GLOBAL BIODATA COALITION

Global Biodata Coalition

Highlights 2022

This publication was produced by :
The Global Biodata Coalition

Cover illustration and brochure design :
Spencer Phillips

For more information about the Global Biodata Coalition please contact :
info@globalbiodata.org

Image credit :
Adobe Stock (unless indicated otherwise)

Printer :
NewPrint

Table of Contents

Who we are

The context & the challenge

The most heavily used biodata resources are accessed by millions of users each month. Despite the crucial importance of biodata resources to the life sciences, many lack stable funding and are at risk. The current global biomedical and life science data funding landscape is fragmented, with aggregate current funding barely supporting existing levels of service and proving to be inadequate for coping with the growing volumes and complexity of data, for example those data generated through new sequencing modalities and imaging technologies. As the researchers' reliance on these data resources grows, the lack of international planning and coordination for their sustainability threatens the availability of these resources to users.

The GBC is working to change the future

The Global Biodata Coalition is the first dedicated group with a remit to coordinate strategic approaches towards establishing and monitoring the health and sustainability of an international set of biomedical and life sciences data resources. By sharing responsibility, effort, experience, and knowledge with the worldwide set of funders, the GBC offers all funders of life science research an opportunity to participate in this vital coordination effort, while enhancing their own activities and increasing the overall impact of their investment.

This coalition of research funders is working, together with resource managers and researchers, to understand the global biodata resource infrastructure and to move towards more internationally coordinated, sustainable, and streamlined mechanisms that support the biodata ecosystem.

What we stand for

Our Aims

- To be a forum for funders of biodata resources to better coordinate and share approaches for efficient management and growth of this infrastructure and share strategies.
- To stabilise and ensure sustainable financial support for the global biodata infrastructure, with a focus on an identified and prioritised set of Global Core Biodata Resources that are crucial for sustaining the broader ecosystem.

Our Mission

To safeguard the world's open life science, biological and biomedical reference data.

Our Vision

The GBC envisions a world in which biodata, essential to science, are freely and openly available to all, with long-term access assured through the tools and services offered by global biodata resources.

We look ahead to global inclusivity, whereby all countries have the opportunity to benefit from and contribute to secure and sustained biodata resources enabled through the alignment and committed support of a breadth and diversity of the world's life science funders.

Our Values

The inherent values of open science and open data underpin the operational plan of the Global Biodata Coalition as it works to ensure free and equitable access to life science, biological and biomedical data for all citizens of the world.

Photo credit : Jeff Dowling

Foreword

The Global Biodata Coalition is a comparatively new organisation with an important mission: to ensure the sustainability of the world's biodata resources. We have all seen for ourselves, during the COVID-19 pandemic, the importance of collaboration and data sharing at the level of governments and scientists in academia and industry. This collaboration, coupled with the preparedness of key elements of the global biodata infrastructure to receive new and emerging data, proved transformative. The task that we face, to ensure that the biodata ecosystem can sustain and scale, is vital for society.

Like all new organisations, we have a lot of work to do but we are supported in our endeavours by our many stakeholders, including our Member funding organisations and Observers, those who give their time generously as Board Members, scientific advisers and expert reviewers and the leaders of biodata resources, among others, working together with our Secretariat to find solutions to ensure that the biodata ecosystem can develop and be safely maintained.

It is a pleasure to present the achievements of the GBC to date, to mark our progress and to look ahead at our future scientific plans.

I look forward to the challenges ahead and to tackling them together with the GBC community across the world.

Dr Guy Cochrane
Executive Director

The importance of biodata sustainability

Users and generators of biodata, often employed in academic research or in industry, rely on easy and immediate access to biodata resources - open life science databases and services - either to consume data or to share data derived from their research. However, in reality, the global biodata resource landscape is fragmented and is largely funded through time-limited grants, often at a national level. Set against a backdrop of rapidly increasing biodata volumes, driven by, for example, high throughput single cell platforms, coupled with demands placed on the system by advances in AI and machine learning, the challenges of maintaining existing databases and developing new ones, is ever more intense.

Sustainability of the biodata resources and the infrastructure as a whole is key to the life sciences research endeavour and its benefit to society. Without access to this infrastructure, life science research activities would slow or stop altogether and the onward impacts of scientific progress, including for example drug discovery and tackling biodiversity loss, would be severely compromised.

Perspectives on the long term sustainability of biodata are broad. Individual users and research funders agree it is critical but the former are often more concerned with the sustainability of a given resource, whilst the latter is preoccupied with the overall health of the global biodata infrastructure. Both agree that funding is vital for the continuity and consistency of services, as is the ability to attract and retain appropriately skilled staff to manage resources.

Research funders and the managers of biodata resources agree that sustainability of biodata is key. At the heart of this must lie sufficient and sustained financing, structured to support biodata resources through development, operational and review cycles, to allow the resources to attract and retain appropriately skilled staff and operate strategically planned procurement of technical assets.

