# Global Core Biodata Resources: Concept and Selection Process

This document has been produced by the Global Biodata Coalition[1].
Contributors to this document are listed in the "Document development" section.
Email address for correspondence: comms@globalbiodata.org.

## Table of Contents



[1] https://globalbiodata.org/

## Purpose of this document

The Global Biodata Coalition has developed a selection process to define Global Core Biodata Resources across biological, life science, and biomedical data resources (biodata resources) worldwide. This document explains the Core Biodata Resource concept and the procedure for their selection, thereby providing interested biodata resources with the opportunity to participate in the process.

## What is the Global Biodata Coalition?

The Global Biodata Coalition[2] (GBC) is composed of biological, biomedical and life science research funders and aims

- to enable funders of biodata resources to better coordinate the global infrastructure of biodata resources and to share approaches and strategies for its efficient management and growth, and
- to stabilize and ensure sustainable mechanisms for financial support for the global biodata infrastructure, with a focus on an identified and prioritized set of Global Core Biodata Resources.

In pursuing these aims, the GBC does not in itself function as a funding agency and will not provide direct funding to support biodata resources. Rather, the GBC serves to identify and illuminate the requirements for long term sustainability of the biological data landscape, working alongside data resource managers and the scientists the data resources serve, to fulfill its aims. A Letter of Understanding[3] describes the rationale, aims, scientific activities and governance of the GBC.

## The need to define Global Core Biodata Resources

Biological and life science data resources have been used extensively in academic research and industry for well over two decades in all fields of life science research, and they now contribute to applications in clinical and industry settings. In aggregate, life science data resources around the world form an immense distributed and interconnected infrastructure, arguably forming the largest infrastructure in biology. These resources are critical for ensuring the reproducibility and integrity of the entire life sciences research enterprise (Bourne et al., 2015).

Despite their importance, almost all biodata resources are supported in whole or in part by short-term grants or other funding mechanisms, and there is little coordination among funders of these resources (Anderson et al., 2017; Berman, 2008; Gabella et al., 2017). This is a precarious situation and poses a threat to the continued provision and development of hard-won data and research insights.

The Global Biodata Coalition was established in recognition of these challenges and to consider a strategy whereby funders might coordinate a response to the threat posed by this instability. Through a series of meetings, consultations, and workshops held from 2016 to 2019, research funders, data resource managers, and other stakeholders recognised the need to identify and prioritize a set of Global Core Biodata Resources, thereby allowing the funding agencies to understand which biodata resources are crucial for the larger infrastructure and develop strategies to coordinate stable funding across those crucial resources.

Within this context, the GBC defines biodata resources as biological, life sciences, and biomedical databases that archive research data generated by scientists, serving as the repositories of record for particular data types; as well as knowledgebases that add value by aggregation, processing, and expert curation. As defined by the GBC biodata resources do not include software or tools with the primary function of analyzing or processing user-supplied data.

The GBC has been able to build on the prior efforts by ELIXIR to define Core Data Resources supported within Europe[4] (Durinx et al 2017), and this has facilitated the development of the concept at the global level. Broadly speaking, the GBC will select Global Core Biodata Resources (GCBRs) based on their demonstrated importance for the global biodata infrastructure and their crucial role in the biological, life science, and biomedical research effort. Funding and research structures differ throughout the world and this is reflected in the

[2] https://globalbiodata.org/

[3] https://globalbiodata.org/wp-content/uploads/GBC_Letter_of_Understanding.pdf

[4] https://elixir-europe.org/platforms/data/core-data-resources

definition of Global Core Biodata Resources. The characteristics that define a GCBR are described in detail in the next section.

Defining GCBRs will provide benefits for funders, for the resources themselves, and for researchers who depend upon biodata resources:

- For researchers selecting repositories to archive their primary data, for example to comply with funders' and publishers' open data requirements, GCBR status will provide confidence in their choices.

- For funding agencies and science publishers, the availability of recognised lists of GCBRs will allow them to recommend suitable data repositories and reliable sources of information to their grantees and authors, with confidence.

- For individual funding agencies faced with applications for support from multiple (and sometimes competing) data resources, a list of GCBRs and the criteria that characterise them will provide insight into the established data resource ecosystem, in turn providing useful context for local decision making.

- For managers of developing data resources, databases that have been identified as GCBRs will provide examples of good practice that can inform the development of their own data resources.

