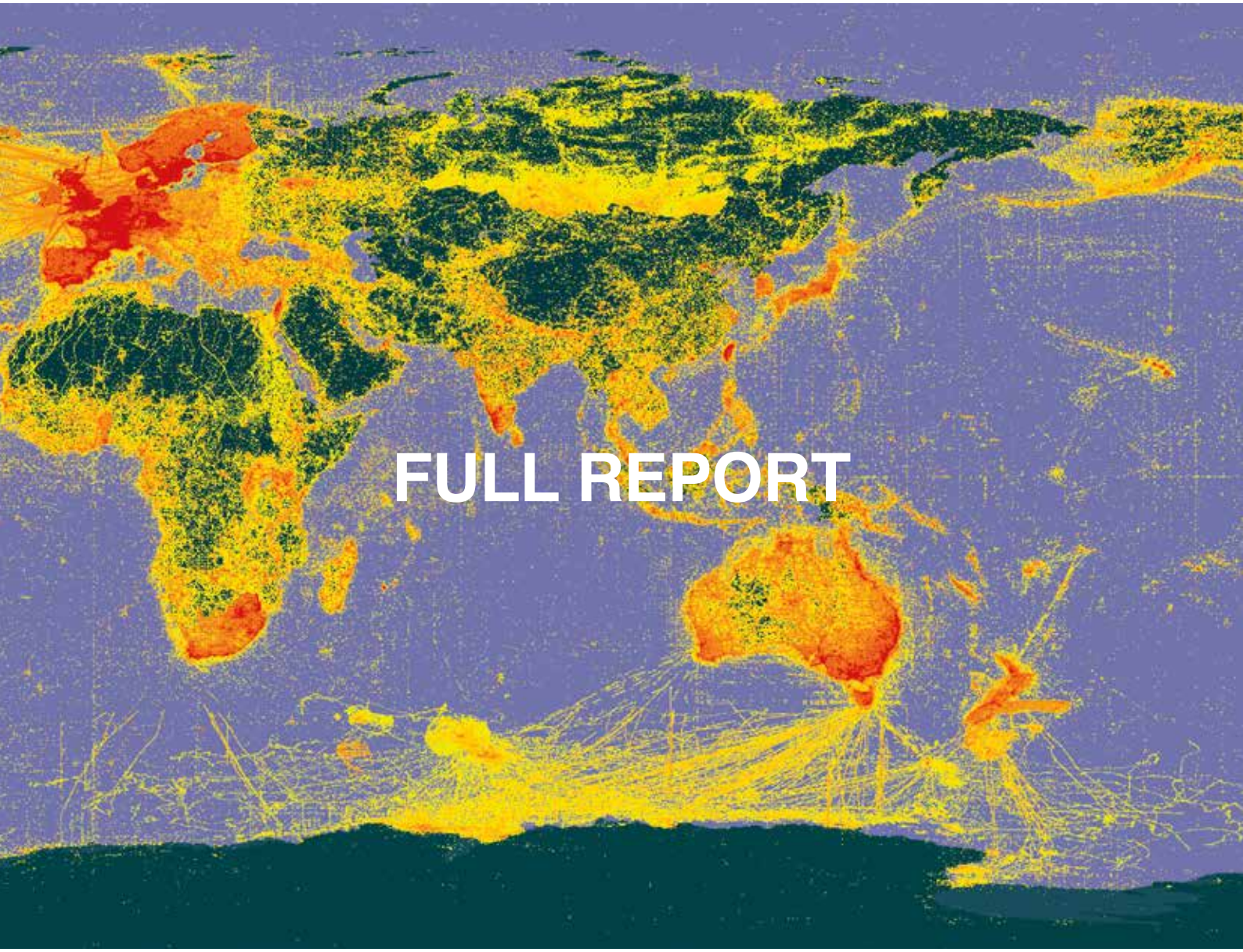
COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL

TWENTY-YEAR REVIEW OF GBIF

GBIF Global Biodiversity Information Facility



FULL REPORT

Please cite this report as:

CODATA, the Committee on Data of the International Science Council; Pfeiffenberger, Hans; Uhlir, Paul; Twenty-Year Review of GBIF (CODATA, Paris, 2020)
https://doi.org/10.35035/ctzm-hz97

The Twenty-Year Review of GBIF comprises two documents:
The EXECUTIVE SUMMARY and the FULL REPORT (this document).[1]

TABLE OF CONTENTS

GBIF Global Biodiversity Information Facility

[1] This independent report was requested from CODATA by the GBIF Governing Board and leadership, who comprise the primary intended audience.

Government decision-makers and others in the biodiversity community may want to read Chapter 9 of the FULL REPORT, which contains the full conclusions and recommendations, and the body of the report for the supporting evidence and rationale. The report is also intended to provide the GBIF team leaders and data Node managers with a neutral view of GBIF by external experts.

1 INTRODUCTION

We live in interesting times. The past two decades of the new century have brought into sharp relief many global trends, both positive and negative, that put the current review of the Global Biodiversity Information Facility (GBIF) in a larger context and will inform relevant activities in the coming years.

We begin with the challenges, not because this is a negative report, but because they provide important reasons for bold action. Our planet is experiencing environmental degradation on a massive scale—whether it is climate warming (IPCC 2018), the impacts on our lands and oceans[2], or the immense loss of biodiversity (IPBES 2019). The latter, which many informed observers have referred to as the beginning of the 6th Great Extinction (Kolbert 2014), is the immediate context for GBIF and this review. Although we are aware of these huge negative trends, there is much that we do not know about biodiversity in many places on the planet and thus are not well equipped to confront the problems comprehensively, even if we mustered the requisite will to do so. And, of course, there are many other forces that work against our ability to respond positively to these crises and that need to be challenged by broadly available factual information and well-informed decision-making. Good information, responsibly used, can be transforming and shine a light on what was previously dark.

These societal challenges provide a compelling rationale and urgent need for data mobilization and collection and support greater action by global information activities such as GBIF. We believe that it is the positive trends, however, that increasingly provide the means to do the biodiversity data work successfully. We therefore focus here principally on the scientific and technical developments, as well as on their governance within the organization and some of the social context in which they have occurred, over the past twenty years.

From a scientific standpoint, there has been an explosion of data and information, and a concomitant paradigm shift to data-driven research and specifically to biodiversity research and its myriad applications. Moreover, there are many new and "non-traditional" sources of information, such as remote sensing technologies and meta-barcoding, and traditional activities that gain new relevance under the umbrella of citizen science, that are being integrated in this changing paradigm. Novel and redesigned scientific data organizations are developing approaches designed specifically to take advantage of the unprecedented data opportunities in a cooperative framework. And innovative open data policies and data infrastructures, including generic Science Clouds, are proliferating.

The inexorable progress of information technologies makes this scientific data revolution possible. Not only is the speed, power, and pervasive nature of computation and data networks extending the realms of possibility, but linked infrastructures for the data and various information management and analysis tools are being rapidly developed. This is also leading to an increasing convergence of the physical and digital manifestations of all matter, including in the life sciences, with profound implications for the future.

As a biodiversity information facility and global infrastructure expressly set up as an international focal point and institution to mediate such data, GBIF is uniquely prepared to take advantage of these opportunities and confront the challenges. This 20-year review and resulting report are intended to help GBIF and its large network make that happen.

---

[2] https://www.ipcc.ch/site/assets/uploads/sites/3/2019/09/SROCC_SPM_HeadlineStatements.pdf

Box 1.1: Ana María Hernández Salgar, chair of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services:

1. GBIF is a key partner to IPBES due the importance of providing the best standardized data available as a basis for building and managing accurate information necessary to take good decisions.

2. Open and accessible data is key to respond to the gaps and needs of knowledge on status, trends and scenarios of direct and indirect drivers of biodiversity loss identified in IPBES deliverables.

1.1. Background and Previous Reviews of GBIF

In 1999, the Organization for Economic Cooperation and Development's (OECD) Working Group on Biological Informatics of the Megascience Forum (now the Global Science Forum) recommended the establishment of an organization such as GBIF. The OECD report concluded that "An international mechanism is needed to make biodiversity data and information accessible worldwide". The report went on to specifically recommend the establishment of a Global Biodiversity Information Facility, to:

> *enable users to navigate and put to use vast quantities of biodiversity information, advancing scientific research … serving the economic and quality-of-life interests of society, and providing a basis from which our knowledge of the natural world can grow rapidly and in a manner that avoids duplication of effort and expenditure. (OECD 1999)*

In summary, the OECD Report expected GBIF to have an impact on a wide range of beneficiaries. We also used these points in our interviews and provide them here as the benchmark for this review. The areas in which the OECD expected benefits from this endeavor included:

- The advancement of science;
- Increased efficiency in R&D spending; and
- A variety of applications including biodiversity conservation, agriculture, health, industry (such as commercial products and informatics tools), the compatibility of protecting ecology and economic development, responsible resource management, and sustainable development.

That recommendation was endorsed by OECD science ministers and GBIF was officially established through a Memorandum of Understanding between participating governments in 2001. It also should be noted that the relevant OECD recommendations are referenced throughout our review report, consistent with our statement of task (see section 1.3).

The original OECD Megascience Forum report laid out substantial expectations on the remit of a Global Biodiversity Information Facility. Today's GBIF is still tasked to fulfil some of these, after a broad reduction of the scope in its work program in 2002 by the Governing Board, when the actual scale of available funding became apparent. We conclude that most of the 1999 expectations, mainly the broad areas of impact recounted above, remain relevant for guiding GBIF today. Other expectations will be introduced throughout the review, where appropriate.

GBIF has now existed for almost 20 years, so it is a good time to examine the past and to look forward. Therefore, at the suggestion of the GBIF Secretariat, the governance bodies of GBIF have asked CODATA to review the organization's activities and accomplishments.

Fifteen years have gone by since the first review, and GBIF and the landscape surrounding it have changed a great deal. The 2005 review (conducted by CODATA and KPMG in 2003-2004) offered some recommendations that still appear valid and may have contributed to the successes of GBIF. We reference them in relevant places throughout the report. For the reader interested in all the initial recommendations and text, please see: "Global Biodiversity: The GBIF 3rd-Year Review" (CODATA 2005).

We should note, however, that when we found that some of the earlier recommendations were still valid and important, it did not show that GBIF has been moving too slowly. Rather, it showed that a successful GBIF depends on many interwoven issues that can only be solved over decades. Such complicating factors include cultural issues (e.g., differing perceptions of biodiversity or data policies), scientific and technological concerns (e.g., harmonizing standards and working habits), and organizational aspects (e.g., establishing a very lean organization as a trusted mediator and source). Another decisive factor has been that the recommendation on the necessary level of funding was never realized. Even today, only about one-quarter to one-third of the funding levels recommended in 2004 is available, depending on whether third-party project funding is included.

GBIF was reviewed again in the late 2000s. The product of that review, the "2010 Forward Look" report, was very science-driven and its recommendations reflected this orientation (GBIF 2010a). The recommendations in that report also were worthy targets – some are still today. But those targets were set for GBIF as a whole, which the authors interpret in a very broad sense, including Participant countries (see the footnote on p.3 of the "2010 Forward Look" report), and thus address a widely heterogeneous set of stakeholders. For example, the digitization of museum resources needs massive national funding, while in many cases data quality issues can only be resolved by involving original contributors or "the scientific community", not just data publishers. We discuss these in turn within the appropriate chapters. For the readers interested in seeing the entire 2010 review report and its recommendations, see "2010 Forward Look" (GBIF 2010a).

1.2. Brief Description of GBIF Today

GBIF—the Global Biodiversity Information Facility—is both an intergovernmental organization governed by a nonbinding Memorandum of Understanding and an international network and research infrastructure. Funded by 40 of the world's governments (as of January 2020), with supplementary support given by other governmental and nongovernmental organizations, it is aimed at providing anyone, anywhere, open access to data about all types of life on Earth.

GBIF has established a network[3] of participating countries and organizations that work through – currently 97 – Participant and Affiliate Nodes (whose roles are detailed in section 7.1.2). It provides data-holding institutions around the world with common standards and open-source tools that allow them to share information about species, now numbering about 1.4 billion (or 1,388 million) occurrence records. This knowledge derives from many sources, including everything from museum specimens collected in previous centuries to recent data provided by "citizen scientists" and environmental DNA samples.

[3] https://www.gbif.org/the-gbif-network

The GBIF network draws all these sources together through the use of data standards, such as the Darwin Core. Biodiversity data publishers provide open access to their datasets, allowing scientists and others to apply the data in a host of peer-reviewed publications and policy papers each year. Many of the analyses that have been done, covering topics such as the impacts of climate change on biodiversity, the spread of invasive species, or priorities for food security, public health, conservation, and the designation of protected areas, would be difficult if not impossible to perform without open access to comprehensive biodiversity data.

The organization is headed by a Governing Board, which has an Executive Committee, and is administered by a Secretariat located in Copenhagen, Denmark. The work is guided by several standing and ad hoc committees. The Secretariat, consisting of the Director and 26 staff (as of January 2020), is currently organized into four teams:

- **Participation and Engagement** is responsible for operating the network of Participants and publishers, recruiting new members and enhancing the capacity of current ones.
- **Data Products** is responsible for the quality and scientific value of the integrated data products produced by the GBIF network.
- **Informatics** is responsible for data management, software development and the overall operation of the GBIF infrastructure.
- **Administration** is responsible for maintaining both the network and the Secretariat's underlying operations and processes.

Additional details about the structure, organization, members, work, and the data that GBIF mediates may be found on its website at www.gbif.org and in the body of the review report that follows.

1.3. Overview of the 20-Year Review Process

The statement of task given by the GBIF Governing Board to the review team specifies:

1. *Review how effective GBIF has been since 2001 in meeting the expectations from the OECD working group*
2. *Review the governance and sustainability of GBIF as a global network and organization (including hosting of the Secretariat in Denmark)*
3. *Review the place of GBIF within the 2018 landscape of biodiversity and research organisations*
4. *Review the technical aspects of GBIF's delivery and its sustainability and trustworthiness (in particular, to researchers) as a research infrastructure*
5. *Consider the challenges in the next 5-10 years that GBIF needs to be prepared to meet*
6. *Provide recommendations on areas needing attention and improvement.*

This task statement guided our work and was the basis for the organization of this report. Our contract also provided a list of stakeholder groups that ideally should be (and were) included. This list detailed the broad classes of:

- Participating GBIF nodes and national agencies from Voting Participant Countries
- Biodiversity-related Multilateral Environmental Agreements (MEAs) and other intergovernmental bodies
- Scientific bodies focused on biodiversity
- Data publishers
- Research users
- Research infrastructure bodies
- Other biodiversity informatics initiatives with missions related to the original OECD Megascience Forum scope

The contract stipulated that there be good geographical coverage as well as broad engagement with experts from a wide range of scientific disciplines and applied domains.

1.4. Review Methods

The review started in January 2019 and ended with the delivery of the report to GBIF at the start of 2020. The review team consisted of the CODATA Executive Director, two lead consultants, and four regional consultants (see Annex 10.1). Much of the information came from telephone and online interviews conducted by consultants in the five major regions in which GBIF operates—Asia/Pacific, Africa, Europe, Latin America, and North America. These interviews were conducted between March and May in 2019.

We conducted 23 interviews in North America (US and Canada—although Mexico is North American, it was included in the Latin American region because of language), 9 in Latin America, 32 in Europe, 17 in Africa, and 10 in Asia/Pacific. These oral interviews were augmented by 17 written responses gathered from all the regions. The particularly high number of interviews *in* Europe corresponds to the high percentage of international organizations being hosted there. Aside from that, the regional distribution broadly reflects the current Participation in GBIF and the strength of biodiversity informatics communities in those countries, but not the importance of regions in terms of their biodiversity. We made an effort to represent regions evenly in SWOTs and tried very hard not to let the authors' "western" origin and work experience distort the conclusions and recommendation. Note that the interviewees from the different regions had diverging emphases on the topics addressed in this report. Although trying and retrying multiple times, we had few responses from Asia.

The experts who were interviewed reflected the broad diversity of the GBIF community and included a mix of data publishers, data users, biodiversity infrastructure providers, museum managers, taxonomists, modelers and researchers, citizen science organizations, funders, and representatives from various governmental and non-governmental biodiversity-related organizations. They were selected from three lists of knowledgeable individuals provided by the GBIF Secretariat, the contacts of the review consultants, and experts associated with CODATA and other international organizations. Several of the people we interviewed had been engaged with GBIF in some fashion since its inception (e.g., as Node managers, Heads of Delegation, Science Committee officers and members, members of the Governing Board, trainers) and a few even participated in the initial planning efforts that led to the creation of GBIF. Interviewees,

especially from the OECD countries, were quite familiar with GBIF services and the GBIF network and broader biodiversity community, although some professed to having only a modest understanding of the underlying technologies and the GBIF governance process.

The interviews followed a consistent format that focused on: (1) understanding the interviewee's relationship with and knowledge of GBIF; (2) ascertaining GBIF's strengths, weaknesses, opportunities, and threats; (3) identifying issues relevant to sustainability of GBIF and other biodiversity infrastructures and initiatives; and (4) soliciting any other pertinent comments and suggestions. The text of the interview template used by the regional consultants is provided in Annex 10.4. All interviewees were highly engaged, valued the many contributions that GBIF had made during its existence, and wished to provide input that would contribute to the future success of GBIF.

The review team also made several visits to high-value committees and individuals in the GBIF community. One of the lead consultants, Hans Pfeiffenberger, attended a joint meeting of the Science Committee and the Nodes Steering Group in Copenhagen on 7-8 February 2019, and interviewed several of the individuals there. The two lead consultants then interviewed key members of the Secretariat on 25-26 February 2019 at the GBIF Headquarters in Copenhagen. Simon Hodson, the Executive Director of CODATA, attended the Executive Committee meeting, also in Copenhagen, on 9-10 April and spoke with members of that group. Finally, Hans Pfeiffenberger attended the European regional Nodes meeting in Oslo, Norway on 22-24 May 2019 and interviewed 10 Node managers there.

All these interviews of members of the GBIF community were accompanied by observations of the Governing Board and related meetings during the week of 21 October 2019 in Leiden, NL by the CODATA Executive Director, Simon Hodson, and the two lead consultants of the GBIF committee meetings noted above. Attendance by the senior members of the review team were augmented by informal background talks with the GBIF Secretariat and other GBIF members, phone calls and emails with the regional consultants, both individually and collectively, two sets of independent expert external reviews, and extensive desk research. Thus, we acknowledge the contributions by many to this report, and list these in Annex 10.1.

It must be kept in mind, however, that the report's prescribed method is based on a relatively small number of interviews, considering the size of the broader biodiversity informatics community and the heterogeneity of stakeholders, and only with individuals who were already engaged with GBIF in one or another role. The review team is quite confident to have caught most of the important aspects, from a qualitative standpoint. However, this method could not uncover all of the issues or, for example, determine the number of potential users that desire specific new products and services, the costs associated with developing individual new products and services, and how much additional funding can reasonably be acquired from the various sources that have been suggested.

1.5. Structure and Organization of the Report

This report is organized to respond to the task statement and to involve the recommended stakeholder groups. The intended audience of this report is, primarily, the GBIF Governing Board and its Executive Committee, and the organization's Secretariat. We hope that the other bodies of GBIF, such as the various standing and ad hoc committees, can benefit from the detailed information and conclusions in Chapters 3 to 8 as well.

The entire review report is summarized in an Executive Summary, which highlights the review process and the principal conclusions and recommendations. We tried to keep the summary as brief as possible.

The body of the report comprises a mix of descriptive, analytical, and advisory approaches. The larger trends and the GBIF review history are presented in this chapter and the "Landscape" setting of GBIF is described in Chapter 2. The Annexes at the end of the report, in particular Annex 10.5 on the references used, augment the descriptive focus of these two chapters.

The analytical approach forms the body of this report. Chapters 3 – 8 respond to the statement of task presented to us by the GBIF Governing Board (see section 1.3) and end with formal conclusions related to these issues. Each of these chapters follows the same structure, with three main sections.

Section 1: The first section focuses on a description of the key elements of the GBIF mission and program in the given area and on some external aspects that we considered relevant.

Section 2: The second contains a listing of the SWOTs (Strengths, Weaknesses, Opportunities, Threats) compiled from the comments made by the expert interviewees during the regional interviews and at the other meetings attended by the lead consultants (see section 1.4).

Note that the compilation of SWOTs does not attempt to correct any factual errors or imprecisions, or contradictions between interviewees and views from different regions. Rather, the SWOT comments directly reflect the views or impressions provided by the interviewed individuals, including those opinions possibly formed in the past or with incomplete knowledge of the complex, evolving landscape. Several of the SWOT sections are also quite long, which is indicative of the interviewees' focus on those issues, and a few of their observations may be repeated in different chapters, which only shows that those comments are relevant to more than one issue area.

The review team came to the conclusion that a full aggregation of the SWOTs named would have averaged out significant differences between regions and stakeholder groups, or destroyed nuances of opinion. Taking this heterogeneity into account and the total number of interviews, we realized as well that it would not be advisable to derive quantitative statements about these SWOTs from a total of over 100 interviews. Where we actually summarize a number of SWOTs, we indicate frequency by qualifiers such as "some" or "many".

Section 3: Each of the Chapters 3 – 8 then has a third section devoted to the conclusions distilled from the first descriptive section and the SWOT comments, and the authors' knowledge of the situation.

The conclusion sections at the end of each of those chapters then lead to the summary conclusions and detailed recommendations proposed in Chapter 9.

1.6. Conventions in this Document

- Throughout this document, with the exception of literal quotes and SWOTs (see 1.5), we use capitalization of a number of terms to indicate a specific definition or meaning in the context of GBIF. The "Secretariat" is that of GBIF. "Nodes" refers to the national GBIF Nodes, while "NODES" is the standing committee representing all Nodes.

Note also that "GBIF" denotes not just the GBIF Secretariat, but also the Nodes and the GBIF governance bodies. "Participants" (Voting or Associate) are countries or organizations that have signed the GBIF MoU. "Participation" is the GBIF term for its formal membership of these Participants and does not include "Affiliates" or "Partners", who are those committed to GBIF by some formal agreement.

"Publishers" denote data publishers endorsed by Nodes, which provide data to the GBIF portal. Of particular importance is the distinction between the mobilization of data – here taken to

mean “making data available for use by others”, as it is performed by Publishers, and the mediation of data – here taken to mean the provision of access to all Publishers’ data through GBIF’s portal.

Literal quotes are either enclosed in “quotation marks” in running text or *formatted in italics*, if comprising one or more paragraphs. Where not explicitly noted, URLs were visited in the timeframe of June to September 2019. Generally, we provide URLs in footnotes, except where the composition of the URL, e.g., the parameterization, adds insight.

Finally, some figures cannot be shown in full size, but only as “thumbnails”, and URLs are provided if one would like to read the entire text. We provide footnotes where this occurs.

2 GBIF AND THE LANDSCAPE OF BIODIVERSITY-RELATED ORGANIZATIONS[4]

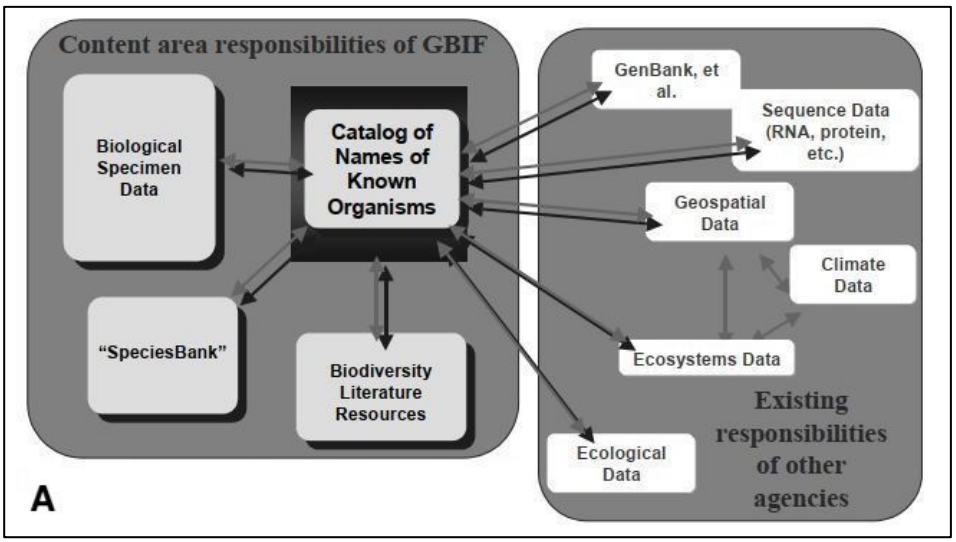
As previously stated, the world has changed a great deal and in many ways over the past two decades. What has not changed much over these years is the predominantly national funding of science and its infrastructures, and that is not likely to change any time soon. It may have been the recognition of this funding approach, which underlies the OECD concept and is still clearly determining the structure of GBIF today, of a nimble Secretariat coordinating national agendas, initiatives, and projects, worldwide. In consequence, GBIF has had to navigate among a "panoply of efforts on the part of so many actors" (GBIF 2010a).

In this chapter, we describe in broad terms those actors who have been the most important thus far, or at least the most visible, in relation to GBIF. One can group these actors into two major categories—first, there are the *users* of the data that GBIF provides and the *contributors* of those data, who frequently are the same, or at least the same types, of people. Second – sometimes intersecting the data contributors and users – are the infrastructure providers who are central to disseminating and organizing the data. Among the users of GBIF are not just scientists and those at the science-policy interface, but also utilitarian players from the economy, health providers, and various governmental bodies. Even the broader public and industry are increasingly large categories of users *and* contributors of biodiversity information.

Among the actors in the biodiversity infrastructure or *informatics* space, some stand out as data providers, such as natural history museums, botanical gardens and other biological specimen collections, while another major and more recent grouping are the genetic information infrastructures. Other data and information infrastructures with an international scope have emerged only over the last 20 years, such as GBIF itself and the Catalog of Life. The purpose of many of these types of organizations is to mediate data from the data providers to the users. All of these groups participate in a range of standard-setting initiatives.

The thinking about biodiversity knowledge and its data and information infrastructure has evolved from the structural picture of the original OECD concept (OECD 1999), to the more recent Global Biodiversity Informatics Outlook (GBIO) framework (Hobern et al. 2012), see Figure 2.1.

[4] This chapter provides an overview to address item 3 of the statement of task, to "Review the place of GBIF within the 2018 landscape of biodiversity and research organizations". The following chapters, in particular 5 and 6, will provide more detail.



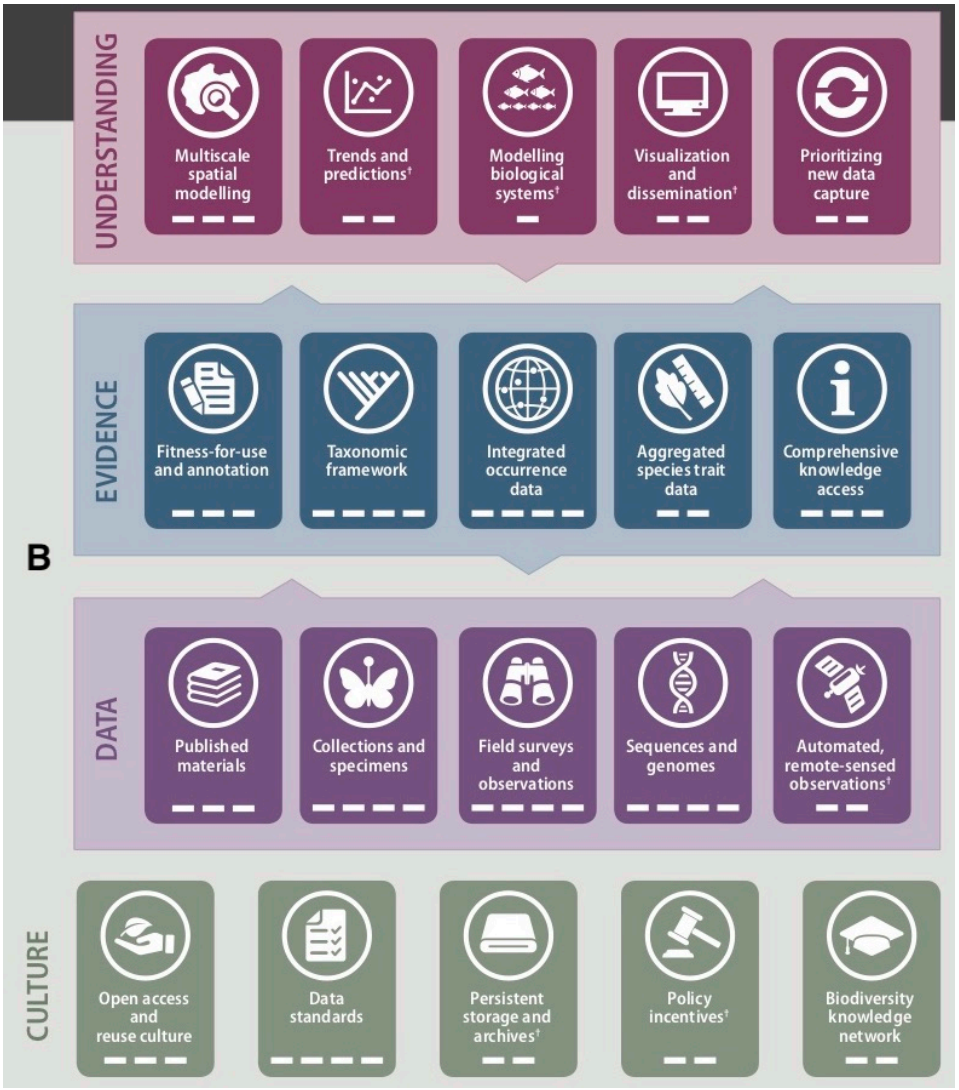


Figure 2.1 (A) Depiction of how "An Electronic Catalogue of Names of Organisms Known to Science will make linkages among many types of biological and non-biological databases possible" (OECD 1999, Fig 1), and (B) The focus areas and components of the "GBIO Framework" (Hobern et al. 2012).

OECD presented a content space, which in a considerable part was to be addressed by "a" GBIF, and assumed links to other data domains within the "responsibility of other agencies", naming one of them (GenBank), whereas GBIO named data types and functions, but no infrastructures or actors. The GBIO authors – participants in the Global Biodiversity Informatics Conference (GBIC) in July 2012 - invited many actors to align their activities, such as projects or funding decisions, to take into account the framework's components.

One element is missing from both diagrams (A and B). Traditional or Indigenous Ecological Knowledge, as recently accentuated[5] in IPBES reporting, will be expected to appear in such frameworks, and indeed ways must be found to integrate and gain knowledge from both spheres. As not all traditional biodiversity information aligns to a "western" taxonomic backbone, GBIF will have to handle this challenge as well.

Both diagrams also have in common that they present biodiversity informatics as one – global – system or concept, presumably employed by all actors. Present structures, however, which we try to capture in this chapter, are not necessarily aligned with the concepts raised in these diagrams (yet), but more strongly determined by asynchronous and nationally or topically focused funding and organizing decisions of the past. Therefore, although this chapter is meant to be descriptive about some biodiversity actors, we need to mention funding sources and issues of geographical or other scope of the resultant actors.

A recent numerical estimate of what some of the actors contribute to the information base about biodiversity has been published by participants of the GBIC2 Conference in the summer of 2018, although it did not address one of the OECD concept's core requirements about "digitizing" museum collections (discussed further in section 2.2):

> *[In] 2018, data published and aggregated through the Global Biodiversity Information Facility (GBIF) surpassed one billion records […] the Catalogue of Life Annual Checklist includes taxonomic information on 1,803,488 living and extinct species (around 75% of known species). The Biodiversity Heritage Library now includes more than 55 million pages of scanned biodiversity literature. The Barcode of Life Data System includes 6,293,000 barcode sequences representing 280,000 species. The recently established Global Genome Biodiversity Network (www.ggbn.org/ggbn_portal) provides access to more than 2 million DNA and tissue samples of 45,000 species. The Encyclopedia of Life Traitbank (https://eol.org/docs/what-is-eol/traitbank) holds structured data on 11 million species traits. (Hobern et al. 2019)*

This description identifies some of the actors currently deemed important. Also, it reveals some of the knowns, but also (implicitly) the known unknowns, namely the current estimate of unknown species. In addition, it shows that for a large percentage of the known species, there is no associated genetic information available or accessible. The article does not discuss whether or to what extent the diverse databases are interlinked, e.g., whether species names are linked to their taxonomic treatment in literature, or whether all items are openly accessible.

The following sections briefly describe some of the actors, including the involvement of GBIF with each, with a focus on their functions and on some of the technical interfaces and data flows.

---

[5] https://www.ipbes.net/deliverables/1c-ilk

We should emphasize that this is not a comprehensive list of stakeholders, Rather, we try to describe some "important" ones, while others are just – hopefully representative - examples (of which there are many, many more).

A summary of the current position of GBIF in this landscape and how it developed from the OECD concept's origins is provided in this chapter's final section, 2.3.

2.1. Users and Contributors of (GBIF-mediated) Biodiversity Data

Science

Per the OECD expectations and its current strategic plan (item 5: "Ensure that GBIF delivers data in the form and completeness required to meet the highest-priority needs of science and, through science, society"), the first line of users to be targeted by GBIF are scientists in academia and scientifically-trained employees of governmental and intergovernmental organizations. This group is also considered the main[6] contributor of published data and its quality assessment. Therefore, this group's general concerns and requirements are dealt with in Chapters 3 to 5.

Academic scientists typically are organized and share knowledge through disciplinary societies, such as the Society for the Study of Amphibians and Reptiles (SSAR)[7] or the European Ecological Federation[8], or through their conferences and journals. There is a plethora of those circles, specializing in certain aspects of biodiversity. Broader areas of biodiversity, but still with a particular scientific interest, are covered by organizations such as the International Commission on Zoological Nomenclature (ICZN), founded in 1895.

In the context of biodiversity data, each group of users and contributors may have very specific requirements, which some try to meet through mostly nationally-funded databases and portals. For example, HerpNET[9] is a database project that was co-funded by GBIF. In 2015, HerpNET was placed under the auspices of VertNet[10], a project funded by the U.S. National Science Foundation since 2011, and VertNet has been a GBIF Associate Participant since 2013.

Considering the field of biodiversity as a whole, there are of course numerous other types of data, corresponding to specific interests, and databases prepared by related scientific communities, such as GloBi[11], on species interaction data. Such databases may assume, over time, the status of an element of the biodiversity data infrastructure.

GBIF occurrence data - and some other infrastructural data elements – provide a reasonably stable baseline in this highly diverse field in view of the many types of data collected.

---

[6] Note that in the case of "citizen science" data, professional guidance on methods and selection of data by scientists is a de-facto requirement.

[7] see https://ssarherps.org/about-ssar/

[8] https://www.europeanecology.org/about-eef/

[9] http://www.herpnet.org

[10] http://www.vertnet.org/about/about.html

[11] https://www.globalbioticinteractions.org/about.html

Organizations at the Science-Policy Interface

Many governmental and non-governmental (NGO), international, regional, and national organizations work at the science-policy interface (Gluckman 2016). They contribute towards analyzing, evaluating, aggregating, and presenting reports to policy-makers based on scientific evidence.

Many of those organizations are guided or inspired by the UN Convention on Biological Diversity (CBD). The convention has 196 governmental members and is represented by bi-annual Conferences of the Parties (COP). The CBD treaty went into effect in 1993 with 168 governmental signatories. In 2010, the CBD adopted a Strategic Plan for Biodiversity 2011-2020, which named the "Aichi Biodiversity Targets"[12] in 2010. CBD also required the parties to develop National Biodiversity Strategies and Action Plans (NBSAPs) - which so far has been done by 190 countries. GBIF is an observer of the Consortium of Scientific Partners (CSP) of the CBD and "contributes directly to Aichi Biodiversity Targets 1, 2, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 17, 19 and 20"[13].

The policy declarations of the CBD and of other intergovernmental bodies, such as the UN Sustainable Development Goals[14], and also regional or national regulations, clearly require a comprehensive and high-quality basis of data inputs. They also need a well-founded and multifaceted aggregation of those data to support indicators related to the targets of policy. The CBD COP and the UN commissioned a number of organizations to develop or provide such data and indicators. The Aichi target 19 requires that "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". Meanwhile, the "Growth in Species Occurrence Records Accessible Through GBIF"[15], is an indicator toward Aichi target 19. (The Biodiversity Indicators Partnership, BIP, is another initiative endorsed by the CBD; GBIF is one of the partners.)

Quite recently, in December 2016, the CBD COP approved a "Guidance to Improve the Accessibility of Biodiversity-Related Data and Information" which, among other measures, endorsed support for GBIF. It stated that: "On a global scale, continued support from Governments for networks such as the Global Biodiversity Information Facility, the Ocean Biogeographic Information System (OBIS) and Group on Earth Observations Biodiversity Observation Network (GEO BON) will help these benefits to consolidate and grow for all Parties." (CBD 2016)

The development of another type of (derived or aggregated) data, the Essential Biodiversity Variables (EBV) "aim to help observation communities harmonize monitoring, by identifying how variables should be sampled and measured." (Pereira et al. 2013). The EBVs will rely on existing data, such as those mediated by GBIF, but also will be influencing what additional data would need to be gathered and how they would need to be described and mediated at a global level. Some of those institutions using EBVs may ask third parties to prepare defined-quality EBV datasets for them, instead of deriving and distilling them from raw data on their own. (As expanded upon below, GBIF works closely with GEO BON, which was invited by the CBD to continue its work on EBVs.)

---

[12] https://www.cbd.int/sp/targets/default.shtml

[13] https://www.cbd.int/cooperation/csp/gbif.shtml

[14] https://www.un.org/sustainabledevelopment/sustainable-development-goals/

[15] https://www.bipindicators.net/indicators/growth-in-species-occurrence-records-accessible-through-gbif

The Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES), formed in 2012, is based on still another multilateral agreement with the mandate to assess "the state of biodiversity and of the ecosystem services it provides to society, in response to requests from decision makers."[16]. It has also been endorsed by the CBD[17]. Some of its outputs and its reliance on GBIF-mediated data are discussed in the later chapters of this review. The most recent report of IPBES is the "Global Assessment Report on Biodiversity and Ecosystem Services" (IPBES 2019). The complexities arising from the goal of developing broad scientific guidance about biodiversity and the desire to develop appropriate indicators were reflected in a *Nature* news article one year before the IPBES 2019 report (Masood 2018). IPBES has been an official Partner of GBIF since 2018, on the basis of a Memorandum of Cooperation.

The International Union for Conservation of Nature (IUCN)[18], a NGO founded in 1948, is best-known for its Red Lists of endangered species. It is a huge organization with 213 countries and government agencies as members, more than 1100 NGOs and indigenous peoples' organizations represented and, above all, more than 15,000 experts organized in six commissions, who "assess the state of the world's natural resources". Their last Red List assessed over 100,000 species. IUCN defined criteria to identify the species that are critically endangered, endangered, and vulnerable, among other categories [19]. Note the "data deficient" category, which for some groups may reach approximately 60% of the assessed species[20]. GBIF has maintained a "long and fairly informal relationship" with IUCN, according to our interview.

Beyond these international actors, there are numerous regional and national agencies, institutes, and groups that provide science-based policy support. These include organizations such as the the Mexican Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO)[21], created 1992, the European Environmental Agency (EEA), a GBIF Associate Participant since 2010, or the Irish National Biodiversity Data Centre, a GBIF data publisher since 2010 and host of the GBIF national Node for Ireland. We could not possibly survey all of them, as there probably are hundreds worldwide, and as they are situated and named differently in each country. (e.g., only in Kenya and Uganda, they go by the appropriate title "National Environment Management Authority" – not to mention that those names would depend on language). But it must be noted that these organizations are a very important class of data users and, at least potentially, contributors of data and for feedback on data quality.

Applied Research

Applied research, in the sense mandated by the OECD concept, would encompass uses of GBIF-mediated data to support applications such as research on human health, food (agriculture, forestry, or fisheries), and invasive species. While the whole field of applications is clearly very broad and deep, some examples drawn from GBIF reporting, particularly from the health area, are highlighted throughout the report. Here, we concentrate on the settings in which applied

---

[16] https://www.ipbes.net

[17] https://www.cbd.int/sbstta/ipbes.shtml

[18] https://www.iucn.org

[19] https://www.iucn.org/resources/conservation-tools/iucn-red-list-threatened-species#RL_categories

[20] See Figure 2 in https://www.iucnredlist.org/resources/summary-statistics

[21] https://www.gob.mx/conabio

work takes place. There are two modes, below, that differ fundamentally regarding the discoverability of GBIF contributions:

Much of this applied research is performed in an academic setting. Its results are most visible in relevant citations of GBIF-mediated data in journal articles (see Chapter 4, Table 4.1). Some of the institutes performing applied research are also visible as GBIF members, e.g., the Faculty of Agricultural Sciences at the University of Abomey-Calavi as host of the GBIF Benin Node.

Non-academic institutions, such as the Food and Agriculture Organization (FAO), an agency of the UN, or the Consultative Group for International Agricultural Research, now known only as CGIAR, ("Home to more than 8,000 scientists, researchers, technicians, and staff"[22]), clearly have an interest in GBIF data as well, as their interaction with GBIF shows. For example, a researcher from CGIAR chaired the GBIF "Task group on data fitness for use in agrobiodiversity"[23], and FAO requested an "Input Paper from Bioversity International, CIAT, CIP and GBIF: Global Information System for In- situ Conservation and On-farm Management of PGRFA"[24] (2015 and 2016, respectively). The outputs of such organizations, however, may or may not appear in public and may or may not clearly indicate which GBIF-mediated data had been used, while their impact on society may be substantial.

It is also clear from the existence of the joint studies that the requirements of such applied research on the "quality" of data may differ considerably from those of basic research. These issues are discussed further in Chapter 5.

Citizen Science

"Citizen Science" is neither a scientific discipline nor is it represented by a single organization. Rather, it is a specific kind of interaction between scientific projects and the public. At the most general level, involving citizens in scientific projects can help create an understanding and trust in scientific methods. Organizations at the science-policy interface of biodiversity, in particular, recognize the potential of citizen science to increase awareness of biodiversity, its value, and the threats to it. They also acknowledge the value of Traditional Environmental Knowledge, which could be considered a specific kind of citizen science – although this kind of information or knowledge cannot easily be mapped to the terms and concepts of scientific data.

Beyond that, science itself can benefit in a narrower sense, in that citizens can help to close gaps in observations and data, which is possible at research-grade quality. As one expert report puts it:

> *Citizen science is a rigorous process of scientific discovery, indistinguishable from conventional science apart from the participation of volunteers. When properly designed, carried out, and evaluated, citizen science can provide sound science, efficiently generate high-quality data, and help solve problems. […] There are not enough professionals (or funding to support them), however, to monitor EBVs at large scale and adequate resolution. Citizen science (CS) offers an additional way to monitor EBVs, and also offers other benefits to conservation through public engagement.* (McKinley et al. 2017)

---

[22] https://www.cgiar.org/research/research-centers/

[23] https://www.gbif.org/news/82398/task-groups-to-help-make-data-more-fit-for-use-in-key-research-areas

[24] http://www.fao.org/3/a-be668e.pdf

For example, amateur ornithologists and other naturalists are a relatively old phenomenon, but they have increasingly sophisticated and broadly available data collection tools. They appear to be still more prevalent in Europe and North America, but in the last decade or so, citizens from other countries have made an impact as contributors to GBIF. For example, citizens from South Africa, supported by the South African National Biodiversity Institute (SANBI), provided their Southern Africa Bird Atlas, which is currently the largest single African data source in GBIF[25]. Citizens from India, supported by eBird, contributed the largest dataset about India, as can be inferred from India's country page at GBIF[26], and iNaturalist supports people from Mexico, Colombia, and Ecuador[27]. Considering the ubiquity of "smart" mobile devices and the technical ease of building global social platforms for communities, citizen science appears to hold a huge potential for all countries.

In the context of GBIF, arming the amateur ornithologist community with a technical platform such as eBird[28] and supporting it with professional and scientific expertise (the staff of the Cornell Lab of Ornithology[29]), did lead to a substantial source of occurrence data. The "eBird Observation Dataset"[30] within GBIF, accounts for almost 562 million occurrences, 42% of all occurrences or 67% of all Aves occurrences mediated by GBIF[31].

This begs the question whether citizen science can contribute beyond birds, or, in general, charismatic species. Fortunately, there are examples not involving such recognizable species: More than 75% of iNaturalist observations[32] are not of birds and a quarter century of collections by the Entomological Society Krefeld, which followed a "standardized protocol of collection", allowed the discovery of a "more than 75 percent decline over 27 years in total flying insect biomass in protected areas" (Hallmann et al. 2017).