However, sustainability goes beyond funding alone. For biodata resource managers, sustainability includes, for example, healthy programmes to engage the user community and respond efficiently to feedback relating to the services, practising good data governance, and providing professional and reliable content through high-quality interfaces. The concerns of institutions who operate portfolios of biodata resources, and research funding bodies who seek the best infrastructure to support their research programmes, are also focused towards the responsiveness of the infrastructure as a whole and its adaptability to the changing needs of science, as new data types emerge and legacy technologies become less prominent. Also of importance to these stakeholders are sufficient redundancy across resources to secure biodata content, strong connectivity across biodata resources and with other data infrastructure beyond the life sciences, such as in the areas of health or the environment, and the neutrality of the infrastructure from policy and political change.

2022 in numbers

11
Member funder organisations

11
Countries represented through GBC membership

3112
Global data resources identified through the GBC inventory

7
Observer organisations

37
Global Core Biodata Resources selected in the first round

57
Reviewers contributed to the Global Core Biodata Resources selection review process

Highlights of the year

March 2022

First Global Core Biodata Resource selection round opens

April 2022

GBC appoints Dr Guy Cochrane as its first Executive Director

June 2022

First in-person meeting of the Scientific Advisory Committee, held in Strasbourg

December 2022

Initial group of 37 Global Core Biodata Resources announced

An open inventory of global biodata resources

Understanding the global biodata landscape

The GBC has a mission to develop mechanisms to more efficiently fund the biodata infrastructure, ensuring sustainability of the ecosystem, but what is this infrastructure: how many biodata resources are there, where are they, and who funds them? To begin to tackle this challenge, the Secretariat worked with Heidi Imker, from the University of Illinois, Urbana-Champaign, who led the project, and Ken Schackart, from the University of Arizona, Tucson, in collaboration with the Chan Zuckerberg Initiative, to achieve a new understanding of the overall global biodata landscape. They developed a reproducible methodology and workflow to produce a comprehensive inventory of biodata resources that can be periodically updated with minimal human curation. This insight provides a fuller picture of the scope of global resources, their number, geographical location and their funding sources.

There have been many efforts to catalogue biodata resources over the years. For example, in each of the past 29 years, the journal Nucleic Acids Research has published its annual database issue that invites papers containing updates from databases previously published in the issue, as well as papers from newly established biodata resources. In this way, it maintains a catalogue of primarily molecular-biology related databases. There are also several registries in existence that encourage registration of data resources.


Photo credit : Heidi J Imker

"We're excited that we were able use open science tools to identify thousands of biodata resources from the literature — and even more excited to see how the inventory will be used to understand the scope and scale of this incredibly important infrastructure."

Heidi J Imker - *University of Illinois Urbana-Champaign*

"We've built reproducibility into our methodology, something we believe will be very valuable for the future."