- For managers of databases defined as GCBRs, who collectively form an expert group critically aware of the need to ensure long-term-sustainability for the research infrastructure to which they contribute, the GCBR community will provide a forum for sharing expertise, driving collaboration, and exploring potential solutions to the challenge posed by their precarious funding.

- For all stakeholders, including private enterprise, open data fires contemporary biological and applied research and allows researchers to access and reuse data, driving discovery. Working toward GCBR status will inspire data resources to implement more permissive open data licenses (Drysdale et al, 2020) so that they more fully reflect the FAIR principles (Wilkinson et al., 2016).

While the GBC is not itself a funding agency, and thus cannot provide direct support to the selected GCBRs, these factors taken together clearly demonstrate the power of defining this infrastructure ecosystem in this way, to the benefit of the full range of stakeholders.

## Characteristics of Global Core Biodata Resources

Global Core Biodata Resources are biodata resources that are of fundamental importance to the wider biological and life sciences community and the long term preservation of biological data.
They:

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

Their operation is based on well-established life-cycle management processes and well-understood dependencies with related data resources. GCBRs have either Terms of Use or specific licences that conform to the Open Definition[5], to enable the reuse and remixing of data (Durinx et al., 2017; Drysdale et al., 2020).

There are of course biodata resources that cannot comply with such broad open data definitions, because, for example, they house confidential human patient data, or sensitive biodiversity data, for which access permission from a Data Access Committee or equivalent oversight arrangement is required. The initial rounds of selection will focus on data resources that provide unrestricted open access worldwide and will not include resources that require permission for access. Resources that require permission for access will be addressed in the future, including via collaboration with other organizations working to develop policies for such data. The requirement for unrestricted open access will be in place for the early rounds of GCBR selection to reduce complexity as the application process and the review of applications are established.

GCBRs may be deposition databases and/or knowledgebases, and may house data from across the biological, biomedical and life sciences, including imaging, molecular, physiological, genomic,

[5] http://opendefinition.org/licenses/

biodiversity, ecological, or other life sciences data. GCBRs are, by definition, part of the fundamental infrastructure of biological and life science research, which also includes a broader range of data resources with diverse motivations, such as those focusing on a particular scientific domain or specialisation. GCBRs are distinct from databases produced as part of a specific research project.

For this initial selection round, it is expected that the GCBRs will be available in English. Some biodata resources may also provide data in other languages and the application process requests information on all languages in which the data resource is made available. Subsequent rounds of selection may expand to include biodata resources that do not provide a user interface in English but nevertheless are used extensively both in terms of the number and distribution of their users.

## The Global Core Biodata Resource Indicators

Global Core Biodata Resources are characterised according to multiple quantitative and qualitative indicators that fall into five categories (Durinx et al., 2017):

| Global Core Biodata Resource Indicator Categories | | |
|---|---|---|
| | **Indicator Category** | **Description** |
| 1 | Scientific focus and quality of science | *This category includes the inherent scientific quality of the data and metadata, the resource's uniqueness and comprehensiveness, whether the resource is a recognised authority, and whether the resource is of fundamental importance to the broad life science community and the long-term preservation of biological data.* |
| 2 | Community served by the resource | *This category reflects the size and the measured demand of the communities that are served by the resource and includes web statistics, user reach, and international use.* |
| 3 | Quality of service | *This category includes service levels, reliability, and technical performance as well as use of data and metadata standards, data availability, provenance compliance and user support.* |
| 4 | Funding, governance and legal infrastructure | *This category covers the funding, governance and legal footing, including open science licensing, and privacy and ethics policy considerations.* |
| 5 | Impact stories | *This category includes evaluation of how effectively the data resource is meeting the needs of the scientific community via counterfactual and accelerating science stories.* |

A complete list of the indicators within each category can be found in the Appendix below. Collectively, these indicators form a profile that characterises the resource and enables identification of those data resources that meet the requirement for GCBR status.

In considering whether or not a resource has the qualities of a GCBR, it is important to recognise that a great many data resources are of exceptional quality, have high technical standards, and may form a crucial component of the research environment for the particular community they serve. However, a resource will not be deemed to be operating in the context of a GCBR as defined by the GBC at this time, unless it:

- forms an essential component of the worldwide fundamental infrastructure of biological, life science and biomedical data resources,

- serves a broad cross-section of research fields, and
- fulfils a purpose towards the long-term preservation of raw or aggregated data internationally.