Among the experiments that are underway, there are some also within the GBIF community, e.g., on how to engage the public on issues that may be considered gaps (see Chapter 5). E.g., at the GBIF European Nodes meeting in 2019, examples of this were shown, namely students observing lichens[33] or the Mosquito Alert[34] in Spain, and how accurately hikers would observe tree lines in Norway[35].

---

[25] https://www.gbif.org/dataset/906e6978-e292-4a8b-9c39-adf6bb0f3323#description

[26] https://www.gbif.org/country/IN/publishing

[27] https://www.naturalista.mx/pages/network

[28] https://ebird.org/

[29] https://www.birds.cornell.edu/home/staff-directory

[30] http://doi.org/10.15468/aomfnb

[31] As of 30 June 2019, https://www.gbif.org/occurrence/search?taxon_key=212 , see facet "publishers"

[32] https://www.gbif.org/occurrence/search?publishing_org=28eb1a3f-1c15-4a95-931a-4af90ecb574d

[33] https://www.gbif.no/events/2019/gbif-eu-2019/day-2--liquencity_2019_cv.pdf

[34] https://www.gbif.org/dataset/1fef1ead-3d02-495e-8ff1-6aeb01123408

[35] https://www.gbif.no/events/2019/gbif-eu-2019/day-2--natur-i-endring-michal-torma.pdf

Commercial Sector

Biodiversity data are used in a variety of private-sector settings, for example, in the financial sector. According to IUCN, the "World Bank Group Performance Standard PS6 uses The IUCN Red List Index to minimize the risk to biodiversity from large-scale infrastructure and natural resource extraction projects"[36]. The Equator Principles (EPs) – "a risk management framework, adopted by financial institutions for determining, assessing and managing environmental and social risk in projects"[37], are agreed upon by an impressive array of commercial banks. The association discussed a "proposal on biodiversity data sharing" for the next version of the EPs at their Annual Meeting 2018[38].

Beyond finance examples, the commercial sector is not yet very visible in GBIF activities, either as a funder of data activities or as a user, but industrial concerns can and sometimes do act as users and contributors to public biodiversity data. Examples of such involvement at both the contributor and the user side are oil companies involved in environmental risk assessment or impact studies, agrobusinesses, forestry and fishing industries, health businesses, and others.

Also at the national level, the GBIF Secretariat and a number of Nodes have been working actively to promote open publication of data generated in the course of Environmental Impact Assessments (EIAs) and post-development monitoring through GBIF by commercial entities, e.g., consultancy companies for engineering and architecture in Norway and a French oil company[39].

2.2. Biodiversity Infrastructures

This section focuses on those actors in the biodiversity infrastructural landscape that implement many of the functional expectations of the OECD working group, today.

Museums and Collections

"Thousands of botanical gardens, natural history museums and universities hold an estimated 3 000 000 000 specimens worldwide" (Wheeler et al. 2012). Museums, botanical gardens, and many other types of collections, such as herbariums, arboretums, and the culture collections of microorganisms and cultured cells, hold specimens of species and make them available for taxonomic and genetic research. Many such specimens or samples can be associated with the location and time of their collection (and in the case of fossils, their approximate date of living) and could thus become the source of GBIF occurrence records.

The OECD concept of 1999 had identified specimen collections as a major source of biodiversity data. This has certainly played out to some extent in the area of taxonomy (see the discussion of the Catalogue of Life, below). However, it had also been expected in the OECD report that today practically all collection specimens should have been digitized. Unfortunately, that is still far from being achieved. An interviewee from one of the most affluent OECD countries reported

---

[36] https://www.iucn.org/resources/conservation-tools/iucn-red-list-threatened-species

[37] https://equator-principles.com/about/

[38] https://equator-principles.com/ep-association-news/ep-association-annual-meeting-2018-outcomes/

[39] https://www.gbif.org/project/2Zik1tfJoh3C92ZslvhDIr/openpsd-promoting-publication-and-use-of-private-sector-data-on-biodiversity, https://www.gbif.org/article/2gnrlnYXNKiuWCoOqi8gQG/mobilization-of-biodiversity-data-from-the-private-sector, and https://www.gbif.org/dataset/5dfd3144-25b0-4a1c-9df6-91b9cc231ccc

that of approximately 150 million specimens held in that country, no more than 10% are digitized by now, and that in a very large project proposal the average cost of digitization is projected to be 2 EUR per specimen.

The scale of the problem is confirmed by one of the most highly regarded museums. "European Natural Science collections contain around 1.5 billion specimens representing an estimated 55% of global collections"[40] (dated 16 May 2018) and "less than 10% of the estimated 1.5 billion specimens [are] digitally accessible"[41] (dated 15 July 2019).

Thus, the cost being in the range of billions of euros or dollars, it is unlikely that the 1999 expectation of close to 100% digitization would be achieved anytime soon. This is true even if the following big projects are taken into account:

- iDigBio in the US is "the National Resource for Advancing Digitization of Biodiversity Collections (ADBC) funded by the National Science Foundation"42. On its homepage, it claims to host more than 120 million digital specimen records.

- DiSSCo, a multinational research infrastructure, funded by the European Commission, is supposed to "work for the digital unification of all European natural science assets" (referring to specimens) and to "represent the largest ever formal agreement between natural history museums, botanical gardens and collection-holding universities in the world."[43]

GBIF currently mediates 164 million occurrence records classified as "preserved specimens"[44] and is considered to be a "thematic international repository" (*sic*) by DiSSCo and perhaps by other collections as well. In 2010 – working along the OECD expectations – GBIF had a significant presence in a special issue of the journal *Biodiversity Informatics*, in particular in the "Summary of Recommendations of the GBIF Task Group on the Global Strategy and Action Plan for the Digitisation of Natural History Collections" (Berendsohn et al. 2010).

Recently, "The Global Registry of Scientific Collections (GRSciColl), a clearinghouse of information about the world's scientific institutions and collections, has found a new home on GBIF.org."[45]. This registry, certainly a key to identify contributors of data, had been developed by the Smithsonian Institution, the Barcode of Life, and other projects. A relevant contribution is the provisioning and maintaining of a stable identifier for each institution and collection (InstitutionCodes and CollectionCodes), allowing to cite - *now unambiguously* - the collection in datasets and publications, and to link data in GBIF to those institutions.

Life science collections, including museums, are also positioning themselves for genetic research through the Global Genome Biodiversity Network (GGBN), "A global network of well-managed collections of genomic samples from across the Tree of Life, benefiting society

---

[40] https://naturalhistorymuseum.blog/2018/05/16/uniting-europes-1-5-billion-specimens-digital-collection-programme/

[41] https://naturalhistorymuseum.blog/2019/07/15/who-uses-collection-data-digital-collections-programme/

[42] https://www.idigbio.org/about-idigbio

[43] https://www.dissco.eu

[44] As of 20 Sep 2019, https://www.gbif.org/occurrence/search?basis_of_record=PRESERVED_SPECIMEN

[45] https://www.gbif.org/news/5kyAslpqTVxYqZTwYn1cub/gbif-provides-new-home-for-the-global-registry-of-scientific-collections

through biodiversity research, development, and conservation."[46]. GBIF staff members work on the GGBN "Task Force on Data Standards and Data Access".

Data Infrastructures

The single most important missing infrastructure identified by the OECD was a "Global Names Architecture that allows for unambiguous and persistent identification of biodiversity entities at the organism level, that is robust and responsive to taxonomic changes, and that does not necessarily depend on prior scientific description of the species in the traditional sense." (GBIF 2010a, p.4)

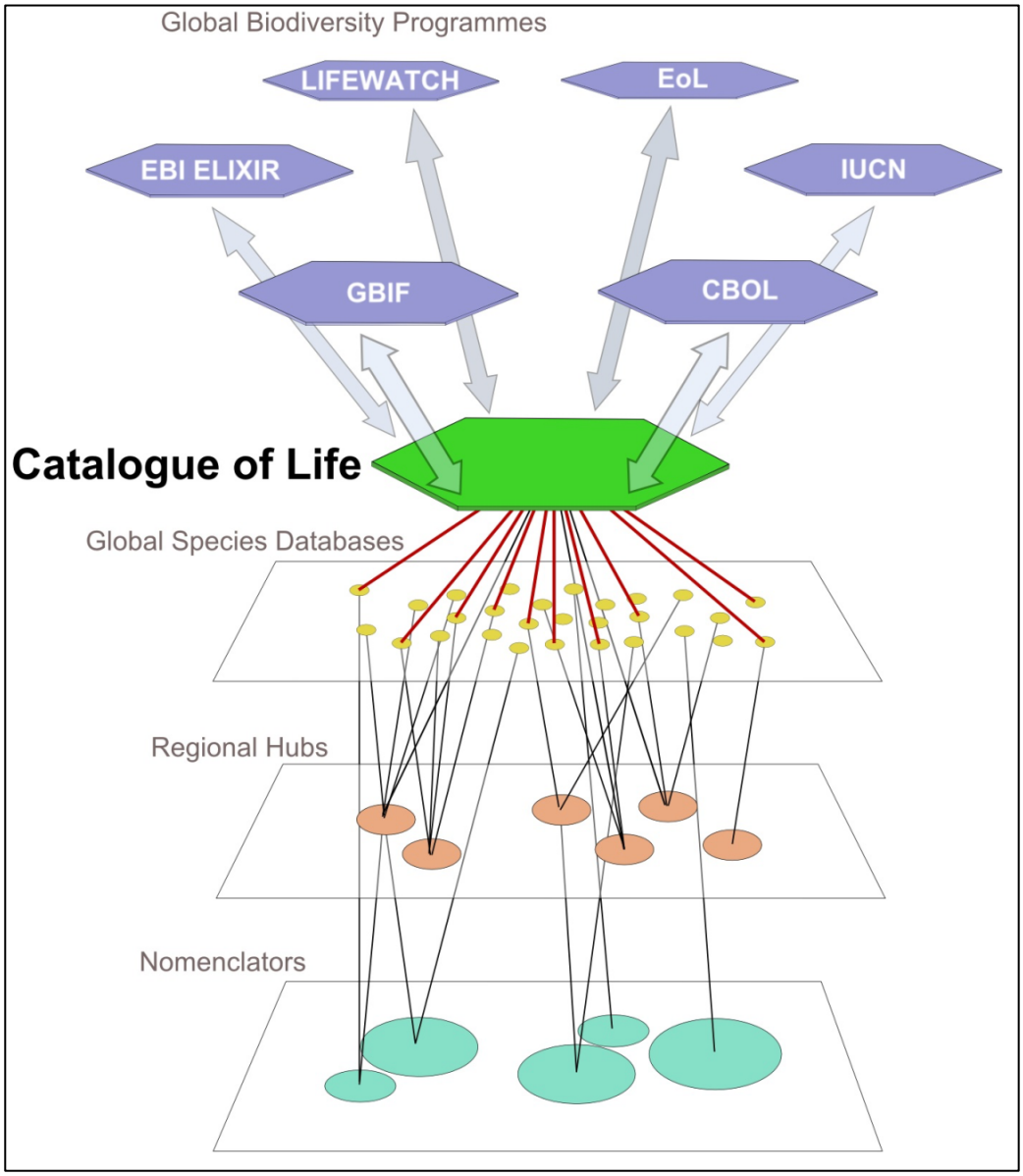


Figure 2.2: The Catalogue of Life is developing as a key partner to the major programs that inform our understanding of global biodiversity (see http://www.catalogueoflife.org/content/about).

[46] https://wiki.ggbn.org/ggbn/About_GGBN

Today, one would certainly identify the Catalogue of Life (CoL) with this goal, and CoL draws a picture of similar centrality of itself (see Figure 2.2, above). CoL merged the interests and data of two large projects, Species 2000 and Integrated Taxonomic Information System (ITIS). A closer description[47] of other funders, users and contributors of data, and the data flows among them, confirms this picture. GBIF appears here as a data user, in that it uses the CoL checklists to form its taxonomic backbone. Both ITIS and Species2000 have been Associate Participants of GBIF since 2001.

Although CoL serves a checklist of over 1.8 million species, there are still downsides. GBIF – just like others – still has had to maintain its own taxonomic backbone, merging the CoL checklist with additional sources. Two of the goals of a new project, Catalogue of Life Plus (CoL+)[48], appear to illustrate the problem:

1) *creating both an extended and a strictly scrutinized taxonomic catalogue to replace the current GBIF Backbone Taxonomy and Catalogue of Life*

2) *separating nomenclature (facts) and taxonomy (opinion) with different identifiers and authorities for names and taxa for better reuse*

GBIF is a partner in CoL+, together with partners of the existing CoL.

The Encyclopedia of Life (EoL) and the Biodiversity Heritage Library (BHL) deliver today on OECD expectations other than taxonomy and occurrences data, namely by EoL holding traits of 11 million species, images, articles and bibliography on species[49], while "the BHL portal provides free access to hundreds of thousands of volumes, comprising over 56 million pages (of an estimated[50] 500 million), from the 15th-21st centuries."[51].

Still another organization that focuses on providing taxonomic data and literature is Plazi, located in Bern, CH. As a comparatively new player in the field, Plazi[52] attempts the automatic extraction of taxonomic treatments and other data particularly from current literature. This is not only a technical and scientific challenge, but many times a legal or economic one as well, as much of the literature resides behind paywalls in a proprietary setting.

A number of other "content areas", not specified in the OECD diagram, are addressed by noteworthy, well maintained databases, such as TRY[53], which holds plant traits. Although not properly an infrastructure, it is maintained by a consortium with clear rules and missions and supported by a membership of scientists working in the field.

---

[47] (http://www.catalogueoflife.org/content/contributors)

[48] https://github.com/Sp2000/colplus

[49] https://eol.org/docs/what-is-eol

[50] See p. presentation by M.R. Kalfatovic, BHL, 2010, https://de.slideshare.net/Kalfatovic/3-years-on-the-biodiversity-heritage-library

[51] https://about.biodiversitylibrary.org

[52] https://www.plazi.org

[53] https://www.try-db.org/

The field of genetic information, which can be linked to species and sometimes even is provided with geolocation data, is of rapidly growing interest. Two organizations holding such data are

Barcode of Life Data Systems (BOLD) with barcodes on more than 300,000 species[54], and the European Nucleotide Archive (ENA) – one of three in the International Nucleotide Sequence Database Collaboration (INSDC)[55]. These Barcoding consortia have been GBIF Affiliate Partners since 2005, and GBIF is working in projects[56] with the European Bioinformatics Institute (EMBL-EBI), host of the ENA, and others to mediate such data as well. This has already resulted in 18 million occurrence records[57].

OBIS was an Associate Participant from the earliest days of GBIF in 2001, and since 2014 has been an Affiliate through the Intergovernmental Oceanographic Commission of UNESCO. Its sphere of action is similar to GBIF's, in that it serves more than 56 million occurrence records on more than 125,000 species, but only from the marine environment, with an international hub and thematic or regional nodes and publishers. OBIS appears to publish[58] just about 10% of its occurrences to GBIF – the reason for which is not clear, but some of it may be due to the fact that some OBIS nodes publish directly to GBIF while for others the international hub does the necessary steps. It is funded by diverse sponsors[59], although some basic funds stem from the UNESCO International Oceanographic Data and Information Exchange (IODE). OBIS originates from the Census of Marine Life (2000-2010)[60], a project funded by the Alfred P. Sloan Foundation and later supported by many others[61].

Information Infrastructures

Information infrastructures are used to deliver data products and information at least one level higher in aggregation than occurrence data. Such products could be species distributions or those kinds of indicators required by the CBD and others at the science-policy interface. Going by the concepts of OECD and GBIO, these infrastructures should be using taxonomy databases such as CoL and occurrence data from GBIF or OBIS to arrive at such products.

Much sought after are visualizations of datasets showing the species distribution or range, either globally or at a geographic scope, tailored to the needs of users. Two projects producing such products are the Map of Life[62] and the Atlas of Living Australia[63], now being rolled out to

---

54 http://www.boldsystems.org

55 http://www.insdc.org

56 https://www.gbif.org/news/6ewyUhBpRYammYWI2CgsM4/biodiversity-infrastructures-to-crosslink-metagenomics-and-species-occurrence-data

57 https://www.gbif.org/dataset/ad43e954-dd79-4986-ae34-9ccdbd8bf568 and https://www.gbif.org/publisher/ab733144-7043-4e88-bd4f-fca7bf858880

58 https://www.gbif.org/network/2b7c7b4f-4d4f-40d3-94de-c28b6fa054a6

59 https://obis.org/about/sponsor/

60 https://obis.org/about/

61 https://www.comlsecretariat.org/about/partners-and-sponsors/

62 https://mol.org

63 https://ala.org.au

many countries, mediated through the GBIF portal[64]. Such portals – and there are many more, especially based on specific data collections[65] – typically draw on a number of data types beyond occurrence records, such as species checklists or expert range maps.

GEO BON, a Partner of GBIF, is a part of GEO, The Group on Earth Observations, and has the mission to be "a global biodiversity observation network that contributes to effective management policies for the world's biodiversity and ecosystem services."[66]. As noted, in 2012[67], CBD "*Invites* the Group on Earth Observation Biodiversity Observation Network (GEO–BON) to continue its work on the identification of essential biodiversity variables and the development of associated data sets".

The UN Environment World Conservation Monitoring Centre (UNEP-WCMP)[68], a GBIF Associate Participant, focuses on data analysis and products with a policy impact. One of its original databases is the World Database on Protected Areas (WDPA). UNEP-WCMP is the official Secretariat of the Biodiversity Indicators Partnership (BIP), mandated[69] by CBD to support the development of indicators of biodiversity change.

Beyond these information infrastructures with an international scope, there is a multitude of similar organizations with a national or regional scope, such as the Mexican CONABIO (already mentioned in the section on science-policy interface) or the European Research Infrastructure LifeWatch[70].

Data Infrastructures as Organizational Entities

It must be noted that actual organizations can assume roles in more than one of the previously described categories and these roles may not be static. In the context of this review, it is important to note that many of the larger organizations, such as IUCN, UNEP-WCMC or CGIAR, all much larger organizations than GBIF, have built and maintained their own database systems and curated large datasets to their specifications. In contrast to many smaller databases that began and ended as projects, these databases – which contain a significant volume of data of interest to GBIF and its users – will probably not fold or merge with existing GBIF publishers.

Actors in Setting Standards

Many, if not most, of the organizations described in this section are involved to some degree in developing and maintaining standards for data, be it the definition of computable indicators, ways to describe species, or to store and transmit such data. Some of the former are already highlighted in previous sections, including GBIF's role in these activities.

---

[64] https://living-atlases.gbif.org

[65] E.g., https://data.nhm.ac.uk/dataset/collection-specimens/resource/

[66] https://geobon.org/about/vision-goals/

[67] UNEP/CBD/COP/DEC/XI/3, 5 December 2012.

[68] https://www.unep-wcmc.org

[69] CBD/COP/14/INF/40, 12 November 2018.

[70] https://www.lifewatch.eu/

However, probably the most influential group in the domain of standards applied by GBIF has been from the beginning and still is the Biodiversity Information Standards (TDWG) – "historically known as the Taxonomic Databases Working Group"[71]. It is a membership organization of expert individuals and organizations with modest dues. It has developed and maintains standards such as the Darwin Core, which is the foundation of the GBIF data mediation technology. Staff of GBIF are members of the Darwin Core Maintenance Group and have contributed, for example, to the DwC-Archive standard[72]. TDWG has been a GBIF Associate Participant since 2002.

GBIF technologies are certainly influenced by a range of more generic standards, such as those for identifiers (the DOI – Digital Object Identifier, and the IGSN – International Geo Sample Number[73]) and for certification of repositories, such as the CoreTrustSeal[74], as developed by World Data System (WDS) of the International Science Council (ISC).

2.3. GBIF's Position in the Global Biodiversity Informatics Landscape

In 2002, the representatives of the first countries that had signed the GBIF Memorandum of Understanding (MoU) decided on a work program, which limited the remit of GBIF considerably, compared with the OECD concept. This was certainly due to the level of the budget they had agreed upon, but also was due to the realization that in the meantime new players had entered the field. Species2000 and ITIS were set to build the "Catalogue of Names of Known Organisms" and had already become Partners of GBIF. The funding of GBIF activities for the "Catalogue of Names", "Species Bank", and "Literature Resources" had been dropped, *de facto*, while a strong role in capacity building, as seen by the OECD, had been confirmed (and upheld, since, see Chapter 6).

The strong role of GBIF in the area of standards, foreseen by OECD and in the first work program, developed over a long period, so that today GBIF can be seen as the promoter of global implementation of standards developed in the "community", in particular as represented in TDWG. Based on such standards, to which the GBIF Secretariat contributed, GBIF switched, around 2008, from the originally envisioned architecture of a loosely coupled federation of databases with distributed searching to the harvesting model and search in a central cache.

It was in the same timeframe that GBIF – based on decisions of its governance bodies – focused its informatics resources on occurrence data.

Today, as has been described in the previous section and will be fleshed out in the following chapters, GBIF inhabits globally visible roles in:

- mobilizing data from national and institutional collections and databases, and mediating a unified access to these,
- capacity development, particularly of the staff of a global network of national Nodes and data publishers,

---

[71] https://www.tdwg.org

[72] http://rs.tdwg.org/dwc/terms/guides/text/

[73] http://www.igsn.org

[74] https://www.coretrustseal.org

- the further development of standards to enable and ease interoperability between biodiversity informatics databases, and

- the creation of a taxonomic backbone for biodiversity informatics, as it sets out to merge its own taxonomic backbone with the CoL+ checklist.

By partnering with the Atlas of Living Australia and forming with them a "Living Atlases" community[75], GBIF is exploring the domain of data products and services at a higher aggregation level. Here it appears to encounter competition from other organizations, such as the Map of Life project[76] (which is, however, a GBIF Partner). Whether or not this and other competitive situations are healthy or should be avoided – either by partnering or yielding areas of activities to each other – needs to be found out on a case-by-case basis, but preferably in a comprehensive community setting (such as the emerging Alliance of Biodiversity Knowledge, see section 5.1.4).

As could be observed in a condensed form at the 2019 *BiodiversityNext* conference[77], there is disruptive potential in a number of developments, which may change the landscape of biodiversity data actors within a few years, perhaps decisively. For example:

- Relatively new scientific methods, such as meta-barcoding, may enable mapping of biodiversity at a much finer resolution and with fewer gaps if pursued uniformly and globally at industrial scale, as P. Hebert, iBOL, advocated in his keynote.

- Likewise, traditional naturalists' observations are being transformed to a new status, being empowered by social-networking paradigms and technologies (and supported systematically by science professionals), as shown in the keynote presentation of C. Seltzer, iNaturalist.

- New informatics technologies, such as machine learning, and the pervasive access to the internet, enable huge "productivity" gains, such as when dealing with the noisy output of camera traps[78], especially when deployed as a globally accessible infrastructure[79].

- Informatics tasks at classical infrastructures, such as the digitization of collection specimens and literature, may need to and appear to be on the way to adopt these modes of industrial and global-scale implementation to retain their relevance.

GBIF will most probably in the near future be affected by these developments. While it is already actively involved in contributing to many of them, it may need to scale up these activities in order to stay relevant.

---

[75] https://living-atlases.gbif.org

[76] https://mol.org

[77] https://biodiversitynext.org/

[78] https://doi.org/10.3897/biss.3.38233

[79] https://www.wildlifeinsights.org

3 USER AND CONTRIBUTOR PERSPECTIVES[80]

As a global biodiversity information infrastructure, GBIF must first of all meet the requirements and serve the purposes of "users" – e.g., just a "dark" archive of data, even if unbiased and representative (see Chapter 5), would not provide much benefit. Then again, a data infrastructure with the best functionality, but a less than representative amount of data is unsuitable, or only meets just a few purposes. Therefore, GBIF must also meet the requirements of "contributors".

In this chapter we explore the broad range of requirements, purposes, interests, and motivations of users and contributors of data, and the constraints guiding their work and their modus operandi, before these issues become conflated in the chapters to follow. Individual issues can combine, e.g., aspects of metrics (Chapter 4), technical issues of data citation (Chapter 5), creating community agreement on citation and the balance of interests (Chapter 6), and convincing Governing Board members to report back positively on openness (Chapter 8).

3.1. Description

There are many different users of biodiversity data, as well as contributors of the data that are made available by GBIF. We have already identified the landscape of these players in the preceding chapter. Users and contributors range from academia to government and industry, applications from pure science to policy, conservation, agriculture and health. In short, user types obviously have different requirements and concerns about the data being made available by GBIF, particularly regarding "fitness for purpose". Similarly, contributors can contribute data to GBIF that are fit for some purposes, but not others. These concerns are shaped by the missions of the institutions as well as by the personal agendas of the acting individuals.

In an era of valuing efficiency and effectiveness almost above all else (even above the quality of results) we believe there are two overarching issues for users and for contributors/publishers of data:

- The **effort** to get at, prepare, validate, submit and make use of data must be as low as possible. This requirement has been partly covered already in the OECD report (OECD 1999, p.15) as "difficult to discover, access, and use biodiversity data … synthesize information from disparate sources" and is most frequently associated with matters such as usefulness and usability of data repositories and portals. In daily practice, for some user types that are important with respect to societal impact, the requirements on data itself boil down to an immediate usability of a dataset, including the "fitness for use", commonly called quality, related to that type's use case. Both the requirements and their implementation by GBIF are addressed in Chapter 5 in detail.

- Data contributors expect that their efforts towards GBIF contribute to their "bottom line" in terms of recognition and in the end, financial rewards. Since this issue is of paramount importance to their long-term commitment, a separate chapter is dedicated to impact and relevant metrics.

---

[80] This and the following chapter on metrics do not address a specific item of the statement of task. They are deemed necessary by the review team, however, to illuminate the "social dynamics" of relationships between GBIF and its stakeholders, including those enumerated and exemplified in Chapter 2. This dynamics drives many of the technical and organizational decisions and measures described in the following chapters.

It must be noted that, in most cases, individual scientists as contributors of data cannot expend more effort on data collection, accuracy, and richness of description (or other quality features) than is strictly required for their research question at hand and for their next publication.

While many government agencies may apply stricter and more consistent quality control than academic institutions (as they may be legally responsible for the validity of their data), they typically collect data only within a very restricted taxonomic and geographic scope (e.g., for purposes of national reporting on a limited list of environmental parameters).

These considerations broadly determine which qualities of a data infrastructure and its data are expected by users and which data are potentially contributed, and the constraints both sides work under. But there is another broad issue, which – under these constraints – potentially widens the gap between the interests of user and contributor groups, namely that of liberal access to data. Scientific and other users of data will certainly appreciate data that are openly accessible and not burdened with restrictive usage licenses. This is an important element of ease of (re)use. When aggregating many datasets from different contributors, it might otherwise be the predominant effort to negotiate licensing and the elapsed time to publication might become unacceptable – to the point of abandoning a research question. While some aspects of Open Access are discussed from the user and contributor perspective in the following paragraphs, other aspects are discussed where appropriate, throughout the report.

In their role as potential scientific or governmental data contributors, it may be the same individuals or institutions that would like to see their interests protected. In the case of governmental agencies (and definitely of commercial entities) this may result in not making data available at all, while scientists could be enticed by citations or co-authorships. Even among scientists, there have been experiments with licensing and contracts to enforce what is today "good scientific practice", required by such actors as science funders, as in a German example[81], national academies, or similarly potent voices.

It is only over the last decade or so that the Open Science and Open Data paradigms have begun to be widely included in accepted rules of scientific conduct[82] (down to the rules of data citation and data policies of journals, see, e.g., the data policy of the journal Nature[83]). Many governments and funders have recognized that open access to publicly funded data provides the best returns for researchers and society in general[84]. Most recently, the concept of FAIR (Findable, Accessible, Interoperable, and Reusable) data has been promoted (Wilkinson et al. 2016) for both fully open and partially restricted datasets.

This process, however, has progressed unevenly across disciplines and even within disciplines. Notable examples have included the Bermuda Principles on the rapid and public release of DNA sequence data in 1996 and the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) commitments rolled out in the last few years among Earth science

---

[81] https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/index.html

[82] See, for example, The Beijing Declaration on Research Data (2019), available at: https://zenodo.org/record/3552330, or Open Data in a Big Data World (2017), available at: https://council.science/cms/2017/04/open-data-in-big-data-world_long.pdf. A compendium of over 800 such statements in this decade is available at: http://tagteam.harvard.edu/hubs/oatp/tag/oa.declarations

[83] https://www.nature.com/documents/nr-data-availability-statements-data-citations.pdf

[84] For a compendium of reports that researched the benefits of such open data policies, especially in the environmental context, see the bibliography in The Value of Open Data Sharing (2015), Group on Earth Observations, available at: http://www.earthobservations.org/documents/dsp/20151130_the_value_of_open_data_sharing.pdf

publishers, infrastructures, and communities (see, e.g., the data policies[85],[86] of Copernicus, the publisher for journals of EGU, the European Geosciences Union, and AGU, the American Geophysical Union). Some biologists have been slower than others to embrace the open sharing of data for a number of reasons but may be on the brink of change[87].

Many publishers of data have interests quite similar to the contributors themselves. Museums and other collections of specimens consider themselves as contributors of data – partly because they do the digitization and partly because of their role as stewards of the specimen. In general, publishers want to see their role in adding value to be visible, a requirement shared by GBIF itself.

Indeed, GBIF has been a leading proponent of open access to data since the organization's inception and it has largely succeeded in convincing its Publishers and other data contributors of the open access model. Nevertheless, many hurdles remain. Most countries still do not have national policies for open access to public data generally or research data specifically, including biodiversity data. There are numerous reasons why researchers and others do not provide open access to their (biodiversity) data or make them available at all (Borgman 2015). Another complicating factor has been the controversy in recent years—still unresolved—in the CBD whether Digital Sequence Information and related biodiversity data are within the scope of coverage within the treaty regime and its various restrictions[88].

In the case of "citizen scientists" as observers in the field, one or more other intermediary layers come into play. One interviewee who leads a long-term citizen science project noted that while these – quite qualified – voluntary observers may make up for 99% of the personal effort, the 1% doing the coordination and quality-related work may need to derive their scientific recognition from these volunteer data. Therefore, these professionals may be reluctant to publish the data beyond closely organized circles of trust, or only after a long period of exclusive use. When volunteers claim (moral) ownership of data this will be better justified than in the case of tax-funded scientists, and so they might, for example, reject uncompensated commercial uses of their data.

It is this environment of interests or even requirements on the observation of – perceived – rights that leads to a meticulous choice of words by GBIF: You shall not speak of "GBIF data", but only of "GBIF-mediated data" so that not GBIF but the original publishers of data appear as such in the credits. In summary, a data publisher and even more so an intermediary such as GBIF needs to walk a fine line between the requirements of users and contributors of data.

---

[85] https://publications.copernicus.org/services/data_policy.html

[86] https://publications.agu.org/author-resource-center/publication-policies/data-policy/

[87] See, for example, the Bouchout Declaration in 2014 for Open Biodiversity Knowledge Management http://www.bouchoutdeclaration.org/ signed by major European and international institutions.

[88] See, https://www.cbd.int/abs/dsi-gr/2019-2020/submissions.shtml

3.2. SWOT Compilation[89]

3.2.1 Strengths

- GBIF has "mediated" an astounding amount of occurrence data (almost 1.4 billion occurrences) from a wide taxonomic and geographic range. It seems to be on track to explosive growth, which is unparalleled by many other infrastructures. For many applications, it has therefore become *the* first stop - and perhaps the only one - that users may need.

> ***Box 3.1: Gerald "Stinger" Guala, Ph.D.; Branch Chief, Eco-Science Synthesis; Director of Biodiversity Information Serving Our Nation (BISON); Director of the Integrated Taxonomic Information System (ITIS); U.S. Geological Survey***
>
> *Early in my academic career, it took me two years to gather the data needed for my dissertation research—an activity that can now be done in two days with GBIF.*

- GBIF-mediated data have been used in a broad range of scientific and other use cases of societal importance (agriculture, health, conservation, and others), which has demonstrated the versatility of GBIF's services.

- GBIF has implemented the use of Digital Object Identifiers (DOIs) in a particularly innovative way to uniquely identify which data have been used in a publication, and tools enabling publishers to track the usage of their data and link the data and themselves to applications of scientific or societal impact.

- All of these data are provided under one of three liberal licensing terms, each within the Creative Commons family of licenses. This is a quite recent development (formal policy since 2016[90]) and GBIF therefore has had and continues to have an important role in the acceptance of open access to biodiversity information.

- GBIF provides relatively easy to use interactive interfaces and tools for users and providers. GBIF even provides training to use their systems and proper biodiversity data practices, where this cannot be provided by national education and research systems.

3.2.2 Weaknesses

- In most cases, data from the GBIF portal is not immediately fit for various applications, such as for scientific or other uses. Careful selection and "data cleaning" by users *is* necessary.

- Scientists acknowledge that the responsibility for quality does not lie with GBIF alone, but there is currently no easy ***and*** effective way for users to give feedback on data, so

---

[89] This section provides the summarized, but also individual SWOTs of GBIF, as provided by the interviewees. It is not an analysis and attempts to convey the original language. For an explanation and rationale of this format, see section 1.5.

[90] https://www.gbif.org/news/82812/licensing-milestone-for-data-access-in-gbiforg

data cleaning efforts do not need to be repeated time and again, as multiple interviewees noted.

- There are data contributors or potential contributors who still have technical, "cultural", or other difficulties making their data available. The technical ones may hinge on capacity bottlenecks of national Nodes, the cultural ones on the slow-moving process of implementing the Open Science paradigm.

- It is not easy for individual contributors to discover their impact metric on the GBIF portal.

- It is still difficult for users to link occurrence data from GBIF to other types of data, such as trait, genomics, and environmental data and it also is hard for data publishers to keep such links intact while publishing the occurrences as part of their data for GBIF.

3.2.3 Opportunities

- There is so much positive feedback from users on GBIF as a useful and reliable infrastructure that this momentum could be converted into quality improvement (see Chapter 5 on technical mechanisms for providing feedback).

- GBIF could "republish" or act as repository for quality controlled, ready to use primary data, reference datasets, or (standard) data products. The underlying processes to determine and assert fitness or quality would have to be fully under the control of the relevant stakeholder communities.

- There are still vast amounts of data to be mediated.

- The preservation of "linking" of occurrence to other types of data is largely a technology and standards issue, which we deal with in Chapter 5. GBIF has built a reputation of successfully implementing and developing standards and technologies about occurrences and linking them to taxonomic databases. Many interviewed users and other biodiversity informatics stakeholders expect GBIF to be able to contribute decisively to the solving of the broader linking issues.

3.2.4 Threats

From the users' and contributors' *abstract requirements* point of view, there was no recognizable immediate threat. See however some relevant associated points about openness of "Digital Sequence Information" under the Convention on Biodiversity, discussed in section 3.1.

3.3. Conclusions

From the user perspective, GBIF may not be ideal, but it is by far the best available general and non-specialized source of global occurrence data, as evidenced by the whole corpus of interviews. The prevailing analysis seems to be: The best approach is to work on the improvement of GBIF and it would be a waste of resources to build a competing system.

This generally positive attitude towards GBIF could be leveraged to address the data quality issue. The existing plans (in the Work Program 2020) of GBIF to implement a feedback path from users to publishers and contributors, see Chapter 5, would need to be implemented and the communities be made well aware of it. It is quite clear that this feedback mechanism would necessitate a kind of social network at the GBIF portal, since contributions of users to quality (such as annotations, suggestions for improvement, flagging of errors) should be acknowledged

and contribute to the reputation of every such person, both as an inducement to perform such work and as a moral right. (For examples, see the WoRMS list of editors[91] or the more differentiated iNaturalist list of contributors[92].)

On the type of data on which to focus, "data products" cannot simply be disregarded as a topic of GBIF – they are clearly sought after by many types of users – and they expect them at the GBIF portal. This kind of contribution can and should be "fostered through community efforts, partnerships, contracts, or outsourcing" (GBIF 2010a). The selection and annotation of such products needs to be a community driven, probably editorial-style process, while it would fall to GBIF to at least provide the infrastructure, with assistance from other organizations that develop biodiversity infrastructures.

Many users from different stakeholder groups expect GBIF to go even a major step further, GBIF could hold analytical code and provide an execution environment for it (also known as a Virtual Research Environment or Science Gateway). This requirement would parallel the thinking emerging in other disciplines: If the quality of the code were provided through similar processes as for quality data, this would provide a giant step not only to ease and speed-of-use but also towards reproducibility of results.

Compared to academic studies, researchers working on many utilitarian use cases are even more restricted as to available capacity or are allowed even less time to a deadline – they might even be emergency driven, such as in tracking disease vectors. Consequently, they can even less afford their own meticulous analysis or finishing of data. Thus, community-endorsed reference data or even processing will be needed for a broad and important range of users. The unresolved question is: Who will provide the platform or platforms for these products and services?

Interviews revealed a large number of suggestions, if not expectations, from users regarding types of biodiversity data other than occurrences to be provided by or to be linked to by GBIF. The implications on the systems and the effort required are so unforeseeable at this point, that a meticulous process of evaluation of priorities will be essential. In order to manage user expectations and to properly balance cost and benefit, this process would need to *concurrently* involve both a broad representation of user communities and of other data infrastructures besides GBIF. This would probably be a worthy topic for the "Alliance for Biodiversity Knowledge" (which we discuss in section 5.1.4) and GBIF appears on a course to use the Alliance in this way. Conceptually, it is this body that could speak authoritatively not just about priorities but also about issues such as which institution should be the long-term steward of which data, standards, and tools, and systems implementing them.

Based on a few interviews it must be suspected that there are still institutions holding large amounts of occurrence data which, while willing to share the data, lack the capacity to implement the GBIF Integrated Publishing Toolkit (IPT), and their underfunded national Node lacks the capacity to support them. In the absence of sufficient perceived benefits - or help from another side than "their" Node - such institutions will need to let the publishing issue languish, since the corresponding GBIF web page[93] does not offer an alternative.

15 years ago, the reviewers of GBIF still had to urge that:

---

[91] http://www.marinespecies.org/aphia.php?p=editors

[92] https://www.inaturalist.org/people

[93] https://www.gbif.org/publishing-data

> *GBIF needs to be much more proactive about explaining and promoting its data policy to its Participants, data providers, organizational partners, and users. GBIF cannot assume that all, or even most, of its potential data providers subscribe to the free and open access ethic. … (CODATA 2005, p.vii)*

Today we can conclude that GBIF has largely succeeded in convincing its existing data publishers and contributors of the open access model. When GBIF took the leap to accept only liberal licenses (e.g., preferably CC0) on datasets in 2016, only a very few data were retracted from GBIF-mediation (as shown in detail in 5.1.1). This has only been achieved by long-term work in convincing all contributors of the low risks and high gains of such an approach. GBIF has been able to gain the trust of many data publishers through enforcing rules for an attribution policy and greater recognition of its data providers and their original data sources.

However, not all data contributors and publishers, or members of the GBIF-community for that matter, are still fully aware of GBIF's methods to track attribution and to measure other impacts, and few institutions seem to have made use of it fully. But still, there are many more requirements, some of them reasonable, for additional, complementary tools addressing this issue.

It is probably unknowable how much *valuable* data is "out there", in desk drawers, on USB-sticks, or institutional servers, but not shared with the world. GBIF is in a position to work with the scientific community and governments as they slowly change their culture and their policies. It can do this by showing off the success stories and benefits of open access to biodiversity data and convincing contributors that their interests will be protected even in an open environment.

Interestingly, although explicitly granting embargo periods has played an important role in many disciplines, this was not a topic in interviews about GBIF. In the long-term, however, most impact on scientific culture – and the timing of data release – may be due to changes in the publication culture, such as journals requiring access to data underlying an article. This practice has begun to be taken up by journals in the field of biodiversity as well. In contrast, governments needed to hear about monetary benefits or lives saved, before they decided to grant open access, e.g., to data from governmental weather services or to satellite imagery. One can see the money and lives arguments in high-level reports such as the pollinator report of the IPBES (IPBES 2016), and also clear indications about gaps in data – unfortunately not the call for more *open* data.

4 GBIF/BIODIVERSITY INFORMATION IMPACT AND ITS METRICS[94]

4.1. Description

It is essential for all organizations, particularly publicly funded ones, to convey their value propositions in a clear and compelling manner to their funders and other stakeholders. How does GBIF – the Secretariat and the national Nodes – benefit the missions of funders, private citizens, researchers, NGOs, trainers, data publishers, other infrastructure providers, and various types of users? A value proposition is most clearly or easily understood when it is expressed as having a **positive impact** on the interests or mission of the stakeholder, and if there is a **metric** which is indicative of that effect.

The 1999 OECD Report expected GBIF to have an impact on a wide range of beneficiaries. The areas in which they expected benefits from this endeavor included:

- The advancement of science;
- Increased efficiency in R&D spending; and
- A variety of applications including biodiversity conservation, agriculture, health, industry (such as commercial products and informatics tools), the compatibility of protecting ecology and economic development, responsible resource management, and sustainable development.

The evidence and analysis can be skewed if an indicator or a metric is chosen just because it is easy to apply. This is obviously the case in science, where article- and citation-counting is common. However, any serious discussion about impact should start with a qualitative discussion of the outputs envisioned and actually delivered, typically using "success stories", particularly as it is claimed that "information [is] 22 times more memorable in narrative form"[95]. Such anecdotes will also show which metric, if any, is appropriate to describe the benefits or the development of impacts over time.

Furthermore, probably all impacts that GBIF may have – beyond supporting individual scientific articles of unknown long-term importance – belong to chains of value creation.

The top level-value chain that we have identified is:

Biosphere » Biodiversity » Impact on society

For example, as implied by the IPBES report on pollinators (IPBES 2016), conservation of biodiversity has a positive impact on society. Thus, the knowledge about changes in biodiversity – and their effect on society – supposedly would allow either to avoid (further) biodiversity losses or to mitigate the impact of the losses.

---

[94] This and the previous chapter do not address a specific item of the statement of task, but they are deemed necessary by the review team to illuminate the "social dynamics" of relationships between GBIF and its stakeholders. This dynamic drives many of the technical and organizational decisions and measures described in the following chapters.

[95] https://dx.doi.org/10.1038/d41586-019-03084-4

Stories at this level, if from enough important facets of our existence, would prove that it is worthwhile to know ever more about our planet – including its biosphere.

When that is established, we need to lay out the value chain of knowledge in detail:

Data » Information » Knowledge » Impact on a sector or application

This value chain is implied by the GBIO framework already (see Figure 2.1B) and one could identify GBIF's and other stakeholders' actual and expected roles from it as well. If this chain and the stakeholders' roles would be made explicit, it would become possible to discuss any numerical measure of contribution to societal benefit.

Many stakeholders, especially funders, need or require easily usable evidence to justify funding. This could be achieved, for example, through success stories about the particular topics of interest or through a numerical metric compatible with other proposals the stakeholder handles, such as the number of citations for a science funder or monetary value for an agency of economic development. In other cases, it may be necessary to lay out the whole value chain through some easily understandable examples.

When considering a long-term commitment to a specific metric, GBIF and other players in biodiversity informatics must observe how their role and impact on value creation may change. Metrics or indicators that are too simple will probably reveal non-trivial developments only after the fact. For example, working backwards from the desired impact, what measures or indicators of biodiversity and, consequently, which data sources are actually used in knowledge creation, forming of policy, activities of public health agencies or the agricultural sector, and the like?

4.1.1 Success Stories

There are many examples of the use of GBIF–mediated data that are immediately relevant to some applications. For example, health and agricultural issues are presented on the GBIF website as a "featured data use" under the "Inside GBIF/News-Outreach" menu, and in the yearly "Science Reviews" by GBIF, such as the current one (GBIF 2019). Many of those already could serve as stand-alone success stories.

"Flying foxes predict Nipah virus transmission risk" (GBIF 2019, p.6) is one such example. As one of the authors stated there:

> *"We were in the process of submitting the final edits when the outbreak in Kerala took place. On the one hand, it's very exciting to see that your models can make accurate predictions, but knowing the consequences of these outbreaks puts an immediate dampener on your excitement."*

"Vulnerability to snakebite envenoming: a global mapping of hotspots" (GBIF 2019, p.8) is another one. The lead author of that article told us that he considers GBIF as the most comprehensive source of biodiversity data for modelling in health research. In their article (Longbottom et al. 2018), authors make the case that:

> *Snakebite envenoming is a frequently overlooked cause of mortality and morbidity, responsible for 81 000–138 000 deaths annually, and between 421,000 and 1.2 million envenomings. Contact from venomous snakes, spiders, and scorpions contribute to 1.2 million years of life lived with disability annually.*

The combination of agriculture and conservation is a recurring theme, which obviously benefits from using GBIF resources. "Conserving genetic diversity of crops in West Africa"[96] and "Identifying gaps to prioritize areas for crop wild relatives' conservation"[97], are both a "featured data use". The article, "Comprehensiveness of conservation of useful wild plants: An operational indicator for biodiversity and sustainable development targets" (GBIF 2019, p.36), is also worth a data story in the GBIF Science Review 2019. In this latter case, the authors did not only use occurrence data but also "GBIF Backbone Taxonomy, using the GBIF Species Lookup Tool … and the GBIF Species API v1", which should have been handy when working with 63 million records on 7000 taxa.