Photo credit : Ken Schackart

Ken Schackart - *University of Arizona, Tucson*

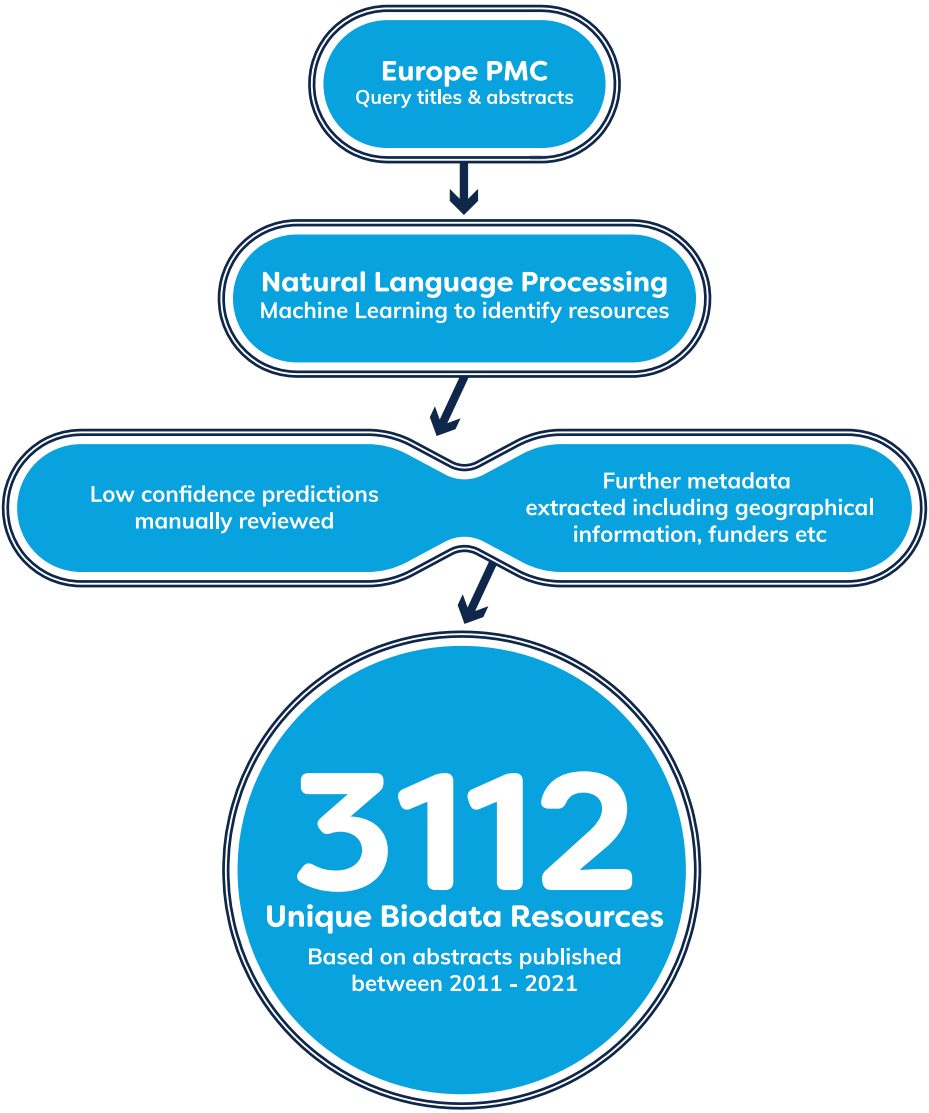


Figure: Methodology and findings for inventory of global biodata resources.

Our approach

Our methodology employed machine learning-enabled, natural language processing techniques on open data from the scientific literature, available in the Europe PMC literature database, to identify data resources by means of associated articles and papers published by the resource owners. This was aimed at informing the broader scientific community about the availability and utility of the biodata resource.

In collaboration with experts at the Chan Zuckerberg Initiative, openly available machine learning models were tested and trained to classify articles and extract resource names from article titles and abstracts in order to build the initial inventory. Metadata from associated articles were then gathered to identify additional information, such as geolocation of the resource and the names of funders.

What we learned

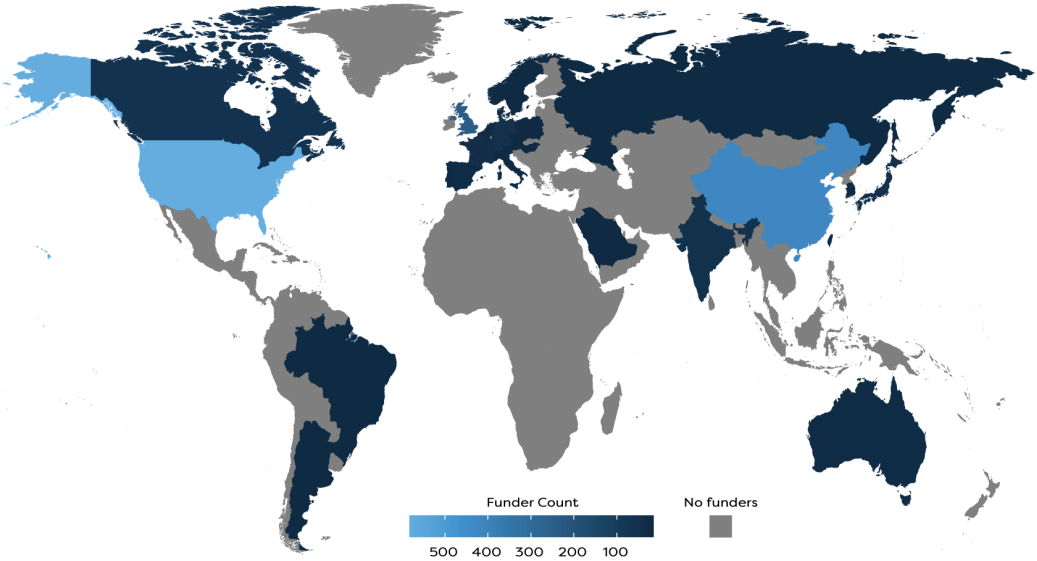
The completed inventory identified 3112 unique biodata resources described in 3705 articles in the Europe PMC data resource for articles published from 2011–2021. For roughly half of these resources it was possible to determine the location of the resource and/or the names of one or more funders. This identified 65 different countries and hundreds of unique funders for the biodata resources in the inventory. Full results are described in a preprint.

In addition to describing the biodata resource infrastructure, a further goal of this work was to create a fully open project with products that could be reused by anyone and to a high level of reproducibility. Accordingly, all code, including a fully reproducible workflow, as well as the final dataset, are available on the GBC's GitHub repository.

Access the inventory preprint

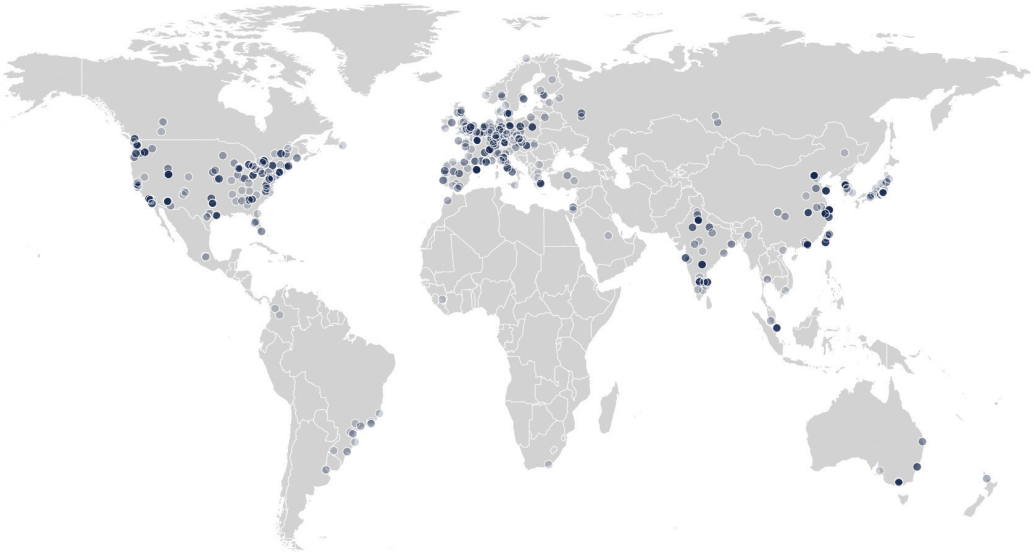
Where are the funders of biodata resources?