Examples of data resources that will not meet the criteria for selection as a Global Core Biodata Resource include:

- early-stage data resources that are primarily used by those who established them,
- data resources that are targeted at a specific research question or scientific field and consequently used by a narrow sector of the research landscape, or
- data resources established to store and represent data results from a specific, time-limited research project.

Data resources have a life cycle, and it may be the case that a data resource that today aligns with one of these examples is at the beginning of a journey to become globally significant over time as its user base and mission grow, expand and mature. Likewise, as techniques evolve, biodata resources that are initially selected as GCBRs may over time become obsolete as new technologies develop and new fields of research emerge.

# The Global Core Biodata Resource selection process

## Overview

This first identification of Global Core Biodata Resources will necessarily be highly selective and will serve as the initial, pilot, selection process: in this first round a limited number of representative GCBRs will be identified. The selection process will be iterative, with future rounds every two or three years to select additional resources as the life science data landscape evolves over time, as biodata resources establish how to meet the data provision requirements for application, and as the principles emerging in this initial round inform future selection rounds.

GCBRs must be freely available for all users worldwide: biodata resources that currently require payment for access, or that require permission for access, will not be selected as GCBRs.

### *Application process - two steps*

Applications will be made in two steps:

1. Expressions of Interest: Biodata resources will submit an expression of interest statement built on a specified template that ensures fulfilment of basic Eligibility Criteria and poses Short Answer questions based on the five categories of indicators for GCBR selection. The expressions of interest will be assessed by a review committee and those selected will be invited to proceed to the full application.
2. Full Application: Those biodata resources shortlisted in the expression of interest step will proceed to a full application including detailed questions about the complete set of 23 indicators used in GCBR selection. The full applications will undergo review, with the review committee having the task of deciding which data resources they recommend to be included in the initial list of GCBRs. The GBC Board of Funders will make the final formal decision on the list of GCBRs, based on the review committee's recommendations.

The two-step application process has been developed for several intersecting reasons, listed below, all relating to the intention of the GBC to benefit the landscape of global biodata resources.

**Focussing impact of GBC mission/strategy/vision:** There are thousands of biodata resources worldwide[6] (Imker, 2018). The GBC has been set up as an international body to consider the entire ecosystem of biodata resources, identifying the GCBRs around which to develop a program to focus attention on the biodata infrastructure needing proper support. It is neither feasible nor necessary for the GBC to focus attention on thousands of biodata resources, but it can consider multiple, possibly as many as 100 ultimately, to make that case, and the GCBR selection will identify that set. The initial round of selection will identify the first tranche of GCBRs, allowing the GBC to begin describing the infrastructure on behalf of research funders. The two step process facilitates the degree of filtering and selection necessary to achieve this goal.

**Expectation setting for applicants:** Related to this high degree of filtering and selection, it is important to set the expectations of applicants so that time and effort are not squandered in preparing applications for data resources that are not likely to be selected as GCBRs in the initial selection round.

[6] http://bigd.big.ac.cn/databasecommons/

Submitting a full application will be a time-consuming effort for the data resource managers and Principal Investigators who will shoulder that task. The more lightweight process of completing an expression of interest form will necessarily alert potential applicants to the rigour of the process upon which they may be embarking, so if they choose to continue they do so from a position of being well informed.

**Optimising reviewer's contribution:** Assessing data resources for inclusion in the initial GCBR set on the basis of the full application will require significant effort on the part of the reviewers and GBC Secretariat. The expression of interest screening will enable controlling the number of full applications to a manageable number, so that at each stage of the process those involved can dedicate the attention that each full application deserves, for its thorough consideration.

### *Review of applications*

Review of applications will be facilitated by the GBC Secretariat, though the Secretariat will have no role in the selection decision for each applicant data resource. At both the expression of interest and the full application stages, applications will be reviewed by a committee comprising the GBC Scientific Advisory Committee, and with additional experts in the scientific domains of the applicant data resources, where this is necessary to provide uniformly high standards of critical review across all applications. In assembling the reviewer committee gender, field of expertise, and geographical/international balance will be considered. The names and affiliations of members of the review committee will be published after the review process is concluded.

## Publicising the selection process

In advance of the opening date for submission of expressions of interest, the GBC will publicise the call for applications widely via its web site[7,8], social media, and via the members of the GBC Steering Group and the GBC Board of Funders. In most cases, individual data resources will choose to submit an application as a consequence of becoming aware of these communications. Additionally, funders, who have an in-depth understanding of the portfolio of biodata resources that they support, may wish to nominate data resources directly to the GBC secretariat[9], and the GBC secretariat will ensure that those data resources are made aware of the opportunity. Irrespective of the route by which they became aware of the opportunity, the application procedure is the same for all interested data resources.