The situation is quite different with highly aggregated, global, all-encompassing assessments and indicators. As Tim Hirsch, deputy director of GBIF, explained:

> *when you look at the sort of indicators that are presented to the public and to the policymakers, they'll see [the] IUCN Red List. They won't necessarily see the dependency on having the underlying data. That's something which we're working on with IUCN as in [other] partnerships to improve.*

Some of the currently most visible outputs of science, at the science-policy-interface, namely the IPCC special report on 1.5°C Global Warming (IPCC 2018) and the IPBES draft report (IPBES 2019), do not disclose data use directly but only through the articles used in the assessment. Sometimes not one but many layers of referencing have to be traversed. This is unavoidable, but it impedes GBIF in presenting its role in these reports in a timely manner, not two years after publication.

The IPCC special report, in an article underlying its chapter on "Impacts of 1.5°C Global Warming on Natural and Human Systems", used 385 million occurrence records from GBIF as its main source, as analyzed by the Secretariat (GBIF 2019, p.28) one year after the IPCC report's publication. In comparison, one of the report's authors, without the benefit of special skills and tools, spent a few hours or so of digging into the references and arrived only at the margins of this fact (see Annex 10.2 of this report).

> ***Box 4.1: Joe Miller, Director of GBIF, on citation tracking and the role of DOIs:***
>
> *GBIF is not mentioned in the IPCC paper. Since it relies on Warren et al. which we can track, we can identify the thousands of datasets and the 385 million occurrences. In the IPBES paper the GBIF data (and other data) has gone through the secondary filter. While it will be harder to track the data back to explicit GBIF dataset and occurrences, we have started preliminary investigation of the IPBES report and have found several likely 'secondary' citations of papers that depended on GBIF-mediated data. (Basically [we] cross referenced every citation in the report to our database of papers that have cited our DOIs and found a couple of dozen candidates for further review to see if that GBIF cited data can be justifiably linked to the report).*
>
> *I think the two reports indicate that we have made great progress in associating the data to the reports but have a long way to go too. First, even*

[96] https://www.gbif.org/data-use/82511/conserving-genetic-diversity-of-crops-in-west-africa

[97] https://www.gbif.org/data-use/59OZhJazMkYisCUg8Y2aks/identifying-gaps-to-prioritize-areas-for-crop-wild-relative-conservation

> *without the cited DOIs we can connect our data to the uses and with that we can show the value of how our DOI approach will work. It is still a manual process. The difficulty will be getting people—not just authors, but journal publishers, editors and reviewers—to buy into the DOI system and use it so that the data will be tracked. […] The DOI portion of [references to GBIF] needs to grow faster and that will be an important outreach component for us […]. Secondly, it also is helpful that this same system can provide [an incentive]. If individual users use their ORCID then attribution can go to the individual not only the dataset of an institution (and eventually also to funder recognition). We need to develop a full-on strategy to get this system in place and communicated to our audiences.*

The IPBES report, not fully analyzed by the Secretariat at the time of this writing, mentions in Chapter 2.2, p.12 about sources that "in conjunction with data aggregators and repositories, such as GBIF (www.gbif.org), OBIS (www.obis.org) and Genbank (Benson et al. 2013), make hundreds of millions of species occurrence records and gene sequences freely available." While clearly stating the importance of GBIF, this sentence does not convey an absolute measure (such as number of records used) or an indication of the share (such as "most of X depend on Z data").

4.1.2 Citation of GBIF-mediated data use

We believe that it is currently beyond reasonable direct influence of GBIF whether or not scientists properly cite data, acknowledge the use of GBIF-mediated data, or at least mention GBIF in the body of the text of an article. The practice of actually citing data in references has gained acceptance in many disciplines only over the last few years, and GBIF could try to help accelerate this process by working with other data infrastructures and journal publishers, as in the COPDESS example in section 3.1.

In those cases where GBIF-minted DOIs have been used, analysis becomes easy. In all other cases, analysis of text searches for "GBIF" involves human effort (the reading and interpretation of articles). The GBIF Secretariat does this at considerable expense, particularly by flagging those articles for the statistic that explicitly describe use of GBIF-mediated data. Although the share of actual, DOI-based citations has grown quickly, perhaps not least due to the fact that GBIF has made it very easy now to identify and cite a selection of data, it currently stands at only 20% of all data uses identified (see Chapter 5).

The *2010 Forward Look* report (GBIF2010a) suggested among the "GBIF targets and goals over the coming 5 years (2011-2015)" a very ambitious "50% yearly increase in citation rates of GBIF data usage in the scientific literature". Actually, articles with references to data obtained through GBIF grew from 147 citations in 2010 to 622 in 2015 (instead of 1116, which would have been 50%, applied five times). But only a year later, the growth from 2011 to 2016 was 185 citations to 2447 – surpassing the goal of 1405 by a significant amount.

Perhaps of even more relevance to the OECD expectations than the mere aggregate numbers is the distribution of use over different disciplines, identified by citations, as analyzed by the GBIF Secretariat in Table 4.1, below:

Agriculture	236	Ecosystem services	38
Biodiversity science	151	Evolution	643
Biogeography	286	Freshwater	39
Citizen science	53	Human health	212
Climate change	509	Invasives	584
Conservation	567	Marine	68
Data management	149	Phylogenetics	269
Data paper	29	Species distributions	150
Ecology	969	Taxonomy	460

Table 4.1: GBIF-mediated data used in peer reviewed journal articles, from the year 2000 until mid-2019 https://www.gbif.org/resource/search?contentType=literature&year=2000,2019&literatureType=journal&relevance=GBIF_USED&relevance=GBIF_CITED&relevance=GBIF_PRIMARY&peerReview=true

It is also possible to search which literature cited a specific dataset, as for example by applying this kind of URL: https://www.gbif.org/resource/search?contentType=literature&gbifDatasetKey=8a863029-f435-446a-821e-275f4f641165 on the datasets provided by the Dutch Publisher "Observation.org", which, among others, turns up a "Bulk IUCN fungal assessment" that uses a (DOI-referenced) collection of 424 datasets. It does not seem possible to find contributions from individuals, as long as these are not expressed as their "own" datasets, at least not at the time of writing (see comment by J. Miller, Box 4.1.). It appears even less possible to name or extract different kinds of contributions, such as those envisioned in the CASRAI "CRediT – Contributor Roles Taxonomy"[98].

In Box 4.2, D. Schigel indirectly makes the case for data citation tracking, not for GBIF's own purposes but for that of data producers – which they need in order to become or remain to be *apparent* as data contributors, as explained in Chapter 3.

Box 4.2: Dmitry Schigel, Scientific Officer at the GBIF Secretariat, on the "value proposition":

GBIF provides an opportunity to diversify the academic portfolio of people and organizations. Museums stop being seen just as physical archives, they start becoming digital data centers, centers of teaching, centers of information access and so on, which protects them, which brings them additional and different kinds of funding.
The same happens with people: 20 years or so ago, it became impossible to publish a list of species as a paper. Very few journals would accept it. Maybe some proceedings of an institution deep in the province can do that. But digital products such as data sets and data papers provide the renaissance of this high- quality research and provide a means to publish it and expose it; you become digital visible as an organization and as a researcher. If you work for the Smithsonian or for the Berlin museum, you may not need this visibility – you have it through other reputation means. But if you work in the deep, deep scientific shadow, this is the way to step out. I coordinate work in the former Soviet Union countries, and this is where all this sounds like a very needed opportunity. Because it's free to publish and free to use, you can do that. You are cited, it's an authored product.
[…]

[98] https://casrai.org/credit/

> *It's easy to be a GBIF representative in scientific conferences. There is a certain kind of positive expectation, oh, here's someone from GBIF how cool and then people want to talk about DOIs and data papers and things they've heard of, but they want more information.*

4.1.3 Other measures of impact

For a number of disciplines, the economic value of research data and cost/benefit relations of data infrastructures have been studied, mostly in and for economically affluent countries. About the narrow issue of efficiencies in research gained by social science data, one study on the UK Economic and Social Data Service (Beagrie et al. 2014) concludes:

> *The contribution of ESDS data and services to its user community can be seen in terms of its impact on their research and teaching efficiency (e.g. in terms of time saved). We found that the total estimated efficiency impacts of ESDS data and services among its non-student user community might be worth as much as £100+ million per annum.*

To our knowledge, no such study has yet been done for biodiversity research at all, certainly not individually for many economically quite different countries and not for the share in value added by an aggregator such as GBIF. It appears quite obvious from testimonies such as in section 3.2.1 – two days of work instead of two years – that all the potential value of local or topical, distributed, and heterogeneous data infrastructures diminish, as long as there is no aggregator. More individual estimates confirming this kind of value creation can be found in the literature:

> *Saving time and money is also a clear advantage using GBIF. We were able to download a complete set of entries for tribe Cinchoneae in less than half an hour. By comparison, obtaining the clean list of our entries for the VD [Verified Dataset] took us almost six months and necessitated visits to major collections in herbaria on several continents. (Maldonado et al. 2015)*

Thinking more broadly, for example, if a crisis in pollination (IPBES 2016) can be avoided, where "$235-577 billion US in annual global crop output is at risk as a result of pollinator loss" (IPBES 2019, summary p.10) – how much of the avoided loss could be attributed to biodiversity research and then to individual players in it, such as GBIF?

Meanwhile, GBIF and individual players resort to further indicators of value creation – most especially crafted towards the goals of their respective, immediate funder (not towards benefits to society at large). For example, the Natural History Museum[99] of the UK operates a thorough analysis of their digital operations. In our interview with him, Vince Smith, their Research Leader, Informatics, told us that GBIF has multiplied the impact of their digitization efforts. Almost two-thirds of all downloads were "via GBIF" (see Figure 4.1, below).

[99] https://www.nhm.ac.uk/our-science.html

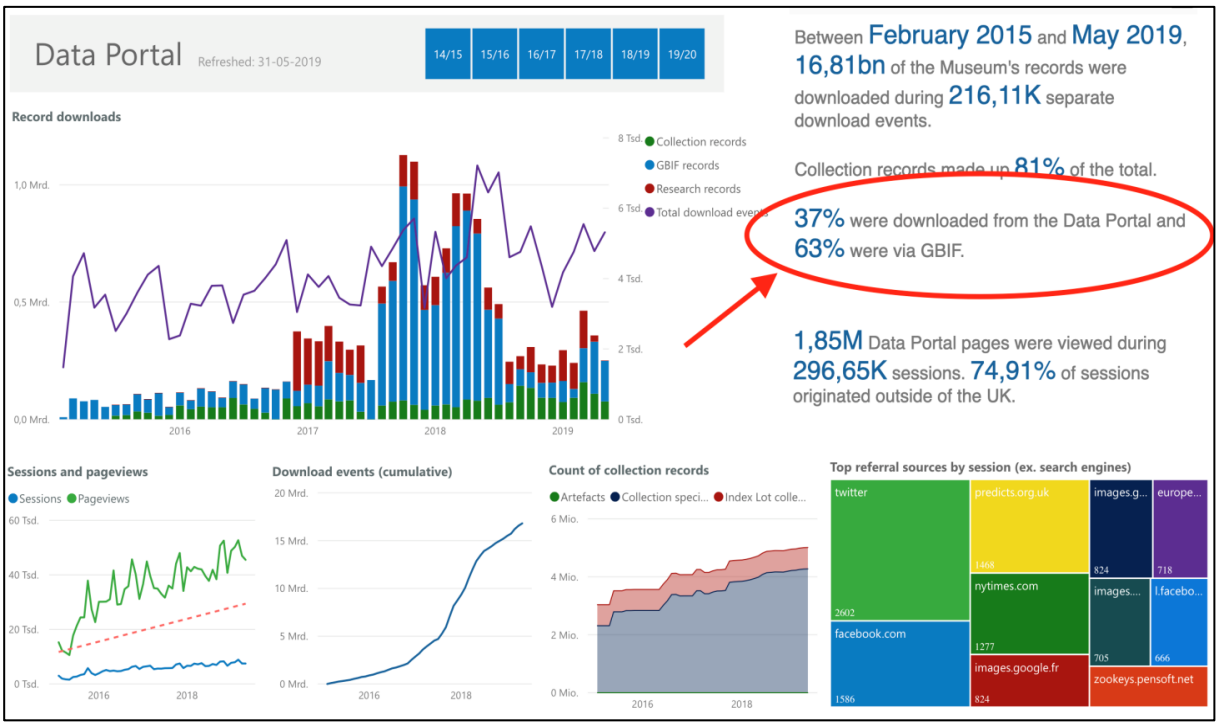


Figure 4.1: Usage statistics of the Natural History Museum (UK) data portal.

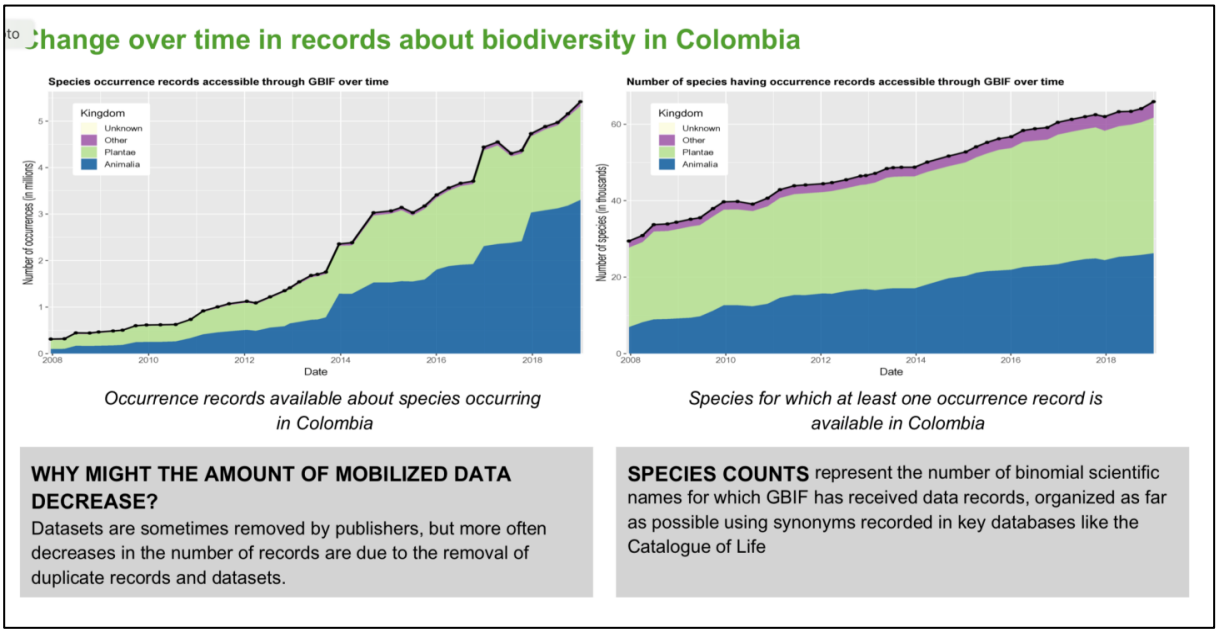


Figure 4.2: Excerpt from a 6-page "Activity Report" about Colombia, 2019[100].
https://www.gbif.org/sites/default/files/gbif_analytics/country/CO/GBIF_CountryReport_CO.pdf

As a proxy for the overall impact of GBIF, including the national Node in each country, GBIF regularly produces "Activity Reports" that provide "a series of summary charts, statistics and other details about the mobilization and use of open-access species data through the GBIF

[100] Figures 4.1 and 4.2 cannot be shown in full size, but only as a "thumbnail", and the URL is provided if one would like to read the entire text.

network, relating to users and participating institutions" in participating countries (see Figure 4.2, below, for an example from Colombia).

In summary, the impacts of research and its infrastructures expected by its funders are fundamentally different among fields of application and so are their respective metrics. With considerable effort, GBIF has achieved much progress in the area of citations (frequently used as a metric of impact in science). But it is even more difficult to prove impact in application areas where impact is based on multiple layers of basic and applied scientific work, and on infrastructural components.

To our knowledge, institutions from GBIF's community have attempted to evaluate and measure impact in terms of monetary benefit only in some very limited cases. Any generalization of that will certainly prove to be even more difficult to implement in daily practice than to find and count citations.

4.2. SWOT Compilation[101]

4.2.1 Strengths

- Data usage reporting for tracking citation, attribution, and impact has been extremely beneficial. The use of DOIs for data so that data providers get attribution has been particularly effective.

- Country-based analytics and pages that highlight a country's contributions can be very helpful in justifying support for both the national Node and GBIF.

- Numerous representatives of collections and other biodiversity organizations commented that the reports on data usage in the scientific literature are extraordinarily valuable. Others noted that "intellectual property rights are a big deal and GBIF has been really clear about who owns the data that are shared".

- GBIF was commended for being behind the creation of data papers so that data contributors get credit for their work.

4.2.2 Weaknesses

- Surprisingly, very few interviewees – in one important region, none – could name a single, major scientific achievement that was enabled by GBIF, or any other "success story".

- Many individuals from all GBIF regions, especially European Node managers, felt that GBIF lacked a clear and understandable value proposition, which made it difficult for GBIF, Nodes, and partners to garner public support and funding.

- There is lack of understanding of the value of biodiversity and data about it by policymakers, which was perhaps most clearly expressed in interviews in the Asian/Oceania region.

---

[101] This section provides the summarized, but also individual SWOTs of GBIF, as provided by the interviewees. It is not an analysis and attempts to convey the original language. For an explanation and rationale of this format, see section 1.5.

  - National governments from those countries failed to perceive the importance and usefulness of biodiversity data provided by GBIF.

  - Interviewees seemed to expect GBIF to take action "to demonstrate how biodiversity data could assist policy making processes, such as environmental planning and environmental impact assessment" and the like. Specific suggestions clearly point at the lack of capacity of national Nodes to perform this on their own (most Asian countries are not yet participants of GBIF; the Nodes operating there, if any, cannot be expected to perform national functions; but even in most Participant countries, Nodes are funded so rudimentarily that there is insufficient capacity in this area to be expected).

- In large countries with high GDP but (relatively) low GDP per capita, GBIF is perceived as requiring a high membership fee but failing to give a cost-benefit demonstration.

- In countries with low GDP there may even not be enough expert users of biodiversity data to potentially create value for that country and who could pitch this potential to funders - thus inhibiting the start of a virtuous cycle.

- Many data producers/publishers seem to be worried that the way GBIF-mediated data are being cited cuts them out of the chain of reputation gain. This may as well be a weakness in communication or providing them with easily applicable tools to analyze their gains, however, rather than a real problem.

4.2.3 Opportunities

- Several individuals from all regions suggested that GBIF really needed to tackle one or a small number of high-profile use cases that would provide a compelling success story (e.g., taking the lead in the creation of one or more Essential Biodiversity Variables (EBVs), examining the decline of bees over the past 50 years, understanding Ebola vectors, or invasive species mitigation).

- Engage the general public. Several individuals felt that GBIF could better connect with and engage the general public as advocates or supporters. It was suggested that GBIF think about new ways to get information about GBIF and biodiversity out to the public. Can individuals "join" GBIF and, if so, how can they contribute and benefit?

- GBIF can host a workshop to identify and further develop biodiversity success stories and to develop a common pathway so that biodiversity organizations can more effectively engage with and support GBIF (and vice versa). The workshop should include "power users and producers" and key organizations such as NatureServe, the African Conservation Center, and the JRS Foundation.

- Educate the general public. GBIF could provide public seminars or webinars that are designed to "fascinate people" with issues like pollinators, exotic species, pests, and other issues that are of interest to the public. Webinars could be done regionally to advertise GBIF capabilities like the Cornell Lab of Ornithology does with bird tools, visualizations, and the like.

- Empower those working in developing countries to use data for policy making and resource management decisions. This could be done via in-person workshops, webinars, and seminars. Interviewees recognized a need for better communication between scientists and policy makers demonstrating the practical value of information (biodiversity plus other sources and types of data). Such training could teach people how to use GBIF infrastructure and may help promote sustainability.

4.2.4 Threats

- Around the globe, science has been challenged in recent years and the investment in better understanding and characterizing biodiversity may be painted as superfluous. A big part of the global population thinks that loss of biodiversity is a distant problem for society or does not know about it at all. Consequently, it is not easy to convince governments or the public of the economic and other values of biodiversity.

4.3. Conclusions

GBIF inhabits a position in at least two types of chains of value creation crucial to its further success:

- As a functioning infrastructure, it has to observe foremost the issues of its users and data contributors (which have been laid out in detail in Chapter 3 and revolve around their own impacts and the metrics *they* are subject to).

- In order to acquire sufficient funding, GBIF also has to prove its own share in value creation to funders – in possibly much longer value chains than is the case for individual users or contributors.

The role of GBIF in each of the long value chains can be occluded to the point of invisibility. Therefore, it will be difficult in any particular case to claim, produce evidence for, and to visibly establish GBIF's *share* in value creation. GBIF and all other contributors in that value chain need to respect and communicate each other's contributions. If they do not, they take the long-term risk of breaking the chain of value creation for future work, if a contributor is no longer funded.

Insofar as individuals or organizations higher up the value chain do not communicate GBIF's share in value, or this communication does not reach GBIF's main stakeholders, GBIF needs to maintain its own communication channels and support them with evidence. This evidence possibly requires substantial effort to be researched or developed, to support each individual claim of value created.

The GBIF leadership, as well as all staff members and most of the Nodes we interviewed, are clearly aware of this. The Nodes mostly lack the capacity (and some may lack the interest) to work effectively on this issue, or to work on more than one aspect, such as the value for basic science and one or more application areas. The Secretariat works hard and with great determination on many individual issues, and on improving solutions. It has received near universal acclaim for its activities in data citation, which is an indispensable metric of academia today.

However, there are still a number of problems to solve in order to be able to track the use of each individual data collector's contribution (e.g., "Has my observation been used in the IPBES report?"). On a quite different front, it is worth mentioning that even the existing analysis appears possible only insofar as GBIF has full control over the data, its extraction, and downloads. In consequence, GBIF could not place a copy of data "in the cloud", or in an Amazon data bucket. In the cloud, it would still have to wrap it with the same functionality as on its own premises, to preserve the proper – DOI-based - citation of data.

The considerable software development and metadata analysis to produce bibliographic metadata and reliable, and stakeholder-specific metrics and analyses has been a significant and ongoing effort at the Secretariat. This area of online (real time / monthly / quarterly) reporting is also *the* big data operation at GBIF. An additional, comprehensive external bibliometric study could help to determine if the Secretariat actually finds all relevant citations.

But the impacts most desired by the stakeholders - and even the metrics employed - can vary from country to country and from region to region (e.g., impact on science vs. impact on issues of agriculture). It is not enough to provide bland metrics, such as citations and downloads or even monetary analytics.

Nodes, in particular, but also the GBIF Secretariat, are in need of targeted success stories and "elevator pitches"[102]. This observation is important as it may be decisive not only for the continued financial contributions from each country but even more so for the funding of each national Node and the possibility of individuals to contribute to GBIF with in-kind or voluntary work (e.g., as trainers).

Many Nodes perceive that there is little systematic support by the Secretariat, and even less capacity at the Nodes, to develop Nodes' individual value propositions as part of GBIF. For example, how have studies based on GBIF data, and particularly data mediated from that country, benefitted that country? How do the data help avoid duplication of work specific to a country, help researchers or specific industries in a country to achieve their goals more effectively, and the like?

Development of best practice in communicating scientific knowledge – and of the contribution of research infrastructures – is now all the more important as in too many countries the relevance of science or of scientific findings is being challenged and knowledge about biodiversity may be painted or perceived as superfluous.

In summary, the theme of "pitching" the value proposition of an intermediary data infrastructure is certainly a difficult one, not only technically, but also regarding the attitudes of all people and institutions involved and it depends to a large extent on continued cooperation and even on changes in the behavior of external stakeholders. The GBIF Secretariat has a full understanding of the issue and uses its limited resources to good effect, but major gaps appear to remain, especially in communicating the success stories effectively to the many types of stakeholders.

[102] The collection of arguments for such "pitches" has already begun at the 2019 meeting of the GBIF Governing Board in Leiden.

5 DATA, STANDARDS, AND TECHNOLOGY[103]

5.1. Description

The implementation of GBIF as a data infrastructure is the organization's central means to fulfil its mission and informs all of its activities. This chapter mainly addresses the "technical aspects of GBIF's delivery and its sustainability and trustworthiness (in particular, to researchers) as a research infrastructure" (statement of work for this report, see section 1.3).

Before laying out the SWOTs compiled from our interviews, we discuss some key characteristics of this technical infrastructure, including: (1) data volume and quality, (2) data gaps and biases, (3) standards and technologies, and (4) current developments of the conceptual frameworks of data infrastructures, on which further development may be built over the next decade.

Data, and their underlying standards and technologies, are most obviously fundamental to a Global Biodiversity ***Information Facility*** (as opposed to a research facility). In order to fully and fairly appreciate what are perceived as the strengths, weaknesses, opportunities, and potential threats of GBIF, we stepped back and considered a most general view of biodiversity science and its data and knowledge sources. What information do their stakeholders actually consider as belonging to the biodiversity domain and which data do they consider necessary and available in sufficient quality?

To inform this view, we considered two studies. On the one hand, the new IPBES assessment report (IPBES 2019, chapter 2.2, p.12) argues that:

> *Synthesising and mapping variations in the state of nature across the globe and over time has been greatly facilitated by major recent advances in remote observation of biodiversity and ecosystems, in modelling and in informatics.", while "Recording of Indigenous and local knowledge [...] can also add relevant information over smaller scales." In this context, it observes that "aggregators and repositories, such as GBIF (www.gbif.org), OBIS (www.iobis.org) and Genbank [...] make hundreds of millions of species occurrence records and gene sequences freely available.*

Most telling here is the evaluation that:

> *Ever-improving metadata mean that such data [...] can increasingly be put to a wide range of uses.*

On the other hand, the IPBES report listed a wide range of gaps in Appendix IV (still marked as draft), and cited "Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity" (Hortal et al. 2015), where authors lay out the limits of the current factual basis of knowledge, which can serve as benchmark for realistic expectations of GBIF data:

> *Faced with the almost overwhelming complexity of the natural world, biologists have always sought to categorize and classify organisms [...] Inevitably, such classifications reflect the goals and interests of the classifiers;*

---

[103] This chapter addresses item 4 of the statement of task: "Review the technical aspects of GBIF's delivery and its sustainability and trustworthiness (in particular, to researchers) as a research infrastructure". It also covers most of item 1: "Review how effective GBIF has been since 2001 in meeting the expectations from the OECD working group".

> *[…]*
> *complete knowledge of any given characteristic of biodiversity is practically unachievable,*
> *[…] This unevenness in survey effort and research infrastructure results in high spatial and temporal variation in the quality and reliability of the data available*
> *[…]*
> *The fundamental and practical limits of biodiversity knowledge mean that scientists must work with incomplete and often unrepresentative data on a limited number of organisms and their characteristics. The gaps, or shortfalls, in knowledge about the identity, distribution, evolution, and dynamics of global biodiversity need to be carefully recognized and quantified.*

Both of these views on the current situation can be reconciled, if in each of the "wide range of uses" the "gaps, or shortfalls are carefully recognized and quantified".

A number of "uses", scientific and otherwise, can thus be prudently performed if the user adheres to the appropriate practices and protocols when using the data – whether their own data or data from other sources, including from global data infrastructures. It may be a matter of good education and training, or of sufficient funding and other factors, whether users will do this well. Data infrastructures, however, can help by reducing the technical effort necessary, or even contribute to increased capacities and awareness of users (which will be addressed in Chapter 6). The OECD 1999 proposal even saw a role of GBIF in shaping the biodiversity informatics curriculum.

In other cases, it remains the responsibility of users to devise more robust analytic procedures for "outlining how robust the conclusions are given the current level of uncertainty" (Hortal et al. 2015). Users may need to gather more or better data before continuing, or even stop the attempt at analysis.

As has already been shown in Chapter 2 on the biodiversity informatics landscape, the primary role of GBIF is to "mediate" occurrence data provided by a range of publishers and to do so for the whole planet and all species (see section 5.1.2). Other stakeholders are expected to provide the taxonomy and other types of biodiversity data, or data related to biodiversity, such as environmental data. This modularization appears reasonable; after all, each of the players can concentrate on the quality issues of their own data (see section 5.1.1).

These advantages have limits, however, as the "duplication issue" discussed at the end of section 5.1.1 indicates. Above all, they devolve most of the burden of linking diverse data types to the user, in that they still need to collect data from a number of sources and merge them into one consistent body of data – starting with making sure that different providers of data use the same taxonomy, geodetic datum, and other usability issues. This requires that each data contributor invest in tools to make such applications easier, especially through thorough adherence to standards (see section 5.1.3), and, finally, to constantly align with each other and work on the conceptual framework of biodiversity data (see section 5.1.4).

The necessity of such linking can easily be demonstrated by a page from a GEO BON meeting, where biodiversity modelling was attempted using occurrence data (of ferns) from GBIF, current and predicted climate data, and land use data, see Figure 5.1.

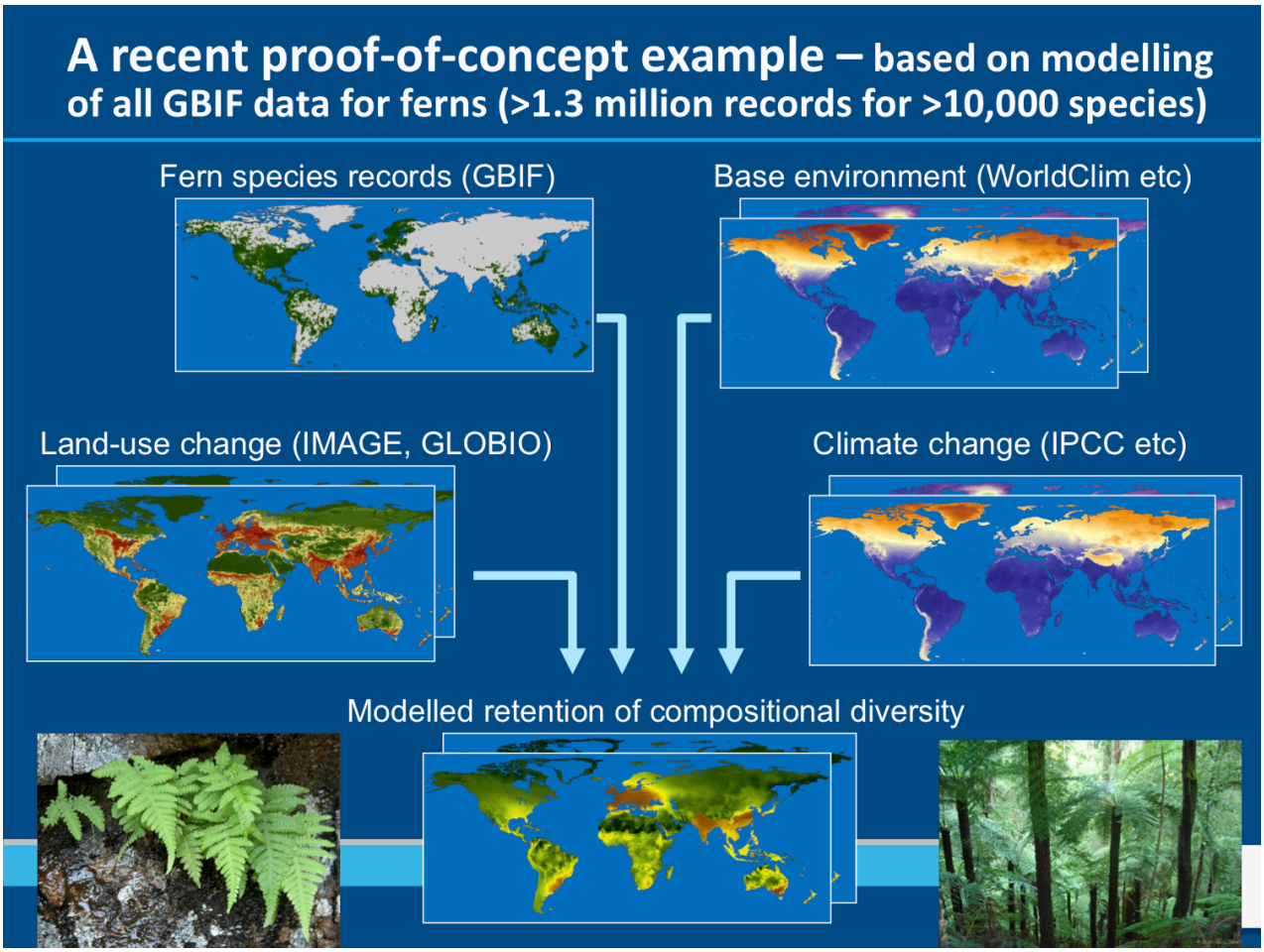


Figure 5.1: Linking GBIF records with other types of data for modelling purposes (Ferrier 2012, "In situ / remote sensing integration through modelling", slide 15).

5.1.1 Data Volume and Quality

Data Volume

The amount of data mediated by GBIF by mid-2019 – about 1.3 billion records – is impressive. For the data practitioner, the more than 1,400 publishing institutions from which these records have been gathered appears even more impressive.

The upper graph from GBIF's "Global data trends", Figure 5.2, below, shows a very substantial growth in the number of records over the last decade. The lower graph shows that there is at least one occurrence record for (almost) each species known to taxonomists – compared with the estimate in (Hortal et al. 2015) and the number provided by the Catalogue of Life, 1.8 million. These numbers have been achieved in spite of the much lower than expected percentage of collection specimens digitized (see section 2.1.1) and the presumed taxonomic biases of observers and collectors (see section 5.1.2). This volume is even more remarkable as these records, and the attending services of GBIF, are openly available.

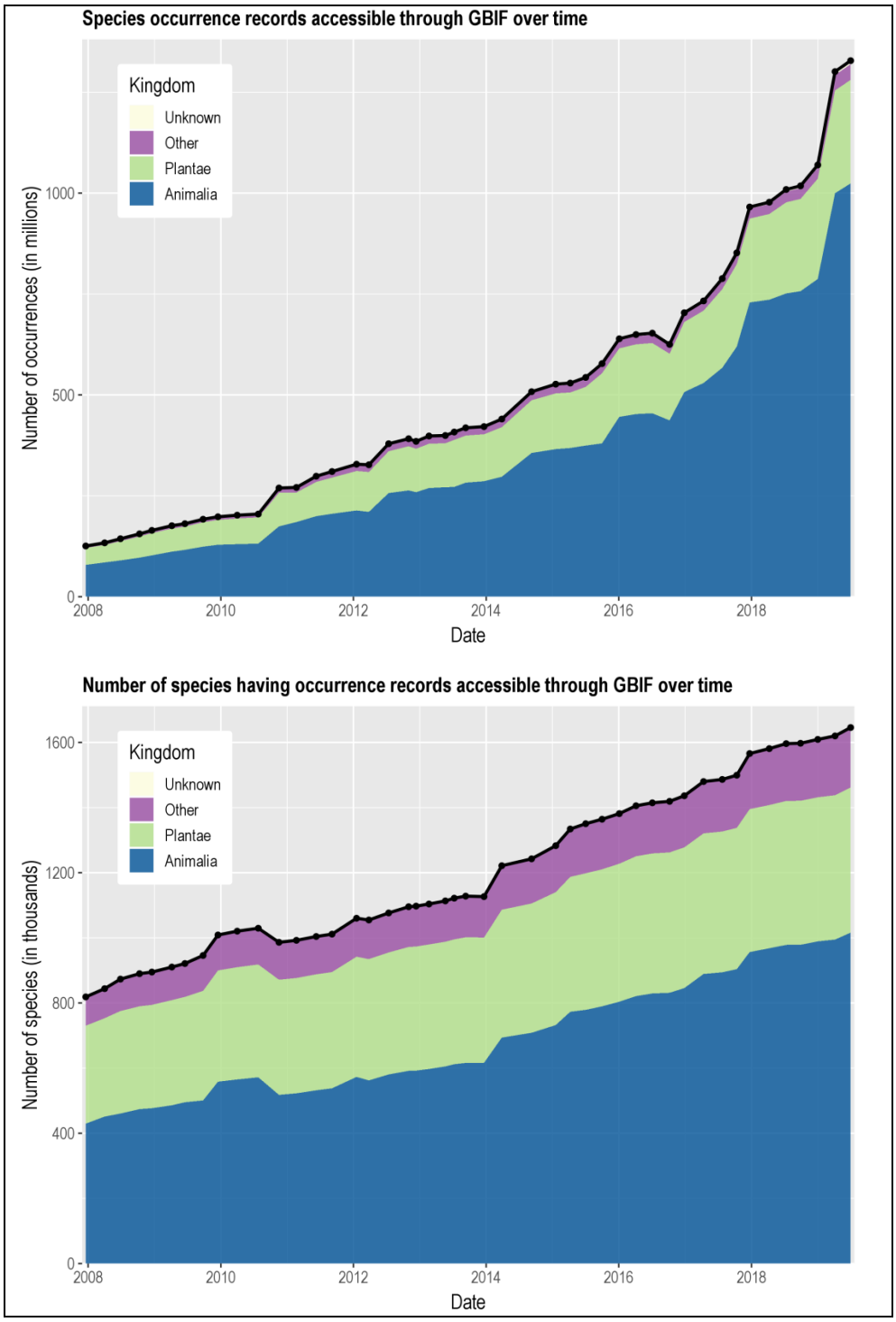


Figure 5.2: From GBIF page "Global data trends", found under the top menu item "Get data /Trends", https://www.gbif.org/analytics/global.

The historic anecdotes in Box 5.1 clearly reveal how open access to data, via a single portal, can streamline the data gathering processes so much (e.g., from two years to two days, as reported in section 3.2.1).

> ***Box 5.1: Dmitry Schigel, Science Officer at the GBIF Secretariat:***
>
> *I worked in systematics and in ecology before and I know the struggle writing multiple emails and waiting for answers, waiting for loans, looking for travel budget to visit the collection. A lot of that is now speeded up by GBIF,*
> *[…]*
> *There is no kind of one stop shopping place for ecologists, if you're outside LTER. So, if you work, say, on Tierra del Fuego and you suddenly need data on the Great Lakes and Lake Baikal, and a lake in India, your life is not easy. You are still in the era before the OECD report. Hundreds of emails: "Can I see the data please?" – "Yes. If you make me a coauthor", all this kind of prehistoric culture.*
> *[…]*
> *There is a growing understanding that ecology needs [open data] too, but, for some unknown reason ecology is 25 years behind museums and citizen science and the molecular people in understanding the potential of open data*

On 18 August 2016, GBIF announced that all species occurrence datasets published on GBIF.org would carry open, standard machine-readable licenses, namely one of the three Creative Commons licenses, CC0, CC-BY, or CC-BY-NC. Although discussion of these issues began around 2006, the implementation was initiated in April 2014 with a community consultation, followed by painstaking communications with all data publishers that was led by GBIF's national nodes[104].

As a result of this decision, only about 7.5% of records were retracted from GBIF. Moreover, many of them were expected to re-appear after extended discussions by some data publishers with their "data partners". Note the small dip in the occurrence records curve, Fig 5.2, above, in the second half of 2016. This small effect, however, may not reflect the potential reaction of individuals, institutions, or whole subdisciplines which had not yet shared their data.

In order to maintain this consent of contributors and publisher, the requirement to cite selected data is communicated quite clearly during the download process, as Figure 5.3 below shows. Designers of the system carefully explain the expectations of data contributors and make it easy for users to comply.

[104] https://www.gbif.org/news/82812/licensing-milestone-for-data-access-in-gbiforg

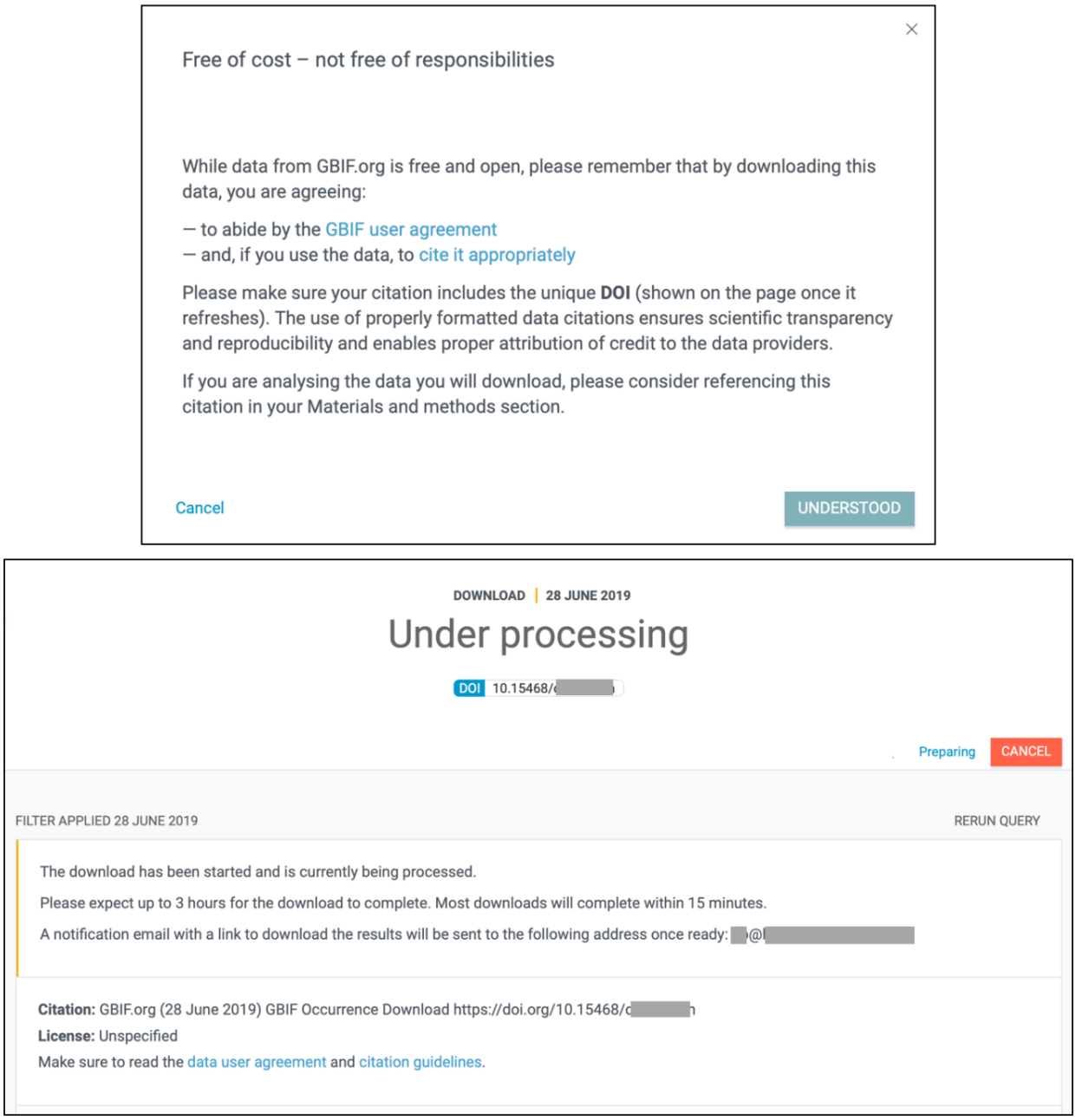


Figure 5.3: Dialog boxes appearing along the download process of the GBIF portal (as of June 2019).

As can be seen in Figure 5.3 above, each selection a user creates is assigned an individual Digital Object Identifier (DOI). Therefore, the selection can be cited unambiguously in articles[105]. This is a very significant contribution to the potential repeatability of conclusions – and for providing credit for data work.

Moreover, if DOIs are employed, tracking of data usage in the literature becomes so much easier, providing the contributors with a much-needed indication of impact. This is nearly

[105] Actually, this is also a very noteworthy implementation of a RDA recommendation on "Data Citation of Evolving Data" (Rauber 2015).

impossible with a mere acknowledgement of "data from GBIF". Today, one can immediately see citations per dataset on in a publisher's page, e.g., for PANGAEA[106].