The map shows the number of biodata resources per country based upon country of origin for the most frequently identified funding agencies.



Where are biodata resources hosted?

The map shows the locations of 1769 biodata resources based upon IP address coordinates. Darker spots indicate overlapping points.



Global Core Biodata Resources

What are the GCBRs ?

In order to focus on its core aim to stabilise and ensure sustainable financial support for an identified and prioritised set of Global Core Biodata Resources, the GBC conducted an open process to determine an initial set of GCBRs. These resources are of fundamental importance to the wider biological and life sciences community and are crucial for sustaining the broader biodata ecosystem.

The GBC has convened the managers of these biodata resources, which are of fundamental importance to the wider biological and life sciences community, to now work together in the GCBR Forum to consider approaches to the challenge of biodata resource sustainability.

Access the list of GCBRs

How were they selected ?

In 2022, the GBC ran a first selection process for the GCBRs, which required them to meet certain key criteria, including :-

- providing free and open access to their data
- being extensively used, in terms of the number and distribution of users
- being mature and comprehensive
- being considered authoritative in their field
- being of high scientific quality
- providing a professional standard of service delivery

From a field of more than 60 applications, an initial group of 37 GCBRs was identified. In 2023, a second GCBR selection round will be held with new GCBRs to be announced by the end of 2023. Their resource managers, in turn, will join those of the existing GCBRs to become Members of the GCBR Forum.

"The investment in the GCBR selection process by biodata resource managers, and the commitment from our expert review panel in support of the GBC's mission, has inspired us all."

Rachel Drysdale - *GBC Scientific Project Manager*

Working collectively as the GCBR Forum

Representatives of each of the GCBRs comprise the GCBR Forum which meets six times a year to share experience and good practice, engaging with the Secretariat to work on initiatives that enhance the sustainability of the global biodata infrastructure.

The GCBR Forum provides an interface between the resource managers of GCBRs and the Global Biodata Coalition's Member funders and opportunities to work together to address shared challenges.

Outputs

By working together as biodata resource managers, the GCBR Forum will co-develop principles and models of biodata sustainability for discussion with funders. The effectiveness of sustainability mechanisms and indicators will be also be evaluated and tested by the Forum.

As the Forum embeds and grows it plans to report core consensus metrics demonstrating the impact and usage of GCBRs, as well as common policy positions and statements of good practice for resources.

Board working groups on Sustainability & Open Data

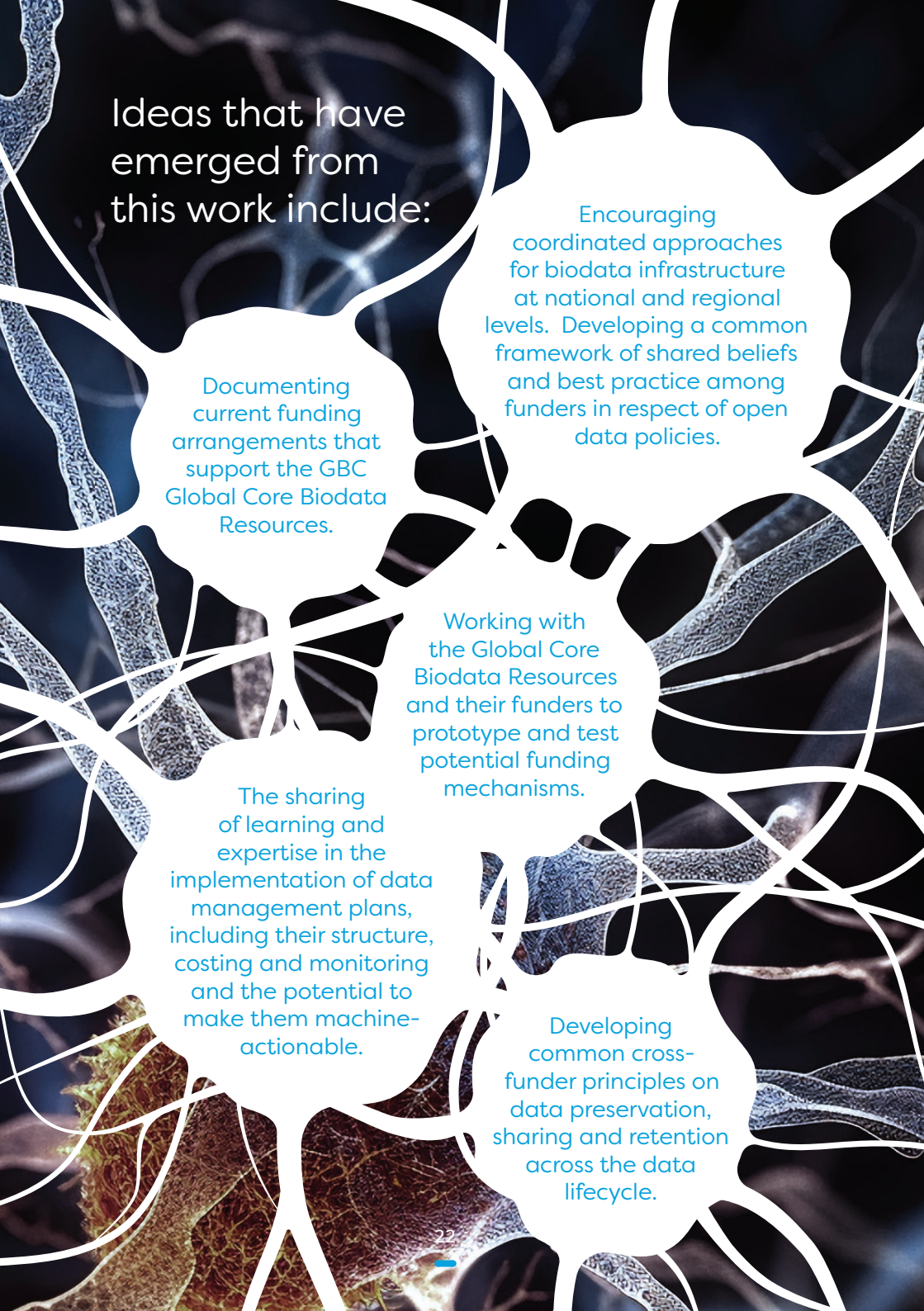
One of the key missions of the Global Biodata Coalition is to bring together funding organisations and stakeholders, who might not ordinarily interact in the course of their work, with the aim of sharing knowledge and proposing cooperative approaches to solving the complex problem of biodata sustainability and open data research strategies.