## Guidance before submission

A series of webinars for applicants will communicate expectations regarding the submissions. Details regarding the webinars will be published on the GBC web site[8] and via social media. The webinars will cover the indicators used for selection, the reasoning for choosing these indicators, the review process that will select the Global Core Biodata Resources, and will allow time for a Q&A session. The presentation (though not the Q&A, for reasons of confidentiality) will be recorded and disseminated via the GBC web site, for those not able to join live. A FAQ document on the GBC web site[10] serves as a point of reference and index for relevant documents.

## Confidentiality

The primary intention of the GBC is to be a beneficial influence on the data resource landscape, and observing confidentiality with regard to the GCBR process will support that intention. In addition, at the full application step, certain data requested in the Application Form may exceed that which the data resources generally make public, for example with respect to quantitative usage statistics. Confidentiality is of the utmost importance here, also, and will be strictly respected. Therefore, all reviewers will be asked to sign a confidentiality declaration to cover their involvement in the selection of the GCBRs.

## Conflicts of interest

All reviewers will be asked to declare any conflicts of interest they have or may have in performing their role, both at the time they accept the invitation to participate and at any point in the process where a possible conflict becomes evident. Considerations of conflicts of interest are important for two reasons:

- To ensure all applications are assessed fairly with due attention to integrity and transparency of the review process
- To ensure the reviewer committee is protected from allegations of bias.

[7] https://globalbiodata.org/

[8] https://globalbiodata.org/scientific-activities/gcbr-selection/

[9] By email to gcbr-selection@globalbiodata.org

[10] https://globalbiodata.org/scientific-activities/gcbr-selection/faq/

Conflicts of interest arise where any of the following conditions apply to the reviewer:

- membership of the Scientific Advisory Board of the candidate biodata resource
- active collaborations with the candidate biodata resource
- membership of the institution or department that hosts the candidate biodata resource
- family relationship with any of the candidate biodata resource's key personnel, Principal Investigator, or equivalent
- signatory for letters of support for funding applications by the candidate biodata resource during the previous three years
- co-authorship of primary research publication with candidate biodata resource key personnel, unless the publication is a mega-multi-author article.

Where a conflict of interest arises the reviewer will be reassigned, to avoid the conflict.

Conflicts of interest might arise where any of the following conditions apply to the reviewer:

- active collaboration with a distinct biodata resource that has also applied for selection, where that biodata resource is a competitor to the resource that is the subject of the review
- personal relationship with any of the candidate biodata resource's key personnel, Principal Investigator, or equivalent
- signatory for letters of support for funding applications by the candidate biodata resource dating back more than three years
- co-authorship of non-research publication (e.g. review, meeting report, news item, commentary) with candidate data resource key personnel

In such cases, the GBC Secretariat will decide whether the conflict will result in reassignment of the reviewer.

## Timeline

The timeline for the GCBR selection process will be made available on the GBC website[11].

## Expressions of Interest

Interested data resources will need to determine a representative, typically a data resource manager, to act as the correspondent for the process. This person will coordinate an expression of interest document and will continue to represent the database for the full application, should the data resource be invited to proceed to the next stage. Those data resources that are selected as GCBRs will be expected to continue to engage with the GBC, as specified further below. This time commitment should be borne in mind when considering who best should act for the data resource in submitting the expression of interest.

When considering whether or not to submit an expression of interest for the GCBR selection process, data resource managers should bear in mind that this pilot selection round will identify a limited number of the most critically important Core Biodata resources. It is likely that around 40–50 biodata resources will be invited to submit full applications, which will require significant investment in terms of preparation by the applicant. While the process is designed to identify Core Biodata resources from across a wide range of biological and life science domains that are critical for researchers in various contexts on a global scale, the intention is not to divert attention away from essential data resource operations and towards premature or hasty applications that are not likely to succeed. Selection of GCBRs will necessarily be iterative, as the life science data landscape evolves over time, and the GCBR list will mature with future selection rounds. Where the data resource manager recognises that the characteristics of their database are firmly in line with the Core Biodata Resource indicators, then a decision to invest in preparation of the expression of interest will be reasonable.