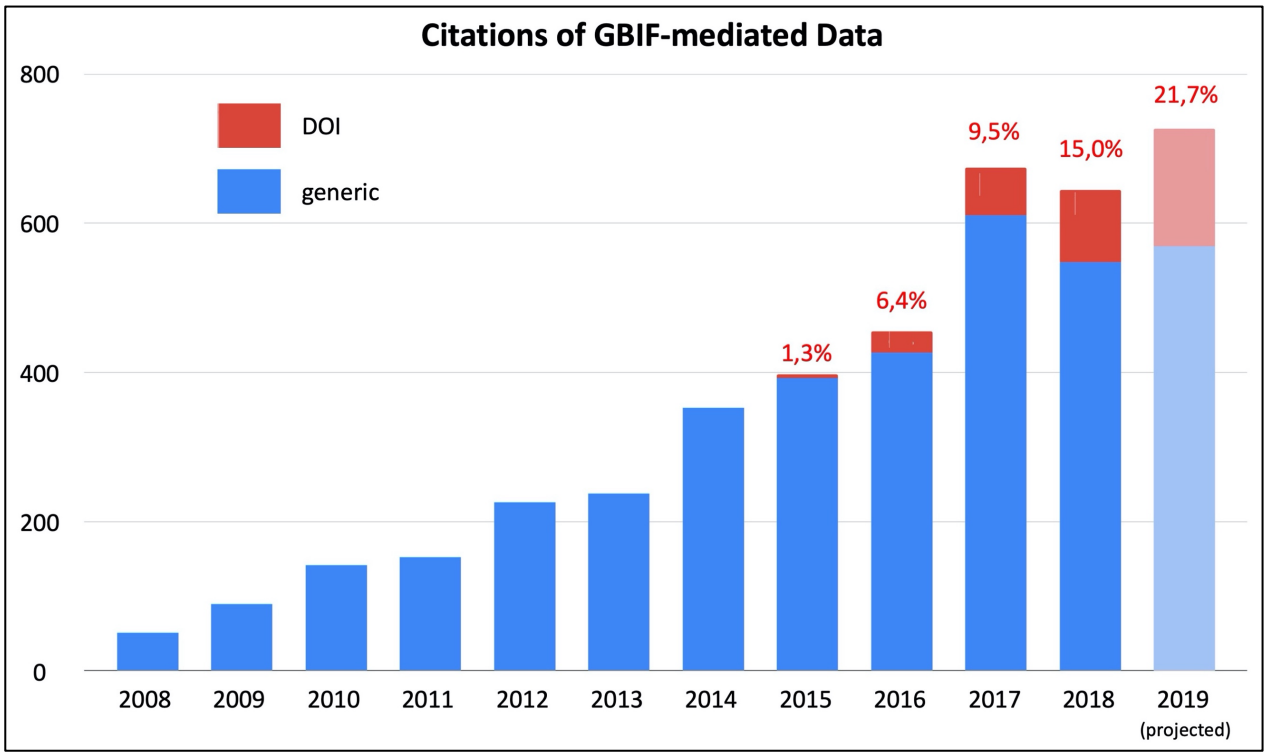


Figure 5.4: Usage of DOIs in citations, as a percentage of all citations of GBIF-mediated data

Since being introduced in 2015, the actual usage of DOIs within the ambit of GBIF has grown to approximately 20% of all citations (see Figure 5.4). The growth rate of DOI-based citation probably reflects more on scientists' habits than on technical or practical difficulties or limits of employing the DOIs, and can be considered a good pace of growth, if it continues – old habits die hard.

Data Quality

The OECD report noted this on data quality (OECD 1999, p.9):

> *Global information systems in biology are of greatest value when they provide data of reliable and known quality.*

Connecting the issues of volume and quality, the report noted also:

> *There is an apparent conflict between the objective to speed up the global availability of biological information and the necessity to improve the quality of that information, which may slow down the process.*

But what does quality mean in the context of biodiversity, specifically? As there are many different use cases in biodiversity research and application, the most general definition of quality, to designate "fitness for purpose", immediately uncovers the fact that there can be no

[106] https://www.gbif.org/dataset/search?publishing_org=d5778510-eb28-11da-8629-b8a03c50a862

simple measure of quality – neither a yes/no, nor a high/low, nor a "5 stars" recommendation system will work.

After stating that there are "myriad uses" of biodiversity data, the "GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network" (GBIF 2010b), finds that:

> *Each user therefore must vet records carefully to determine their fitness-for-use: often, a time-consuming task. Although user vetting will always happen, the key discussion point in this white paper is what can be done prior to user access of data to enhance and better report the data's fitness-for-use. (p.4)*

This paper from 2010 concentrated on geospatial issues and recommended employing more plausibility checks to detect them, to flag implausible records, and to employ, eventually, automated corrections, without overwriting original data. It goes on to say that: "In order to know if a piece of information is actually correct, there is no choice but to ask the information owner." (p.13) These general recommendations can certainly be applied to other than geospatial issues, such as taxonomic ones.

Comparing the types of issues identified today in the GBIF portal, Table 5.1 below identifies more types of geolocation issues, and beyond that dating, taxonomy, counting and other miscellaneous issues, such as invalid links.

Issue	# of Records affected
Zero coordinate	2,085,834
Coordinate out of range	48,410
Coordinate invalid	378,031
Coordinate rounded	55,311,390
Geodetic datum invalid	1,118,152
Geodetic datum assumed WGS84	108,197,990
Coordinate reprojected	2,094,543
Coordinate reprojection failed	0
Coordinate reprojection suspicious	2,844
Coordinate accuracy invalid	0
Coordinate precision invalid	5,040,179
Coordinate uncertainty meters invalid	171,732
Coordinate precision uncertainty mismatch	0
Country coordinate mismatch	2,296,721
Country mismatch	150,966
Country invalid	811,120
Country derived from coordinates	16,441,546
Continent country mismatch	0
Continent invalid	0
Continent derived from coordinates	0
Presumed swapped coordinate	122,962
Presumed negated longitude	169,150
Presumed negated latitude	43,605
Recorded date mismatch	4,232,812

Recorded date invalid	3,613,830
Recorded date unlikely	186,251
Taxon match fuzzy	1,309,956
Taxon match higher rank	5,431,498
Taxon match none	0
Depth not metric	0
Depth unlikely	28
Depth min/max swapped	166
Depth non-numeric	983
Elevation unlikely	0
Elevation min/max swapped	1,641,240
Elevation not metric	1,011
Elevation non-numeric	82,845
Modified date invalid	0
Modified date unlikely	196
Identified date unlikely	405,977
Identified date invalid	0
Basis of record invalid	3,633,317
Type status invalid	0
Multimedia date invalid	536
Multimedia uri invalid	278,845
References uri invalid	10,652,270
Interpretation error	10,091
Individual count invalid	21,971,866

Table 5.1: "Issues and flags" on 247,821,632 occurrence records of Tracheophyta, extracted from https://www.gbif.org/occurrence/search?taxon_key=7707728

What may be even more important is that, as Figure 5.5 shows, the GBIF portal enables the user to filter "raw" records by issues in a very finely grained way and offers basic statistics and graphs ("metrics") on the selected records, This approach could help detect gross anomalies or guide further selection based on, say, the month of occurrence or publisher.

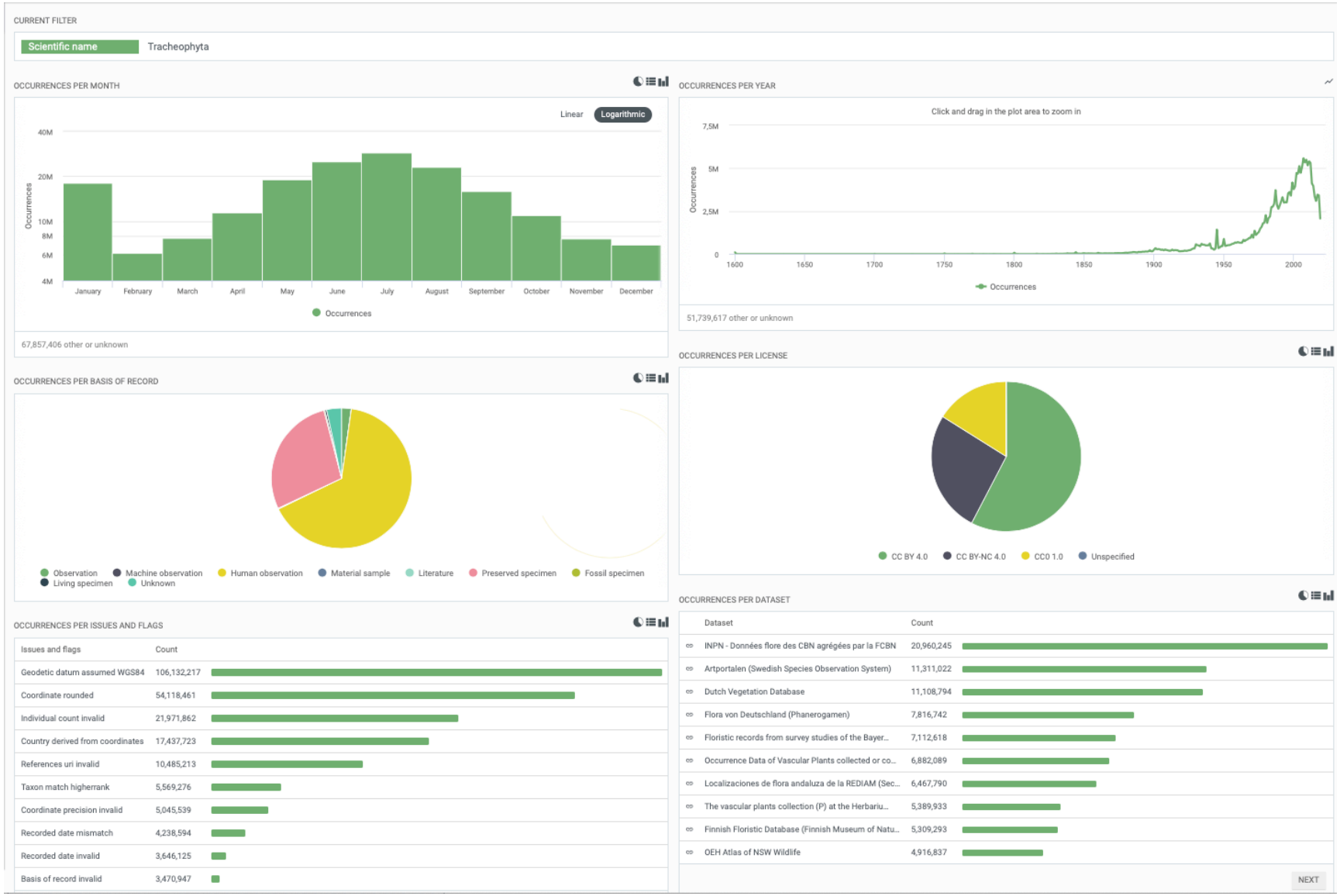


Figure 5.5: GBIF portal “metrics” on a selection of Tracheophyta, tab “Metrics” of https://www.gbif.org/occurrence/search?taxon_key=7707728.

While this kind of detection of implausibility or missing information is certainly helpful, a more detailed requirement would be to quantify accuracy or certainty (as indicated by Hortal et al. 2015). Figure 5.6, below, implies that an "expert opinion range" does not provide the required geolocation precision for certain types of modelling since it does not reflect the suitable land cover and elevation, in this case, for Hartlaub's Turaco.

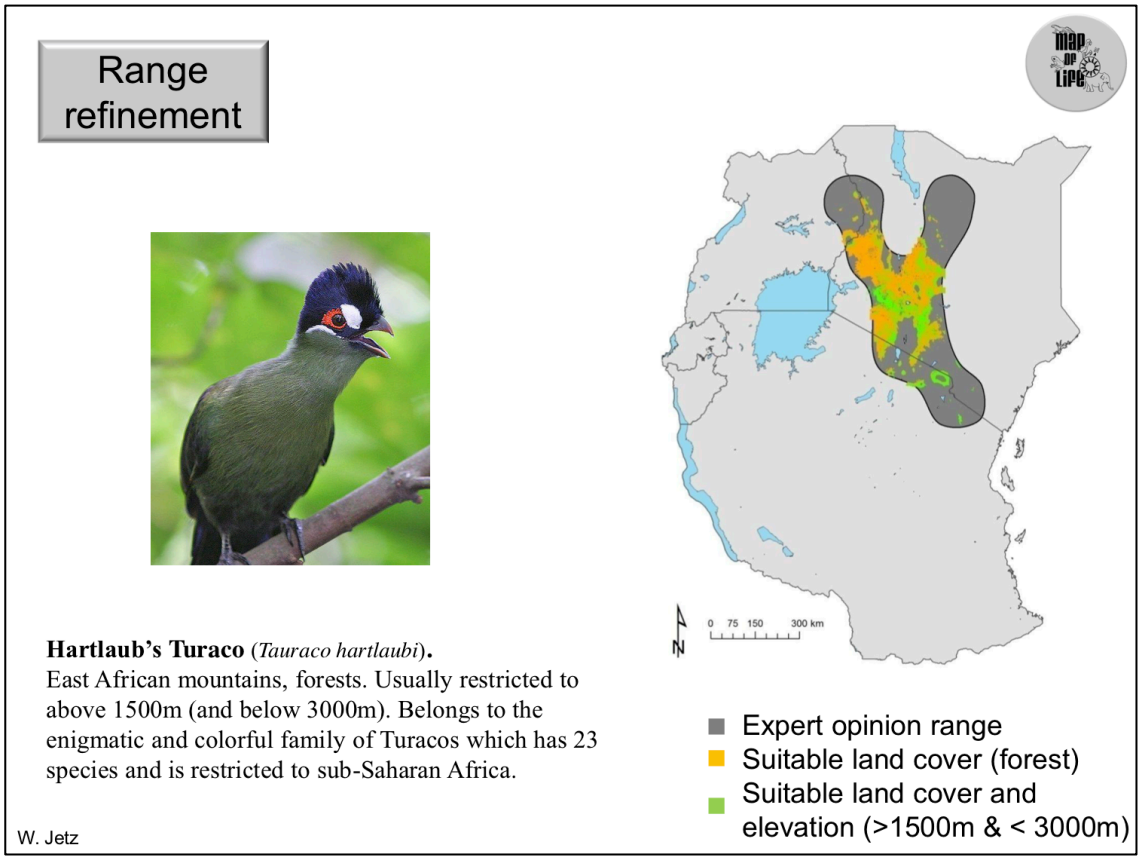


Figure 5.6: Expert opinion range of Hartlaub's Turaco vs. suitable landcover and elevation, slide 10 from In situ / remote sensing integration through modelling (Ferrier 2012).

Some explicit methods of *a posteriori* determination of accuracy (or detection of errors) exist. Interpolation and modeling methods frequently employ statistical reasoning about accuracy, implicitly. Both however will be flawed if a significant fraction of the records are duplicates (although it is better to have duplicates than missing data). GBIF has employed duplicate detection methods since at least 2012[107], but this effort can be undermined if users feel compelled to use more sources besides GBIF and do not employ sufficient duplicate detection themselves.

Thus, for example, the authors of "Big data of tree species distributions: how big and how good?" (Serra-Diaz et al. 2018) found that, combining data from "five major aggregators of occurrence data", duplicates alone comprised 63.17% of 23.18 million records. Whether or not this is the reason in that article, there is potentially an increasing risk that data are duplicated as publishers deliver data both to GBIF and to users directly (as promoted by their own interests and resulting metrics of impact - see Chapters 3 and 4), and as cleaned up, "vetted", or even aggregated versions are published elsewhere by users.

[107] https://www.gbif.org/news/82302/duplicate-and-redundant-records-cleaned-up-in-gbif-data-portal

5.1.2 Gaps and Biases

The OECD report (OECD 1999) noted that:

> *Biodiversity itself is distributed all over the Earth, with concentrations primarily in developing countries. In contrast, scientific biodiversity knowledge is concentrated in major centers in developed countries.*

A relatively recent assessment of data availability in the IPBES summary report on pollinators (IPBES 2016, pp.18 and 21) confirms that the situation persists today, in that occurrence data about bees and butterflies allow for an "established but incomplete" assessment of their decline in North-West Europe and North America, but "data for other regions and pollinator groups are currently insufficient to draw general conclusions".

The reasons for the disparity between the geographic concentrations of biodiversity and the amount of knowledge, in particular the underlying data, can be explained by various obstacles, such as wars, a lack of experts or technologies, basic infrastructure, and probably many others. The main reason, however, is frequently financial. Given limited resources and the high cost of logistics, individuals as well as institutions choose topics of research or surveys closer to the workplace (e.g., Speed et al. 2018). The overall occurrences map of GBIF (see Figure 5.7) seems to illustrate this point in, for example, the central Amazon and Siberia, which are areas of low observation density (green). Other areas are not accessible to more cost-effective observation methods. For example, it is not possible to survey the ocean biodiversity (benthic or pelagic) using indicators from remote sensing, except for the uppermost meters.

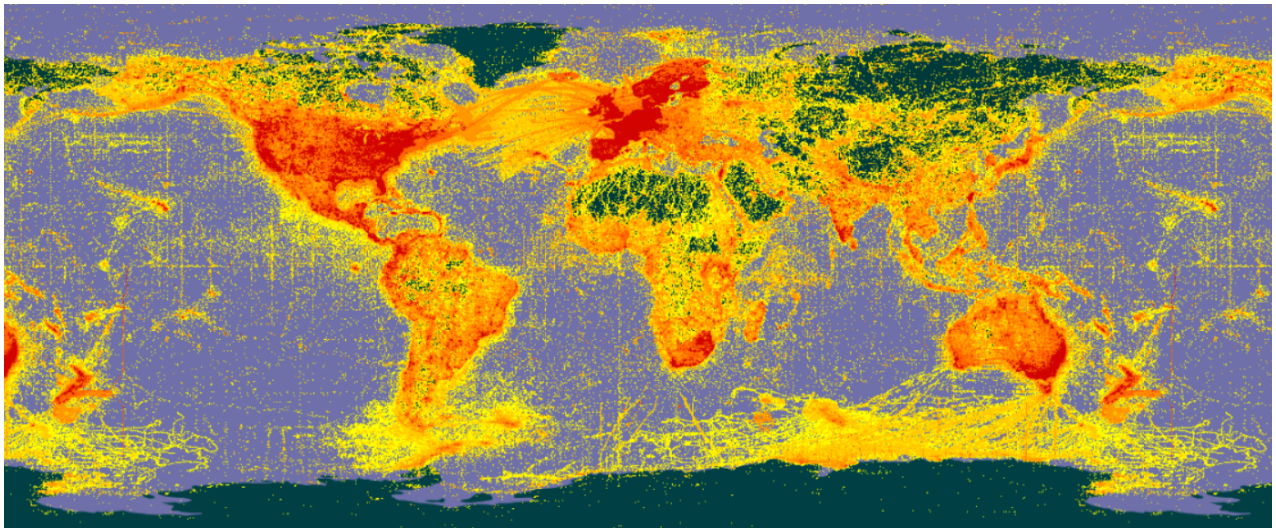


Figure 5.7: GBIF global occurrences map for all species, copied from www.gbif.org, 2019-08-09

The influence of the resource limits and pressure on research to produce should not be underestimated, as an analysis from a wealthy OECD country shows: "Both specimen and observation records were concentrated in regions of Norway with high human population density [...] records thus differ in taxonomic, temporal, spatial and environmental coverage" (Speed et al. 2018).

In between the objective and the financial reasons may be the relative lack of capacities in terms of qualified observers, data curators, and data infrastructures in certain regions of the world, as reflected by maps of GBIF publishers, conference attendance, and existence of data repositories, by country, in Figure 5.8, below. Unsurprisingly, as Chapter 4 demonstrates, any one of these maps may be misleading in one way or another, and be insufficient as a predictor of gaps or completeness of data.

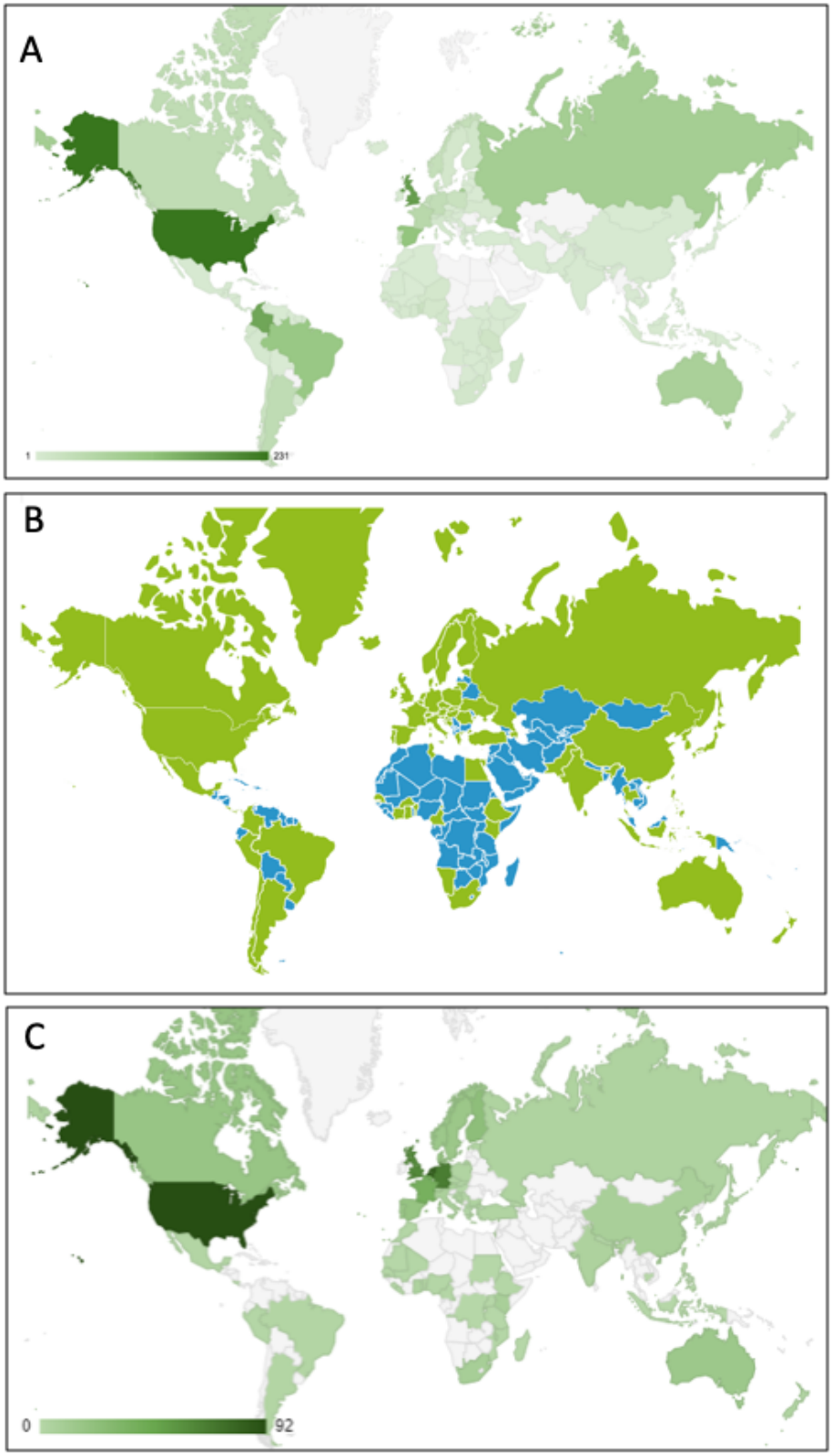


Fig 5.8: (A) number of institutions registered as GBIF data publishers, by country, as of 21 August 2019, (B) countries with at least one data repository registered in re3data (depicted in green), https://www.re3data.org/browse/by-country/, and (C) attendance in the BiodiversityNext conference, (personal communication, Arturo Ariño Plana, November 2019).

For instance, the good number of GBIF publishers in Russia and Asia leaves out the fact that there is scarce formal Participation in GBIF; re3data[108], "a registry of research data repositories", provides just a yes/no classification whether a country hosts at least one repository, and attendance at the *BiodiversityNext* conferences may be more related to funding issues than to the existence of professionals. Currently there may be no better indicators, however.

Taxonomic gaps or biases, in biodiversity research in general and concerning GBIF data specifically, are brought about in a variety of ways. One major bias is mostly due to the high activity of citizen scientists who are bird watching aficionados: GBIF search finds 833 million results for the class Aves, 63% of the total of 1.3 billion records[109]. Individual scientists and research organizations follow their own agendas as well, however. They may survey for predetermined lists of endangered species, plants useful for agriculture, or record of fish caught using industrial fisheries' nets. These surveys typically also focus on species and individuals easy to detect, e.g., due to their size or habits.

Some geographical or taxonomic biases probably rest on political or structural funding issues. Impoverished countries may, for example, concentrate their resources on agriculture, fisheries, or tourism (the later favoring "charismatic" or "iconic" species). In the context of the GBIF organizational structure, the fundamental role of national Nodes in mediating data may be changed or modified locally by which ministry (e.g., science or environment) is funding it and by the agenda of the hosting institution – in particular, if the Node itself is underfunded.

Most of such gaps and biases cannot be influenced by a data infrastructure (short of stopping to accept data from some providers until others, with compensating biases, have caught up in delivering their data). In the case of GBIF, setting other priorities at the Nodes might help somewhat, but only to the point of the biases and resources of biodiversity research experienced in each Node's home country. (GBIF cannot order specific data to be collected to fill gaps.)

5.1.3 Standards and Technologies

The Darwin Core (DwC) metadata standard[110], and related tools, such as the Integrated Publishing Toolkit (IPT), as well as the GBIF portal itself, have been fundamental in GBIF's acquisition of a huge amount of data from a large and heterogeneous group of publishers. However, as a speaker at the Oslo regional European Nodes Meeting, 22-24 May 2019, noted, DwC may be superseded by other standards in the coming decade.

Following this discussion in Norway and at other meetings and in interviews, Darwin Core – used without proprietary extensions or specific understandings of generic attributes (the "event-core" being such an extension) – can make it hard, in general, to link biodiversity data to other types of data. For instance, it cannot include or transport environmental data at the time and location of observation, which are critical for aquatic biodiversity research, or easily enable the mapping of co-occurrence of species in a specific soil sample or observed in one survey of a patch of land. Under the same restriction to the basic Darwin Core standard, tracking data of one individual through time and space, and other new types of observational data, particularly automated ones such as data from camera traps, pose challenges.

---

[108] https://www.re3data.org/

[109] https://www.gbif.org/occurrence/search?taxon_key=212, accessed 2019-08-09

[110] https://www.tdwg.org/standards/dwc/

The success of standardization largely depends on the ability and willingness of those applying it. In this case, the 1,500 publishers and their unknown number of "data partners" need to apply a standard using the same interpretation as the others. This is hard to do even for a simpler standard, such as Dublin Core for bibliography that has just 15 attributes. In the case of more detailed standards, such as ISO 19115 with about 1000 attributes, this has led to the use of "profiles" by individual repositories, which makes it difficult for data aggregators such as GBIF to map individual interpretations of the same attribute or fill in the blanks for non-overlapping attributes. Moreover, choosing just the common minimum will merely result in an attribute set equivalent to Dublin or Darwin Core metadata standard.

Because the GBIF community is fully aware of the limits and fragmentation of the current standards and their implementations, the GBIF staff are working with standardization groups, such as TDWG members and other stakeholders, on a next generation of metadata standards for contributing and using data (see section 5.1.4). Considering how much time was needed and how much effort was spent by publishers on the implementation of DwC and the IPT, it is prudent to assume that the implementation of a necessarily more complex new standard will require even more effort, and may therefore take another decade or longer before reaching most of the data publishers.

GBIF supports its stakeholders with tools and the portal to ease contribution and use of biodiversity data. It does not just collect datasets from Publishers but performs extensive plausibility checks and some automatic corrections (flagging corrected and uncorrected implausibilities, and missing information, and designating them as "issues"). GBIF comes back to publishers if they detect systematic issues, such as consistently swapped latitude and longitude. It offers some quality-related tools, such as the "data validator tool" for data Publishers and a feedback mechanism through GitHub for individual users to report issues or problems with their downloads or bugs in the system, and to offer suggestions for improvement. Many features offered to users have already been discussed in sections 5.1.1 and 5.1.2, and there are certainly more features that are not obvious.

It is important to emphasize that the GBIF portal is a big data operation in the technological sense, even though the pure data volume is not huge in comparison with some other disciplines or applications. While some of the regular reporting, such as providing a country report, can be done in batch mode, user demands for on-demand statistics or maps to be produced immediately have required the use of advanced database and visualization technologies, many sponsored by prominent cloud-providers and the Apache Software Foundation.

A tool such as that in Figure 5.9, below – and fast response times – allow for the detection of issues and anomalies while working with the portal's functions, before downloading selected data.

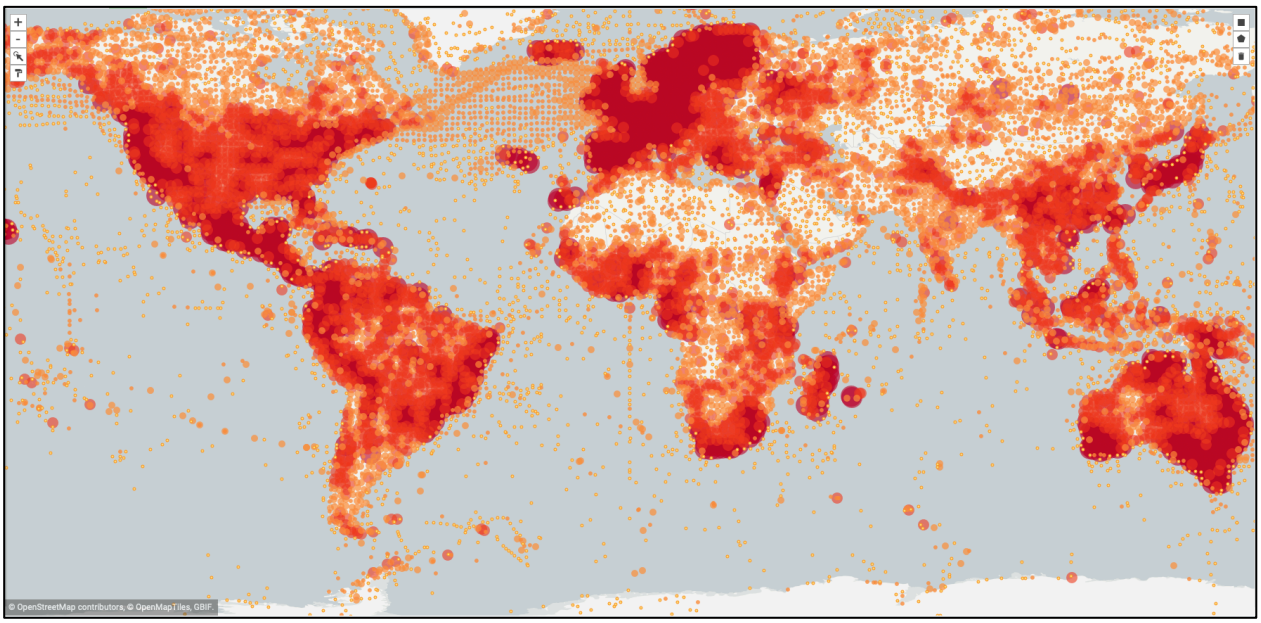


Figure 5.9: Visualization of the geographic density of 231 million occurrence records for plantae. https://www.gbif.org/occurrence/map?has_coordinate=true&has_geospatial_issue=false&taxon_key=6.

The Secretariat's informatics team has evaluated options of deploying its whole system to a cloud platform but found no evidence of an economic benefit at this time. It is currently exploring the application of machine learning to "Identify Species Using GBIF-mediated Datasets" (Robertson et al. 2019) and may well find important new purposes for GBIF data, but also new ways to improve volume and quality of data mediated by GBIF.

This will most probably not result simply in another software module of the portal or tool for users or Publishers, but is a long-term endeavor and major effort. Reliable or at least helpful (semi-) automation of species recognition relies on skillfully tagged, carefully managed, and specific training datasets (Schmid et al. 2016) as a minimum prerequisite, and in the case of natural language analysis[111] to analyze biodiversity literature, application of huge computational power.

5.1.4 The Alliance for Biodiversity Knowledge

The preceding chapters showed that the world of biodiversity informatics is complex and full of unknowns. In order to navigate this landscape, its inhabitants need to share a common conceptual framework as much as possible. Such a framework would have implications for information and data models, standards and interfaces, efficient sharing of tasks, agreement on priorities of further technical developments, and consistent communication among each other as well as externally (e.g., with funders).

The second GBIC conference in 2018[112], noted that "Stakeholders have been able to develop a common vision for aligning their activities [...]. Nevertheless, the complexity of the domain [and a number of other factors] has limited progress in making such alignment." (Hobern et al. 2019) The conference therefore focused on "a coordination mechanism for developing shared

[111] "Microsoft is investing $1 billion in OpenAI to support us building artificial general intelligence (AGI)" https://openai.com/blog/microsoft/

[112] https://www.biodiversityinformatics.org

roadmaps for biodiversity informatics. GBIC2 attendees reached consensus on the need for a global alliance for biodiversity knowledge". GBIF has been asked to lead this coordination effort.

However, the views on the *concepts* underlying the biodiversity data challenge differ considerably between the scientific and the bioinformatics community. Compare, for example, in Figure 5.10, next page, the graphs from the "Forward Look" report (GBIF 2010a) and the Global Biodiversity Informatics Outlook (GBIO) framework (Hobern et al. 2012). (The latter has already been offered for comparison with a graph from OECD 1999, in Chapter 2, Figure 2.1.)

The figure from the "Forward Look" report, here at the top of Figure 5.10, revolves around scientific and methodological terms and mentions two tools and one application. The GBIO framework, Figure 5.10, bottom, builds on some of the scientific and methodological concepts and adds concepts from the information and informatics space, and identifies some socio-cultural issues.

Both views – and possibly even more – will need to be taken into account by the emerging Alliance to ensure the completeness and usefulness of the future biodiversity information infrastructure. However, when trying to predict the possible speed and impact of this Alliance one needs to consider the inertial forces slowing institutional change and transformation of scientific culture. For example, as it is cultural change that needs more time than, for example, technical developments, it is not too surprising that a single global system for "data cleanup", as it had been envisioned in 2012[113], has not materialized even today.

The most sensitive issue, one that appears addressable in the context of such an Alliance, is that of reducing the duplication of effort and *unproductive* competition. This would include a consensus on which institution should contribute what kinds of data products and analysis services to the various recipients, and who should build and maintain the information systems' building blocks providing the basis for comprehensive, global "biodiversity knowledge". This discussion would also lend itself to discussing and prioritizing activities regarding the security of this infrastructure and a determination of which of its elements cannot be allowed to fail.

The Alliance would also have to be a clearing house or at least provide a coordinating mechanism for much needed clarity about such issues as which objects are needed to be citable and how, or to be identifiable as unique – objects such as observations, persons, taxa, and countries, that all come with variants of spellings, synonyms, and multiple submissions of the same term from different sources. However, the Alliance would also have to design and drive social instruments to enable and expedite such technologies, e.g., by seeking the contribution and commitment of data infrastructures and journal publishers, and other organized expert groups. The example of the adoption of data DOIs shows clearly that a consistent and complete adoption of all these identifiers might be a matter of decades.

Considering the so-called replication crisis, one of the potential aims of the Alliance, which would be most valuable and necessary for good scientific practice, may be to establish full traceability of the provenance of data, from knowledge and data products back to its sources as specimen or camera pictures, including all processing steps performed and contributors involved.

---

[113] "The next step will be to agree [...] how data cleanup efforts can be recognized and valued, [and a number of other measures] will be the first step towards making distributed data curation the norm." (Hobern et al. 2012, p. 16/17)

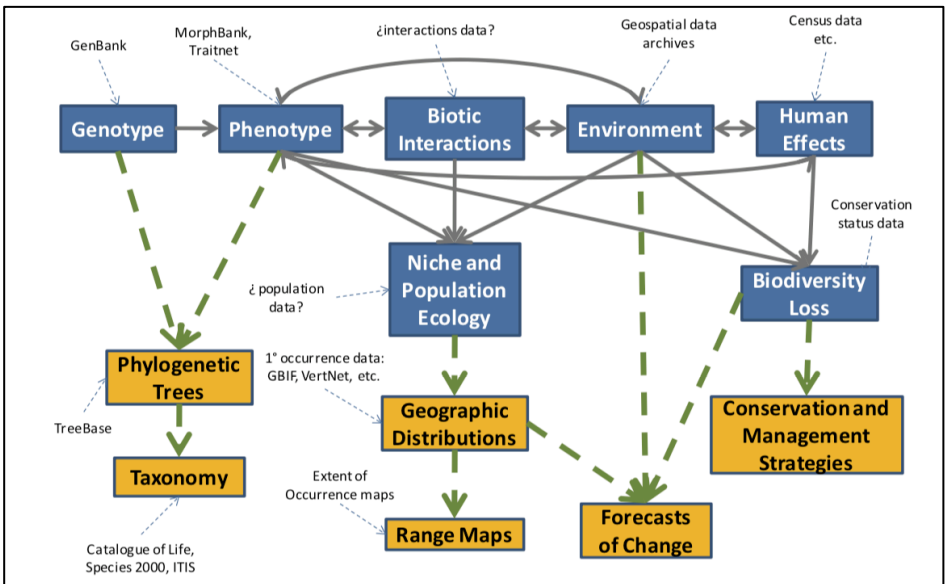


FIGURE 1. PARTIAL SUMMARY OF THE INFORMATION WORLD THAT SURROUNDS GBIF AND ITS EFFORTS TO SERVE BIODIVERSITY DATA. THE BLUE BOXES INDICATE ASPECTS OF BIOLOGY AND BIOLOGICAL PROCESSES. THE GOLD-COLORED BOXES REPRESENT INFORMATION PRODUCTS THAT GBIF AND OTHERS MANAGE. FINALLY, THE BLACK-LETTERED LABELS PROVIDE PARTIAL SUMMARIES OF INITIATIVES ASSOCIATED WITH EACH OF THE ISSUES, FROM PETERSON ET AL. (2010).

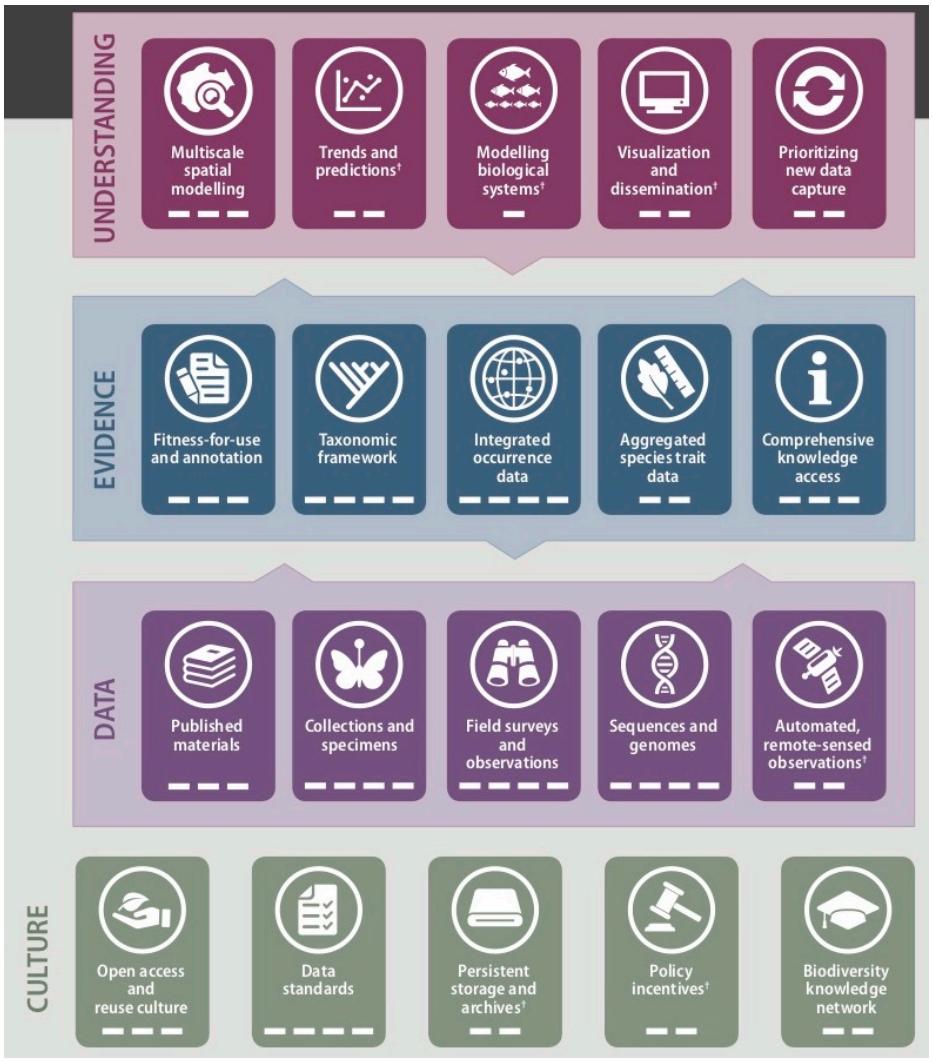


Figure 5.10: Conceptual diagrams from (GBIF 2010a), top, and (Hobern et al. 2012), bottom.

5.2. SWOT Compilation[114]

5.2.1 Strengths

- The most frequently identified strength was the *unique* and rapidly growing volume of occurrence records and their global coverage, both in terms of geolocation as well as taxonomic groups, sources, and potential to serve many kinds of interests. In many cases, this was described as unrivalled or even without a plausible competitor. This global reach was also identified as a means to repatriate data, such as from European museums to the biodiversity-rich countries.

- GBIF has created a stable, reliable, robust, and open data infrastructure resource that is widely used, globally. It has become an indispensable element for research and applications that use biodiversity data.

- GBIF's role has evolved over its nearly two decades of existence. The organization is now broadly viewed as doing an excellent job as an aggregator and enabler of discovery of biodiversity data.

- The portal and discovery tools are effective, reasonable, and up-to-date. They present data in a way that make them useful to many stakeholders and many individuals noted that they were incredibly impressed with the technologies GBIF has incorporated and the functionality included.

- Globally among key stakeholders, particularly with data producers and publishers, GBIF has gained a reputation as a neutral facilitator and mediator of diverging interests and opinions. This was instrumental in uniting the biodiversity community on the Darwin Core metadata standard (developed and maintained by TDWG) and is the background for the expectation of many people in the community that GBIF will need to lead the next steps on standardization of the data technologies.

- Importantly, GBIF data, the portal itself, and the accompanying software tools are free and openly accessible. As one interviewee said:

  > *Centralization of the information and its free access, bringing together ever more institutions into this principle is one of the greatest victories of GBIF, earning the respect of the whole community, positioning itself within the community as the reference institution on standards about how to publish: License of data, implementing of DOI.*

Most interviewees who expressed a view on GBIF's implementation of data policies and technologies, perceived that its role in its current scope is secure. More detailed views on strengths were:

- Technically, the web portal and the APIs are solidly constructed and are being widely used. The IPT, technology pioneered by GBIF, represents a major advance over earlier software tools and enables nodes, museums, non-governmental organizations and others to easily publish biodiversity data and make them more readily available to an array of users in support of research, conservation, and decision-making. Interviewees

---

[114] This section provides the summarized, but also individual SWOTs of GBIF, as provided by the interviewees. It is not an analysis and attempts to convey the original language. For an explanation and rationale of this format, see section 1.5.

reported that other biodiversity institutions and infrastructures have adopted the IPT for internal use.

- Multiple interviewees named other tools, such as the data validator tool and species matching tool, as very helpful for quality control of the data. Documentation of these APIs is well-written, logically organized, and user-friendly.

- GBIF enables deep and broad access to biodiversity specimen and occurrence data and information. The aggregation of museum data has been very successful and an immense amount of comprehensive and "*mostly high-quality*" data has been coalesced by GBIF. GBIF has also incorporated a large amount of observational data in addition to specimen data. The organization is especially strong with respect to citizen science observations (e.g., eBird and iNaturalist).

- GBIF data have been particularly valuable for supporting the modeling of species occurrence data, leading to many scientific papers and influential reports. One researcher observed that *"The field of niche modelling has been transformed by GBIF."* It was noted that "*members of the NSF research community use GBIF information widely".* A number of researchers stated that GBIF is unparalleled as an access point. Individual interviewees praised GBIF with comments such as: "*GBIF is the first stop for biodiversity data and information*", one individual noted that "*GBIF is the internationally trusted biodiversity authority serving researchers, educators, government, and the general public.*" A couple of individuals noted that GBIF holdings are not as comprehensive with respect to rare species, but do a good job with everything else, including invasive species.

- GBIF is expanding its support for biodiversity data to include links to sequence data and environmental samples. One museum representative noted that *"the idea of getting all biodiversity data consolidated is a dream and that dream is GBIF's strength".* Thus, several interviewees commented positively on GBIF's recent efforts to include further biodiversity data types. However, particularly so in Northern America, the expansion of GBIF's focus was also somewhat controversial and some interviewees were concerned about perceived scope creep. They suggested that GBIF only continue doing what it does well.