To this end, two working groups of the Board of Funders and Observers were established in 2022, on the topics of Sustainability and Open Data Strategies. These groups have worked to produce consultation papers which are now ready for consideration by the wider biodata stakeholder community, before being published as white papers. The purpose of these papers is to inform and build alignment among Member funders for their respective policies as they relate to the long term funding of biodata.


Photo credit: David Carr

"Our Working Groups are enabling cross-funder dialogue on cooperative approaches to sustain the global biodata resource infrastructure and maximise the value of open research data."

David Carr *- Global Core Biodata Resource Selection Process Manager*



Ideas that have emerged from this work include:

Encouraging coordinated approaches for biodata infrastructure at national and regional levels. Developing a common framework of shared beliefs and best practice among funders in respect of open data policies.

Documenting current funding arrangements that support the GBC Global Core Biodata Resources.

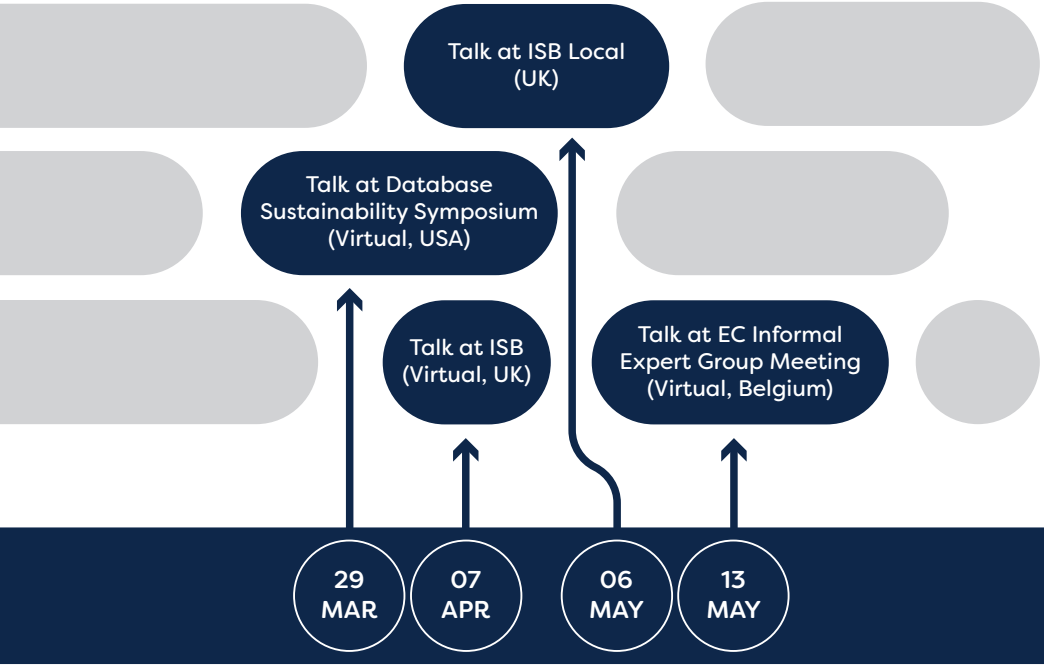
Working with the Global Core Biodata Resources and their funders to prototype and test potential funding mechanisms.