### Preparing submissions for joint resources

Managers of data resources that are components of larger established collaborations or consortia will need to consider the context in which they lodge their application: the unit of application may be an individual resource or the larger consortium. Consultation between members of the consortium, and possibly their funders, will be necessary so that all parties are aware of, and in agreement regarding, the strategy being adopted for application. The expression of interest template includes a short answer question where the reasoning behind the application strategy can be explained.

### Expression of Interest data required

A complete listing of the information requested at the expression of interest stage can be found in the "Data_Required_GCBR_Expression_of_Interest_Suppl_Data" file supplied with this article, here[12].

[11] https://globalbiodata.org/scientific-activities/gcbr-selection/

[12] https://doi.org/10.5281/zenodo.5846742

### Eligibility Criteria

The Eligibility Criteria take the form of yes/no questions regarding

- data resource maturity and stability
- ability to provide quantitative usage data
- metadata and standards
- data availability and format
- governance
- open data policy

### Short Answer Questions

The Expression of Interest template also includes short answer questions relating to the alignment of the data resource with the goals of the GBC and the characteristics of the resource—the indicators introduced above—that qualify it to be considered a Core Biodata resource.

The combination of the Eligibility Criteria and the Short Answer questions will ensure that only those data resources most likely to meet the requirements for the initial GCBR list will be invited to complete a full application.

Instructions for submission of an expression of interest will be publicised on the GBC website[13].

### Expressions of Interest review process

Once the deadline has passed the expressions of interest will be screened initially by the GBC Secretariat for administrative completion. A review committee will screen the expressions of interest and will decide which resources will be invited to submit a full application.

It is likely that there will be a high level of interest in this initiative, with many data resources submitting an expression of interest. Consequently it will not be possible to provide detailed feedback. Those that will proceed to the next stage will be provided with detailed instructions for the full application.

## Full Application

### Full Application data required

Each candidate data resource that clears the expression of interest step will be invited to submit the Full Application, which includes questions covering all the indicators used to identify the Global Core Biodata resources.

The indicators used for GCBR selection are grouped into the five categories introduced above:

1. Scientific focus and quality of science
2. Community served by the resource
3. Quality of service
4. Funding, governance and legal infrastructure
5. Impact and translational stories

The data requested in the Full Application reflect these five categories, covering the specific indicators within each category. The full list of specific indicators is presented in the Appendix below: some of the information requested is quantitative and some qualitative. Details of the information requested at the full application stage can be found in the "Data_Required_GCBR_Full_ Application_Suppl_Data" file supplied in association with this article, here[14].

### Full Application review process

The full application review process will have two phases. Initially, each full application will be reviewed by three members of the review committee who will comment and assign a score for each of the five indicator categories:

[13] https://globalbiodata.org/scientific-activities/gcbr-selection/

[14] https://doi.org/10.5281/zenodo.5846758

| Full Application Review Scoring Scheme | |
|---|---|
| **Score** | **Description** |
| 4 | Comprehensively meets criteria for a Global Core Biodata Resource across all indicators within the indicator category |
| 3 | Meets criteria for a Global Core Biodata Resource across indicators within the indicator category, though less comprehensively for some indicators than others |
| 2 | Does not meet criteria for this indicator category due to weaknesses in one or more specific indicators; could become a Global Core Biodata Resource in the future if the weaknesses were addressed |
| 1 | Does not meet criteria due to major weaknesses, not suitable as a Global Core Biodata Resource |

Reviews for each application will be compiled and reported back to the applicants. Applicants will be invited to raise any concerns or objections regarding, for example, the accuracy of the reviewers' remarks, or to identify misconceptions or misunderstandings that might have arisen as part of the review process. Applicants can submit a response to reviewers' comments within two weeks of receiving the compiled reviews. Responses to reviewers comments arriving after the two week period will not be carried forward to the second phase of the full application review process.

In the second phase of the review process, a specific panel meeting will be convened for the review committee. The panel meeting will be facilitated by the GBC Secretariat, though the Secretariat will have no role in the selection decision for each applicant data resource. Review committee members will be assigned data resource applications for which they will act as rapporteur. In this role, the review committee member will introduce each application for which they are responsible, for discussion by the committee as a whole. Any responses to reviewers' comments received from the applicant will be included for consideration at this stage. After discussion, the review committee will agree on recommendations regarding which of the applicant biodata resources to include on the initial list of Global Core Biodata Resources.