- Data usage reporting for tracking citation, attribution, and impact has been extremely beneficial. Several interviewees commented positively on recent efforts with respect to tracking citation and attribution, and, in particular, using DOIs for data so that providers get attribution. For example, the country-based analytics and pages that highlight Canada's contributions are viewed as being very helpful in justifying support for both the national node and GBIF. Numerous representatives of collections and other biodiversity organizations commented that the reports on data usage in the scientific literature are extraordinarily valuable. Others noted that "intellectual property rights are a big deal and GBIF has been really clear about who owns the data that are shared". GBIF was commended for being behind the creation of data papers so that data contributors get credit for their work. It was noted that "metrics demonstrating the impact of GBIF on scientific research will increase exponentially in value over time."

5.2.2 Weaknesses

- Most named were "quality and gap issues" of all kinds: from faulty or imprecise taxonomy or geolocation, to bias due to the "range" of observers (e.g., near cities) and their favorite observation targets (e.g., birds), and to gaps in (national) participation. It was acknowledged that the root of these quality issues does not reside with GBIF, but

with the data producing (e.g., scientific or museum) communities. However, many regarded it overdue that GBIF build better means to enable feedbacks from data users to publishers and contributors, so that at least errors can be corrected. It would mostly be up to scientific communities to correct biases and close gaps.

- Darwin Core (and the IPT) provide just for a bare minimum of functionality and descriptive depths; it might even create opportunity for faulty interpretation due to ambiguities or lack of precise definitions. The problem of course is that different communities require different enhancements to the standard. For example, marine biodiversity requires intimate association of environmental data with the occurrence data (see OBIS extensions[115] to Darwin Core), while some from the museum and collection community mentioned that GBIF should engage in consolidating the landscape of collection data standards.

- Many individuals expressed the "far too many holes" or gaps in the coverage of taxa, regions, and countries as a weakness of GBIF. This regards the paucity of data for various taxonomic groups, as well as the lack of coverage for Brazil, Russia, India, China (BRICs) and parts of Africa, South America, and Asia in general.

- In the summary from Asia and Oceania, it was noted that current GBIF data only covers a narrow scope of biodiversity data (occurrences). GBIF had only raw and un-curated data or data products, and it had not yet devoted efforts to link its data with other types of biodiversity data, such as remote sensing, genetic, evolutionary and trait information, and data collected from ecological surveys and other environmental initiatives.

- Some interviewees pointed out that there were unexploited sources, such as from governmental agencies (e.g., environmental, fishery, and forestry data). Some Node managers explained this absence of relevant data as rooted in insufficient resources in combination with the lack of networking links of their academically oriented hosting institution.

- There is a wealth of raw biodiversity data confined to biological collections of different institutions, yet they are not digitized.

- Several collections experts noted that taxonomy has always been a real weakness but recognized that GBIF is now gaining expertise in this area, although there is still room for improvement. The ITIS/COL consortium was viewed as having potential to improve the situation. Some claimed there is a key need to develop a standard taxonomy instead of attempting to crosswalk all taxonomies, but interviewees particularly from Latin America noted that existing taxonomies do not fit their countries' needs.

- Many interviewees questioned how GBIF can better push information back to the collections (e.g., cleansed data, controlled vocabulary, usage metrics, and the like) in a way that supports, helps sustain, and improve sources. Some frustration showed through about missing reintegration back into node databases so that "data cleansing efforts do not need to be repeated time and again". In particular, GBIF does not have automated infrastructure for providing feedback to data publishers; someone from GBIF must manually provide feedback. The problem applies to both specimen and observational data.

---

[115] https://obis.org/manual/dataformat/

- Some interviewees from North America noted that the IPT could use a more robust coding base and that it would be beneficial to transition to a more open, community-based system.

- Some individuals conceded that there are concerns about data quality, but that the responsibility for data quality lies largely elsewhere. Others complained that "GBIF tends to consider that the data publishers have the responsibility to verify the data they upload". One such interviewee provided this example:

  > *Moving into the future better algorithms are needed to check for errors, for example the entire dataset of a Tasmanian research organization is erroneously relocated to the Antarctic, and specimens stated to come from Fish collections, and with a Fish Family are labelled as Birds, and many species are inaccurately georeferenced or have old synonyms, which can erroneously inflate species richness in analysis. Some of these errors result directly from algorithms designed to prevent them, and more care and attention is needed to ensure these are filtered out before the data reaches users.*

- There may be pronounced issues, especially in underfunded institutions or regions, that those who collected data at the source are not trained to observe data quality standards, causing frequent and common data quality problems.

- Others noted that GBIF is working hard to promote and provide tools to data providers so that they can clean the data, but also recognized the need for even better QA/QC tools.

- Some data providers, especially those who were pressed for resources, noted that it is still hard work – despite the availability of the IPT - to prepare and publish data. It was pointed out that it is not possible to incorporate certain aspects of the data "under the GBIF format" (e.g., environmental or physiochemical parameters). Stakeholders with this type of data now have to store and publish them somewhere else.

- The data standards and the requirements for data formats are too professional and complicated for beginners to master. This introduces difficulties to potential data contributors, especially those from developing countries, where advanced data infrastructures are not in place.

- A few individuals expressed concern over the quality of the increasing volume of citizen science data that is available through GBIF.

- Extensions to Darwin Core are managed by GBIF in ways that are mysterious – ideally, this would go through a certification methodology and community vetting.

- Some researchers complained that in several cases the metadata is too verbose with no easy entry point and, in other cases, it is not detailed enough.

- Improved georeferencing technologies are needed so we can get better error bars on locations in order to extract the most information from the specimens.

- A few individuals commented that they would like to see the GBIF infrastructure replicated more internationally (such as the Atlas of Living Australia), support customization down to countries, ecosystems, states, and counties, and include more environmental data as well as non-traditional natural history collections (e.g., cultural objects, such as objects made with bird feathers).

5.2.3 Opportunities

- GBIF starts from a unique point regarding volume and coverage of occurrence data. While going for even more (only 10% or so of life is known), it could leverage its standing by helping the community to address the quality issues through feedback channels and to broaden the depth of information on each occurrence.

- Various suggestions about what GBIF should hold in addition to occurrence data or even to produce, were, for example: precomputed species distributions; automated reporting for local or regional governmental requirements (e.g., for impact assessments); quality controlled subsets of data; one globally authoritative listing for all taxonomies; information of interest to the public, such as images; and other suggestions.

  In particular, a considerable number of interviewees thought that GBIF must be involved in defining and producing EBVs, or even [the UN Sustainable Development Goals] SDG indicators, in order to provide critical solutions to applied global challenges. In this way, GBIF would serve as "the most inclusive network of biodiversity datasets".

- GBIF would also need to lead the discussion on standards. Using its good reputation, GBIF should lead the discussion of the Alliance on (meta)data: ecologists need co-occurrences and marine ecologists need environmental data associated with occurrence. Of general interest were "abundance type information", genetic information, and taxonomic checklists.

- A number of interviewees suggested that GBIF should move beyond the role of passive collector of data to be more strategically proactive in working with data providers to set priorities for data collection, digitization, and aggregation globally. For example, where are the geographic biodiversity hotspots – the pollinators, bats, or invasive species?

- A number of interviewees wanted to see GBIF directly tackling more scientific research problems but recognized that it would need more staff to address that and provide transparency, following international best practices. For example, GBIF could focus community efforts on one or a few high-profile, high-impact biodiversity science use cases that have real-world applications. Across the board, interviewees recognized that GBIF was supporting the aggregation and discovery of data for research, conservation and other uses, yet only one or two people were aware of specific successes.

More detailed suggestions included:

- Support broader data integration: Provide linkages from specimens (and observations) to other types of data—e.g., molecular, ecological, and geospatial data and information.

  - One researcher noted that "It may be beyond the scope of GBIF to provide resources to co-locate some of the data, but it would be really amazing to see species-level data (GBIF) joined with, say, trait data (TRY, FunGUILD, AMPHIBIO, and the like), species-interaction data (GLOBI), phenology (USDA phenology database), etc."

  - There is a need for GBIF to integrate microbial taxonomic and genomic data, especially given the importance of microbes for human health and ecosystems. It may prove useful to form an alliance with the Genomic Standards Consortium (GSC).

  - There is a need to link to lots of other types of ecological and environmental data, and to partner with corresponding infrastructures and projects; much of the data is collected by citizen scientists and environmental observatories and research networks. GBIF has the opportunity to bring all kinds of different types of data into a single infrastructure.

  - Add middle-level data integration services so that collections can be managed online collectively and one can relate specimens (organisms, parasites, or blood) between collections. This would assign a GUI at the organism level and different objects could be related to a particular organism.

  - Add a location service that levels the playing field for all types of data management (shape files, grids, etc.) including spreadsheets. This could lead to better spatial data integration across the network, including supporting the repatriation of specimens.

- Some individuals suggested that GBIF lead a targeted gap analysis to ascertain what data gaps exist and where and how we could get better data to fill in major gaps. Comments and suggestions related to this issue included:

  - *"You can make huge progress if you get 95% of the literature. A possible solution is to get floras of published records and scan and capture these data. You will then have the world's knowledge handy."* This interviewee was especially interested in ecology and conservation and thought that much information (species info at fine spatial and temporal resolution) is in parks, nature centers, and other such venues; and would like to see GBIF visit all of these sites and empower/encourage them to contribute their checklists of plants, bees, vertebrates, and other classes that are available. Ideally, the interviewee would like to see GBIF have the network of people contributing both literature and specimen data; maybe a million sites (thousands of sites, millions of people) around the world so that you could determine loss of species in real time and address questions such as: "Why are the insects disappearing?"

- Develop a mechanism of collecting data gathered for theses by university students.

- Work with universities and research institutions to ensure high-quality, accurate biodiversity data.

- Media collection and high-resolution scanning of specimens and fossils are generating massive volumes of data; if GBIF can promote or mediate shared storage across participants, that would be a win to bridge the "feast or famine" times.

- Take leadership in creating access to a global, authoritative taxonomy. Story: Lots of organizations are trying to create a global taxonomy and build an authoritative taxonomic backbone ("Catalog of Life Plus" - EOL, Barcode consortium, and others). Eventually, we will get to the point where groups can access one portal for taxonomic authority. GBIF has long-term support and can do this sort of thing.

  - Possible solutions included getting more taxonomists involved in GBIF and enhancing the capacity to separate out erroneous spellings.

  - However, some individuals felt very strongly that GBIF is needed, but that it does not replace taxon-based networks that ideally should feed into something like GBIF. Concerns were expressed that taxon-based networks have fallen or are falling by the wayside and yet they are critical for enhancing the quality of data

in the network; taxonomists are increasingly being left out of the picture; and GBIF cannot maintain taxonomic data quality and authority. A possible solution is to reconstitute taxon-based networks such as MANIS or HERPNET. and then have them submit data to GBIF, possibly through intermediaries such as VertNet.

- Take the lead on planning the future of biodiversity informatics. Ascertain community needs and challenges to be solved in the future and develop a community-wide strategic plan for doing so.

- GBIF could even help the museum collections community to define a common platform; even DiSSCo can take advice from GBIF.

- Digitize key data kept in colonial museums, make it available to the countries of origin, and support biodiversity-rich countries in their collection digitization projects.

- GBIF could improve data quality via automation and by pushing clean data back to the community. Lots of work goes into cleaning data, but this work is done independently and repetitively by researchers and others. Several interviewees suggested that GBIF re-think the data publication process.

- It was also noted that satellite data and new tools could be used to improve the accuracy and precision of georeferencing, which is a big part of data quality improvement.

- Develop a global schema/ontology. Some of the more technically adept interviewees noted that Darwin Core (DwC) is a least common core and key semantics are often removed when using DwC. They would like to see a global schema/ontology created so that each community could publish their data and requisite metadata would not be lost and retain its full richness. A global schema would allow data contributors to "publish once, harvest many times".

- Enhance stability of APIs. A couple of individuals noted that GBIF's APIs could be improved and made more stable.

- Open up GBIF infrastructure for community development. Several data content providers and thematic aggregators noted that they would like to see GBIF infrastructure open to community development.

- It was suggested that it would be great if GBIF could accommodate related community CI [corporate identity?] that supports particular themes and services (e.g., replacing VertNet, enabling trait search, etc.). The reasoning is that if different views could be added to what the GBIF base infrastructure already does, then there would be fewer siloes, and more data QA tools and other advanced services available.

- Provide support for high-powered visualization. A couple of individuals thought it would be great if every time someone requests data, data were automatically analyzed and served as an ecological niche model; GBIF could partner with LifeMapper or Map of Life to do this. In supporting such analyses and visualizations, the interviewees felt that this would promote more open access to data and interoperability. One individual suggested that visualization tools could be provided for additional specific purposes such as animal telemetry visualizations.

- Support citizen science data. Getting a digitized collection of image data via a crowd-sourced platform would be a major opportunity. One individual noted that iNaturalist and eBird account for a large and growing percentage of the new data coming into GBIF.

Also, most "new" specimen data is coming from the digitization of old data. These observations raised questions about how users can verify the quality of both the new citizen science data and the newly digitized "old" specimen data.

- A couple of individuals noted that there were opportunities to present data to researchers and other users so that the data could be more easily understood and used.

- A couple of individuals thought that GBIF could mint DOIs for a variety of purposes, such as geological sample numbers (IGSN).

- Tailor the GBIF portal to meet local to regional to continental and thematic needs. Develop a way to flexibly tailor the portal to address local, national, regional needs or thematic needs or university/state needs. In this fashion, the GBIF brand could benefit local and other entities. GBIF would essentially provide a potential home for more integrated databases for ecoregions, biomes, and systematic reviews. The hope is to evolve by bringing together technical folks to look at how separate CI [corporate identity?] can be more closely matched, using the same APIs, software stacks, and controlled vocabularies. One interviewee, for example, questioned if there is a way to link with GBIF to create a portal or research hub for legumes? If so, taxon-specific researchers could benefit and contribute to higher quality data overall.

5.2.4 Threats

- There is a huge and increasing amount of observational data in the GBIF database; roughly 60% are bird observations. GBIF has not thought strategically about how much of its holdings should be specimen-based versus observations, nor the impact of explosive growth of citizen science on GBIF products and services, funding needs, or staffing.

- There will be a major change in technology that rapidly captures all data. Although no one could conceive of what this technology might be at present, this threat was perceived as a potential issue. One interviewee noted that "Google Life" would be a huge competitor if Google were to move in that direction, although it was also recognized that Google would have a serious taxonomic impediment to overcome.

- Threats related to competing infrastructures:

    - Not advancing beyond Darwin Core and not being able to solve connectivity among biodiversity data types (e.g., genomics and metagenomics data) and identifiers other than DOIs. For example, the Map of Life has already taken initiatives and worked on developing data standards to deal with survey data and inventory data.

    - Appearance of *national* initiatives of free access data, which might compete and prevail over sending the data through the GBIF portal (and might in turn also weaken the national nodes).

- Cyber-attacks, e.g., of the simple greedy type (extortion / encryption trojans) or more elaborate attacks, which would undermine the credibility of GBIF.

5.3. Conclusions

The GBIF data infrastructure is viewed broadly as being a major success and greatly preferable to alternative solutions. The 1999 OECD expectations that are still relevant to GBIF (see Chapter 1), insofar as their fulfillment can be influenced by GBIF at the current level of resources, have broadly been met by GBIF. It appears as the most comprehensive, global, application-agnostic (most unbiased), easiest-to-use, and modern access point to known digital occurrence data.

The main quantitative target set by the OECD regarding the percentage of digitized specimens has not yet been met. However, this is and will most likely remain outside the responsibility of today's GBIF (see Chapter 2). Also, it cannot be ruled out that major quantitative and qualitative gaps exist because sub-communities or institutions still reject sharing their data or are unable to share their databases (fully) due to resource issues.

Of course, there are caveats or perceived weaknesses and areas for improvement, but some of these can also be assigned to unrealistic expectations at the current funding level (which is probably unknown to most interviewees). GBIF may even be a victim of its own achievements. The sum of all "opportunities" named does not only encompass most or all expectations named in the OECD report of 1999, but even more. Also, many of those who supposedly "knew" GBIF asked for functionality which is already there or ascribed responsibility to GBIF where it (currently) does not belong, e.g., on the taxonomy.

The two most frequently named, and – in our assessment – most difficult and most important areas of urgently needed *visible* improvement were:

- better feedback mechanisms, not just to publishers but to those scientific communities that could rectify issues (this is only partially a standards and technology issue, but to a larger extent a community and funding issue),

and

- closing gaps in occurrence data, perhaps even through tapping *different* sources of *existing* data. (Within today's remit of GBIF, this is mainly a Participation and communities' issue, see Chapter 6. Considering the OECD expectations, there is the – still huge – potential of collection specimen and the "historical" biodiversity literature, both of which are standing at only about 10% digitization!)

Actually, the GBIF Secretariat, but also some Nodes, are working on many of the opportunities named by interviewees (see the GBIF Strategic Plan 2017-2021, reproduced in Annex 10.3, and GBIF workplan 2019 / Implementation Plan 2017-2021 – Google search term: "site:GBIF.org workplan 2019"). However, most of these activities receive a minimum of resources (at the Secretariat, equivalent to a few person-months per year, for each activity) and depend in many cases on co-funding by third parties and on the co-operation of other organizations – some of which are even unfunded initiatives, such as TDWG.

It is debatable whether some of the threats named by interviewees are to be taken as such. If properly handled, most of the developments singled out may actually be transformed into opportunities, or even weaknesses to be addressed. The threat of cyberattacks, however, is very real and on the rise. Unfortunately, in most decisions about priorities, security issues are deferred to the next round of funding … . It must be quite clear to everyone in a responsible position that as infrastructural elements of science such as GBIF become crucial, such infrastructures become responsible to maintain the integrity of data and the availability of services.

In the following sub-sections we discuss additional conclusions in detail, following the classification provided in section 5.1.

5.3.1 Data Volume and Quality

GBIF has certainly succeeded in becoming *the* global access point to occurrence data, by volume, accessibility, and usability. The volume of data is unrivalled. The standards employed and the infrastructure built offer many options for sufficiently rich metadata and selection mechanisms to enable quality assessment for many – but certainly not all – biodiversity-related scientific and applied purposes.

For reasons already analyzed in Chapter 3, anything other than working with open data (data with one of a few liberal, machine-readable licenses) is very impractical if not impossible. We can state that GBIF has succeeded in this area by making it unambiguously clear that only open data are worth considering for GBIF and, by implication, the biodiversity informatics landscape as a whole.

It appears quite certain that, as all previous reports about GBIF insist, it will be able to mediate ever more data. This would come about through a number of factors such as increasing Participation (in particular, from the still underrepresented biodiversity-rich countries), adding more national contributors and potential Publishers endorsed by existing Nodes, inclusion of occurrence data from genetic databases (GenBank and BOL), further increases of citizen science programs, quantitative successes in digitization of collections, and automated analysis of the literature. In view of this growth, it appears even more urgent to address the questions about quality as soon as possible, whether through actually improving quality, publishing quality-controlled datasets or products, easing the task of assessing data quality, or working on the awareness of users.

There are "quality issues" with a considerable proportion of the data volume. This means that there are errors, inaccuracies or missing information in individual fields or whole datasets, which would render these records or datasets unfit for specific purposes. In the case of some, but certainly not all records or datasets, they might be unfit for most purposes. Considering that problems of quality are unavoidable in any data collection – for reasons fundamental as well as practical – the major questions are: Is there enough ancillary information? Are there helpful tools to enable a user to assess the fitness of data for an individual purpose? Do users know about this information and the tools (and how to use them)?

Though there is certainly a way to assess quality for many purposes, as proven by actual usage (see Chapter 3), it is out of scope for this review to provide a systematic analysis of what is actually missing for which specific purposes – although it is well known that a number of large institutions, such as CGIAR, continue to do much of their work with their own curated datasets.

GBIF itself has commissioned reports and contributed to workshops and initiatives on the quality issue in general (e.g., the TDWG interest group on "Biodiversity Data Quality") and as related to specific purposes (e.g., the "Report of the Task Group on GBIF Data Fitness for Use in Agrobiodiversity"). The portal now displays a plethora of facets enabling selections based on quality-related criteria (and downloaded datasets contain such criteria for "local" cleanup by users). A large number of these criteria are composed of flags resulting from plausibility checks devised after these inquiries with user groups and with the data community.

Two important types of data stand out in the quality discussion. First, the credibility of citizen science data – comprising a huge percentage of all GBIF-mediated data – is being questioned by a number of interviewees, although there is evidence (see section 2.1) pointing out that such data can be research-grade and GBIF appears to work with Publishers providing scientific guidance to citizens. Second, it is obvious from GBIF "Trends" that a very large percentage of

specimen occurrence records – historically deemed very important – do not provide machine-readable geolocation data at all (insofar they may have location information, georeferencing seems to have not been performed).

5.3.2 Gaps and Biases

> ***Box 5.2: Wim Hugo, Chief Data and Information Officer, South African Environmental Observation Network***
>
> *GBIF does an excellent job of curating the data gathered by multiple channels, but has little power to influence biases or deficiencies in spatial, taxonomic, or temporal cover.*

A pure growth in data volume is not sufficient for many purposes, as has been shown in Chapter 3 and section 5.1. While GBIF cannot mediate data that do not exist, it could contribute in many ways to address gaps and biases in data. Much of this would need to be done through better Participation and the instruments of "Capacity Enhancement" and other community-related activities, which are dealt with in Chapter 6. We enumerate some of the data gaps and biases here as well.

Addressing the conspicuous empty areas in the GBIF Participation map might bring about a significant reduction in uncertainty for many purposes. For example, it might help mediating data on pollinators, which were deemed "insufficient to draw general conclusions" (IPBES 2016).

In Africa, the capacity enhancement programs, especially BID, have brought about a remarkable increase in national Participation. This is just a stepping-stone into African biodiversity; it has laid a seed for access to existing data and perhaps even increased primary data capture. Although it would certainly go beyond the means of GBIF to handle large monitoring or survey initiatives, it could help establish networks between African Nodes and funders (e.g., from OECD countries) and help prioritize their activities by leveraging their good standing with these networks.

In contrast, many Asian countries may already have the human capacities and even the data but are not yet convinced of the merits of having their data "mediated". Here, as in Russian speaking and in Latin American countries, the language barrier pertains not only to the GBIF portal, and its manual and tools, but also to much of the existing (meta-)data in these countries itself. Some countries may need funding to make them available to others (in particular, to richer countries), whereas others may need to be convinced of the benefits before bearing the cost themselves.

It may be asking too much of GBIF to fill true gaps within participating countries – see Box 5.2. But it might prove worthwhile – even monetarily – to enable each of the GBIF Nodes to reach beyond their intellectual and disciplinary "homes" to find more data contributors and publishers from their country.

Finally, since any one of such activities would not only require considerable funding, but also take up valuable capacity in the Secretariat (if only to acquire the funding), a thorough analysis of priorities would certainly be necessary. The authors of this report are not convinced that a more or less technical analysis of the existing data (an "Implementation of gap analysis modules for coverage in terms of taxonomy, geography, and time for GBIF-mediated data resources", GBIF 2010a) is a sufficient or even a viable first step. For example, is there any

metric available to determine sufficient coverage in all these dimensions? Would one metric be sufficient for all purposes?

5.3.3 Standards and Technology

GBIF has done an excellent job of turning a community agreement on a standard, Darwin Core, into an operational service and supporting it with related tools – most importantly the IPT and DOIs for data. The staff of GBIF have done this so well, technically as well as in the way they pursued this as members of a community, that there is the almost universal expectation among interviewees that they could repeat this feat by extending the semantics of Darwin Core or for a range of other standards, vocabularies, and ontologies.

The usability and other features of the GBIF portal and the appreciation of these features have grown substantially over the last years and can be expected to continue to do so as the GBIF informatics group has many new features in its pipeline. The expectations on more than incremental improvements – namely, on access beyond "raw" occurrence data – are numerous and far reaching, however. A specific and probably very noteworthy proposal from our interviews is laid out in Box 5.3.

> ***Box 5.3: Dr. Simon Ferrier, Australia, Senior Principal Research Scientist at CSIRO Land and Water, and co-lead of the GEO BON Essential Biodiversity Variables task force:***
>
> *Some people complain that GBIF has failed to show the importance and usefulness of biodiversity data. Scientific teams like ours who are working on global biodiversity modeling and developing biodiversity indicators, can help GBIF to demonstrate the usefulness of its datasets. Instead of modeling the distribution of individual species (in which data gaps in geographical coverage can be problematic), our models focus on collective properties of regional biodiversity, which are more robust in face of data gaps. We have developed indicators at fine spatial resolution across the entire land surface of the planet, which assess the representativeness of biodiversity conservation areas and the impacts of habitat transformation. These can help policymakers set priorities for biodiversity conservation and monitor regional biodiversity losses and gains.*
>
> *Regarding the use of biodiversity data, I think the biggest current problem is not poor data but limited communication between scientists and policymakers. Rather than promoting data in its raw form to policymakers, GBIF should instead promote the use of such data for deriving policy-relevant biodiversity indicators. People like us can help, but to do so we need to forge stronger formal connections or partnerships with GBIF. GBIF should partner with other initiatives, such as GEO BON, so that they can go hand-in-hand in talking with policymakers.*

Some interviewees explicitly and in our view, rightly, warned against "scope creep", i.e., unplanned or unmanaged growth. Aside from the resource issue, this could potentially and inadvertently intrude on other organizations' turfs. Others feared, obviously in anticipation of successful growth of funding, more bureaucratization; that is, they seem to advocate healthy growth that sustains the current spirit of GBIF. We concur that avoiding these traps is a critical factor of success, but conclude from the intensity and wide range of suggestions that the area of standards and technology will be under the biggest pressure for extension.

Some interviewees suggested that GBIF delegate some of the development or open it up to the community. GBIF currently works closely with programs such as the Atlas of Living Australia and with the Catalogue of Life Plus consortium. We also see, however, that the activity of unfunded initiative groups frequently fades, after first meetings conclude optimistically. And finally, one cannot expect scientists, or even infrastructures funded as ("research" or "innovation") projects, to build and maintain infrastructure-grade software, tools, or standards – just prototypes, which then would have to be refactored and maintained.

There are certainly enough fundamental issues that could keep GBIF, at its current size, busy for years. Examples include the matter of duplicates, planned to be addressed by unique identifiers for observations, specimens and the like, and other issues of versioning. As new concepts are turned into standards, tools, and functions, and then implemented over time by the data publishers and contributors, the portal will have to deal with multiple versions and implementations of the standard, and still produce meaningful and understandable results for users.

This leads to a last remark about the area of standards and functionality. Many suggestions invite the conclusion that users do not know what is possible right now. Their SWOTs may be based on past experiences, they may not have invested reasonable time, lately, or the user interfaces of the portal and tools could actually be better. In the latter case, documentation or handbooks are very important – but seldom used.

All the previous detailed conclusions in this section may, however, pale in comparison with the impact of some accelerating technical and scientific developments, as described in section 2.3. GBIF will need to observe and adapt to these developments in parallel to the daily struggle with technical, cultural, and organizational problems.

Finally, in the area of basic IT and software frameworks, GBIF does a more than solid job of reliable development and operations of its databases and access portal. It is currently based on housing the organization's servers at the University of Copenhagen, with very good internet connectivity. At the current scale, there are no economic or performance benefits in moving GBIF's systems to "the cloud". This picture may change, however, if GBIF's functionality and systems were to be expanded into large-scale data storage or analysis, particularly if that were based on machine learning with the large demands of these methods on computational resources. This is a field where even well-funded global players in "Artificial Intelligence" research cannot succeed without very substantial help[116] from big cloud platform providers. A dependency of GBIF on these methods might force the organization's hand about deploying its services to the cloud and even about which provider to choose.

In summary about item 1 of the statement of task, we find that:

- GBIF has effectively contributed to the advancement of science, as evident from fast growing number of citations, but also from individual "success stories" and testimonials from researchers. However, its effectiveness could grow further and substantially, if the quality, or rather the assessment of the quality of the data mediated by GBIF would be improved, so that it could serve more purposes with greater certainty.

- GBIF has increased the efficiency in R&D spending, as evident from the large acceleration of individual scientists' work and its contribution to the adoption of standards, tools, and best practices, thus avoiding duplicate work in national biodiversity informatics infrastructures. However, while there are reasonably monetary benefits – potentially even huge benefits – to society, these have not been evaluated

[116] https://openai.com/blog/microsoft/

systematically – neither for GBIF, nor, as it appears, for biodiversity informatics as a whole.

- GBIF has enabled a variety of important applications beyond fundamental research, as evident from examples at the science-policy interface, health, and agriculture. GBIF's data growth has explicitly been named as a desideratum in Aichi target 19, and CBD confirms GBIF's direct contribution to Aichi Targets 1, 2, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 17, 19 and 20, and thus its value to protecting the environment and sustainable development. Due to the difficulty of discovering these use cases outside academic research (with its growing culture of data citation), there currently is no means available to evaluate this finding quantitatively.

In summary about item 4 of the statement of task, we find that:

- The "technical aspects of GBIF's delivery" are soundly and professionally handled – with the caveat that GBIF will need to retain the quality of its informatics staff, which will certainly remain non-trivial for the time being. More emphasis on IT security is advisable.

- The effort to sustain GBIF's operation as a research data infrastructure appears manageable and scales well with Participation – with the caveat that the capacities of Nodes need to match their conceptual roles, especially to support Publishers with respect to technical questions and to manage data quality.

- GBIF as an infrastructure is trusted by researchers and their infrastructural peers. There is a technical basis for this impression, but at least as important, a basis in GBIF's longevity and the professional conduct of its staff.

6 COMMUNITY, OUTREACH, CAPACITY ENHANCEMENT, AND TRAINING[117]

6.1. Description

The original OECD concept for a Global Biodiversity Information Facility foresaw the main role of a secretariat to co-ordinate participants' activities, from the development and agreement on standards to training activities, including the development of a curriculum. That vision, however, did not include GBIF providing a technical infrastructure and its development.

Although the mission of the GBIF of today has been adapted considerably (see section 2.3), this original concept may still shine through when we consider the first priority of the current Strategic Plan 2017-2021 (see Annex 10.3), the first activity of the corresponding Implementation Plan, and the 2019 Workplan all have a "focus on people", not just on data and technology. Even beyond that, all priorities and many work items are concerned with GBIF's relationships with the people and institutions involved with GBIF.

GBIF maintains a number of relationships with quite different groups of stakeholders, from its formal Participants and other funders to users of the GBIF-mediated data. Frequently, institutions, initiatives, and individuals who are first considered to represent one stakeholder group, are members of more than one of those groups. Some should actually represent all of these entities, vis-a-vis GBIF – as the first GBIF review recommended (CODATA 2005, p. xi). Thus, the loose term "GBIF community" may designate all these entities in the aggregate. It is still necessary, when looking at activities such as community outreach, to focus on the specific roles, interests, and needs of stakeholders (CODATA 2005, p. xi), and to establish their size.

Regarding their interactions, the stakeholders – most of whom have already been enumerated or described in detail in the previous chapters – can broadly be categorized in four clusters for the purposes of this review:

1. The GBIF core: Governance bodies and committees, Nodes, and the Secretariat.

2. Funders and influencers (both actual and potential): Various national ministries and agencies, (public) science funders and councils, intergovernmental bodies and organizations, foundations and charities, and industry.

3. Users, data contributors, and data publishers: Individuals as well as organizations – such as the intergovernmental bodies and others from cluster 2 – and projects (scientific or other).

4. Biodiversity informatics: Other infrastructures, IT departments of institutions that use, contribute, or publish data, standardization bodies, projects, and other initiatives.

There are some individuals or institutions that inhabit roles in all four clusters; the number of all individuals and organizations in each cluster ranges from the order of hundreds in the GBIF core to many more. Although there appear to be no figures, we have estimated several 10,000,

---

[117] This chapter addresses the non-technical issues of item 1 of the statement of task ("Review how effective GBIF has been since 2001 in meeting the expectations from the OECD working group"). It contributes to item 2 ("Review the governance and sustainability of GBIF as a global network and organization (including hosting of the Secretariat in Denmark).

or even more than 100,000 individuals (even excluding citizen scientists) and several thousand institutions of users and contributors, whether actual or potential. For example, IUCN alone claims "We have Member organisations and State Members in more than 160 countries and a network of over 10,000 voluntary scientists and experts spanning the globe."[118]

GBIF itself seems to define the term "GBIF network" to include our first cluster, plus the data publishers, when they say in the Implementation Plan 2017-2021: "GBIF is the result of work by thousands of people in agencies and institutions worldwide. This network's long-term sustainability … ", referring to the GBIF Network[119]. It may serve GBIF well in many cases to be not too specific when using the terms "GBIF community" and "GBIF network". But where appropriate, we will use "GBIF core community" for the first cluster, and "the wider (GBIF) community" for all four, or the latter three. (Note that some institutions, particularly funders, may object to being co-opted into a community explicitly and publicly.)

Of interest to GBIF are not just the relationships between the GBIF core and the latter three clusters, but also some bilateral interactions between entities from the three external clusters (e.g., science leaders or funding agencies to induce other institutions or projects to co-operate with GBIF). Each of these separate relationships may need to be taken into account for GBIF's strategies and activities. We deal with most of the specific relationships in the appropriate chapters of this report but need to delve into them in this chapter when discussing outreach strategies and activities.

In its Strategic Plan and the Implementation Plan, aside from its immediate mission to provide an infrastructure to the user communities, GBIF uses verbs such as "empower" and "support" for the GBIF network, while it plans to "co-ordinate", "engage with", "promote", or even "provide leadership" within the wider GBIF community, throughout all five major priorities of the plan. This planning thus aligns with the OECD concept (OECD 1999) as well as with various recommendations of the previous reviews (CODATA 2005, GBIF 2010a). Whether or not this also includes or represents "a strategic marketing approach" – one each for user groups and institutions (CODATA 2005)—may be debatable. There are very heterogeneous interests, requirements, and changing priorities within the GBIF community, and marked differences between regions and even countries. Therefore, the feasibility or utility of a single approach might be problematic.

Interviews with the GBIF Secretariat staff and the other experts indicate that the topics of this chapter are not just words in plans, but are very much on the minds of the Secretariat staff and the community. "People" issues are supported by roughly the same staff effort at the Secretariat as are data infrastructures, standards, and technologies.

Most of the effort by the Secretariat staff and from core funding concentrates on the existing GBIF network, or even on its core community. This must then reflect these members' needs and biases, and implicitly the gaps in Participation. The much-lauded Biodiversity Information for Development (BID) program – which provided an inroad on one of the most conspicuous taxonomic and geographic gaps in data and of GBIF Participation – is funded by a third party, the European Commission. This BID funding alone provides about a quarter of the yearly core Secretariat budget in addition to the regular Participant contributions. This amount thus appears to establish a benchmark for the cost of increasing Participation beyond OECD countries.

---

[118] https://www.iucn.org/régions, last accessed 2019-08-12.

[119] https://www.gbif.org/the-gbif-network

According to our interviews and observations, GBIF has successfully implemented an early recommendation (CODATA 2005, p. ix) to leverage capacities of national Nodes:

> *3. In capacity building, there needs to be much more emphasis on having Participants and nodes help each other instead of having the Secretariat as the focal point.*

In GBIF's capacity building programs, mentors and trainers from other Nodes, but also volunteers from other institutions are recognized to play a significant role. This seems to work both within and across regions. Beyond that, however, national Nodes have supported and embraced capacity building and training programs, as well as infrastructure for data publishing, outside GBIF and even outside their region, and acquired additional funding for this activity. (There are, of course, many noteworthy training activities by other stakeholders.)

For the remainder of this chapter we deal mainly with all the community interactions of GBIF besides those concerning data, biodiversity informatics, and standards, which have been dealt with in Chapter 5. However, it is worth noting some of the community interactions directly related to data, to be able to compare. As mentioned in previous chapters, GBIF commissioned reports in the 2010s, conducted workshops resulting in such reports, or participated in activities that led in a call for action by GBIF on a wide range of topics. These included data quality in general and in agrobiodiversity in particular, prioritization of digitization efforts, GBIC conferences, the newly formed Alliance for Biodiversity Knowledge, and "Living Atlases – an open community created around the Atlas of Living Australia platform."[120]. Many of these activities were preceded by or resulted in a call for leadership by GBIF.

For a group of GBIF Participants that contribute a large or even dominant share of the Secretariat budget, namely from Northern America and Europe, there is an issue easy to overlook or not to prioritize—the language barrier. Within those two regions, and besides the countries with a majority of native speakers of English, 22 of 32 European countries rank high or very high on the Education First (EF) English Proficiency Index 2018[121]. However, considering both the proficiency index and a recent statistic on the language of internet users, from Table 6.1, below, it appears obvious that any activity – marketing, capacity building, training, or technical support – with inadequate material and personal proficiencies in languages other than English may fail to reach a critical mass of stakeholders in those countries and even in whole regions.

[120] https://living-atlases.gbif.org

[121] https://www.ef.com/wwen/epi/, last accessed 2019-08-11

TOP TEN LANGUAGES IN THE INTERNET	World Population for this Language (2019 Estimate)	Internet Users by Language	Internet Penetration (% Population)
English	1,485,300,217	1,105,919,154	74.5 %
Chinese	1,457,821,239	863,230,794	59.2 %
Spanish	520,777,464	344,448,932	66.1 %
Arabic	444,016,517	226,595,470	51.0 %
Portuguese	289,923,583	171,583,004	59.2 %
Indonesian / Malaysian	302,430,273	169,685,798	56.1 %
French	422,308,112	144,695,288	34.3 %
Japanese	126,854,745	118,626,672	93.5 %
Russian	143,895,551	109,552,842	76.1 %
German	97,025,201	92,304,792	95.1 %
TOP 10 LANGUAGES	5,135,270,101	3,209,122,400	62.5 %

Table 6.1: "Top Ten Languages Used in the Web - April 30, 2019 (Number of Internet Users by Language)", from https://www.internetworldstats.com/stats7.htm, accessed 2019-07-27.

At the time of this writing, the GBIF.org user interface and key content pages are available in all six UN languages except Arabic, as well as Traditional Chinese, Japanese and Portuguese. Simplified and Traditional Chinese, and Russian have just been completed, while Arabic should be expected within a few months, according to Tim Hirsch, deputy director of GBIF. Some training material has been translated, for example in the context of the BID program, into languages other than English, but it would probably be financially impossible to expect all documents and services to be available in all major languages, anytime soon.

6.2. SWOT Compilation[122]

6.2.1 Strengths

- GBIF is the broadest association of the biodiversity data community—from a scientific and political perspective—with representation from different regions, institutions, and stakeholders. Convincing governments of so many countries to participate in and support this initiative is a manifestation of its global reach and good functioning.

- The GBIF framework has allowed national nodes to gain national and regional recognition for their successes and for the organization to be recognized as a reliable provider of a phenomenal amount of data to the community.

- The GBIF Secretariat has built good communication with its members and initiated mutual support and capacity building *among* national and institutional Nodes. This helps many underfunded Nodes considerably and may help build credentials for supportive Nodes, even nationally. This activity within GBIF is strongly enmeshed with capacity- building efforts in non-participating countries (e.g., in Europe, the EC-funded BID program in Africa, Caribbean, and the Pacific; BIODATA, a Norway-funded program in eastern Europe, Caucasus, and Central Asia; the German GIZ-funded

[122] This section provides the summarized, but also individual SWOTs of GBIF, as provided by the interviewees. It is not an analysis and attempts to convey the original language. For an explanation and rationale of this format, see section 1.5

program in the Balkans; and the Japan-funded Biodiversity Information Fund for Asia (BIFA)).

- GBIF helped new participants in Africa (and in the Caribbean and Pacific regions) to build the capacities for local infrastructures, communities, and Nodes. Moreover, GBIF has actively provided systematic guidance on data standards and tools, which could be otherwise rather complicated for beginners to adopt. It has collaborated with universities to develop courses related to biodiversity and encouraging them to offer bioinformatics training. The conferences also helped promote cooperation among biodiversity players. Thus, GBIF empowered whole national communities and helped create awareness about them and within them about international linkages.

- A few interviewees from a wealthy region felt that there was still too much unnecessary competition among partners and that more collaboration and coordination were needed to expand biodiversity data more generally and also to fill in key data gaps. Several individuals thought that GBIF was ideally poised to assume this leadership position.

- Outside of GBIF, it has built partnerships and understanding with many organizations, some much larger. Thus, it maintains one of the most comprehensive global networks in the area of biodiversity infrastructures. Some interviewees noted that there were workshops and other instruments of awareness creation and training. The scope ranged from data gathering and cleaning, to mobilization and sharing with other interested stakeholders and data use.

- Advances in information infrastructure have allowed GBIF to take a lead role in supporting collections needs as a global registry and clearinghouse, thereby enabling users to address broader scientific and societal challenges. GBIF is widely recognized as being a leader and promoter of useful partnerships.

- Although GBIF has a broad mandate and relies on a coordinated informatics approach to biodiversity data, it has succeeded without competing with its partnering organizations. Several individuals noted that *"GBIF has put the community first and foremost"* and GBIF has matured as an initiative to reflect the needs of the community. It has been an arbiter for the community with respect to occurrence records.

- The way that GBIF works in the community helps Nodes to maintain an active presence by co-support as a user and contributor, and a Node can remain very relevant despite limited resources. A commonly expressed view is that *"GBIF provides global context to the national data"* and *"GBIF does a great job of engaging data providers and countries, all of whom have different data, museum, and research cultures."*

- GBIF was broadly recognized as being committed to educating and improving the community via tool creation and training, documentation and education materials, and capacity-building.

6.2.2 Weaknesses

- Some individuals felt that GBIF and the biodiversity community have become very big, and that sometimes Nodes do not get as much credit and feedback as they deserve, especially since Nodes contribute significant volunteer effort.

- Many interviewees claimed that there should be more support from the GBIF Secretariat in formulating and proving the value proposition for *their* GBIF activity and

involvement, providing professional "marketing" brochures targeted specifically at their leadership and funders.

- Some individuals noted that GBIF is perceived as an informatics infrastructure and does not do enough to lead or support science directly. A couple interviewees felt that GBIF generally was not using the Science Committee as effectively as it could.

- A couple people noted that GBIF empanels task forces, but sometimes little seems to happen as a result, providing only the illusion of an impact. One example given was the digitization panel to prioritize threats to biodiversity and how to go about it, which published a report that was apparently never acted upon.

- There were questions raised about whether GBIF has the staff to engage with citizen science programs and whether or not this is too far outside their main mission.

- Training for the use of the tools should be regular (i.e., at least once per year to give an introductory training in order to add more new users and publishers). So far, however, it is very limited, especially beyond the academic space, and it is missing particularly for the environmental management type of stakeholders.

- Recorded lectures about the specific functions within the GBIF portal, which could mitigate the lack of periodic training, are not consistently available in multiple (major) languages.

- Language is still a big barrier, in general for non-English speakers, which not only makes it difficult for local staff to manage biodiversity data but also limits potential data users. For example, there has been a huge collection of biodiversity records developed in China (the largest dataset in Asia) but most of those records are in Chinese. Translating those Chinese records into English would require a great effort and commitment, but who should do that?

- In some of the Asia-Pacific region, national governments may not fully perceive the importance and usefulness of biodiversity data provided by GBIF.

- In many African countries, the capacity building programs in place are still inadequate. Even available biodiversity data are not fully utilized due to a lack of skills and there is inadequate training on the production of data products. The research community may not be sufficiently engaged and there is no strategy to change this. Biodiversity stakeholders are working in isolation due to poor coordination and GBIF could assist member countries with a strategy to address this.

6.2.3 Opportunities

- A national Node in a country with a large and scientifically strong biodiversity community has difficulties engaging all stakeholders. GBIF is asked to accept direct interaction with many of the strongest community players.

- GBIF could facilitate better community outreach to Caribbean and South American countries by taking advantage of North American Nodes and partnering organizations.

- There may be opportunities with relevant scientific domains (including the social sciences or computer science) and with citizen science and professional societies (e.g., in the United States' organizations such as ESIP, AGU, and ESA) to broaden boundaries of engagement.

- GBIF should host a workshop to identify and further develop biodiversity success stories and to develop a common pathway so that biodiversity organizations can more effectively engage with and support GBIF (and vice versa). The workshop should include "power users and producers" and key organizations such as NatureServe, the African Conservation Center, and the JRS Foundation.

- GBIF could think about new ways to get information about GBIF and biodiversity out to the public. For instance, could individuals "join" GBIF and, if so, how could they contribute and benefit?

- The types of people involved in GBIF are perceived as being either informatics experts or specimen collectors, taxonomists, and systematists, or technical experts and scientists, and there remains a disconnect between the two communities. Is there a way to reconnect the two? iDigBio has done this to some extent and it was very successful.