The sharing of learning and expertise in the implementation of data management plans, including their structure, costing and monitoring and the potential to make them machine-actionable.

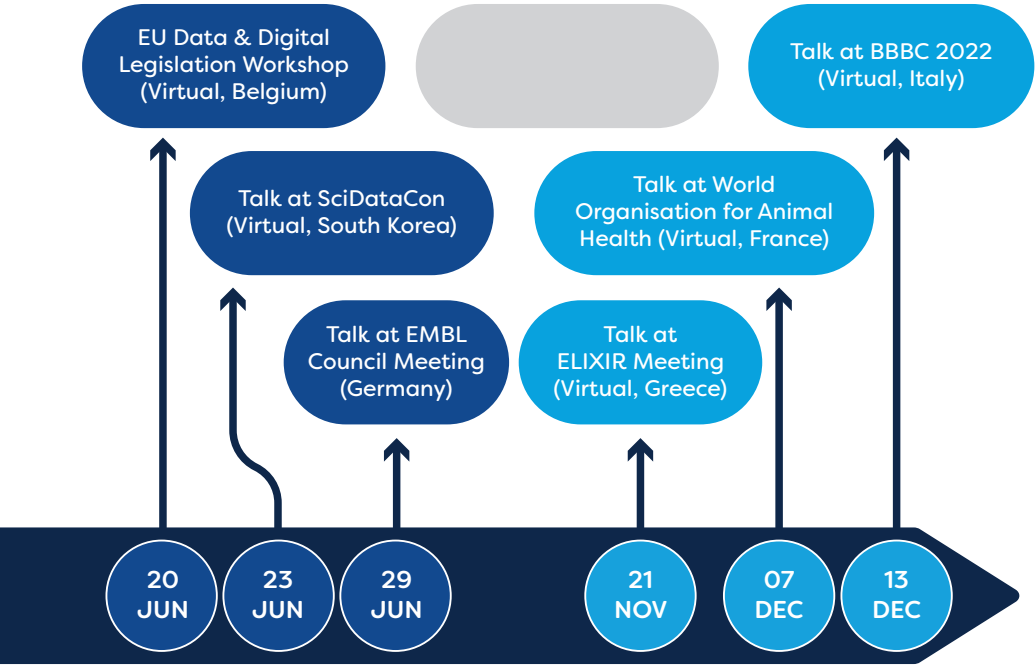
Developing common cross-funder principles on data preservation, sharing and retention across the data lifecycle.

The GBC out & about

Our Executive Director, Guy Cochrane and our Program Manager, Chuck Cook spend a lot of time explaining the work of the GBC to a range of audiences around the world and engaging with life science funding organisations and other stakeholders to encourage their membership of, and participation in, the activities of the GBC.



This engagement takes the form of remote participation in global scientific conferences, meetings and workshops, both hybrid and in-person. We are conscious of our carbon footprint and our small Secretariat was able to cover the world in 2022 using the power of video conferencing.



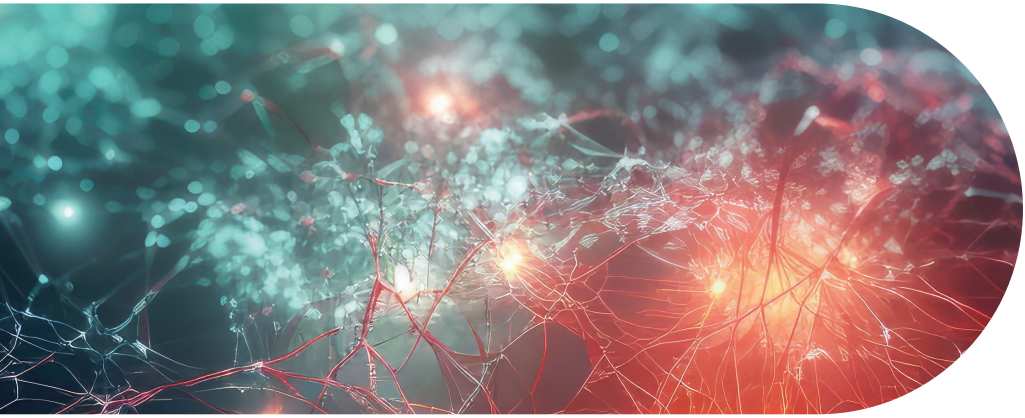
In 2022 our virtual meeting attendance included meetings hosted in the USA, Korea, Greece and Italy and beamed to audiences in multiple countries across the globe.

Forward look

2023 & Beyond

In the coming year and beyond, the work of the GBC will be centered on developing and driving forward a comprehensive scientific programme that provides a framework for collaboration amongst our Member funders, biodata resource managers and wider stakeholder community. Our clear focus remains the long term sustainability of the global biodata infrastructure and its component resources.

Later in 2023, we will announce the second tranche of Global Core Biodata Resources. Managers of resources selected in this second round will join managers from the first round in the GCBR Forum, which meets regularly with the aim of building a strong community of data resource managers who will share best practice in resource sustainability, while developing and piloting models for sustainable funding.