### Approval of recommendations

The final formal decision regarding the applicant biodata resources to be included in the initial GCBR set will be made by the GBC Funders Board. The review committee will prepare the recommendation for the Board, with the administrative assistance of the GBC secretariat. The Secretariat will forward the list of GCBRs proposed by the review committee, together with a summary statement for each recommended application and a brief report describing the committee's ranking process, to the Board, for consideration and confirmation.

### Reporting the outcome of review to applicants

Each data resource that submitted a full application will receive the summary statement generated at the full application review committee meeting as well as a notification of the decision made by the GBC Funders Board.

The complete list of applicants, all submitted application documents, and all reviewer feedback, scores and summary statements will be held by the GBC Secretariat. None of these documents will be made publicly available.

## Announcing the Global Core Biodata Resource list

After the review process has reached completion, the outcome in terms of the initial list of Global Core Biodata Resources will be announced publicly via the GBC web site, a news announcement, and social media.

# Responsibilities of the identified Global Core Biodata Resources

In undertaking this identification of the Global Core Biodata Resources, the Global Biodata Coalition is seeking to support its aims towards better coordination of the data resource infrastructure and

developing sustainable mechanisms for its long term financial support. Although the GBC is not itself a funding agency and therefore will not provide direct support to the selected GCBRs, the data resources so identified will be invited to continue their engagement with the GBC towards its goals. For example:

- The Full formal application procedure will require applicant data resources to supply detailed information including technical, usage, citation and staffing data. The GBC will periodically request updates to this information from the GCBRs, in order to support the efforts of the GBC in developing long term support from funders for the Core Biodata infrastructure.
- The managers of the GCBRs will be key stakeholders in the GBC and form an expert group. For them, the GCBR community will provide a forum for sharing expertise, driving collaboration, and exploring potential solutions to the challenges they all face. It is likely that participation in the consequently established GCBR Forum will be periodically requested.
- The biological and life sciences undergo constant innovation and development, and the data resource landscape will continue to evolve to reflect that. Some data resources that are nascent and of limited reach at this time may mature into GCBRs. Conversely other data resources may become less pivotal, as techniques evolve and they are superseded because of newer, emerging methodologies. In recognition of these life cycle factors, and the need to keep the GCBR collection relevant, focussed, and aligned with the principles on which it was selected, it will be necessary to review the GCBR list periodically. All GCBRs can expect to participate in periodic review, anticipated to be run on a three- to five-yearly cycle.
- Each data resource named as a GCBR will be expected to support the mission of the GBC, by, for example, reflecting GCBR status on the data resource web site, and including relevant logos in promotional materials and slide presentations.

In all these activities, the GBC and the GCBRs will take a global perspective, and work to foster data resources throughout the world. While the focus of this document is the definition of the initial set of Global Core Biodata Resources, the principles established in and as a consequence of the selection process should be understood to be aspirational, and inspire good practice and optimal adoption of the FAIR principles across the data landscape.

## Document development

This document has been approved by the Global Biodata Coalition Board of Funders[15].

This document was developed under the leadership of **Rachel Drysdale** (ORCiD 0000-0003-3037-0216, GBC Secretariat, UK) working with **Charles E. Cook** (ORCiD 0000-0002-4145-8048, GBC Secretariat, UK) and **Warwick P. Anderson** (ORCiD 0000-0001-7130-3530, GBC Secretariat and Monash University, Australia).

The work was further supported by the GBC Steering Group (SG) with specialist technical advice from a Task Force (TF) convened for this purpose. Listed alphabetically, these contributors were:

**Ron D. Appel**, SIB Swiss Institute of Bioinformatics, Switzerland (SG)
**Rolf Apweiler** (ORCiD 0000-0001-7078-200X), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), UK (SG)
**Guntram Bauer** (ORCiD 0000-0002-7464-2773), Human Frontier Science Program, Strasbourg, France (TF, SG)
**Niklas Blomberg** (ORCiD 0000-0003-4155-5910), ELIXIR, UK (SG)
**David Carr** (ORCiD 0000-0003-1435-307X), Independent Consultant, UK
**James Deshler**, Division of Biological Infrastructure, NSF, USA (SG)
**Valentina di Francesco**, Office of Genomic Data Science, NHGRI, NIH, USA (TF)
**Michael Dunn** (ORCiD 0000-0002-2472-7779), Wellcome, UK (SG)
**Christine Durinx** (ORCiD 0000-0003-4237-8899), SIB Swiss Institute of Bioinformatics, Switzerland (TF)
**Eric Green** (ORCiD 0000-0002-2974-3823), National Human Genome Research Institute, National Institutes of Health, USA (SG)
**Inge Jonassen** (ORCiD 0000-0003-4110-0748), Department of Informatics, University of Bergen, Norway (TF, SG)
**Corinne S. Martin** (ORCiD 0000-0002-5428-2766), ELIXIR, UK (SG)
**Rowan McKibbin** (ORCiD 0000-0003-1445-8269), UKRI BBSRC, UK (TF)
**Saurabh Raghuvanshi** (ORCiD 0000-0002-8349-4290), Indian Biological Data Centre, Regional Centre for Biotechnology, Faridabad, Haryana, India (TF)
**Dario Taraborelli** (ORCiD 0000-0002-0082-8508), Chan Zuckerberg Initiative, USA (SG)
**Yik Ying Teo**, Saw Swee Hock School of Public Health, National University of Singapore, Singapore (SG)
**Alex D. Wade** (ORCiD 0000-0002-9366-1507), Open Science, Chan Zuckerberg Initiative, USA (SG)

[15] https://globalbiodata.org/governance-and-support/

## References

**Anderson W**, Apweiler R, Bateman A *et al.* (2017) Towards Coordinated International Support of Core Biodata Resources for the Life Sciences, bioRxiv, 110825 (doi.org/10.1101/110825).

**Berman HM** (2008) The Protein Data Bank: a historical perspective. Acta Crystallogr A , 64, 88-95 (doi.org/10.1107/S0108767307035623).

**Bourne PE**, Lorsch JR and Green ED (2015) Perspective: Sustaining the big-data ecosystem. Nature , 527, S16-17 (doi.org/10.1038/527S16a).

**Drysdale R**, Cook CE, Petryszak, P *et al.* (2020) The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences, Bioinformatics, btz959 (doi.org/10.1093/bioinformatics/btz959).

**Durinx C**, McEntyre J, Appel R *et al.* (2017) Identifying ELIXIR Core Data Resources [version 2; peer review: 2 approved]. *F1000Research*, 5 (ELIXIR):2422 (doi.org/10.12688/f1000research.9656.2).

**Gabella C**, Durinx C and Appel R (2017) Funding knowledgebases: Towards a sustainable funding model for the UniProt use case. F1000Res , 6(ELIXIR), 2051 (doi.org/10.12688/f1000research.12989.2).

**Imker, HJ** (2018) 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance.Front. Res. Metr. Anal., 29 May 2018 (doi.org/10.3389/frma.2018.00018).

**Wilkinson MD**, Dumontier M, Aalbersberg IJ *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data, 3, 160018 (doi.org/10.1038/sdata.2016.18).

# Appendix: Indicators for Global Core Biodata Resource identification

| **1. Scientific focus and quality** | |
|---|---|
| 1a. Deposition database and/or Knowledgebase | *The distinction between deposition databases and knowledgebases is important contextual information. Deposition or archival databases receive and archive de novo data sets and well-structured metadata deposited by scientists. Knowledgebases are added-value databases which are built on archival data and add substantial value through expert curation, annotation of metadata, sophisticated data processing and/or data integration. It is possible to be both a deposition database and a knowledgebase.* |
| 1b. Scope statement | *Describes the scientific focus/domain covered by the resource, including factors such as nature of the primary data item (e.g., nucleic acid family, species occurrence, protein interaction, gene, biological sample, image, metabolite, dataset), experimental methods represented and characteristics that distinguish the resource from other biodata resources of related focus.* |
| 1c. Global dimension<br>1c i. Operation<br>1c ii. Users/contributors | *Describes the global characteristics of the biodata resource with respect to:*<br>*(i) the operation of the resource e.g., is it run via an international consortium? Is it funded by agencies in different countries?*<br>*(ii) the geographical distribution of both its users (including the basis on which this is known) and contributors e.g., does it take data submissions from a globally distributed range of nations?* |
| 1d. Staff effort | *Describes the staff effort in terms of Full Time Equivalents required to support and run the biodata resource. This might include curators (covering both providing support for submission adherence to metadata requirements and extraction/indexing of data from the primary scientific literature), bioinformaticians, technical staff and helpdesk staff.* |
| **2. Community** | |
| 2a. Data resource usage - quantitative data | *Describes usage of the biodata resource over time, including access via a web browser in terms of number of visits/hits per month, unique IP addresses (which are a proxy for unique visitors) and sessions/page views, as well as data downloads per month in terms of hits/requests/requesters (unique IP addresses), data transfer volume, and global distribution of users.* |
| 2b. Usage in research as measured through data resource citation in the scientific literature | *Describes citation in the scientific literature in terms of the frequency of citation of the resource name, and the resource-specific data items (for example via data accession numbers) over time.* |
| 2c. Citation of key publications describing the data resource | *Key publications that describe the biodata resource, for example articles in the Nucleic Acid Research "Database" issue, with the numbers of times they have been cited in the scientific literature.* |
| 2d. Connections to other data resources | *Describes how the biodata resource is embedded in the ecosystem of biological, life science, and biomedical data resources. Includes data exchanges between biodata resources and the direction and nature of those exchanges.* |