- Similarly, it was noted that there is room to improve the relationship with TDWG. Implementation plans have not been fully transparent. GBIF needs to better coordinate identification of needs and plans for implementation.

- GBIF could conduct an annual community (members) survey to look for successes, identify needs, and provide other feedback.

- Numerous individuals noted that GBIF has an opportunity to significantly increase community capacity via education materials, workshops, and webinars that provide guidance with respect to publishing, data reporting, biodiversity data analyses, using data for policy-making, resource management decisions, georeferencing, data quality improvement, and an array of other topics. Webinars, for example, provide an effective mechanism for broadening and strengthening the community by focusing on concepts, tools, and services that address user needs.

- GBIF would benefit from giving more presentations at domain-based meetings and in soliciting feedback from attendees, but it was noted that "this requires an army of the willing."

- GBIF staff and partners should attend professional society meetings and give posters, workshops, training sessions, and the like that empower individual researchers and show GBIF use cases and products and tools. This may be a great way to engage big names in the biodiversity community as champions.

- Respondents were very supportive and complimentary of GBIF's role in education and training, and some expressed their wishes that even more training could be supported. But some urged GBIF to better recognize trainers for their contributions to the biodiversity community.

- GBIF trainers and mentors noted that they valued the opportunities provided to them by GBIF to train the next generation of GBIF data publishers and users and felt that, in doing so, they were having a positive impact on science and society by sharing their expertise and knowledge with others. Virtually every trainer or mentor offered some suggestions for improvement, including:

    - offering workshops/webinars so that trainers/mentors could expand their own skillsets in some way, such as new tools and services, but also new and more effective ways to teach a concept or skill;

  - offering a node staff exchange program (sabbatical, node pioneering) by invitation so that trainers and mentors could gain new professional expertise (perhaps like a Fulbright scholarship);
  - having experienced trainers introduce GBIF to a region by having them work in a new country (perhaps a sabbatical-like arrangement); and
  - providing opportunities for trainers/mentors to network with one another via periodic in-person or online meetings.

- GBIF could provide seminars or webinars for the general public that are designed to "fascinate people" with issues such as pollinators, exotic species, pests, and other issues that are of interest to the public. Webinars could be done regionally to advertise GBIF capabilities like the Cornell Lab of Ornithology does with bird tools and visualizations.

- North American interviewees suggested that GBIF could try to empower those working in developing countries to use data for policy making and resource management decisions. This could be done via in-person workshops, webinars, or seminars. Interviewees recognized a need for better communication between scientists and policymakers, demonstrating the practical value of information (biodiversity plus other sources and types of data). Such training could teach how to use GBIF infrastructure and ultimately help promote sustainability.

- GBIF needs to find out about and address IPBES needs. It should continue to work with other major players, in particular "consumers" of information. GBIF could compute or provide regular reports and indicators/metrics such as for SDGs, biodiversity targets, and species distributions.

- GBIF could take a more direct role in supporting research that leads to great science successes in understanding biodiversity functioning.

- One interviewee mentioned the report *Global Biodiversity Information Outlook* (Hobern et al. 2012), which has provided a valuable framework and guidance for developing local biodiversity-management projects. GBIF could strengthen this part and demonstrate how biodiversity information can be used to solve practical challenges, such as environmental planning and environmental impact assessment.

- More efforts are needed to demonstrate how biodiversity data could assist policy making processes, such as environmental planning and environmental impact assessments. Interviewees made several suggestions:
  - organizing discussion forums, especially focusing on the science-policy interface;
  - inviting policymakers to join group discussions and international fora; and
  - working with local Nodes to launch localized public communication initiatives.

- GBIF should work with national-level bodies to mandate data sharing as an integral part of grants and funding solicitations. Although good examples exist, there appears to be a special need for this in the Asian region.

We provide the following individual opportunities from one region as an example of the wide-ranging expectations placed on GBIF:

- Enable nodes to capture and mediate data from citizen science.

- Work with local universities to develop and establish Bioinformatics postgraduate courses that would contribute to capacity building.
- Host regular and scheduled trainings that raise awareness and empower those interested in promoting GBIF activities.
- Organize a workshop on how to interlink biodiversity data repositories of NGOs, universities, and research institutions.
- Review existing data policies and develop new ones where necessary.
- Collaborate with JRS [Biodiversity Foundation] and CBD on biodiversity data collection and mobilization.
- Establish more awareness of GBIF programs among stakeholders through publicity, species working groups, and workshops.
- Encourage more diverse participation, especially from the global South.

6.2.4 Threats

- The Nodes concept – which includes capacities and training roles – is not fully understood, implemented, or accepted (especially where there is strong competition with respect to some of these roles).
- Insufficient collaboration with OBIS was observed. The level of collaboration between GBIF and OBIS is not what many had hoped for; this may be due to competition for resources between the two.
- The presence of many new biodiversity initiatives may be confusing to users and data contributors. GBIF's role in the landscape might become unclear.
    - For example, the DiSSCo.eu initiative in the EU was viewed by a couple of individuals from another region as being either a major threat or a great collaborator. Similarly, in Africa, the CBD and the JRS Biodiversity Foundation undertake activities very similar to GBIF's.
- Many "opportunities" might be seen as an attempt of hostile takeover on the turf of other "bioinformatics initiatives", science, or organizations at the science-policy interface.
    - For instance, a major data infrastructure sees the role of GBIF in assembling raw (or perhaps quality controlled) data, while its own role was limited to the provision of higher-level information.
- Not advancing cautiously might endanger the perceived "neutrality" of GBIF (which led to the tentative assignment of the role to lead or mediate the "Alliance").
- Difficulties in promoting open-data culture and with the protection of indigenous intellectual property rights were mentioned, particularly in the Asia/Oceania region.

6.3. Conclusions

6.3.1 General Observations

In our interviews with the GBIF community, there were perhaps the most and strongest opinions voiced about GBIF's SWOTs in this topical area of the organization's interaction with its community. This reflects on the perceived importance of GBIF's community relations, as had been foreseen in the original OECD concept and confirmed by previous reviews. Moreover, the focal points of each of the SWOT categories varied considerably among regions, more than on the data topics. Even the evaluation of current or potential activities as either a strength or weakness, or an opportunity or threat varied among regions.

The large number of national Nodes and the Secretariat's location in Europe may account for an overall impression, there, of a good balance between expectations and results. This is in line with the considerable number of international bodies located in Europe. The strong US biodiversity community, however, seems to drown the single US Node under a broad range of expectations that cannot be easily met. African and Latin American interviewees appreciated what GBIF has achieved, but urgently called for more Secretariat activities – quantitively and qualitatively. In Asia, there appears to be no noticeable or effective footprint of GBIF within the biodiversity communities.

Much of this can be attributed to hugely different levels and structures of funding, the resulting sizes of the local biodiversity communities, and competition for resources. Where competition becomes too fierce, it may preclude collaboration, which is a necessity for GBIF's whole cooperative paradigm and the organizational concept of a national Node.

Perhaps as important as the national funding issues – in Asia, but also in Latin America and Eastern Europe, and to some extent in Africa – the language barrier came up very prominently. In Asia, it may even be decisive. In Africa, the support by the French Node in the context of the BID program may have been enough to mitigate the problem, for the time being.

In the case of Asia there appears to be a "chicken and egg" problem. In the GBIF model, community engagement depends a great deal on the existence of enough Nodes and of some strong Nodes, but Participation – and thus, the existence of Nodes – depends on the favorable opinion of national biodiversity communities.

GBIF maintains an impressive list of institutional partnerships[123], with different levels of involvement, from an observer position to council membership and formal memoranda of cooperation. Yet, in interviews, a number of relationships were deemed important to establish, intensify, or improve (to the – debatable – point of naming one relation as being threatened).

The long-term success of GBIF may critically depend on this kind of institutional partnership, especially to avoid unnecessary and unproductive competition or duplication of effort. However, even to sustain the existing links to biodiversity infrastructures and organizations may be a stretch for Secretariat staff, let alone if the number or intensity of relations were to be extended, particularly with scientific disciplines and their representation through societies and at conferences. (It is outside the scope of this review – if it is possible at all with a perspective of ten years or more – to suggest a ranking or priorities on which organizations or disciplines to engage with. Prioritizing needs to be a regular process.)

[123] https://www.gbif.org/partners

6.3.2 Capacity Enhancement and Training

This is a clear strength, a source of high reputation of GBIF and appreciation of its activities – comparable to its achievements as a data infrastructure and promoter of standards. Even in North America, where active capacity enhancement would appear superfluous, GBIF is obviously regarded as a capable organizer and suggestions of opportunities and recommendations on education abound. Many of these suggestions refer to the opportunity of leveraging this region's capacities to support capacity building elsewhere, such as in Latin America, by voluntary trainers (their time provided in-kind by other stakeholders).

In Europe, the main focus seemed to be on the support to Nodes – by the Secretariat but also by mutual support among Nodes (as advised by earlier reviews). This could be observed to be at work and confirmed by informal talks at the European Nodes meeting in May 2019 and by formal interviews conducted previously. The networking of regional Nodes appears to be successful here, just as it was most apparent that existing Nodes supported capacity building in not-yet-participating countries in Europe and the BID program in Africa, Caribbean, and the Pacific. Again, trainers – other than from Nodes' staff – are assigned an important role.

The impressions in Africa about GBIF's strengths, weaknesses, and opportunities seem to have been formed by the BID program. The opportunities listed perhaps mainly reflect too optimistic assumptions about available resources, but they nonetheless clearly call for a continuation and even an extension of this program. The qualitative expectations in Africa range from support of data acquisition and quality control to education about scientifically appropriate use of data. GBIF is further expected to create public awareness, provide support for increased local funding, and help universities establish post-graduate biodiversity courses.

SWOT assessments and expectations in Latin America seem to be less pronounced – hinting at higher capacities of the local biodiversity communities – and more focused on the support to local Nodes. The language barrier was highlighted as well. Asian interviewees voiced no recognizable opinion on the capacity building and training activities by GBIF, but the emphasis on the language barrier in other areas would predict major problems in performing capacity enhancement and training there in the future, while capacity enhancement is considered a major strength of GBIF in the other regions.

In summary, GBIF's strengths and reputation in the area of capacity enhancement and training, within the Nodes network and beyond, have led to very high expectations (with the exception of Asia). Even if GBIF would simply handle programs funded by third parties, it would need more relevant staff at the Secretariat and considerable resources contributed by national Nodes and voluntary trainers.

6.3.3 Relations with Funders

Once Participation (or other funding) is established, relations with these funders need to be cultivated, particularly to increase support of a national Node. It will fall mostly to members of the Governing Board and to the Nodes to employ the necessary instruments, from the "elevator pitch" and the one-page leaflet to dedicated workshops, to create and maintain local awareness of the scientific opportunities, the utility to policy making, economic benefits, and health outcomes of using and contributing GBIF-mediated data – and, consequently, of funding GBIF.

The Secretariat was universally deemed responsible for and capable of providing these instruments, or large parts of it, including appropriate translations. The portal pages and yearly science report were brought up in this regard in interviews with the Secretariat, but were barely mentioned, if at all, in other interviews.

6.3.4 Relations with Science

This is an area of weaknesses, but also of large opportunities. Many scientific users do not know enough about GBIF data – neither their strengths nor their weaknesses nor opportunities for them (see Chapter 5). Direct engagement with the many different disciplines – in particular, the applied disciplines – and with (sub-) communities has recently been taken up by the Secretariat through a science officer, but it would appear too much of a drain on the Secretariat and the Nodes to put this into effect on their own resources alone. It appears that the new GBIF Open Data Ambassadors initiative[124] of the Secretariat is designed to address this.

As with funders, the existence of exemplary use cases in portal pages and the annual science report do not seem to have much of an impact. The quite recent GBIF data blog might develop more impact, as it may align somewhat with requests for analytic support or even data products, without intruding on someone's turf. It should be noted, however, that it is still well-hidden in the portal menu, as the last entry below Tools on the GBIF Labs sub-menu.

6.3.5 Partnerships

Many of the relations of GBIF with other institutions or their mechanisms and outcomes are unknown to most interviewees. Consequently, where addressed, this leads to suggestions of weaknesses or opportunities (or even a threat – which may be disproportionate). Among representatives of the international bodies interviewed, a positive to very positive balance of strengths and weaknesses was observed, with opportunities to be addressed. Still, this area deserves persistent and close attention.

In one region, more direct partnerships with large local initiatives or institutions – instead of a relation mediated by the Node – were deemed necessary. Still, GBIF's reputation in the same region is high enough so that its potential leadership was invoked.

6.3.6 Open Data Advocacy

In Europe and North America, open data may appear as a largely settled matter, policy-wise, and on a slow but good trajectory culturally, among scientific disciplines. This is not so clear in other regions. Interviewees who themselves appear to see the benefits and who do not have objections themselves, claim the need for activity by GBIF in establishing open data as a policy in their countries. Some just see the need to justify the effort – in particular insofar as it is necessary to expose existing data in foreign languages. But there were also a few hints – even as a threat to GBIF – that the knowledge of indigenous people could not be shared in various contexts. Most of these interviewees therefore saw the need for help with explicit open data advocacy by GBIF.

We regard this area mainly as one to be addressed by such organizations as journal publishers, scientific societies and – at the level of societal benefits – IPBES and CBD. But GBIF can certainly provide ample evidence on the benefits and feasibility of open access to biodiversity data. The Secretariat can and does do so at international levels, while the Nodes could and some do so at national levels.

---

[124] https://www.gbif.org/news/1EL1oQa7ZmWtIZlTsNHblx/programme-seeks-biodiversity-open-data-ambassadors-to-expand-best-practices, dated 12 July 2019, last accessed 2019-08-12)

7 GBIF'S PARTICIPATION, ORGANIZATION, AND FUNDING[125]

7.1. Description

As this chapter will show, the three issues of Participation, organization, and funding are fundamentally intertwined. Core funding of the Secretariat depends on the Participation of many countries. The actual success of GBIF in mediating data depends on the activities of national Nodes, and thus, their funding. However, third-party funding of Secretariat activities and nationally funded outreach activities of Nodes can contribute decisively. A sufficiently funded and successful organization in turn is likely to gain more Participation.

The recent joining of Angola and Belarus (March and July 2019, respectively) as Participants in GBIF illustrates one kind of this interworking, as Box 7.1 shows.

> ***Box 7.1: Supplementary funding and the Nodes network as enablers of Participation***
>
> "Angola's formal membership has its roots in a 2015 Capacity Enhancement Support Program project led by GBIF Portugal aimed at promoting the network in the Portuguese-speaking countries of Africa. The continent's seventh-largest country is also the tenth member to join as a direct result the Biodiversity Information for Development (BID) program funded by the European Union and led by GBIF."[126]
>
> "Belarus has taken the step of becoming a formal GBIF Participant while engaged in BioDATA, a two-year Norwegian project aimed at helping undergraduate and postgraduate students from Tajikistan, Belarus, Ukraine, and Armenia to develop skills in biodiversity data management and data publishing. Coordinated by the University of Oslo with support from the GBIF network, BioDATA is funded by the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (DIKU)." [127]

7.1.1 Participation

By definition, Voting Participants supply the core GBIF funding and formal Participation is a foothold to mediate biodiversity data from that country, region, or organization. Data mediation is the part of GBIF's mission that provides the most obvious benchmark of GBIF's

---

[125] This chapter addresses the organizational and funding aspects of item 2 of the statement of task ("Review the governance and sustainability of GBIF as a global network and organization (including hosting of the Secretariat in Denmark)").

[126] https://www.gbif.org/news/65lsvpPbsAGGawC4gEGOoq/angola-becomes-the-newest-member-of-the-gbif-network

[127] https://www.gbif.org/news/5prC9d8GDRCS8HtoPSon8n/belarus-extends-gbifs-european-membership-map-eastward

success and potential benefits. Thus, formal Participation is essential for GBIF's role as a global organization.

Touching on both issues, a previous review (GBIF 2010a) found: "After 10 years, of the 30 OECD member countries, 21 are voting members, 4 are associate members, and 5 have not joined GBIF." Now, out of now 36 OECD member countries, 26 are Voting Participants, 2 are Associate Participants and 8 have not joined. Of the megadiverse countries that were not part of GBIF in 2010, Brazil, Ecuador, and the Democratic Republic of the Congo are now Associate Participants, for example.

While GBIF has made progress in the western part of the map, Figure 7.1, it has almost no formal Participation by countries in Asia. However, there are noteworthy Other Associate Participants from Asia which do not appear on this map, and there is already a good membership of Publishers in Russia and Asia, as shown in Figure 5.8 (A). Both could be stepping-stones to building greater Participation in GBIF.

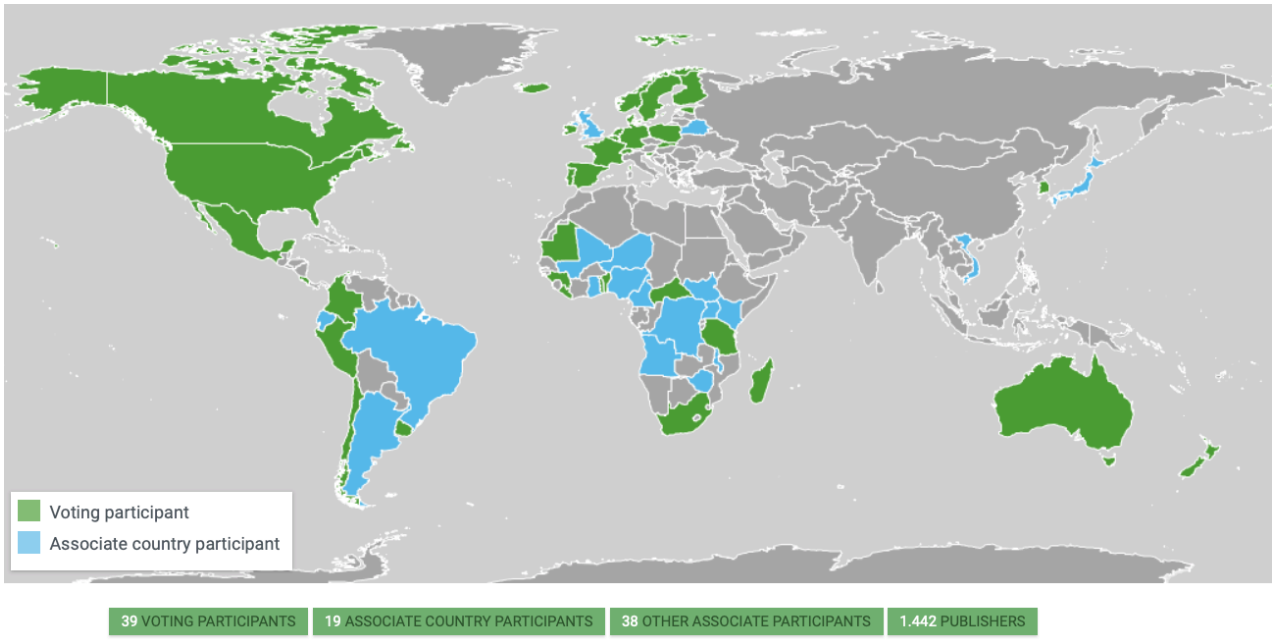


Figure 7.1: The GBIF Network, https://www.gbif.org/the-gbif-network, last accessed 2019-08-14

7.1.2 Organization

The main organizational feature of GBIF is that its operation is tied to its many Nodes. Conceptually, their crucial importance lies in being the major pathway to data Publishers, and thus data, and the local communities and organizations. The new GBIF document "Establishing an Effective GBIF Participant Node" (GBIF 2019b), published in July 2019, explains:

> *Since GBIF's founding in 2001, the participating countries and organizations have been testing and developing models for coordinating the mobilization, management and reuse of biodiversity data at the national level or within an organization's scope. The formation of Participant 'nodes' has been central to these efforts. Designated by each Participant, these teams coordinate the needs and interests of the many stakeholders involved.*

The foremost function of a national Node described in (GBIF 2019b) as "Coordinating a community of initiatives relating to biodiversity, including making connections to the international GBIF network" and only then supporting data mediation, reuse, and data

management. It identifies four very distinct roles in a "Node team" that require different skill sets.

It appears that this is meant to implement a principle of subsidiarity – to solve issues most closely to where they reside. Thus, the effort for some of the Node's services, provided as examples (GBIF 2019a, Chapter 4), would obviously scale with the size of the country's community and the amount of biodiversity data. As two of the key characteristics expected by GBIF of a Node are organizational and scientific neutrality and the ability to provide leadership and initiative, the potential for conflicts of interest – perceived or real – and other advantages and disadvantages regarding the hosting institution, are clearly spelled out.

Building the relationships with other international organizations falls into the domain of the Secretariat. GBIF has managed to become connected to many of the essential international players, through memoranda, memberships and even governance positions, or participation in and organization of workshops (see Chapter 6). Some of these organizations have even joined GBIF as "other Associate Participants". There are more than 20 intergovernmental organizations and international societies as Participants and about 10 of another type, as well as 16 other Partners[129], including crucial non-Participants such as CBD, IPBES, and GEO BON. (GBIF committees and other bodies are discussed in Chapter 8, as part of its governance.)

Table 7.1: Relationships maintained by GBIF as of August 2019.

39	Voting Participants (countries)
19	Associate Country Participants
36	(other) Associate Participants
4	Affiliates
16	Partners
114	**Institutional Relationships**
96	**Nodes**
1,916	**[128]Publishers**

7.1.3 Funding

The OECD concept (OECD 1999), while providing some estimates of the cost, had no specific plan for funding the organization. It probably could not do so accurately, because GBIF encompassed almost all of the biodiversity informatics landscape. But among other funding opportunities, it mentioned OECD countries as sources to fund a secretariat. National data nodes, which were to be invested in by OECD members, were already mentioned in the concept.

Today, the scope of GBIF activities that are to be funded has become clear and so have the contributors – namely voting Participants (countries). The distribution of the financial burden has been adjusted to a scale that reflects fairness regarding the means of each country, but it also reduces the dependency on a few big payers. Principally, dues are proportional to the Gross Domestic Product (GDP) of each Participant. Nevertheless, as approved by the 22nd meeting of the Governing Board in 2015 and effective for the years 2017-2021, the "Approved Funding Model" reduced the cap to limit the maximum share of a country to 15% of GBIF's core budget and upheld the reduction by 50% for countries with a GDP per capita less than US$ 13,500.

In financial relations with each country, GBIF faces the idiosyncrasies of their funding mechanisms. In many cases, the sum expected and agreed upon to come from one country will not be directly transferred from a ministry or governmental department, but through a secondary transfer from another government agency or institution. In a number of important cases, some institutions have to pool money from different ministries or departments. Each

[128] Note that this is the number of "registered" Publishers, as of 3 Oct 2019. The number of Publisher that had already published data was 1,517 at that time.

[129] https://www.gbif.org/partners

type of ministry, and possibly even the type of intermediary institution, may require a different specific expression of the cost-benefit relation of funding GBIF.

As the Memorandum of Understanding (MoU) that all Participants sign is legally non-binding, the Secretariat cannot responsibly build its financial operations entirely on the expectation of payment of dues. Indeed, the payments received have deviated considerably from long-term models and typically have been transferred over more than the first half of the year in a variable schedule. Thus, the Secretariat manages liquidity in a way that enables it to fulfil legal and contractual obligations under the assumption of average or "worst-case" scenarios, to avoid running the risk of disrupting services or losing valuable competencies.

Not too much is known about the actual funding of most national Nodes, except anecdotally. From our interviews it became obvious that the levels of funding vary considerably among Participants but is difficult to measure anyway: In most cases, the Node will be supported infrastructurally and administratively by the hosting institution. Beyond that, Node staffing varies from fractional time of one person to about four full-time staff. In most cases, this must be considered severe underfunding.

Box 7.2, below, highlights some of the Secretariat's thinking about funding issues:

> ***Box 7.2: Tim Hirsch, GBIF Deputy Director***
>
> There are significant positives and significant negatives in us keeping the governmental model as our core participation. I would say that the positive from my perspective is that it provides us with credibility and a sense of continuity that you don't get with [other models] and I have experienced, that it is important for us to be able to say we are an intergovernmental organization. The downside is that we are very subject to those inevitable, political fluctuations.
> [...]
> In order for the network to function effectively we need national investments in the node activities to be at least as much as and hopefully considerably more than they're paying into the global infrastructure. [...] And often this will be not necessarily investing in lots of fancy equipment, but it is dedicating people to this activity.

7.2. SWOT Compilation[130]

7.2.1 Strengths

- Quite a number of interviewees (in particular in Europe) spoke of the distributed nature of GBIF funding and good practices of the Secretariat in handling it as a strength, in that the dropping out of any single funder can be accommodated. Also, commitment and direct funding by governments was seen as more stable than project funding or other short-term initiatives, and more ample than, say, indirect funding through a single intergovernmental organization.

---

[130] This section provides the summarized, but also individual SWOTs of GBIF, as provided by the interviewees. It is not an analysis and attempts to convey the original language. For an explanation and rationale of this format, see section 1.5.

- The overall strength of GBIF is its global reach and scope. This is achieved by its dynamic policies of collaboration and interaction between the different members of the network.

- The global character of the initiative has grown in a robust and consistent manner. It provides a feeling of community, generating confidence, credibility and motivation by the publishing partners, and countering some of the stigma of sharing data.

- GBIF has been very well thought through from its initiation, from the governance through the technical component within the network and associating the best persons for each topic.

- GBIF's success is also founded on the collaborative spirit of the network, working with nodes of several countries for knowledge transfer and sharing experiences. More mature nodes help and mentor new nodes entering the network, assisting new nodes in resolution of issues that they have already faced and resolved.

- Collaborative work is especially promoted through the capacity enhancement support program.

- In Africa, direct funding and support for Nodes and data mobilization were seen as a GBIF strength.

- The partnership with government ministries of environment helped to build trust with biodiversity stakeholders.

7.2.2 Weaknesses

- GBIF's structure is confusing to some as GBIF has grown and evolved. New individuals find that it can be a problem to understand the work of all the committees and projects supported by GBIF.

- In North America, several individuals expressed an interest in seeing GBIF's strategy and plans for attracting and retaining members. Some questioned why some big herbaria and botanical gardens have seemingly been hesitant to publish with GBIF.

- Numerous individuals commented that it is difficult for people to understand the value proposition of GBIF and biodiversity data. A couple of individuals wanted to know "What is the elevator speech for GBIF success?"

- Funding in individual countries may depend strongly on political winds, good connections with single individuals who come and go, and relationships with a variety of science ministries and environmental departments – each of which creates different issues. For example, science funders prefer funding short-lived projects or single investments, while environmental agencies have a focus on funding for national issues or governmental reporting requirements. Each of these funding sources requires the provision of different metrics of value.

- There is no core financial support for the maintenance and digitization of specimen collections.

- Usually, resources and mechanisms of support are allocated towards regions that have the majority of voting countries (the wealthy North). Such regions have very different

needs from regions of the South where the highest levels of biodiversity are typically concentrated.

- Until recently, the GBIF portal was only in English and its translation was done on a voluntary basis. The information in the portal is the basis for the Nodes to provide different stakeholders with information about the tools. It would be in GBIF's interest to allocate a budget to carry out such translation tasks.

- It is claimed by a few interviewees that documentation for many GBIF tools, and the tools' interfaces themselves, are only provided in English, making them difficult for certain stakeholders to use.

- Most Node managers do not have full-time positions and therefore no exclusive dedication to the Node activities. Many of them also have no dedicated team to work with, which makes carrying out data integration activities and other important tasks very challenging.

- Capacity enhancement programs are great opportunities that help to advance Nodes without much experience. However, such grants favor activities involving several Nodes, and mentoring of one Node by another, which makes success dependent on the existence of strong Nodes.

- The GBIF budget to establish capabilities regarding the use of standards and tools at the national Nodes is limited in comparison with the BID program, which allows a more efficient implementation. This is a critical point as in addition to technical implementation, scientists, environmental managers, and all the different stakeholders need to be convinced to use GBIF's standards and tools.

- Standards platforms and tools such as Darwin Core are not taught at the university level and hence such technologies need a budget, resources, and logistics to disseminate them effectively.

- Active support to assist developing countries in obtaining resources to carry out activities such as knowledge dissemination, digitization of specimens, and training to use biodiversity data is crucial, but missing.

- Currently all the staff and data-management infrastructure are based in Europe where labor costs are high. GBIF may consider outsourcing part of its management activities to developing countries and shifting part of its data management activities there.

- There appear to be no clear funding strategies for some Nodes. Nodes are expected to pay annual subscription fees based on GDP, which most countries struggle to pay.

- GBIF is viewed by some biodiversity stakeholders as a closed system that needs to open up.

- Funds required to support GBIF activities are insufficient. What many of the governments provide is inadequate and there is a need to look elsewhere for funding.

- Nodes are often trapped in government bureaucracy.

- Often there is a lack of funding for data collection at the source.

- Inadequate sponsorships to attend annual GBIF meetings that are frequently held in Europe (about half the time). Government priorities do not include funding GBIF meetings that would benefit key biodiversity stakeholders and Node managers. Failure to attend meetings leads to a loss of research project collaboration opportunities, the sharing of experiences and best practices, and general networking.

- The approach to monitoring performance of country Nodes is not robust.

- There is no common platform for interaction between Node managers and the heads of delegations.

- Donor fatigue is an issue. For example, resources are needed in cases of governments that might not value the importance of cooperation, the sharing of data, and good quality data.

- One Asian interviewee claimed,

> *"currently there is no support from the organization for maintaining the team for running national Nodes and the Nodes have to find financial supports on their own. It is difficult to secure long-term and stable funding for many national Nodes".*

7.2.3 Opportunities

- GBIF could examine the potential roles of corporations (e.g., pharma or agriculture) that depend on the continued existence of highly varied biodiversity and a robust natural environment and foundations (e.g., a 10-year "moonshot" for our own planet). In the US, the National Laboratories, NASA (which historically has had a large interest in biodiversity), and the US Department of State should be interested in supporting and promoting GBIF. In short, there needs to be a mechanism developed through which a much broader group of stakeholders can be involved with GBIF.

- No compelling suggestions surfaced about additional or more stable funding in Europe; actually, some suggestions were contradictory (seek more scientific funding or more environmental ministries funding). The frequent statement that the current funding scheme is actually a strength, suggests that just stabilizing the distributed nature of funding could be regarded as an opportunity.

- If GBIF would help to create more full-time positions it might help to meet its aims, functions, and activities.

- Guidelines about specific solicitations to finance regional projects are sometimes difficult to meet without the help of a tutor. Reliable training for new Nodes during the writing phase of projects would be helpful.

- The Capacity Enhancement Support Program (CESP) facilitates knowledge transfer and collaboration between Nodes and provides the possibility for interested countries to participate in activities even before they are officially part of the network.

- GBIF resources could be used to finance and sponsor the digitization of natural history collections, especially in those countries harboring most of the biodiversity. Most of the time it is not really asking for that much money, but it can make all the difference between being able to publish them or not. The creation of funds to digitize biological

collections in regions where biodiversity is the richest could provide a large amount of data.

- Getting support from GBIF for the creation of full-time positions for Node managers might allow for continuity of the activities.

- Data quality workshops across the region have been carried out lately, but it still is necessary to curate and clean data more. It would be very helpful to get human and financial resources to carry out the data cleaning and having a dedicated group within each Node.

- Resources for activities regarding implementation of a Node (policies and lineaments) are very scarce. In order to increase awareness, and the attraction of new data users and publishers, it would be helpful to launch programs to offer and disseminate information, "spreading the word" about GBIF, as well as workshops about usability of GBIF data. Develop an incentive program showing to the data providers and publishers how their data have been used, published, and recognized. Even though actions on stakeholder training and capture should be local (i.e., at the national Node) it is important to have the support and back up of GBIF to valorize even more these actions.

- For GBIF to be sustainable in the future it might be important to think about alliances and associations with institutions such as the World Bank or the Inter-American Development Bank (IADB).

- Having a visiting commission from the GBIF Secretariat every five years to each region to see and evaluate the reality of the participant countries might mitigate the feeling of isolation and of being distant from the GBIF governance.

- Develop an induction program for newcomer Nodes, strategic plans for Nodes, a policy guide on collaborations, and instruments for engaging with stakeholders.

- Develop a Business Continuity plan or mechanism to minimize disruption of services in the event of transfers of officers, government reshuffling, retirement, and natural attrition.

- Budget for all key GBIF annual conferences and meetings for Node managers.

- Work with non-governmental organisations (NGOs) to increase awareness and visibility of GBIF and its impacts on individuals and institutions.

- Cultivate buy-in by governments and explore collaboration with the JRS Foundation.

- Work closely with the scientific community.

- Sensitize governments, members of parliament, and universities to fund biodiversity data activities at Nodes.

- Develop an engagement strategy where biodiversity is endangered to [help] ensure preservation of endangered species.

- Establish a funding program for supporting newly created GBIF Nodes.

- Some researchers from universities attend trainings and workshops organized by GBIF but have no linkage or formal collaboration mechanism with GBIF. These institutions could be asked to establish a GBIF desk to provide the link to the GIBIF Node.

- Capture the attention of most data-source institutions, e.g., through an Environmental Protection Institute. Assigning an officer at such an institution to link with GBIF would be desirable.

- Pay for data publishing.

- Strengthen and empower the Nodes to look for funds, for instance, by writing proposals to seek funding. This will help them become sustainable as they continue to collect and mediate data.

- Consider Africa critically and particularly biodiversity-rich regions that lack resources to mediate data and support them.

7.2.4 Threats

- There is confusion about the US Node and competition among US partnering organizations. Some interviewees expressed concerns that the US Node was not as effective as other national Nodes, despite the size of the US biodiversity community. A few respondents reported that they felt that another institution in the US has acted more like the US Node than has the actual one. A couple respondents thought that iDigBio was moving into the realm of activities supported by GBIF and was competing with both the US Node and GBIF. Although there was disagreement about which of the US biodiversity organizations were most effective, there was general agreement that there is a lot of perceived undue competition and lack of collaboration within the US biodiversity community, some of which was thought to be related to competition for scarce resources. One collections representative stated that with respect to biodiversity infrastructure in the US, "The whole should be greater than the sum of the parts, but this is not currently the case."

- It was suspected by a few interviewees that the Secretariat might be destabilized, should the biggest funder or funders drop out. The GBIF budget depends entirely on the willingness of every government of each country, and threats might come with changes in those governments.

- Limited participation of developing countries due to high membership fees was seen as especially critical for China and India. One interviewee from a large country mentioned that currently there is no funding for the national Node, which leads to limited commitment to the GBIF activities at the national Node.
  At another Node in a similar situation, it was emphasized that it is necessary for GBIF to work together with the national-Node team to get in touch with high-level government officials and to demonstrate the usefulness of biodiversity data to policymakers.

- An interviewee from another large country mentioned that GBIF required a high membership fee but failed to give a cost-benefit demonstration. If the membership fee remained high, this country would be unlikely to become a voting Member.

- There are inadequate human resources in many developing countries.

> In particular, there is a high turnover of human resources due to new appointments, transfers and reshuffling, while there are no clear structures and arrangement for retaining persons with knowledge and experience.

7.3. Conclusions

Perhaps not surprisingly, a large number of threats, tied to insufficient or volatile funding – albeit at vastly different scales – were named in interviews and we report them as such. However, in our assessment most can be regarded as weaknesses, as there has been long-term stability for GBIF as a whole (putting fears into perspective), and, more important, as there are relatively straightforward though not effortless solutions that can be read from each threat that would actually be addressable by GBIF.

7.3.1 Participation

Considering the legal construct of Participation, the first review of GBIF (CODATA 2005, p. v) found:

> *With regard to the question of legal instruments on which to base GBIF, we are convinced that the choice of an MoU instead of a treaty was correct and explains why GBIF was formed rather quickly. A non-binding, voluntary MoU is not only sufficient and appropriate; we also believe that it will not be possible to find support for elevating this legal status to a binding agreement.*

This review supports the conclusion of the 2005 review. GBIF has now operated for 20 years and has proved robust across decades including through fluctuating political and financial circumstances (e.g., the financial crisis after 2008). However, neither this non-threatening contractual regime nor a newly agreed fair sharing of the costs has led to breakthroughs in Participation, most visibly in Asia. The recent positive development in Africa is largely attributed to the effect of the capacity building program BID, funded by the European Commission – 10 of the currently 21 national Participants joined "as a direct result" [131].

Considering the actual successes of GBIF, the testimonials of experts to the effect that GBIF is necessary, and its ways are actually the most effective to achieve the mediation of biodiversity data and its socially beneficial re-use, one must wonder why Participation is not near universal. Some hypotheses can be derived from interviews and desk research:

- The new funding model approved by Participants in 2015 is widely unknown and, whatever is really known about it, it may still be considered not fair enough.
- Either the actual benefits or their description have been too indirect to convince new Participants, in particular:
    - In cases where the responsible authority employs a purely economic metric ("return on investment", cost-benefit analysis purely in terms of money).

[131] https://www.gbif.org/news/65lsvpPbsAGGawC4gEGOoq/angola-becomes-the-newest-member-of-the-gbif-network

  - Where the national Node and Delegation, which would have to maintain the motivation of participating, do not have the capacities to communicate the benefits in the terms of their many stakeholders.

  - Where data mediated by others are re-used without contributing funds or data.

  - A country assumes it could gain more advantages by keeping biodiversity information to itself, rather than by sharing, whether or not that country has agreed to the Aichi targets, including the target to share data.

- A country does not actually want to disclose its biodiversity details for considerations other than about benefits.

- Or – as with many other community relations – a language or cultural barrier is at work.

In summary, a new approach to more and stable Participation may need further detailed analysis, based on knowledge of individual countries and regions, a strengthening of tools such as cost-benefit analyses and capacity building of Nodes, and still more advocacy, especially about the sharing of data.

7.3.2 Organization

GBIF is built on a scalable architecture, which delegates many of GBIF's important functions to the Nodes, closer to the data users and contributors. It was first proposed by the OECD as a concept and again recommended by the first review (CODATA 2005, p. x) as a "distributed support structure". Considering its success in mediating data, GBIF has fared well with this structure. Much fine tuning has been implemented over the years, such as the regional grouping or engaging trainers from the wider community in each country, which increases the efficiency and effectiveness of GBIF and its architecture.

However, success of the Nodes in performing all the roles laid out in (GBIF 2019b) appears to be uneven. With some, this may result from a habitual inability to break out of their hosting institution's community, while others might lack support or guidance, such as on the justification of the Node's funding and other local fund raising. In a majority of cases, however, we find that the Nodes are simply underfunded, as the many functions and roles (facing a plurality of stakeholders) cannot reasonably be mapped to the skillset of a single person, and definitely cannot be taken on by one person assigned just part-time to these tasks.

A well-connected network of about 20 Nodes in Europe may be able to remedy some of the understaffing of each Node, but the roughly comparable North American region is supposed to be coordinated by just two Nodes. This may explain some misgiving about GBIF's organization there.

North America is a wealthy Participant that has established a number of GBIF essentials, such as Publishers' knowledge about standards and networking of the national biodiversity informatics community, without too much need for a coordinating Node. The situation has been different in Africa. There, the build-up of a network of Nodes, comparable in number to Europe, has been achieved only through substantial external funding and support from European Nodes, and may depend on that for some rounds of such funding and other support.

The Latin American and Asian regions appear to some degree less dependent on external support than Africa. The situation in those regions, regarding the number of Nodes and their potential to build and profit from a regional network, might be more similar to the North American than the European region. The number of countries in Latin America is lower than in the European GBIF region, distances there and in Asia are longer, and in some cases, countries

are huge in terms of population and number of institutions to coordinate. Thus, the concept of a national Node may fail in countries like China and India, especially if it is underfunded.

The Secretariat maintains and extends the relationships with other international organizations. Many of these currently about 50 relationships would need more than fractional engagement, be it at the coordinating or more technical levels. Considering that these relationships do not even comprise many of the scientific societies and other representations of the biodiversity disciplines, and considering the necessary efforts to mitigate weaknesses (particularly those close to being a threat), the Secretariat is clearly at or beyond its limits.

In summary, we conclude that the organizational principle of subsidiarity is well chosen, as it is scalable to ever more and diverse Participants. However, it may be necessary to rethink if or how one Node per country and regional assemblies of Nodes can actually work well in large countries or at huge distances between Nodes. Finally, most of the Nodes and the Secretariat are presently not sufficiently funded to perform their nominal duties in a fully satisfactory manner, which brings us to the conclusions regarding the funding.

7.3.3 Funding

The large number of participating countries and the institutional, long-term commitments expected from them – rather than relying on chains of short-term project funding – has proven successful in providing a relatively stable core funding; namely, that of the Secretariat, its work programs, and the GBIF governance. Interviews with and documents from the Secretariat show that the short- and long-term volatility that remains, has been managed by the Secretariat operationally and in terms of human capacities in non-disruptive and financially prudent ways. The organization also has acquired substantial supplementary project funding, which has augmented its regular work program.

There nonetheless is apparent underfunding of even current activities, coupled with the huge expectations to extend GBIF's activities, generated by its successes. This would include tackling more than just data on occurrences, capacity building on the scale of the BID program, building and leading the Alliance, and the mandatory construction of the GBIF Asian region. This gap between current resources and expectations would require either a strict curtailment of expectations or substantial additional funding.

It is noteworthy that interviewees from different regions had quite different suggestions regarding from whom to expect additional funding. Europeans typically expected more from their governments, including the European Commission, while North Americans looked more towards philanthropic organizations and industry. In Latin America, intergovernmental institutions such as the World Bank were brought up, while African interviewees either looked directly towards GBIF or to philanthropic organizations, such as JRS, which are already active there.

Considering the core funding of GBIF, a number of observers recommended looking for an entirely different "business model". However, the most cost-intensive steps of curation-related and publishing-related efforts are performed by Publishers and Nodes in the existing GBIF model. By contrast then, the cost of most tasks at the Secretariat scale less than linearly with the amount of data. Whether any of this cost can or should be recouped – by Publishers, Nodes, or the Secretariat – through fees attached to publishing or using data, needs to be considered very carefully and address the needs and resources in all GBIF regions. In particular, we believe such fees would compromise the openness and the comprehensiveness of GBIF's data, which would clearly undermine its mission.

One of the OECD expectations (OECD 1999), repeated by the Forward Looking Team in 2010, was the digitization of the "information resources residing in major natural history museum

and herbarium collections across North America and Europe" (GBIF 2010a). A quick estimate of the cost level (see Chapter 2) shows that this is completely out of the current financial scope of GBIF, making it unlikely to be added again to its remit. It would seem that some of this is now partially on the way with the DiSSCo initiative (at an estimated level of a few hundred million EUR)[132]. National funders in regions other than Europe, however, may not have the resources necessary, or set other priorities. Nonetheless, it is obvious that there are very important collections outside North America and Europe that should be digitized with equal attention, if not urgency – one might think of the National Museum of Brazil, which was destroyed by a fire in 2018.

In summary, the current Participation-based mode of funding the core GBIF budget provides a stable platform from which GBIF has been able to reliably provide many useful services and to manage or coordinate additional projects based on supplementary funding. The expectation of the OECD concept (OECD 1999) that all countries would contribute according to their economic means, has been met only partially. However, even considering the slowly moving process of agreed funding models, GBIF will need to seek increased funding from participants – for the core budget as well as for the national Nodes – to stabilize and improve services within the current scope. The additional expectations of GBIF's communities would require a substantial expansion of GBIF's activities. It may not be possible to fund all of these additional activities from the core budget, even if increased. Therefore, any move in these directions would need to be carefully considered to ensure that the organization's existing strengths would not be sacrificed.

[132] Personal communication by a member of the DiSSCo consortium, October 2019.

8 GOVERNANCE, WORK PROGRAMS, AND THE SECRETARIAT[133]

8.1. Description

As previous chapters have shown, GBIF has successfully organized and built a global network of official Participants, Partners and Publishers, and established good working relationships in the landscape of important players around it. It has also become clear that the core funding, though not adequate to support further qualitative or quantitative improvement, can be considered quite stable. This chapter assesses the success of the governance structures, work programs, and the Secretariat in implementing the spirit and the terms of the GBIF MoU (signed by all Participants), and if they are adequate to support core functions of GBIF's remit, and their further development.

Before going into detail, it should be noted that there is a page on the GBIF portal, well-written for general consumption, on GBIF governance[134], a directory of people involved[135], and the strategic plan[136] – all openly available. These pages link to the relevant documents discussed below.

8.1.1 Governance

There are three dominant characteristics of the revised Memorandum of Understanding (MoU 2011):

- The "Voting Participants" are countries that make financial contributions. While countries and international organizations may be Associate Participants (and have GBIF Nodes) as well, only dues-paying countries can have voting status.