The ideas that have been generated this year by the cross-funder Board Working Groups on Sustainability and Open Data Strategies will be shared in a stakeholder consultation process, and inform the development of collaboratively-authored and published white papers for how GBC will work with our communities to address these complex challenges.

We will be continuing to engage globally with funding bodies through bilateral conversations and speaking opportunities at conferences, meetings and workshops which are central to the interests of our Members. One of the key aims of raising the profile of the GBC is to encourage new Member funders to join our organisation and lend their expertise to developing principles for biodata resource sustainability globally.

We are always very pleased to hear from stakeholders who would like to discuss membership or indeed any aspect of the GBC's scientific programme.

Please contact us at : **info@globalbiodata.org**

GBC Members

We would like to thank our Members for their support for the establishment and ongoing development of the GBC.

- National Institutes of Health, USA
- National Science Foundation, USA
- Wellcome, UK
- SERI, Switzerland
- NHMRC, Australia
- Biotechnology and Biological Sciences Research Council, UK
- The Research Council of Norway
- Genome Canada
- Chan Zuckerberg Initiative, USA
- The European Molecular Biology Laboratory
- The African Academy of Sciences

Our observers

- European Commission
- Indian Council of Medical Research
- Institut National de la Santé et de la Recherche Médicale, France
- Canadian Institutes of Health Research
- Natural Sciences and Engineering Research Council, Canada
- South African Medical Research Council
- European Research Council
- National Research Foundation of Korea
- Royal Society Te Apārangi, New Zealand

In-kind support

We are grateful for in-kind support for the GBC from :

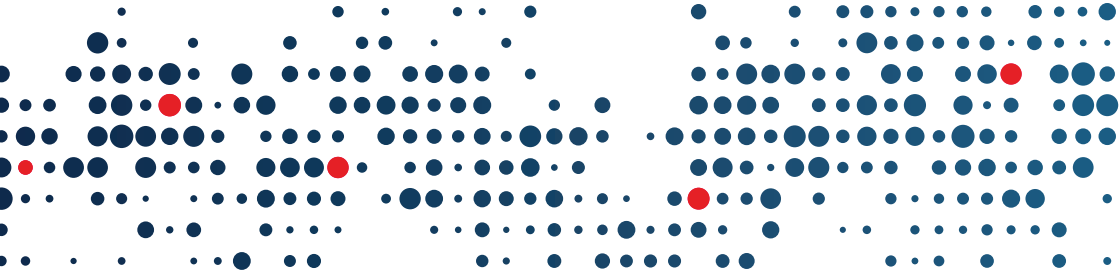
- The Human Frontier Science Program Organization
- The Swiss Institute of Bioinformatics

Our governance

The Global Biodata Coalition is a non-profit member organisation, receiving financial support for its Secretariat from its Member funders. Our funds are managed by the Human Frontier Science Program Organization, based in Strasbourg, France.

The strategic direction of the GBC is set by our Board of Funders, comprising representatives of the Members. Our Executive Director, Guy Cochrane, leads our small Secretariat in implementing strategy and developing our scientific programme.

The GBC Scientific Advisory Committee (SAC) provides advice to the GBC Board of Funders and the GBC Executive Director on all aspects of the development and implementation of the GBC's scientific activities, and on the broader international life science and biomedical science research landscape.



With thanks

Board Members

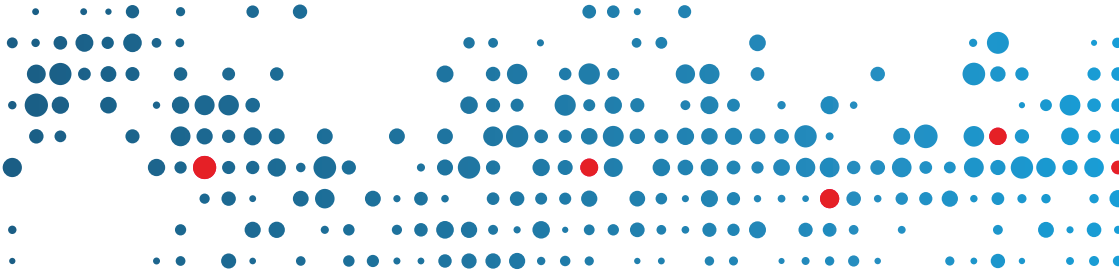
We would like to thank our Board Members for their support for the strategic direction of the GBC and Professor Warwick Anderson for his work as Chair of the Board.

Scientific Advisory Committee Members

We would like to thank our Scientific Advisory Committee Members for their support for the scientific endeavours of the GBC.

GCBR Selection Reviewer Committee

We would like to thank the experts who volunteered their time to participate in the independent review and selection process for the Global Core Biodata Resources.