| **3. Quality of service** | |
|---|---|
| 3a. Identifier use | *Describes the system used to generate and implement persistent and unique identifiers, with identifier resolution services/mechanisms employed, if relevant.* |
| 3b. Data volume | *Describes in quantitative terms the cumulative total number of entries, records processed, depositions, assays, etc as relevant to the biodata resource, as well as and total data volume, in gigabytes etc., over time.* |
| 3c. Technical performance:<br>3c i. Uptime<br>3c ii. Response times of key web pages<br>3c iii. Back-up and disaster recovery | *Describes:*<br>*i. percentage availability per month for a sample of indicative web pages and/or search functions over the past 12 months*<br>*ii. response times for web pages that represent the typical web-based use case*<br>*iii. the strategy for ensuring adequate back-up/disaster recovery for the data housed within the data resource.* |
| 3d. Use of standards | *Describes community interoperability standards used for metadata and data housed in the biodata resource, and/or requested as part of a data submission protocol.* |
| 3e. Documentation<br>3e i. Data Curation<br>3e ii. Provenance and Evidence<br>3e iii. Quality Assurance | *Describes the provision of*<br>*i. documentation of the data curation process/deposition workflow*<br>*Ii. links to the primary scientific literature for provenance of and/or evidence for data statements or biological context*<br>*ii. versioning and/or evidence trails for modifications to datasets or data/metadata statements* |
| 3f. Data availability<br>3f i. Data sharing services<br>3f ii. Data sharing formats | *Describes the options in place for sharing data from the biodata resource in terms of*<br>*i. the services that facilitate sharing*<br>*ii. the formats in which the data is made available.* |
| 3g. User support<br>3g i. Helpdesk<br>3g ii. User feedback<br>3g iii. Training<br>3g iv. Communications<br>3g v. Language | *Describes support to users in terms of*<br>*i. helpdesk provision/access*<br>*ii. opportunities provided for user feedback*<br>*iii. training materials/opportunities*<br>*iv. notification methods employed for updates and announcements*<br>*v. language(s) in which the resource is made available* |
| **4. Funding, governance and legal infrastructure** | |
| 4a. Funding | *Describes funding secured by the biodata resource over the previous five years, current funding, and future committed funding.* |
| 4b. Scientific Advisory Board | *Describes the composition, function and activities of the Scientific Advisory Board, or other equivalent advisory body.* |
| 4c. Data preservation | *Describes the planning by the biodata resource for data preservation in the long term.* |
| 4d. Open Science | *Describes the licensing arrangements in place for the biodata resource that support open science.* |

| **4. Funding, governance and legal infrastructure (*continued*)** | |
|---|---|
| 4e. Privacy policy | *Describes the policy under which user personal data is collected and employed in the provision of the biodata resource services to the user, and how security around that data is managed.* |
| 4f. Ethics policy | *Describes any ethics policies adopted by the biodata resource in the context of relevant international standards and best practices. These might include, for example, policies regarding data use, data about users, data within the data resource, or of the research that generated the data.* |
| **5. Impact stories** | |
| 5a. Accelerating science | *Describes the ways in which the biodata resource has made specific contributions that have potentiated scientific progress or discovery, or facilitated scientific methodologies. This may include for example setting and promoting the use of metadata standards, actively promoting re-use of data or software, extending technical products.* |
| 5b. Counterfactual | *Describes the consequences for the biodata resource ecosystem, the scientific community and primary scientific research were the biodata resource to cease to exist and its data, services, and functions not be replaced.* |