- The Understandings of the MoU declare that it is "open-ended" – referring to paragraph 11, that does not specify a termination date. But it is "legally non-binding" on signatories, and all financial contributions are "voluntary". Consequently, the MoU, the Rules of Procedure (RoP) and the Terms of Reference (ToR) of the GBIF bodies explicitly emphasize consensus decision-making and an array of recommendation mechanisms between bodies.

- The fundamental organizational and operational principles are that "GBIF:

    a) be shared and distributed, while encouraging co-operation and coherence;

    b) be global in scale, though implemented nationally and regionally …".

The MoU establishes a Governing Board (GB), which meets annually, as the central decision-making body and an Executive Committee (EC) for intersessional decision-making. It requires

---

[133] This chapter addresses the aspects of governance and the hosting of the Secretariat from item 2 of the statement of task ("Review the governance and sustainability of GBIF as a global network and organization (including hosting of the Secretariat in Denmark)").

[134] https://www.gbif.org/governance

[135] https://www.gbif.org/contact-us/directory

[136] https://www.gbif.org/strategic-plan

the GB to "establish its Rules of Procedure and such subsidiary bodies as it sees necessary". It institutes the office of the GBIF Executive Secretary, who will simultaneously hold the office of Executive Director of the GBIF Secretariat.

According to the MoU, the GB will decide on the Guidelines for calculating Basic Financial Contributions, so that the organization is "funded primarily by those that have the greatest financial capabilities", and, according to the ToR of the EC, on the "Strategic Plan, the Work Programme, and the budget" of GBIF. The GB also decides on its own RoP and – according to these RoP – on the ToRs of the EC and any other bodies.

The GB has approved the establishment of three "standing committees", namely the Budget, Science, and Nodes Committees. The latter is amended by a Nodes Steering Group and regional Nodes subcommittees – all of which are confirmed by the GB RoP version of 2018. The committees are advisory in nature and can make "recommendations to the Governing Board, the Executive Committee and the Secretariat".

According to the GB RoP, "The Executive Committee will be comprised [sic] of the Chair, Vice-Chairs, Executive Secretary (ex officio member), and the Chairs of all standing committees that it has constituted". According to the standing committees' ToRs, "The following individuals serve as ex-officio members of the Science Committee: Chair and Vice-Chairs of the Governing Board [and the] Executive Secretary." This clause appears, analogously, in the Budget Committee's ToR.

Some of the key roles of the EC are to:

- "Oversee the Supplementary Fund and decide on which contributions can be accepted into the Supplementary Fund when contributors have stipulated particular uses of the contributed funds" – under the guideline of the GB RoP that such particular uses must be "consistent with the strategic goals of GBIF".

- "Decide

- on whether a petition to become an Associate or Voting Participants meet the criteria for becoming a GBIF member;

- on Participants or non-members to be observers to GBIF;

- on petitions for affiliation to GBIF;

- […]

- on members to be appointed to the Science Committee and Budget Committee" (the Nodes committee consists of all Node managers appointed by the participants)."

While the EC and the Budget and Science Committees meet more frequently than the GB, the Nodes Committee, for the practical reason of its size, meets only every other year. This meeting schedule is compensated for by the Nodes Steering Group (NSG) and regional Nodes groups, which meet more frequently.

8.1.2 Work Program

Having defined "Participant Biodiversity Information Facility (BIF or Node)", the MoU states that "The Participant BIF has the overall objective of promoting, coordinating and facilitating the mobilisation and use of biodiversity data among all the relevant stakeholders within the Participant's domain, …", thus distributing much of the work of GBIF.

GBIF's current 5-year implementation plan and the annual work programs, see, "GBIF Implementation Plan 2017-2021 and Annual Work Programme 2019" (GBIF 2018), are structured according to the organization's Strategic Plan (see Annex 10.3). It emphasizes the supporting role of the Secretariat at the "Participant level" towards biodiversity experts, namely, to empower the GBIF network, its stakeholders, and partners, especially to enhance the capacities of the Participants' BIFs or Nodes and to enable their networking.

At the "global level", GBIF serves the broad biodiversity informatics community by providing the standards, tools, and the portal and stimulating the mediation of quality assured data, particularly for filling gaps. The implementation plan also recognizes a third level, the "Publishers level", as "Data-holding institutions, agencies and individual researchers [that] are the foundations on which GBIF depends".

The execution of these programs is monitored by the EC. The three standing committees' main role is to develop recommendations on the Strategic Plan and Work Programs.

8.1.3 The Secretariat

The Secretariat performs the execution and coordination of the community- and informatics-related elements of the Work Programs, and the Participation and financial management. The Secretariat is currently organized into four teams, as expanded upon in section 1.2.

The publicly available work plan for 2019 (GBIF 2018) discloses the non-salary costs and the staffing of the Secretariat's teams (also available through the directory):

- Participation and Engagement (eight staff).
- Data Products (four staff).
- Informatics (eight staff, with one additional position to be filled).
- Administration (five staff, plus student support).

The gender balance at the Secretariat may be considered reasonable, especially for an informatics-heavy organization (10 female staff and 17 male staff), but this will require ongoing attention.

8.2. SWOT Compilation[137]

8.2.1 Strengths

- GBIF has built a sustainable organization. There was universal sentiment that the fact that GBIF has been sustained for 19+ years is a noteworthy achievement! It has been able to be flexible and responsive and manage to get the job done responding to stakeholder needs. Several individuals noted that there are not enough resources in general, but GBIF is agile, uses its resources wisely and strategically, and has contingency plans in place. Numerous individuals commented that there is no

[137] This section provides the summarized, but also individual SWOTs of GBIF, as provided by the interviewees. It is not an analysis and attempts to convey the original language. For an explanation and rationale of this format, see section 1.5.

alternative to GBIF and that it is essential for understanding and conserving biodiversity.

- GBIF is staffed with exceptional people. Several individuals noted that GBIF is not top-heavy in any way and staff have continually learned and improved. Others noted that GBIF staff members are “great”, “not defensive”, and “always looking for opportunities”.

- Not a few interviewees named the enthusiasm and commitment of Secretariat personnel as key to the recent success and impact of GBIF.

- Many individuals noted that GBIF had a terrific leader the last few years and they were very pleased with the direction GBIF is following. Several stated that they were delighted that Joe Miller is taking over, and it would be great if Donald Hobern could stay engaged for a year to two afterwards to ensure continuity.

8.2.2 Weaknesses

- Size versus agility. As GBIF has grown it has necessarily operated more like a government agency, which makes it less nimble and more bureaucratic in nature. At the same time, GBIF has been agile and sometimes will move forward to get things done without fully engaging the standards process (e.g., ad hoc vocabularies that are not community vetted). This can be problematic with respect to promoting standards or teaching.

- Although summary information is posted, some individuals found it difficult to understand the true role of each committee.

- Gender disparity in staff and participants: Several individuals asserted that greater gender balance was needed within the organization and the community.

- Regional representation within the different committees of GBIF is biased toward the wealthy countries, including in the Scientific Committee, Budget Committee, and Regions [Nodes sub-] Committee. Most of the representatives that assist the Governing Board belong to European countries, United States, and Canada.

- Payment for the trips to meetings are reimbursed. However, for many countries in the Latin American region it is sometimes difficult to advance the money to be able to travel.
  Despite the fact that there is always the possibility to connect via skype, it is not the same as face-to-face interactions.

- A few individuals noted that like any large and growing enterprise, GBIF is spread too thin and is trying to do too many things. For example, GBIF is attempting to incorporate genomics data, but has a lack of understanding about it (i.e., little internal expertise) and how to effectively incorporate such data. A couple of individuals questioned how much of an education role GBIF should take, as well as the suitability of its role as a funder of projects.

- A few interviewees pointed out that the current success and standing of GBIF might depend on the committed and competent individuals at the Secretariat (and also at individual Nodes) and thus on the success of leadership in retaining staff.

- One software developer noted that he/she would like to program applications that are useful across GBIF, but to be a developer that does not sit in Copenhagen. An

administrator questioned whether there is a way that GBIF programmers could be located in other countries with half support from GBIF and half support from the Node.

- There is a singular, unreliable funding model. One interviewee asserted that the GBIF funding model is based on funding from national governments, but GBIF products and services could and should also be supported through non-governmental sources. Government funding can be bad because governments are always changing and, therefore, are not reliable. GBIF has few fund-raising efforts besides governmental MOUs and no fund-raisers per se. Key questions come to mind that require deliberation and action: How sustainable is GBIF? Is their funding model sustainable? Can it fall apart? What can be done to develop funding that will "float all (biodiversity) boats".

8.2.3 Opportunities

- GBIF could consider developing a strategy for engaging and seeking support from business and industry, private foundations, universities, museums, governmental organizations, NGOs, and the public. Several individuals felt that GBIF should broadly diversify its revenue stream and suggested holding an initial workshop with thought leaders and entrepreneurs to develop such a strategy. A workshop might examine: (1) roles of engagement for private-sector funding and other partnerships (e.g., Google, Microsoft, Bill and Melinda Gates Foundation, Sloan Foundation); (2) implementation of different types of memberships and sponsorship categories (e.g., corporate memberships, tiered donor memberships, and others), as well as the underlying value proposition; (3) creation of a professional GBIF society for organizational and individual members; and (4) identifying new potential funding sources that have not been considered.

- One individual noted that agencies in the US such as the USGS [United States Geological Survey] have supported internal investments in biodiversity informatics, but it may prove more beneficial and cost-effective to look at co-supporting GBIF activities that also directly benefit the US. It was also noted that the US will host the GBIF annual meeting in 2021, and this will be an opportunity to further enhance US involvement by agencies, biodiversity collections, and private foundations.

- Keep the focus on what you do well. Some individuals felt that GBIF's current role (as a data aggregator and connector among organizations) and membership model were probably adequate and GBIF only needs to intensify its efforts to grow and retain members.

- Identify the current and possible future leadership role(s) and value proposition for GBIF. Several individuals suggested that GBIF host a workshop that would bring thought leaders and biodiversity experts together to: (1) perform a targeted gap analysis of where [the community] could get better data and identify what is missing currently; (2) establish targets for specimen digitization efforts; (3) assess the role of GBIF in supporting observational data, such as citizen science data; (4) define a clear value proposition so that countries and other stakeholders want to be a part of GBIF; (5) examine how GBIF could work with National Museums around the world (e.g., see the Research Museum Forum that was recently led by the Nature Museum in Berlin – a 2nd meeting is scheduled to occur soon in Washington, DC); and (6) provide leadership to the biodiversity community with respect to sustaining in-country infrastructure.

- One individual noted that "Organizations like GBIF are very complex; it needs a clear message and a strategy with the various players in the field to turn competitors into collaborators and supporters."

- Sensitize governments and universities to fund biodiversity informatics activities at the national Nodes.

- Encourage developing country representatives to be more involved in the GBIF organization.

- Strengthen participation of the Science Committee in GBIF activities.

- Maintain a global presence. GBIF should continue working with major international organizations such as the UN, CBD, GEO, and GEO BON to ensure that GBIF is integrated. Making data free to governments, scientists, and non-profits is a notable achievement and provides GBIF a place at the table.

- Task and resource sharing. Consider new models for how GBIF can leverage and share resources (e.g., personnel, infrastructure) with and among key stakeholders such as ALA, iDigBio, DiSSCo, TDWG, and the Society for the Preservation of Natural History Collections).

8.2.4 Threats

- Instability of the funding model. The GBIF revenue stream is capped based on the number of countries. Although most interviewees thought that an order of magnitude increase in funding for GBIF could easily be justified, a majority of respondents felt that GBIF is currently not threatened financially but did question whether the current funding model was sustainable given today's political climate. This commonly held concern was accompanied by a number of suggestions: grow the number of countries supporting GBIF (especially BRICs); add a community development officer; develop a clear and concise value proposition, along with accompanying success stories related to topics that resonate with the public and funders (e.g., agriculture, invasive species, or public health); create an endowment for GBIF; and develop a more decentralized funding model.

- Commercial enterprises that add value to GBIF data. A few individuals noted that some existing organizations will (or could) take GBIF data, add 5% value to the data, and then sell the data at a premium to others. Examples include organizations such as Google, ESRI, and others. Such groups assume GBIF will be there and there is money to be made off the GBIF-provided data. A possible mitigation strategy is to partner with potential competitors but make it clear that GBIF is central to the collaboration because it is the data provider. This, of course, is somewhat of a double-edged sword, but GBIF needs someone at the table to represent its interests with respect to these competitors and potential collaborators.

- Growth of current collaborators into competitors. A few individuals noted that organizations like GEO, DiSSCo, and iDigBio are currently collaborators. However, they noted the potential for these organizations to expand their scope purposely or indirectly (via scope creep) and become direct competitors with GBIF.

- Bureaucratization of GBIF. Several individuals noted that large national and international organizations are largely ineffective or become that way as they grow. Most felt that GBIF had struck a good balance thus far but cautioned that GBIF should continue to pay attention to this issue as they hoped that GBIF would retain agility, flexibility, and adaptability. Some did express the concern that the "internationalized" GBIF structure may inevitably lead to non-adaptiveness down the road.

8.3. Conclusions

Public pages on governance, work programs, and staffing make GBIF very transparent. It is relatively easy to establish contact with the appropriate function or person in GBIF, despite its global nature and somewhat complex operations.

This does not mean that all information can be found and understood quickly or that its presentation could not be further improved. A new or potential Node manager, for example, might not fully understand the rules and functions about his or her position in relation to the governance of GBIF.

Some of the highest praise of GBIF can be attributed to a functioning governance model, although most of it is explicitly assigned to the staff and Executive Director of the Secretariat.

8.3.1 Governance

The governance system is appropriate, considering the number of entities with which relationships have to be maintained, see Table 7.1, and their global distribution.

In particular, it is scalable to even more Participants, through the functions and powers delegated to the Executive Committee (EC), and through the tiered organization of Nodes, via regional subcommittees and the Nodes Steering Group (NSG).

Furthermore, the EC – represented by all GB chairpersons and the Executive Secretary – is well coordinated with the standing committees, which should ensure good communication among them and with the GB itself. The exception may be the NODES committee, which is just represented by their chair in the EC; other EC members do not participate in the deliberations of NODES or the NSG.

The central position of the EC ensures the ability of GBIF to be organizationally responsive. Besides being represented in the EC, all standing committees are authorized (or even required, as in the cases of NODES/NSG) to offer their advice and recommendations not only to the EC, but also to the GB and the Secretariat directly.

Thus, formally, the representation of individual Nodes is coupled less directly to the decision-making bodies. However, according to our observation, the actual interaction between NSG and a regional Nodes meeting, and Secretariat staff, the Secretary/Director, and with SC and GB chairpersons, is cordial and productive. How effective this important interaction is clearly depends on thoughtful co-location and design of the agenda of such meetings, as facilitated by the Secretariat. The feasibility of such designs may, however, depend on the region(s) involved.

The physical distance between a region and the Secretariat and the average distance between Nodes are important problems specific to less than wealthy countries. They have problems in securing the funds necessary to attend GBIF meetings and this affects the participation in the organization, including the regional coordination of Nodes. If unchecked, this reduces the effectiveness of governance structures to work as designed in terms of the representation and influence on decision making of individual Nodes.

The Terms of Reference of the Science Committee are, as the name implies, heavily focused towards biodiversity science. This leaves a gap in the governance arrangements in the representation of other users, especially at the science-policy interface. These users are represented as Other Associates or partners with observer status in the GB, at best. By the rules, their representation in committees is typically restricted to those also operating a GBIF Node. Note that CBD, IPBES, and IPCC are partners of GBIF but do not have Nodes. Thus, while there are some good working relationships between the Secretariat and these organizations – see

Chapter 2 – they are not actively represented in governance. Their under-representation may result in less than optimal commitment to work with GBIF and also may result in the longer-term in opportunities missed for greater impact.

According to the spirit and terms of the MoU and governance rules, national users and institutions with no international organization representing them must resort to the capacities of the national Nodes and their established ways to reach out to GBIF governance. Tim Hirsch, the Deputy Director and Head of Participation and Engagement at the Secretariat acknowledges: "We do have a challenge in making sure that particularly the user base of GBIF is well represented in our governance". The SWOTs seem to reflect this weakness, in that all kinds of outreach to organizations and communities are suggested, in order to establish their requirements and priorities, and to prove the value of GBIF to them.

The following conclusions are centered on the three major threats identified in the SWOTs, which require attention for improving GBIF's governance.

The first big threat and corresponding opportunities – brought up in this way from North America – relates to a potential catastrophic loss of multiple big financial contributors and the general perception of governments as (possibly?) unreliable funders. The alternative, in particular to seek substantial contributions from philanthropic organizations and commercial entities, which could offset even the withdrawal of large governmental funders, would certainly require intense deliberations of the governance bodies (and, eventually, effective marketing). Currently, there appears to be only the status quo, with reasonably stable funding enabling a roughly steady activity, supported by the existing management of Participation and funding, or a plan X, about how to shut GBIF down in an orderly manner.

The second threat, about undermining the value proposition and visibility of GBIF by a commercial activity making use of freely available data sourced by GBIF, and its possible mitigation by cooperation with these same companies, needs at least close observation. Considering the nature of these players, the Executive Committee would need to be prepared with a well-considered[138] policy and authorized to respond quickly.

The third potential threat suggests that good current collaborators may turn into competitors. At the same time, there are enough high-level advocates from outside GBIF who could testify that it would be foolish to duplicate GBIF's work. The very good placement of GBIF within the CBD recommendations – see Chapter 2 – is certainly a good sign in this area of the science-policy interface. Still, aside from the idea of keeping enough of a competitive edge, it might be wise to develop a strategic approach on how to keep those potential competitors convinced of continued cooperation.

8.3.2 Work Program

Overall, the structure of the work program reflects the goals of GBIF and the requirements of its stakeholders. It is successfully executed, as proven by the successes of GBIF, the general satisfaction of stakeholders and, most of all, repeated calls for GBIF to deliver more of what it has already delivered, or to assume leadership about an issue, all of which were reported in previous chapters.

---

[138] In some countries, it is government policy for publicly-funded data to be used freely by commercial entities (e.g., data are encouraged to be used by commercial entities for value added products and services), while elsewhere, there are (implicit) policies against the private sector using public data freely for profit-making applications.

Some non-structural weaknesses and opportunities about the work program, however, are showing, beyond the thematic ones discussed in previous chapters. Interviewees had diverging, and even contradictory suggestions, however, about how the work program should develop.

There are clear indicators of the Secretariat being spread thin on many fronts. As an example from the work program, some diverse critiques about the management of relationships were already contained in the SWOTs of some previous chapters. It is therefore enough here to conclude that there appear to be insufficient communications with some Other Associate Participants, Affiliates, or Partners.

That there are some occasional complaints in a background of universal praise of the Secretariat staff should come as no surprise, however, when one recognizes that the work program allocates just eight staff to the task of maintaining over 100 institutional relationships and also providing capacity enhancement to and coordinating the work of almost 100 GBIF Nodes (see Table 7.1). Similarly, 12 staff for informatics and data work may seem comfortable until one recognizes that the Secretariat has to deal with the technical and data-management idiosyncrasies of some 1,400 Publishers and is expected to establish, prioritize, and implement the standards and functional requirements of dozens of stakeholder groups. Likewise, the budgets allocated to the many individual work items in the annual plan are marginal, compared to the size of the problems. For example, the budget to "revise the GBIF data standards and to begin modernization of the Integrated Publishing Toolkit (IPT)" includes just 50,000 EUR "pass through" money.

There needs to be a strategic decision, then, whether:

- the current level of addressing individual work items is actually "good enough", considering the long-term satisfaction of stakeholders, or
- there should be a reduction in scope of the work matching the resources, or
- there could be an increase in resources to improve quality at the current scope, or
- funding should be sought for an extension of the scope of the work program to accommodate a few or many of the "opportunities", some of which may be necessary to maintain its current standing, and others, to strengthen it.

8.3.3 The Secretariat

It is worth repeating that there was universal praise of the individuals working at the Secretariat, citing their competence, dedication and enthusiasm, among other positive traits. It must be considered a success of the Secretariat leadership to have attracted and retained such an exemplary workforce.

Some factors of the successful work and the assembly of this staff are contributed by the hosting of the Secretariat at the University of Copenhagen, from stable and inexpensive infrastructure and administrative support to a family-friendly and safe social environment. Downsides have been named, such as of Copenhagen being not as easily reachable as some international air traffic hubs, or Denmark not being a native English-speaking country, and the city being expensive. But if one were to find a location combining Copenhagen's benefits with other ones, it would certainly not be an easy move. Thus, there might be comparably or more attractive hosting situations, but only in theory. The potential of losing staff over a relocation would have to be factored in as well.

It has also been suggested earlier in this report, particularly about the informatics work, that some or even much of the work could be spread out over a number of institutions, globally.

However, our interviews with the Secretariat staff revealed that – after having developed the current constellation of teams – the intensive, on-site interaction between informatics, data, and Participation teams has led to the much-improved development of GBIF. Effective pure teleworking or even effective distributed teams may be a myth – at least in a complex and dynamic environment such as GBIF's. Therefore, plans for distributing work would need to be considered very carefully.

As stated in the previous section, the Secretariat staff is spread thin over a large scope of work. This is not a new situation, as even the first review (CODATA 2005) stated, that "Each component of the Work Programme depends almost entirely on the work of one key staff member, […]". The appreciation of this made interviewees and reviewers think about redundancy and plans for the loss of key staff. Leadership at the Secretariat has handled this situation well, so far. It would most probably not go away even if the overall budget were to increase substantially, since the increase would be coupled with expectations of an extension of scope.

In summary, we find on item 2 of the statement of task, that:

- The governance of GBIF is well adapted to its foundational premise as an international, distributed organization, agreed upon by many countries, using a light-weight contractual arrangement.

- Its bodies provide the structure and means for a high degree of consensus building while maintaining enough flexibility for the operations of GBIF to adapt to rapid changes of the landscape.

- However, the representation of users and other stakeholders in the landscape appears in need of further strengthening.

- The internal organization of the Secretariat and the organization of national Nodes, particularly the relatively recent introduction of regional organization of Nodes, is adequate at GBIF's current scope, but may need adaptation to accommodate future growth of Participation and GBIF's activities.

- Funding of GBIF has been reasonably stable in the past and its handling by the Secretariat prudent. However, it is insufficient to allow for many of the foreseeable paths into the future, which GBIF may need to follow. Increases in Participation, of the contributions from each Participant, and of third-party funding may all be needed to account for this.

- The Secretariat is well hosted and supported by the University of Copenhagen. We see no evidence of a need to relocate it. Secretariat staff is excellent and a major source of GBIF's reputation.

- However, both the governance bodies and the Secretariat are strained to their limits at the current size of GBIF. They will need to be further adapted and made scalable to accommodate additional Participation, e.g., a doubling of the number of participating countries.

Beyond our conclusions in this chapter, which already provide some suggestions for improvement, the major challenges for the future and associated recommendations are provided in Chapter 9, in the context of all other topics of this review.

9 THE FUTURE OF GBIF[139]

In the body of this review report, consisting of Chapters 2 to 8, we looked at many crucial aspects and functions of GBIF. Chapter 2 illuminated the landscape of biodiversity and its actors in research, policy-setting, and informatics infrastructures, and how GBIF interacts with each. Chapters 3 and 4 analyzed the most general requirements, which are independent of GBIF, by data users and data contributors, and the demands placed on them and all other actors, including GBIF, to show the impact and effectiveness of their activities on scientific and societal progress. Chapters 5 to 8 provided and analyzed in depth the facts and SWOTs reported about GBIF, in the areas of data, standards and technologies, its support for and outreach to the stakeholders and actors, and GBIF's organization, funding, and governance. Chapters 3 to 8 also reported our conclusions in detail for each topic.

This chapter delivers on the remaining task of this report, to "consider the challenges in the next 5-10 years that GBIF needs to be prepared to meet" and "provide recommendations on areas needing attention and improvement." Keeping the limitations of our method in mind, we offer a summary of our conclusions in the first section of this chapter. The second section then sets forth our recommendations.

9.1. Summary of Conclusions

First of all, our findings show that GBIF is the most comprehensive, openly available, application-agnostic (most unbiased), easiest-to-use, and modern access point to known digital occurrence data. Consequently, as a global, distributed platform, GBIF is viewed widely as being a major success and a great improvement over alternative solutions; this relates equally to the data it provides and to its capacity building activities. There is now a broadening range of applications of GBIF-mediated data that matches the OECD expectations of 1999.

> *"No one comes close to offering what GBIF does and no one else should; we should all get behind GBIF and help make it thrive."*
> Healy Hamilton, Ph.D., Chief Scientist, NatureServe[140]

This success would not have been possible without GBIF's foundational principle of a distributed organization and its pursuit of a dependable funding structure. Its 20-year existence alone – and the way it is funded – creates much of the trust on which most of the current, extremely high expectations depend.

Under the heading of Capacity Enhancement, GBIF is well on its way to create methods to increase the strength and stability of its network of Nodes and Publishers, and to establish a global network of well-connected experts. GBIF has built and maintains relations with a large and very heterogeneous array of other actors in the fields of biodiversity informatics and conservation, and at the science-policy interface. It has established itself variously as a member, partner, and coordinator of biodiversity-related activities.

---

[139] This chapter addresses items 5 and 6 of the statement of task ("Consider the challenges in the next 5-10 years that GBIF needs to be prepared to meet" and "Provide recommendations on areas needing attention and improvement."). The findings of this chapter are further summarized in the Executive Summary, which precedes Chapter 1 of this report.

[140] http://www.natureserve.org/

The successes would not have been possible without the much-lauded staff at the Secretariat, which was uniformly praised as competent, dedicated, responsive, and enthusiastic. We found the Secretariat, and GBIF as a whole to be an agile, learning, and evolving organization. Few of the SWOT comments will come as a surprise to the Secretariat. In particular, its leadership, which is fully aware at least of the most important issues raised, has actually addressed many of them in the current strategy and work program.

The governance of GBIF – its bodies and rules – has contributed particularly to the essential agility and evolution of GBIF. While it is broadly inclusive of Participants and Affiliates, it has a core, consisting of the Executive Committee and the standing committees supporting it, which ensures that GBIF remains adaptable and flexible.

Still, the analysis of SWOTs indicates that major challenges lie ahead, and that proceeding as before may not be an option for the continued success of GBIF. The most obvious challenges are the issues of data quality – in all its facets --and formal Participation. Beyond that, the SWOTs delivered a multitude of views and suggestions, from which we provide a synthesis here, but which cannot cover every one of them. Future planners may need to go back to the individual chapters for more detail.

It does actually belong to the successes of GBIF that most experts interviewed appear to see GBIF able to address or contribute to the solution to the challenges identified, even the challenges that are beyond GBIF's current remit.

9.1.1 Successes

This description of successes is built on the strengths identified in previous chapters, but also on the opportunities brought forward, with the expectation that GBIF would be able to turn them into reality if it has not already done so, by the time this report has been published.

S1, **Biodiversity Informatics**: GBIF is successful in the global biodiversity informatics landscape in that it:

- mediates an unrivalled number of *openly available* occurrence records;
- which are used in numerous academic studies, in applied research, and for reports at the science-policy interface[141];
- makes it possible to track their uses in science, numerically and qualitatively;
- enables Publishers and Nodes to claim their contribution to science;
- built a geographically inclusive organizational and technical network of Nodes and Publishers;
- maintains an intimate connection to taxonomy, museums, and other collections;
- has built a standing in this community, particularly through the staff at the Secretariat, which is viewed as being highly competent;
- has developed excellent technology and systems;

[141] Importantly, it has been acknowledged by CBD that GBIF "contributes directly to Aichi Biodiversity Targets 1, 2, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 17, 19 and 20", https://www.cbd.int/cooperation/csp/gbif.shtml

- is trusted as a neutral broker;

- and has established its volume of data as an indicator towards Aichi target 19 and is involved in defining further indicators at the science-policy interface.

S2, Capacity Enhancement: Supported by supplementary funding, GBIF has rapidly increased its ability to perform and manage Capacity Enhancement activities:

- for existing Nodes and for new ones,

- in countries that are not yet Participants.

This has been done particularly by

- focusing on under-served countries,

- leveraging the capacity of strong Nodes, Publishers, and other community members, such as voluntary trainers,

thereby

- creating an emerging global network of experts, and

- laying the groundwork for further Participation outside North America and (western) Europe.

S3, Networking: GBIF has built a very broad network of relations with numerous and heterogeneous actors in the biodiversity landscape, which

- is largely representative of globally organized biodiversity informatics, and

- creates links for GBIF to important governmental and other policy actors.

9.1.2 Challenges

Nonetheless, our analysis also indicates that major challenges lie ahead and that proceeding as before may not be an option for the continued success of GBIF. Some of the challenges are fundamental and will certainly needing a decade to be resolved. Other challenges are practical in nature and can be addressed in the short term, i.e., within one to three years at the current core budget levels. The overarching challenges have been distilled from the conclusions in Chapters 3 to 8 (which provide much more detail).

In the data and technologies domain, two major issues loom large.

C1, Fitness for Purpose: First, the fitness for purpose of current GBIF-mediated data needs to be made even more transparent and further improved. Transparency relates to indications of accuracy and best practices in the usage of data. Improvement depends strongly on better and more effective and efficient quality assurance methods, especially in employing and engaging user and community feedbacks, which may be addressable in the short term, and on closing taxonomic and geographic gaps, which is a longer-term task, mostly outside GBIF's control.

C2, Non-traditional Data Sources: The second major issue is the expected rise in the quantity and importance of non-traditional biodiversity data, such as from remote sensing, genomic analysis methods, other automated observations, and contributions by citizens – including indigenous knowledge. Some of these developments appear to accelerate dramatically

through the recent application of machine learning technologies and the pervasive availability of networks of people and smart devices. Here, the question cannot be whether, but rather how and how deeply GBIF should be engaged. Many of the interviewees and the Secretariat consider any of a number of engagements necessary for GBIF to remain relevant, while other interviewees warn about "scope creep" or quality issues with specific data sources.

Consequently, GBIF will need to make some well-considered choices when further increasing its data holdings, which has been a persistent and crucial goal since GBIF's inception. In some cases, the technical volume of data may become a resultant issue, as it requires either more of GBIF's own hardware or cooperation with big data repositories or cloud providers. The volume of data from "traditional" sources, such as digitized specimen data and records derived from scholarly taxonomic publications, will probably increase substantially as well.

As uncovered in the preceding chapters, there are additional major issues **in the domain of GBIF's organization and community**.

C3, Gaps in Formal Participation: The weakness of biodiversity data gaps is not specific to GBIF, but a universal problem of biodiversity knowledge (see section 5.1): Apparently all major actors in biodiversity – including IPBES – experience the most pronounced data gaps outside the US and Europe. This weakness can, in GBIF's context, partially be attributed to the first major issue: gaps in formal Participation from biodiversity-rich countries and, particularly, from Asia as a whole. But the identified strengths as well as opportunities indicate that GBIF is structurally better positioned than most to close this gap "on the ground": GBIF has been rather successful in Latin America and newly in Africa, the latter mainly based on a specific capacity-enhancement program, BID. Here in particular the principle of regional networks of Nodes has demonstrated opportunities, but also limits, which are mainly due to the shortfalls in funding (see below). In Asia, there is no disagreement in principle on the strengths of GBIF, but the expectation that a much better communication of a cost-benefit analysis would help build formal Participation.

C4, Varied Strengths of Nodes: In GBIF's organizational principle, fully functional Nodes are fundamental to address almost all issues, from closing gaps through their outreach to the full national biodiversity community, to improving other quality traits of data (by working with those Publishers that the Nodes have endorsed), or to organizing, performing, and supporting capacity enhancement through training and education.

However, the second major issue in this domain is that even the concept of national Nodes (or national Biodiversity Information Facilities) has its limits. There are 20 national Nodes in Europe, serving it well. But there is general skepticism about the Nodes concept in the comparably sized US, with just one Node serving it.

A number of well-funded national Nodes show that the GBIF concept can be implemented. However, in most countries the Nodes are not staffed to achieve full functionality. This can be compensated to some extent by a strong hosting institution, but it does potentially introduce the risk of an unwanted thematic bias. Financially strong hosting institutions exist, for example, in some European countries, but hosting institutions in most countries outside Europe lack such resources.

In Asia, a fundamental principle of GBIF funding currently precludes a dense Nodes network to be built at all: Nodes, and their support from core funding, are tied to formal Participation. In all GBIF regions except in North America and western Europe, the language barrier appears to be a real impediment to capacity building and also in the ability of GBIF's Publishers to contribute data.

The major challenges in the previous domains inevitably lead to a consideration of **the domain of funding.**

C5, Insufficient Funding: Here the first major issue is that the current level of GBIF funding – including the national funding of Nodes and supplementary funds – is barely sufficient to perform its activities at the current scope of its strategy and work programs. An increase in formal Participation may compensate for part or all of the Secretariat's cost increase associated with growth in all forms of Participation, Affiliations, and Partnerships. But it definitely would not be sufficient to undertake big strategic moves, as they may be determined necessary in the long term. Still, the mostly hidden cost of adequately staffing Nodes may be in need of the biggest increase. National funding for this is, and may remain, unavailable in most countries.

C6, Funding Instability: Due to developments in the political realm, some observers expect a second major funding issue in that GBIF may face increased instability of core funding over the next years (see Chapter 7). GBIF is prepared to absorb major fluctuations to at least maintain stability in the short term and to avoid premature sacrifices in its work program. Some major successes have been achieved using supplementary funding, but GBIF's governance functions are largely adapted to managing these only as moderate contributions to its existing work programs. Consequently, substantial third-party funding, for example by philanthropic organizations or industry, cannot bolster the core of GBIF's work program without major changes in governance, and possibly, strategy.

C7, **Management of Expectations:** The final major challenge lies in the **management of expectations**[142]**.** The weaknesses, opportunities and threats reported in the previous chapters' SWOT sections reflect the expectations of GBIF's stakeholders. The fact that its community assumes that GBIF could meet great expectations, is, above all, a very good sign about its reputation, but it reveals in many cases fundamental misconceptions about GBIF's resources or the effort required. Consequently, while addressing those challenges that are in fact within its remit and its limits of resources, GBIF will urgently need to manage expectations about what it can achieve and when. If the organization allows overoptimistic expectations to stand, it might experience an unnecessary decrease in its reputation – most probably in those regions where it needs to close gaps in data.

9.2. Recommendations

The high expectations on GBIF as well as the technical and scientific challenges ahead strongly suggest that, while GBIF is well positioned to meet them, a scenario of "business as usual" may fail to preserve GBIF's relevance. Strong growth should be considered seriously, with implications for GBIF's technology, services, organization and funding.

We present our recommendations in three sub-sections: The first one presents generic guidelines mainly formulated to maintain a focus on what we have identified as the main factors in GBIF's success and reputation. Specific short-term recommendations in the second section aim to encourage progress on the most urgent challenges, which could be achieved at the current funding level and structure, and without revising the organization's strategy. The third section provides recommendations related to all of the challenges, but which probably can only be approached by a longer-term activity or by a substantial expansion of funding – and probably funding sources – in the next two funding cycles (of five years, each). That third set of

[142] "management of expectations" is a technical term used in software development with an aim to reduce unrealistic expectations of costumer or users about what can be achieved, and when it can be achieved.

recommendations thus might require an explicit expression in strategy, organization, or even governance.

Our specific recommendations are generally ordered in a sequence that reflects a plausible phasing of related activities, particularly as they may depend on intermediate results from the lower-numbered recommendations. The final recommendation relates to the potential level and structure of funding that may be necessary in the medium and longer term, to meet most of the current expectations of GBIF's stakeholders. Obviously, some decisions about this will need to be made prior to the start of the corresponding funding cycles, namely in 2020 and five years later.

Consequently, we separate short-term recommendations, to be implemented within the next two to three years, and longer-term recommendations, for the next two funding cycles (the first one occasionally referred to as medium term). Beyond this, we do not specify target dates. Also, we mostly refrain from specifying numerical targets which – while evidently well-suited as indicators for internal benchmarking – might be misleading, if pursued to meet outside evaluation (e.g., number of records) or too dependent on the volatile political situation (e.g., number of Participants in Asia). Also, the time to achieve a target certainly depends on the overall funding of GBIF, whose development may lead to unpredictable acceleration or delay on the timescale of 10+ years. Note that many of the specific examples of how to implement each recommendation – mostly formatted as lists – may appear somewhat arbitrary but are directly derived from the interviewees' SWOTs.

Finally, because GBIF is navigating a highly dynamic environment, it needs to retain the ability to adapt as quickly as possible. Therefore, many recommendations deal with the evaluation of opportunities and threats, but leave the planning and decisions to the appropriate GBIF bodies – which has worked well in previous years. However, we propose some guidelines for this planning in the next section (9.2.1), and specific suggestions in the following two sections (9.3.2 and 9.3.3).

9.2.1 General Recommendations

In order to maintain and strengthen GBIF's relevance and standing, we recommend the consideration of growth paths in a number of dimensions. Such growth will enable GBIF to support biodiversity research broadly in the future. To do so, the organization will probably also need to participate in and contribute to the rapid development of scientific data methods and standards.

An important first guideline in this dynamic environment of requirements and opportunities is to maintain a strategic and operational focus, while working with limited resources and in partnership with others.

One focal area is to carefully select the data types GBIF needs to embrace, and the technologies and services to mediate and provide them to users, while maintaining and improving their fitness for diverse purposes. The other focal area should be to fully extend the global network of Nodes and Publishers, while keeping it well managed and organized, and technologically supported.

"May all your problems be technical"
Jim Gray, Microsoft Research

Each data, standards, and technology issue will, in hindsight and in the abstract, probably seem straightforward or even obvious. But, along the way we recommend:

R1a, Data Quality and Quantity: GBIF needs to build on and maintain its reputation as the most comprehensive source of openly available global occurrence data. This means that it should continue the trajectory of growth of the data that it mediates, in quantity as in quality, by extending and deepening relations within its network, and by supporting non-traditional types of biodiversity data.

R1b, Technology and standards: GBIF needs to maintain or attain leadership in essential technological and standardization areas related to biodiversity informatics. In order to do so, it should continue to work actively with other stakeholders in the "landscape", such as the researchers at the forefront of such fields as metagenomics, remote sensing, and observation and cloud technologies, to keep abreast of developing data sources, standards, and technologies. GBIF should add a long-term focus on IT security.

Among the technical and diplomatic issues are the ongoing creation of *one* authoritative taxonomy (or a system supporting multiple taxonomies), the linking of taxonomic and genetic concepts, their implementation in nomenclatures and identifiers, the determination of computable Essential Biodiversity Variables (EBVs), and indicators for use at the science-policy interface. All of these examples are of major importance to maintain and improve the quality and impact of GBIF's services.

As social architecture is an essential component of any information infrastructure undertaking, the essential obstacles on the way forward will be cultural, diplomatic, and organizational rather than technical. Therefore, at the core, most of the following recommendations are concerned with community communication abilities and activities.

R2, Networking: GBIF should maintain and even strengthen its capabilities to network its stakeholders and to lead them to cooperate and build consensus.
GBIF, and in particular the Secretariat, should continue to be seen as a neutral broker. In general, it should not take sides where consensus has not yet been achieved. Neither should it be a passive observer, however, as it can help mediate or accelerate processes that can lead either to a decision or a mitigation of continued disagreement. Where it is essential to implement its strategy, in a few carefully chosen areas only, GBIF may need to break through some remaining disagreement with a solution, as it did with committing to fully open data.

9.2.2 Short-Term Recommendations

Most major challenges (e.g., gaps and biases in data, see section 5.1) are not resolvable by GBIF on its own or by simple measures. Therefore, many of those can only be solved in the long-term. Still, it is necessary to address them somehow even in the short-term, if only to reassure stakeholders of GBIF's awareness and of progress that the organization is making. GBIF should show that it is working on them or even provide what could be seen as short-term workarounds.

Some of the weaknesses, opportunities, and even threats described in the interviews appear to be based on incomplete or outdated knowledge about GBIF, its data, or the functions of its portal, or about GBIF's work program or the limits of its resources. Therefore, many of our short-term recommendations are related to improved communication and GBIF's visibility and to managing stakeholders' expectations. Most of this could and should be done in the next two to three years, and probably can be done by employing current resources.

We recommend that GBIF, in the next two to three years:

R3, Visibility: Establish greater visibility among decision-makers and the policy community in the biodiversity arena, as well as throughout the scientific community relevant to biodiversity data:

- Develop a strategy to identify key ("important") individuals at the global stage and make them thematic champions for GBIF (such as lead authors of relevant reports, society presidents, retired senior staff of governmental agencies or departments).

- Likewise, identify and support [a network of] regional or national champions, in addition to Nodes' managers and Heads of Delegation.

- Recruit those champions and help them to make GBIF's points not only within the scientific community, but also with funders, including national governments, philanthropists, development banks, and other potential funding entities.

R4a, Value Proposition: Develop compelling value propositions, especially as a tool for the champions and the Nodes:

- Develop, enrich, translate, and actively distribute high-level success stories.

- Develop "elevator pitches" and one-page summaries for each kind of stakeholder.

- Review the GBIF portal: Are stakeholder groups led directly to what they need without searching? (But do not compromise the portal's excellent functionality for experts).

- Commission, or suggest Participants to commission, a study or targeted studies on the monetary value of sharing biodiversity data and of the added value that data infrastructures provide to it.

R4b, Grand Challenge: Choose a "grand" but doable real-world challenge (i.e., low-hanging fruit). In this regard, GBIF should identify associated data gaps and focus community effort on filling those gaps (where data are known to exist), work with scientists to perform the analysis, and actively disseminate how the collections, countries, and GBIF are coordinated to solve or contribute to the solution of such a significant real-world problem.

R5, Eastern Participation: Build Participation in the East. This is a long-term activity, but flexibility and creativity should be employed for short-term measures, which would show progress and build good will for full Participation in GBIF, particularly in Asia. Some measures come to mind, such as:

- Empower and visibly embrace "Other" Affiliates and Publishers, especially those from countries that are not Participants, such as the ASEAN Centre for Biodiversity, the World Data Center for Microorganisms (WDCM), and the Asian microbiology association. Let them appear in maps of the GBIF network.

- Help organize and stabilize – with supplementary funding only – a network of people and organizations that could become Nodes in Russia, India, and Asia. This should be a foothold for future Participation.

- Find more ways to address the language – and perhaps cultural – issues, beyond the mere translation of the portal, documents, and tools.

R6, Strengthen Nodes: Make the full Nodes concept work in all regions. Each GBIF region has different needs and issues – therefore, each needs their own engagement strategy:

- Make the Nodes concept work in North America – perhaps by reconstituting it as a network of sub-Nodes, as in Germany. Success here may be essential to satisfy a large community, but also to provide templates for other countries with large populations, such as China and India.

- Stabilize and extend regional networks in Africa, Latin America/the Caribbean, and Oceania. There, regional networks of mutually empowering Nodes are – and probably will be for a long time – dependent on third-party funding.

- Systematically secure the motivation and recognition of voluntary mentors, trainers, and educators in all regions. Consider how to extend formal recognition to them.

- Make the importance of Nodes known and increase their standing, e.g., by explicitly empowering NODES and the regional subcommittees to nominate candidates for the Ebbe Nielsen Challenge and Young Researchers Awards, recipients of GBIF Badges, and similar recognition.

- Consider new awards for Publishers (or for a Node and a Publisher together), for trainers, and for citizen science projects.

R7, Communicate Successes: Discuss with stakeholders and show prominently and appealingly what has and can be done with GBIF-mediated data in a scientifically sound way. Focus on high-value products (see R4):

- Collect good examples, particularly their data and software codes and their explanatory text, and build a collection to publish them.

- Further develop the data blog and turn it into a community.

- Explain how to avoid fallacies (e.g., through tests of fitness of data).

- Put some effort into effective, scientific visualization.

- Present these actions at scientific and other conferences, such as at the CBD and other UN fora.

R8, Improve Data Fitness: Develop criteria and metrics to measure data quality and demonstrate improvements in the short term:

- Make sure that key concepts and products meet the needs and consensus of communities, including local communities in biodiversity-rich regions (e.g., the taxonomic backbone/future taxonomy of GBIF. Consider whether the "Catalogue of Life Plus" project consortium[143] is inclusive enough?).

- Make sure that filtering/flagging criteria and methods in the portal are sufficient, well documented, and visible.

---

[143] https://github.com/Sp2000/colplus

- Develop and display quantitative indicators of overall quality.

- Extend a targeted quality tool chest and education resources and training to Nodes, Publishers, and data contributors and users.

R9, Publish Methods and Applications: Continue exploring and publish (e.g., in a prominent data blog) how each new type of biodiversity data is relevant in the context of "traditional" data and vice versa, and explain the problems and compromises involved. For example:

- Explore how GBIF data can be used as ground truth for remote sensing.

- Explain how the metagenomics of the contents of a malaise trap, or even a network of such traps, could be linked with traditional GBIF data or be mediated as GBIF data.

