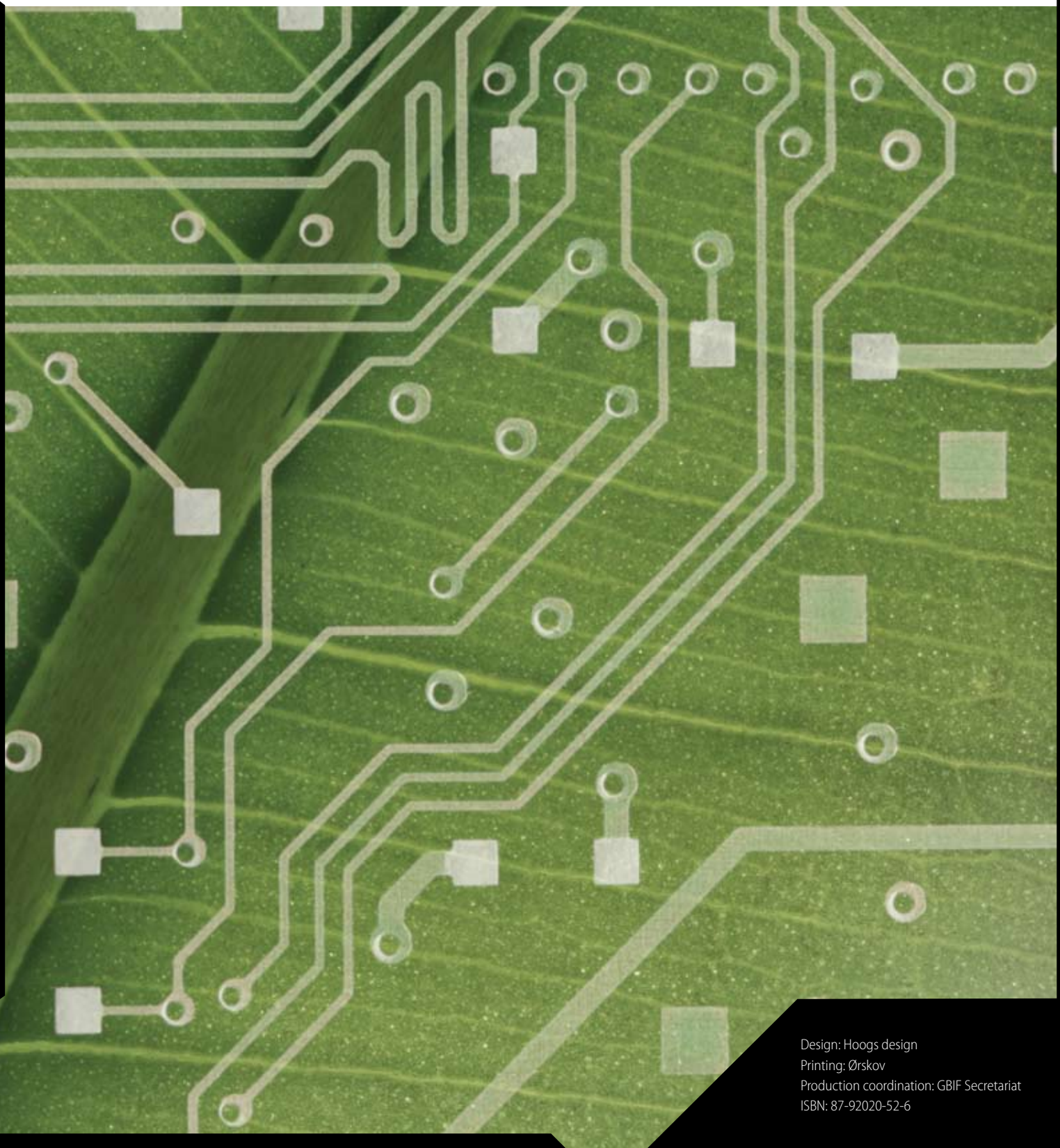
https://www.facebook.com/biodiversityinformatics

@biodiversityinf

info@biodiversityinformatics.org

Global Biodiversity Informatics Outlook

www.biodiversityinformatics.org



Design: Hoogs design
Printing: Ørskov
Production coordination: GBIF Secretariat
ISBN: 87-92020-52-6