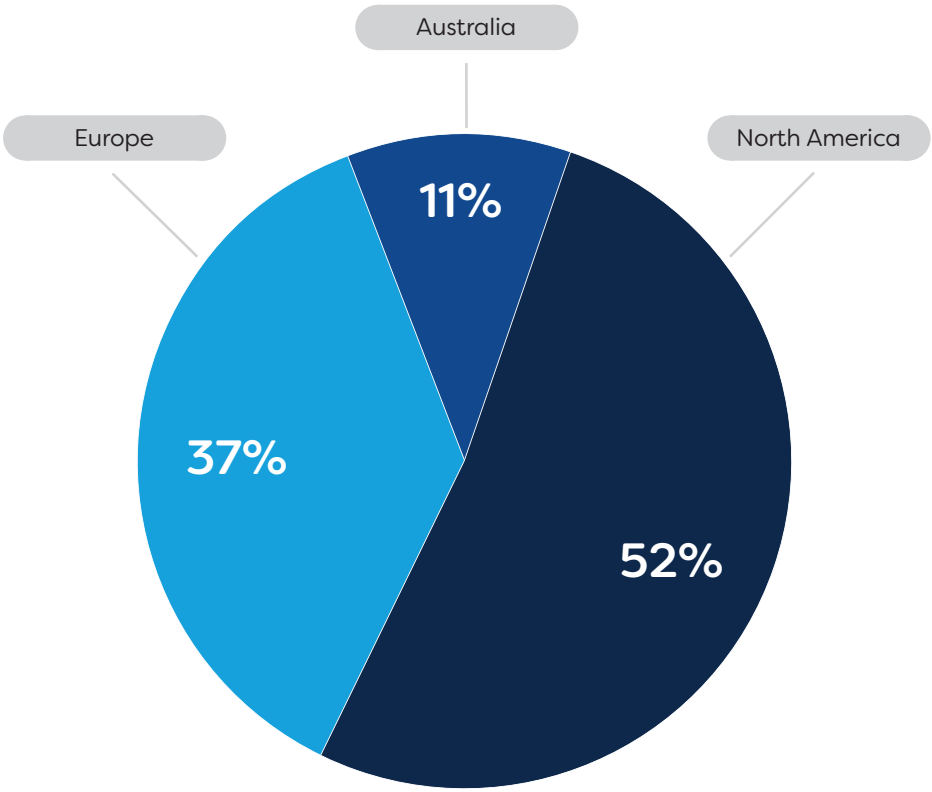


Financial figures

Funding sources / Operating expenditure

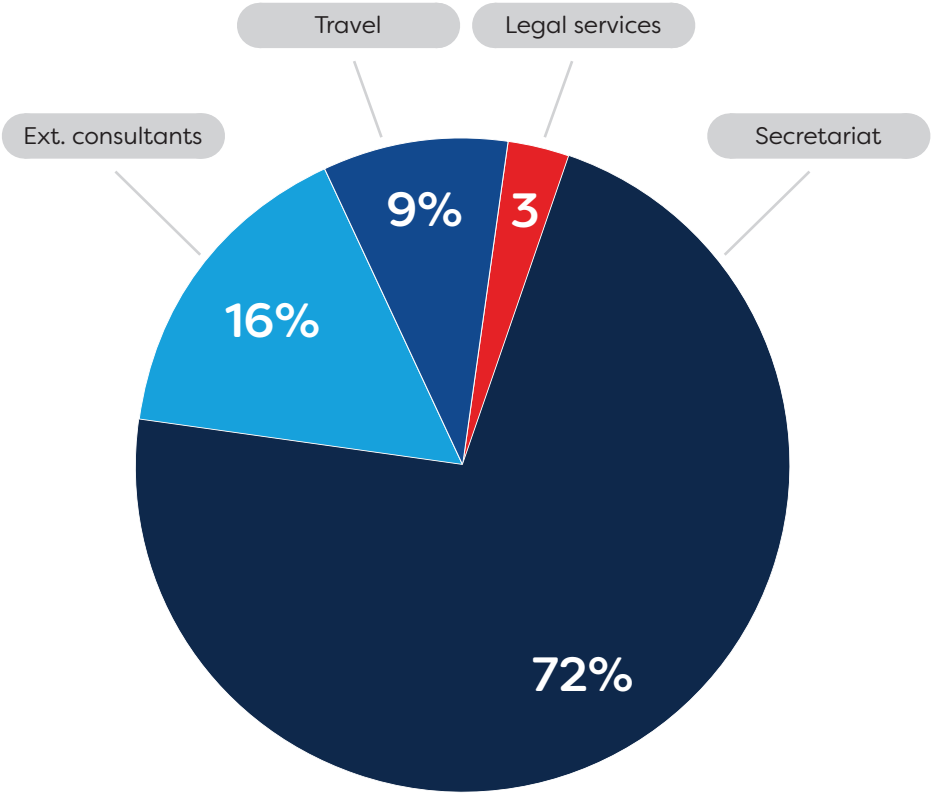
Funding

The GBC Secretariat is funded by contributions from its Members who are public and charitable research funding organisations around the world. In 2022, 52% of our funding came from North America, 37% from Europe, and 11% from Australia.



Operating expenditure

GBC's fiscal year runs from April to March. In the 2022-2023 fiscal year, GBC's total expenditure was €582,983. Of this, 72% was for support of the Secretariat, 16% for external consultants, 9% for travel (including travel for staff, SAC members, and consultants), and 3% for legal services and administrative costs.



Global Biodata Coalition

www.globalbiodata.org

info@globalbiodata.org

@globalbiodata

@globalbiodatacoalition2030

A digital version of this publication is available at :

www.globalbiodata.org/newsandmedia/