- Publish results and methods of experimental citizen science projects by GBIF network members and successful projects from the broader community – showing how research-grade quality can be achieved.

R10a, Manage Expectations: In managing expectations about what GBIF can or cannot deliver, the organization's leaders should do more to communicate its mission, the magnitude of its funding, as well as scalability issues in general. For example, GBIF probably cannot – and should not:

- fund or manage the digitization of museum collections;

- build or manage citizen science projects;

- fund or manage field work (such as data collection); or

- develop biodiversity curricula for universities.

R10b, Leverage the Network: But GBIF should identify in a strategic planning process where it **could leverage its network** and could succeed in brokering support for such urgently requested activities as:

- Follow-up activities to initial capacity-building projects and their funding.

- Financial support from wealthy countries' big digitization projects to others.

- Collection of best scientific practices and working technologies for citizen science.

- Challenge organizations such as the League of European Research Universities to support universities in less economically developed countries, such as African universities, in building their biodiversity informatics curricula (this would probably need a champion from academia).

To be able to execute some or all of the following medium-term to long-term recommendations below, we recommend the following action in the short-term:

R11, Urgent Deliberations and Decisions of the Governing Board: Discuss and establish guidelines regarding the following issues:

- Follow up on the finding that, considering the importance of mobilizing data and closing gaps, the most important role within national Nodes is that of building the community

of data users, contributors, and Publishers. (However, the role of managing the data is vital for data quality).

- Determine how realistic is it to have more than one staff member (including the Node manager) at each Node. When could that be achieved?

- Identify which of GBIF's core strategies and work programs could or should be (co-) funded by intergovernmental or governmental third parties (e.g., the European Commission, World Bank).

- Decide for which activities is it advisable to look for large or small, one-time or long-term contributions from philanthropic organizations or industry.

- Explore how core funding from Participants could be increased.

9.2.3 Longer-Term Recommendations

In the following, we use "medium term" to mean no longer than the next five-year GBIF funding interval and "long term" to mean up to 10 or more years. Note that some of the recommendations, particularly R15 to R20, may need significant additional funding.

In an era of rapidly growing societal awareness of the importance of biodiversity for daily life, or even survival, of fast development of related observations, and of analytical methods and technologies, a cycle of five or more years to evaluate strategies and strategic workplans may not be sufficient. There are critical issues already visible today, as discussed below, and there probably will be other ones in the future. Many of the Secretariat staff and leadership, members of the Governing Board, and its standing committees have shown their awareness of these issues. Nonetheless, we recommend a strengthening of regular observation, advice, and support from the outside:

R12, Create Advisory Board: The Executive Committee create a small, but inclusive, high-level advisory body (consisting of external experts, doubling as GBIF champions) to advise the Governing Board on strategies, particularly on matters of fundamental importance to the mission, organizational principles, governance, and funding of GBIF.

This new body as well as the other GBIF bodies will benefit from the outputs of some short-term recommendations, but they all will also need regular input about the strategic status, progress, and positioning of GBIF. To this end, we recommend that:

R13, Foresight and Monitoring: The Executive Committee should, with the support of the Secretariat and its standing committees, establish an explicit, strategic foresight and monitoring process, including landscape analysis updates and estimates of the penetration of the potential user base, with the aim to adjust priorities within the work programs and suggest adaptation of strategic approaches and directions to the Governing Board, in a timely manner.

- This process will need to be underpinned by systematic, on-going collection of reliable evidence, whether by carefully targeted surveys, from observations at conferences, meetings, and workshops attended and publications studied, and usage patterns observed at the portal.

- Evaluation of the evidence about expectations cannot simply rely on numbers of downloads or votes for features, but must take into account the actual value created by committing resources (building on the results from R4a).

- Observation of the landscape will result in identification of any dynamic changes (or lack thereof) of the contributions, and realistic or unrealistic aspirations of collaborators and competitors.

- Evaluation of this evidence should happen regularly, that is, annually or at least bi-annually.

There are and certainly will continue to be so many issues to resolve that GBIF and the other stakeholders in the biodiversity informatics landscape could be overwhelmed trying to address them in the many disparate workshops and consultations. Currently, the most promising approach appears to be through the emerging Alliance of Biodiversity Knowledge ("the Alliance"—see section 5.1.4). We therefore recommend that:

R14, Leadership: In the spirit of moderation and cooperation as set out in R1 and R2, GBIF should assume leadership of the Alliance, as requested by major voices in the community, in order to create consensus and collaboration at a much broader scale, including on these issues:

- Work out consensus on extension and integration of existing and new standards, methods, and components at the technical level;

- explore and establish reliable ways of cooperative and distributed software development;

- embrace (some of) the collaborators and competitors and establish relatively uncontested fields of work for all to reduce unproductive competition for funding and the reinvention of wheels;

- make sure that international organizations at the science-policy interface are aware of the Alliance members' strength and capabilities and have them endorsed accordingly;

- create agreement on who should be the provider of, and who should publish, aggregated, analytical data, indicators, and visualization; and

- agree on a set of infrastructures and services that cannot be allowed to fail.

We offer the following three specific recommendations to be implemented by GBIF in the medium term, informed by previous explorations and, as far as possible, by outputs of the work of the Alliance, but not waiting for its conclusion:

As the major technical and organizational activity with the highest priority, there needs to be a more effective and efficient, community-supported, quality assurance feedback from users to Publishers or even to the original contributors of data. Even the current publication process and its supporting standards may need to be adapted or changed. This is a major, long-term challenge, probably involving the need for cooperation and activities from the majority of the 1,500 Publishers and their support by Nodes, or even to create or join a global social network involving biodiversity data users and contributors. Therefore,

R15, Feedback and Quality Assurance: With increased funding, GBIF should speed up the work on feedback and quality assurance, and their implementation, so that major results are shown already in the medium term, that is, three to five years.

It would certainly be useful and will likely be necessary to work with partners in science and elsewhere to help provide critical solutions to applied global challenges in order to prove the value and quality of GBIF-mediated data. But GBIF should seek community consensus on

whether GBIF itself should provide those, particularly indicators or even reports, systematically, as some have suggested.

R16, Full Service Portal: It is important to follow up on expectations for a "full service" portal. We recommend GBIF not to wait for a consensus, but to *prototype and demonstrate* a concept for a "full service" portal, perhaps in some partnership, which would provide data and services such as:

- the customary access to occurrence data;
- quality controlled reference datasets;
- software codes implementing methods that are considered to be best practice (e.g., according to peer review); and
- an execution environment (a GBIF cloud).

R17, Long-Term Strategy: The medium-term explorations of new data types and technology, and, if achievable, consensus from the Alliance should enable GBIF to formulate strategic plans for the long term, underpinned by cost estimates, on:

- whether to host, or link to, data other than occurrences;
- whether and how to perform standard analytics and visualizations,
- and/or to offer a platform to execute analytic code provided by users; and
- how to host GBIF services in the future.

The preceding recommendations do not reflect the whole list of challenges identified in the SWOT analysis, just those we consider of the greatest importance and difficulty concerning the quality and quantity of data and services to be addressed by GBIF. Many of the challenges, especially those related to gaps in data and, again, their quality cannot solely be addressed technically, but will require progress elsewhere.

The Participation issue – and by implication, much of the gap issue – rests to a large extent on GBIF's success in Asia. There, a number of impediments to effective organization coalesce. Therefore, we recommend:

R18, Asian Participation: In order to establish a permanent foothold in Asia, GBIF should consider a branch office in an Asian country which is, like Denmark, small but wealthy enough to support hosting, politically (relatively) neutral, multilingual, and safe. This office would have a focus on nurturing Participation and creating and supporting a regional Nodes network (but it may contribute technical talent as well).

In many regions, Participation may strongly depend on demonstrating a compelling value proposition (see R4), on the support by local champions, and on a thriving regional network, which would be attractive to join. The former needs a powerful engagement team; the latter, in many regions, much more than the Capacity Enhancement Support Program could deliver today from core funding (see R6).

Secretariat staffing for the engagement with the GBIF network and other actors and stakeholders, particularly the scientific communities, is insufficient. There would definitely not be the capacity to perform additional activity types, e.g., such as advisable regular surveys, including their follow-up activities. Therefore, we recommend that:

R19, Staffing Increase: Staffing at the Secretariat in support of the various biodiversity communities' engagement should be doubled in the medium term.

But a few staff, even if doubled, at the Secretariat will never be enough to compensate if major weaknesses in the implementation of the distributed GBIF organization are not resolved. Conceptually, GBIF defers much of the community relations management and data quality issues to the Nodes, while in practice, these are not adequately staffed. Results from the short-term recommendations on empowering and funding Nodes would give indications whether the organizational concept succeeds, if it is actually sufficiently funded. We thus recommend:

R20, Organization and Funding of Nodes: If success in implementing fully functional Nodes in most, if not all, of the Participant countries, or developing their sustainable funding cannot be achieved in the medium term, a major adaptation of either the funding of Nodes or the organizational concept of GBIF should be considered by the Governing Board, supported by its newly established high-level advisory board.

Our conclusions – from Chapters 3 to 9, underpinned by many of the weaknesses pointed out by interviewees – have shown that even GBIF's current activities are underfunded. Certainly, there is further need for increased funds when taking into consideration the necessary growth to take advantage of the opportunities and to meet the major challenges. To realize the preceding recommendations of this review, the level of funding for GBIF and for technical developments performed or procured by the Secretariat for the operation and support to its Nodes, needs to be increased substantially. We therefore recommend that:

R21a, Funding Guidelines and Ability to Execute: GBIF should be prepared with guidelines (from the Governing Board, see R11) to fund new strategic initiatives and be organizationally able to absorb the workload or distribute it to reliable partners.

R21b, Increased Core Funding: The Governing Board should prepare to increase core funding, perhaps doubling it in the medium term (the next funding period of five years). For the longer term – the following five years – we confirm the recommendation of a previous review (CODATA 2005) to arrive at 12 million EUR per year, in 2019 Euros (which would constitute another doubling).

If – as we may expect – this level of funding cannot be achieved by the collection of reasonable and fair dues from national Participants, our last two recommendations are about seeking other funding sources. It takes into account that whatever the monetary effort is agreed upon by the Participants, the further development of a distributed global biodiversity data infrastructure and the mediation of access to data from so many contributors cannot be anything other than a continuous long-term endeavor. The work of GBIF cannot be executed successfully based on short-term planning and project funding (RFII 2016, p.2) at its core:

R21c, Core Funding Sources: The Governing Board should consider whether to invite third parties (non-Participants) to contribute to the funding of the core program. Such funding should involve predominantly longer-term commitments, i.e., for a minimum of 5 years.

R21d, Node Funding Sources: The long-term funding of many Nodes will very likely need to be supplemented by third parties. Both this and eventual third-party contributions to the core funding should be strictly monitored, however, to exclude undue influence on the "neutrality" of Nodes or on that of GBIF as a whole.

In summary, we see the organization as being able to rise to the challenges and today's expectations. To succeed, it must be enabled – financially, structurally, and with good leadership – to continue to perform the requisite exploratory work and implement the necessary steps in a timely manner.

We hope to have provided some recommendations that will enable GBIF to sustain its high standing in the biodiversity informatics landscape and to further continue to improve its support of biodiversity research and societal welfare. These recommendations should help to face the challenges of as yet unforeseeable demands from the science-policy interface, or applications in the health, agricultural, or commercial domains. Such developing expectations require the ability to adapt strategies and technologies flexibly and quickly, particularly as methods of systematic biodiversity data acquisition and scientific analysis may undergo some radical changes in the near future, and the pace of technological development in informatics and biotechnology accelerates.

10 ANNEXES

10.1. Review Team Members and Acknowledgements

Hans Pfeiffenberger

Dr. Hans Pfeiffenberger is an independent consultant on scientific data infrastructures and policies. Until recently, he led the IT Infrastructure Department at the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research in Germany. Hans has been chair of the Helmholtz Open Science working group, has advised the Knowledge Exchange and chaired the Science Europe policy working group on Access to Research Data. He is also founder and now member of the advisory board of the journal Earth System Science Data, an early data journal providing quality assurance to published data through peer review. He received his doctoral degree in Physics from the Technical University of Hannover - now Leibniz University Hannover - (1987). (see also: www.hans-pfeiffenberger.de)

Paul Uhlir

Paul Uhlir, JD is a consultant in information policy and management. He was Scholar at the National Academy of Sciences (NAS) in Washington, DC in 2015-2016, and founding Director of the Board on Research Data and Information at the NAS, 2008-2015. Paul was employed at the NAS from 1985-2015, first as a senior staff officer for the Space Studies Board, where he worked on solar system exploration and environmental remote sensing studies for NASA, and then as associate executive director of the Commission on Physical Sciences, Mathematics, and Applications. He directed the Office of International S&T Information for eight years after that, where he organized projects and meetings on scientific data throughout the world, and from 1992 to 2015 he was director of the US CODATA at the NAS. Before joining the NAS, he worked in the general counsel's office at the National Oceanic and Atmospheric Administration of the Department of Commerce in Washington, DC.

Paul has a B.A. degree in world history from the University of Oregon (1977), and a Master's degree in foreign relations and a Juris Doctor degree from the University of San Diego (1983, 1984). For more detailed information about his professional activities, see his website at: www.paulfuhlir.com

Aitong Li

Dr. LI has a postdoctoral appointment at the Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, and at the Graduate School of Public Policy, The University of Tokyo, Japan. She has degrees from Peking University, Yale University, and the University of Tokyo and speaks fluent Chinese, Japanese, and English. She focuses on environmental sustainability policy. (see also https://www.researchgate.net/profile/Aitong_Li)

Joseph Muliaro Wafula:

Dr. Wafula is Professor at the Jomo Kenyatta University of Agriculture and Technology (JKUAT) and Director of the ICT Center of Excellence and Open Data at JKUAT. He is also the chair of the Kenya CODATA Committee and a member of the CODATA

Executive Committee. He focuses on IT infrastructure and scientific data policy. (see http://ddsn.group.shef.ac.uk/members/joseph-muliaro-wafula/)

Liliana Ballesteros-Mejia

Dr. Ballesteros-Mejia is a Postdoctoral Fellow Researcher at the Institut de Systématique, Evolution, et Biodiversité in the Muséum National d'Histoire Naturelle, in Paris. She has a PhD in environmental sciences and her current work focuses on biodiversity and ecological research. (see https://www.researchgate.net/profile/Liliana_Ballesteros-Mejia)

William Michener

Dr. Michener is professor and director of eScience Initiatives at the University of New Mexico in Albuquerque. He has a PhD in biological oceanography, and degrees in entomology and zoology. His current research interests include environmental data, information and knowledge management; and cyberinfrastructure design and specification. (see https://ulls.unm.edu/people/directory/curriculum-vitae/michener-bill.pdf)

Simon Hodson

Dr. Hodson is Executive Director of CODATA in Paris, France, and an expert in research data management and policy. He is responsible for the primary editorial oversight and for project management and logistics for the GBIF review project. (see also http://www.codata.org/about-codata/secretariat)

We acknowledge the valuable contributions by many experts from all GBIF regions. In chronological order, we received information, opinion and advice from: Members of GBIF Secretariat staff and Directors, 108 interviewees, 10 external reviewers, members of the GBIF Governing Board, and members of the CODATA Executive Committee.

Without their input, this report could not have been compiled, and mistakes or ambiguities would have remained. However, any remaining error or bias is entirely the lead authors' fault.

10.2. Identifying GBIF policy impact through citation analysis

Chapter 3, "Impacts of 1.5°C Global Warming on Natural and Human Systems" of the 2018 IPCC Special Report "Global Warming of 1.5°C" (IPCC2018) relies heavily on one "Science" article (referenced in Chapter 3 on p.218 and p.309) (Warren et al. 2018) – R.Warren is a lead author both of the Science article and the IPCC report.

Unfortunately, GBIF is not referenced or even mentioned either in the IPCC report or in the Science *article* itself, but *only in the supplement* to the Science article, where GBIF is mentioned 5 times on 5 pages of text, including in the very first paragraph:

> *"Materials and Methods*
>
> *Data, Materials and Methods are as described in (4) with differences noted below. This paragraph provides a brief summary of the method. Current species distribution data were obtained from the Global Biodiversity Information Facility (GBIF; (34)).*"

The reference 34 here, ostensibly on or about GBIF, points to an "almost" third-party article from *2007* (Yesson et al. 2007) "How Global Is the Global Biodiversity Information Facility?", which discusses data quality and gaps *at that time* – one wonders what that analysis would show today, if repeated! Besides that, this analysis is performed only on legumes and seems to assume another database, ILDIS, as faultless:

> *"We have explored the global point data provided by GBIF using the International Legume Database & Information Service (ILDIS) to validate point data, both taxonomically and spatially.*
>
> *[…]*
>
> *ILDIS is a global species database providing expert taxonomic and area occurrence data for the twenty thousand species of Leguminosae [15], one of the largest families of flowering plants, often considered as representative of global plant biodiversity [16].*"

The *actual* methods used for the Science article are described in reference 4 of the supplement to the Science article, which is a letter to Nature Climate Change (Warren et al. 2013):

> *"Biodiversity records were sourced from the Global Biodiversity Information Facility[29] (GBIF) and vetted for locational reliability (see Supplementary Information). We used MaxEnt[27,28] to create statistical relationships between the vetted species occurrence records and present (1961–1990) climate, and to calculate the present geographic distribution of each species[27,30]. To eliminate potential omission and commission biases, distributions were then clipped to the bio-geographic zone(s)[31] from which the species information was derived and to a conservative 2000 km buffer around the species' outermost occurrence records.*"

Note that here, the reference 29 for GBIF does not refer to an aging article, but to data.gbif.org.

We had to stop our analysis at this point and await that one by GBIF, now published (GBIF2019).

10.3. GBIF Strategic Plan 2017-2021

GBIF Strategic Plan 2017-2021 (copied from https://www.gbif.org/strategic-plan)

Priorities

1. Empower global network

Ensure that governments, researchers and users are equipped and supported to share, improve and use data through the GBIF network, regardless of geography, language or institutional affiliation.

Remove barriers to participation

Increase benefits associated with publishing biodiversity data

Address capacity needs

2. Enhance biodiversity information infrastructure

Provide leadership, expertise and tools to support the integration of all biodiversity information as an interconnected digital knowledgebase.

Coordinate vision and strengthen partnerships with major biodiversity informatics initiatives

Promote standardization and common mechanisms for exchange of biodiversity data

Provide stable and persistent data infrastructure to support research

3. Fill data gaps

Prioritize and promote mobilization of new data resources which combine with existing resources to maximize the coverage, completeness and resolution of GBIF data, particularly with respect to taxonomy, geography and time.

Expand checklists to cover all taxonomic groups

Identify and prioritize gaps in spatial and temporal data

Engage institutions and researchers with complementary data

4. Improve data quality

Ensure that all data within the GBIF network are of the highest-possible quality and associated with clear indicators enabling users to assess their origin, relevance and usefulness for any application.

Enhance automated data validation

Implement tools for expert curation

Provide clear quality indicators for all data

5. Deliver relevant data

Ensure that GBIF delivers data in the form and completeness required to meet the highest-priority needs of science and, through science, society.

Engage with expert communities to manage data to the highest quality possible

Deliver well-organized and validated data to support key applications

Critical challenges addressed by this plan

Widening and strengthening participation

Many countries in all regions remain unconnected with GBIF, and progress toward establishing and strengthening national capacity to mobilize and access relevant data varies significantly among existing participants. Funding remains a challenge, with costs shared among a relatively small number of countries and disproportionate impact when one or more of these is unable to contribute at agreed levels. More conspicuously, data coverage is highly variable between regions and even between adjacent countries.

Building trust in data products

Improvements are required in the quality and fitness-for-use of aggregated data and metadata within the network. More work is needed to ensure that all data are sufficiently documented and catalogued in ways that assist users in filtering according to their needs, and greater precision and accuracy is necessary particularly in representing taxonomy. Mechanisms and incentives are needed to engage expert communities in validation and correction of this global data resource.

Filling data gaps and broadening evidence base

Opportunities exist to make significant advances both in the completeness and coverage of GBIF data and of the richness of available information. GBIF must identify and understand where gaps and biases in existing data make it inadequate to meet user needs and must prioritize effective responses to address these issues.

All relevant sources of data must be incorporated, including sample-based data sets, ecogenomics and other molecular research, remote-sensing, literature records, local and regional checklists, and expert knowledge. These resources should be used to establish GBIF not only as a source of occurrence information but also as an effective tool to discover and access data on species abundance and community composition, and related genetic data.

Scaling up infrastructure

Integrating growing volumes of data will bring new challenges in efficient storage, management, presentation and access of these data. GBIF will face related challenges as it engages with more countries and organizations and as its services become more mission-critical for many stakeholders. During 2017-2021, GBIF must accordingly continue to innovate and to review all processes to ensure smooth future growth.

10.4. Interview Guidelines

Regional consultants were supplied with a mail template to invite potential interviewees and an interview script to use during the actual interviews:

Mail Template

Dear [personalize]:

As you may know, the Global Biodiversity Information Facility (GBIF) recently contracted with CODATA to perform a "20-Year Review" of the GBIF organization. I am one of the consultants for this review and there are other consultants for each of the major regions of the world. I have copied Simon Hodson, the Executive Director of CODATA, on this message to keep him informed of the process.

We are collecting information for the review now and conducting interviews with a sampling of stakeholders in the organization. It is not possible to do that with all the GBIF stakeholders, but you have been selected as one of the key contacts in the region for such an interview.

I will use a short questionnaire to guide our conversation. Other representatives of GBIF who were not selected for interviews will receive a separate written questionnaire to make sure we offer an opportunity to all members to provide comments.

Please respond to me at your earliest convenience if you are willing to participate in a one-hour (or shorter) interview. If you agree, I can provide a doodle poll with a number of possible days and times; based on your availability, I would then provide conference call information for a mutually suitable time. We would like to complete all the interviews by the end of April.

If you have any questions about the review, please do not hesitate to contact me. I look forward to your response!

Interview Script

We suggest to use the interview template as follows: While in the interview, you might type (write) notes in the document (printout) just below the questions, while you talk. After the interview, you would copy the essence of each answer into the SWOT table.

<salutation>

To be able to capture your valuable input most precisely, I would like to record our conversation. I would use it for only one purpose, writing up our conversation and not share it with anyone else. Do you agree to recording?

If we would like, later on, to attribute anything you said to you by name in our report, we would come back to you and ask permission.

We chose to interview you because of your role as A/B/C/D of GBIF (← true?)

When answering the following question, please consider them within a time range of up to 5 or 10 years into the future:

Who should be asked	Question / Notes
All groups	1. **(S) How has GBIF met your expectations, your needs most successfully? What sets it apart, where is it indispensable?**
	Could you relate a most telling story about that?
	2. **(W) Where / in which way did GBIF less than you expected?**
	Who or what did supply what you needed, in that case or cases
	3. **(O) What could be a big win for GBIF and its community, something no one else does or could do (a system feature, a dataset or …)**
	4. **(T) Do you believe there are any threats to the existence of the organization or its mission?**
Funders, Government agencies, GBIF Governance and Node Managers only	5. **Do you consider the funding of the GBIF Secretariat and your national Node (and other national biodiversity initiatives) sufficient and stable to fulfill expectations and needs?**
Other bioinformatics initiatives	6. **How do you see your relationship to GBIF, how do you see it evolve?**
All groups	7. Are there any other issues you wish to raise? •

RESULTS

Consultant, date	You, 2019-mm-dd
Interview partner	Name / email
Role	A/B/C/D
Country	

Please fill in answers as SWOT as you see fit; this relates to the relationship to competitors / collaborating initiatives in the "landscape" as well.

STRENGTHS	**WEAKNESSES**
• [brief, bulleted list of affordances that currently exist and affect you personally or relate to your individual work]	• [brief, bulleted list of limitations that currently exist and affect you personally or relate to your individual work]
OPPORTUNITIES	**THREATS**
• [brief, bulleted list of prospective opportunities that may come from external sources or have the potential to impact your wider community]	• [brief, bulleted list of potential challenges that may arise from external sources or have the potential to impact your wider community]

Please add any stories and non-/attributable quotes here.

10.5. References

Beagrie 2014, "The Value and Impact of Data Sharing and Curation", Neil Beagrie, John Houghton, 2014, http://repository.jisc.ac.uk/id/eprint/5568

Berendsohn 2010, "Summary of Recommendations of the GBIF Task Group on the Global Strategy and Action Plan for the Digitisation of Natural History Collections", Walter G. Berendsohn, Vishwas Chavan, James Macklin, Biodiversity Informatics: Vol 7 No 2 (2010), https://dx.doi.org/10.17161/bi.v7i2.3989

Borgman 2015, "Big Data, Little Data, No Data: scholarship in the networked world", Borgman, Christine L., (2015) Cambridge, Massachusetts: The MIT Press, ISBN: 978-0-262-52991-4, also available at: https://mitpress.mit.edu/books/big-data-little-data-no-data

CODATA 2005, "Global Biodiversity The GBIF 3rd-Year Review", Report from the Review Committee, 210 pp. Feb 2005, http://www.codata.info/archives/2005/GBIF2005Rept.pdf

CBD 2016, "Key scientific and technical needs related to the implementation of the Strategic Plan for Biodiversity 2011-2020 and related research" (CBD/COP/DEC/XIII/31) https://www.cbd.int/doc/decisions/cop-13/cop-13-dec-31-en.pdf

Ferrier 2012, "In situ / remote sensing integration through modelling", WG7 plenary presentation 3 Dec 2012, Asilomar http://www.earthobservations.org/documents/meetings/201212_geobon_allhands/7%20WG7%20plenary%20presentation%203Dec2012%20Simon%20Ferrier.pdf

GBIF 2010a, "2010 Forward Look Report", Peterson, A.T., Canhos, D., Gärdenfors, U., Scholes, R.J. & Shirayama, Y. (2010) GBIF Forward Look Team. 45 pp. Global Biodiversity Information Facility, Copenhagen, https://www.gbif.org/document/80683/gbif-2010-forward-look-report

GBIF 2010b, "GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network", version 1.0. authored by Hill, A. W., Otegui, J., Ariño, A. H., and R. P. Guralnick. 2010. Copenhagen: Global Biodiversity Information Facility, 25 pp. ISBN: 87-92020-11-9. Accessible on-line at http://www.gbif.org/

GBIF 2016a "Final Report of the Task Group on GBIF Data Fitness for Use in Agrobiodiversity" Elizabeth Arnaud, Nora Patricia Castañeda-Álvarez, Jean Ganglo Cossi, Dag Endresen, Ebrahim Jahanshiri, Yves Vigouroux, 2016, https://www.gbif.org/document/82283/report-of-the-task-group-on-gbif-data-fitness-for-use-in-agrobiodiversity

GBIF 2018, "GBIF Implementation Plan 2017-2021 and Annual Work Programme 2019", https://www.gbif.org/document/4Pcu4jX68UE2iMOMQgQEE8/gbif-implementation-plan-2017-2021-and-annual-work-programme-2019, last accessed 2019-08-20

GBIF 2019, "GBIF Science Review 2019", GBIF Secretariat. (2019), https://dx.doi.org/10.15468/QXXG-7K93

GBIF 2019b, "Establishing an Effective GBIF Participant Node: Concepts and general considerations". GBIF Secretariat 2019, Copenhagen, https://dx.doi.org/10.15468/doc-z79c-sa53

Gluckman 2016, „The science–policy interface", Peter Gluckman, Science 2016, Vol. 353, Issue 6303, pp. 969, https://dx.doi.org/10.1126/science.aai8837

Hallmann 2017, "More than 75 percent decline over 27 years in total flying insect biomass in protected areas", Hallmann CA, Sorg M, Jongejans E, Siepel H, Hofland N, Schwan H, et al. (2017) PLoS ONE 12 (10): e0185809. https://dx.doi.org/10.1371/journal. pone.0185809

Hobern 2012, "Global Biodiversity Informatics Outlook: Delivering biodiversity knowledge in the information age", Hobern D, Apostolico A, Arnaud E, Bello JC, Canhos D, Dubois G, Field D, Alonso Garcia E, Hardisty A, Harrison J, Heidorn B, Krishtalka L, Mata E, Page RDM, Parr C, Price J & Willoughby S, (2012) Copenhagen: Global Biodiversity Information Facility, https://dx.doi.org/10.15468/6jxa-yb44

Hobern 2019, "Connecting data and expertise: a new alliance for biodiversity knowledge", Hobern D, Baptiste B, Copas K, Guralnick R, Hahn A, van Huis E, Kim E, McGeoch M, Naicker I, Navarro L, Noesgaard D, Price M, Rodrigues A, Schigel D, Sheffield C, Wieczorek J (2019). Biodiversity Data Journal 7: e33679. https://dx.doi.org/10.3897/BDJ.7.e33679

Hortal 2015, "Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity", Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Annual Review of Ecology, Evolution, and Systematics, 46(1), 523–549. https://dx.doi.org/10.1146/annurev-ecolsys-112414-054400

IPBES 2016, "Summary for Policymakers of the Assessment Report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) on Pollinators, Pollination and Food Production", 2016, Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES), ISBN: 978-92-807-3568-0

IPBES 2019, "Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services". S. Díaz, J. Settele, E. S. Brondizio E.S., H. T. Ngo, M. Guèze, J. Agard, A. Arneth, P. Balvanera, K. A. Brauman, S. H. M. Butchart, K. M. A. Chan, L. A. Garibaldi, K. Ichii, J. Liu, S. M. Subramanian, G. F. Midgley, P. Miloslavich, Z. Molnár, D. Obura, A. Pfaff, S. Polasky, A. Purvis, J. Razzaque, B. Reyers, R. Roy Chowdhury, Y. J. Shin, I. J. Visseren-Hamakers, K. J. Willis, and C. N. Zayas (eds.). IPBES secretariat, Bonn, Germany; https://www.ipbes.net/global-assessment-report-biodiversity-ecosystem-services, last accessed 2019-08-01

IPCC 2018, "Impacts of 1.5°C Global Warming on Natural and Human Systems. In: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty", Hoegh-Guldberg, O., D. Jacob, M. Taylor, M. Bindi, S. Brown, I. Camilloni, A. Diedhiou, R. Djalante, K.L. Ebi, F. Engelbrecht, J. Guiot, Y. Hijioka, S. Mehrotra, A. Payne, S.I. Seneviratne, A. Thomas, R. Warren, and G. Zhou, 2018 [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. In Press. (accessed 2019-04-23 through https://www.ipcc.ch/site/assets/uploads/sites/2/2019/02/SR15_Chapter3_Low_Res.pdf)

IWS 2019 "Top Ten Languages Used in the Web", https://www.internetworldstats.com/stats7.htm and "World Internet Usage and Population Statistics", https://www.internetworldstats.com/stats.htm, Internet World Stats, accessed 2019-07-27

Kolbert 2014, "The Sixth Extinction: An Unnatural History", Kolbert, Elizabeth. 2014. New York, NY: Henry Holt and Co.

Longbottom 2018, “Vulnerability to snakebite envenoming: a global mapping of hotspots”, Joshua Longbottom, Freya M Shearer, Maria Devine, Gabriel Alcoba, Francois Chappuis, Daniel J Weiss, Sarah E Ray, Nicolas Ray, David A Warrell, Rafael Ruiz de Castañeda, David J Williams, Simon I Hay, David M Pigott, Lancet 2018; 392: 673–84, http://dx.doi.org/10.1016/S0140-6736(18)31224-8

Maldonado 2015, “Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases?”, Carla Maldonado, Carlos I. Molina, Alexander Zizka, Claes Persson, Charlotte M. Taylor, Joaquina Albán, Eder Chilquillo, Nina Rønsted, Alexandre Antonelli, Global Ecology and Biogeography (2015) 24, 973–984, https://dx.doi.org/10.1111/geb.12326

Masood 2018, “The battle for the soul of biodiversity”, Ehsan Masood, Nature 560, 423-425 (2018), https://dx.doi.org/10.1038/d41586-018-05984-3

McKinley 2017, “Citizen science can improve conservation science, natural resource management, and environmental protection”, Duncan C. McKinley, Abe J.Miller-Rushing, Heidi L. Ballard, Rick Bonney, Hutch Brown, Susan C. Cook-Patton, Daniel M. Evans, Rebecca A. French, Julia K. Parrish, Tina B. Phillips, Sean F. Ryan, Lea A. Shanley, Jennifer L. Shirk, Kristine F. Stepenuck, Jake F. Weltzin, Andrea Wiggins, Owen D. Boyle, Russell D. Briggs, Stuart F. Chapin III, David A. Hewitt, Peter W. Preuss, Michael A. Soukup, Biological Conservation (208), 2017, 15-28, https://dx.doi.org/10.1016/j.biocon.2016.05.015 (Elsevier - full text inaccessible)

OECD 1999, “OECD Megascience Forum. Final report of the working group on biological informatics.”, Organisation for Economic Co-operation and Development (1999), Paris. https://dx.doi.org/10.5281/zenodo.3383187

Pereira 2013, „Essential Biodiversity Variables“, H. M. Pereira, S. Ferrier, M. Walters, G. N. Geller, R. H. G. Jongman, R. J. Scholes3, M. W. Bruford, N. Brummitt, S. H. M. Butchart, A. C. Cardoso, N. C. Coops, E. Dulloo, D. P. Faith, J. Freyhof, R. D. Gregory, C. Heip, R. Höft, G. Hurtt, W. Jetz, D. S. Karp, M. A. McGeoch, D. Obura, Y. Onoda, N. Pettorelli, B. Reyers, R. Sayre, J. P. W. Scharlemann, S. N. Stuart, E. Turak, M. Walpole, M. Wegmann, Science 2013, Vol. 339, (6117), pp. 277-278, https://dx.doi.org/10.1126/science.1229931

Rauber 2015, “Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC)”, Rauber, Andreas, Asmi, Ari, van Uytvanck, Dieter, & Proell, Stefan. (2015). http://dx.doi.org/10.15497/RDA00016

RFII 2016, “Enhancing Research Data Management: Performance through Diversity. Recommendations regarding structures, processes, and financing for research data management in Germany”, German Council for Scientific Information Infrastructures, Göttingen 2016, http://www.rfii.de/?p=2075

Robertson 2019, “Training Machines to Identify Species using GBIF-mediated Datasets”, Robertson T, Belongie S, Adam H, Kaeser-Chen C, Zhang C, Chuan Tan K, Liu Y, Brulé D, Deltheil C, Loarie S, Van Horn G, Mac Aodha O, Beery S, Perona P, Copas K, Waller J, Biodiversity Information Science and Standards (2019) 3: e37230. https://dx.doi.org/10.3897/biss.3.37230

Schmid 2016, “The LOKI underwater imaging system and an automatic identification model for the detection of zooplankton taxa in the Arctic Ocean”, Moritz Sebastian Schmid, Cyril Aubry, Jordan Grigor, Louis Fortier, Methods in Oceanography (2016), https://dx.doi.org/10.1016/j.mio.2016.03.003

Speed 2018, “Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data”, James D. M. Speed, Mika Bendiksby, Anders G. Finstad, Kristian Hassel, Anders L. Kolstad, Tommy Prestø, PLoS ONE 13(4) (2018). https://dx.doi.org/10.1371/journal.pone.0196417

Warren 2013, “Quantifying the benefit of early climate change mitigation in avoiding biodiversity loss” R. Warren, J. VanDerWal, J. Price, J. A. Welbergen, I. Atkinson, J. Ramirez-Villegas, T. J. Osborn, A. Jarvis, L. P. Shoo, S. E. Williams & J. Lowe, Nature Climate Change (2013) https://dx.doi.org/10.1038/nclimate1887

Warren 2018, “The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5°C rather than 2°C”, R. Warren, J. Price, E. Graham, N. Forstenhaeusler, J. VanDerWal, Science (2018), https://dx.doi.org/10.1126/science.aar3646

Wheeler 2012, “Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity”, Q. D. Wheeler , S. Knapp , D. W. Stevenson , J. Stevenson , S. D. Blum , B. M. Boom , G. G. Borisy , J. L. Buizer , M. R. De Carvalho , A. Cibrian , M. J. Donoghue , V. Doyle , E. M. Gerson , C. H. Graham , P. Graves , S. J. Graves , R. P. Guralnick , A. L. Hamilton , J. Hanken , W. Law , D. L. Lipscomb , T. E. Lovejoy , H. Miller , J. S. Miller , S. Naeem , M. J. Novacek , L. M. Page , N. I. Platnick , H. Porter-Morgan , P. H. Raven , M. A. Solis , A. G. Valdecasas , S. Van Der Leeuw , A. Vasco , N. Vermeulen , J. Vogel , R. L. Walls , E. O. Wilson & J. B. Woolley, Systematics and Biodiversity (10) 2012, https://dx.doi.org/10.1080/14772000.2012.665095

Wilkinson 2016, “The FAIR Guiding Principles for scientific data management and stewardship”, Wilkinson, M., Dumontier, M., Aalbersberg, I. et al., Scientific Data (3), 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Serra-Diaz 2018, “Big data of tree species distributions: how big and how good?” , Josep M. Serra-Diaz, Brian J. Enquist, Brian Maitner, Cory Merow and Jens-C. Svenning, Forest Ecosystems 2018, https://dx.doi.org/10.1186/s40663-017-0120-0

Yesson 2007 “How Global Is the Global Biodiversity Information Facility?”, Chris Yesson, Peter W. Brewer, Tim Sutton, Neil Caithness, Jaspreet S. Pahwa, Mikhaila Burgess, W. Alec Gray, Richard J. White, Andrew C. Jones, Frank A. Bisby, Alastair Culham, November 7, 2007 https://dx.doi.org/10.1371/journal.pone.0001124

10.6. Glossary

GBIF-specific acronyms and terms:

BID	Biodiversity Information for Development, an EU funded program, managed by GBIF, to "support capacity enhancement activities and projects to mobilize biodiversity data and strengthen national or regional biodiversity information facilities in these regions" (sub-Saharan Africa, the Caribbean and the Pacific)
CESP	Capacity Enhancement Support Programme, a program of GBIF
Data Publishers	In the GBIF context, a data publisher is "A custodian of data making it technically available. This may or may not be the data owner. If not they will have declared to GBIF that they have permission to make the data available." https://www.gbif.org/terms/data-publisher
IPT	Integrated Publishing Toolkit, https://www.gbif.org/ipt

General acronyms, names and terms

ALA	Atlas of Living Australia, https://www.ala.org.au
BoL	Barcode of Life, see International Barcode of Life Consortium, https://ibol.org and Barcode of Life Data System, http://www.boldsystems.org/
BHL	Biodiversity Heritage Library, "a consortium of major natural history, botanical, and research libraries that cooperate to digitize and make accessible the literature of biodiversity held in their collections"
CBD	Convention on Biological Diversity (1993), https://www.cbd.int/intro/ , a ➔ MEA, "has 3 main objectives: The conservation of biological diversity. The sustainable use of the components of biological diversity. The fair and equitable sharing of the benefits arising out of the utilization of genetic resources."
CETAF	Consortium of European Taxonomic Facilities, 33 members, "aims to promote training, research and understanding in systematic biology and palaeobiology, and facilitate access to information (collections) and the expertise of its member institutions" https://cetaf.org/about-us/what-cetaf
CoL	Catalogue of Life, https://www.catalogueoflife.org
CMS	Convention on the Conservation of Migratory Species of Wild Animals, "an environmental treaty under the aegis of the United Nations Environment Programme, CMS provides a global platform for the conservation and sustainable use of migratory animals and their habitats." https://www.cms.int (more acronyms here: https://www.cms.int/en/about/partnerships)
EBV	Essential Biodiversity Variables, "whose development by GEO BON has been endorsed by the CBD" "aim to help observation communities harmonize monitoring, by identifying how variables should be sampled and measured." Science (2013) https://dx.doi.org/10.1126/science.1229931

EoL	Encyclopedia of Life, https://eol.org
FAO	Food and Agriculture Organization (1945) "is a specialized agency of the United Nations that leads international efforts to defeat hunger" http://www.fao.org/
GBIC2	2nd Global Biodiversity Information Conference, organized by GBIF, https://www.biodiversityinformatics.org
GEO	The Group on Earth Observations, "GEO is a partnership of more than 100 national governments and in excess of 100 Participating Organizations that envisions a future where decisions and actions for the benefit of humankind are informed by coordinated, comprehensive and sustained Earth observations." https://www.earthobservations.org/geo_community.php "GEO community is creating a Global Earth Observation System of Systems (GEOSS) to better integrate observing systems and share data by connecting existing infrastructures using common standards." https://www.earthobservations.org
GEO BON	The Group on Earth Observations Biodiversity Observation Network "Mission: Improve the acquisition, coordination and delivery of biodiversity observations and related services to users including decision makers and the scientific community." https://geobon.org/about/vision-goals/
GGBN	"A global network of well-managed collections of genomic samples from across the Tree of Life", established about 2013/2014 https://wiki.ggbn.org/ggbn/About_GGBN
ICSU WDS	World Data System is an Interdisciplinary Body of the International Science Council (ISC; formerly ICSU) (2008; 1958) "coordinating and supporting trusted scientific data services for the provision, use, and preservation of relevant datasets" https://www.icsu-wds.org/organization
ICZN	International Commission on Zoological Nomenclature, (1895), "provides and regulates a uniform system of zoological nomenclature ensuring that every animal has a unique and universally accepted scientific name.", https://www.iczn.org
IPBES	Intergovernmental Platform on Biodiversity and Ecosystem Services (2012), https://www.ipbes.net , a ➔ MEA, "assesses the state of biodiversity and of the ecosystem services it provides to society, in response to requests from decision makers." "a younger sibling to the Nobel-prizewinning Intergovernmental Panel on Climate Change (IPCC)" "… brought IPBES to life in 2012. The new panel, which has cost US$31 million so far …" (Nature 2018, https://www.nature.com/articles/d41586-018-05984-3)
IUCN	International Union for Conservation of Nature (1948) "is a membership Union uniquely composed of both government and civil society organisations. It provides public, private and non-governmental organisations with the knowledge and tools that enable human progress, economic development and nature conservation to take place together.", https://www.iucn.org
OBIS	Ocean Biogeographic Information System (2002) "is a global open-access data and information clearing-house on marine biodiversity for science,

	conservation and sustainable development" "More than 20 OBIS nodes around the world connect 500 institutions from 56 countries. Collectively, they have provided over 45 million observations of nearly 120 000 marine species" https://obis.org "developed as the information management component of the ten year Census of Marine Life" (a 650 M$ scientific initiative) https://en.wikipedia.org/wiki/Ocean_Biogeographic_Information_System
MEA	Multilateral Environmental Agreements (here: related to biodiversity) https://www.informea.org/en/topics/biological-diversity
TDWG	"Historically known as the Taxonomic Databases Working Group, today's Biodiversity Information Standards (TDWG) is a not-for-profit, scientific and educational association formed to establish international collaboration among the creators, managers and users of biodiversity information" (1985) https://www.tdwg.org
UNEP	UN Environment Programme https://www.unenvironment.org/
UNEP WCMC	The UNEP World Conservation Monitoring Centre "works with scientists and policy makers worldwide to place biodiversity at the heart of environment and development decision-making" "UNEP-WCMC is a collaboration between UN Environment and the UK charity, WCMC", staff about 100 , https://www.unep-wcmc.org/about-us
UNEP WCMC /BIP	"The Biodiversity Indicators Partnership (BIP) is a global initiative that has operated since 2007. Mandated by the Convention on Biological Diversity (CBD), it promotes and coordinates the development of indicators of biodiversity change … UNEP-WCMC is the official Secretariat of the BIP." (Wild Bird Index Guidance – no mention of GBIF) https://www.unep-wcmc.org/resources-and-data/biodiversity-indicators-partnership-global
WDPA	The World Database on Protected Areas (WDPA) is the only global database of protected areas, underpinning Protected Planet. Protected Planet® is a joint product of UN Environment and IUCN, managed by UNEP-WCMC and the IUCN working with governments, communities and collaborating partners. The WDPA can be viewed and downloaded at https://www.protectedplanet.net/, where it is integrated with other relevant information.
WFCC	The World Federation for Culture Collections is an international body formed under the umbrella of the International Union of Biological Sciences and a Federation within the International Union of Microbiological Societies. The WFCC operates as a clearing house for information on collections of microbiological specimens. It supports the development, maintenance and establishment of culture collections. The WFCC bylaws were published in 1972 in the International Journal of Systematic Biology (Int. J. Syst. Bacteriol., 22: 406-409, 1972) and updated several times since. https://en.wikipedia.org/wiki/World_Federation_for_Culture_Collections

COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL

CODATA

The Committee on Data of the
International Science Council

Connect with us at:

www.codata.org
simon@codata.org

International Science Council
5 rue Auguste Vacquerie
75016 Paris, France

www.twitter.com/CODATANews

www.facebook.com/codata.org

Cover image: Species occurrences,
www.gbif.org