Global Core Biodata Resource Selection: Data Required for Full Application

Purpose of this document:

This document is one of two supplementary data files that accompany the Global Biodata Coalition's article: "Global Core Biodata Resources: Concept and Selection Process" (https://doi.org/10.5281/zenodo.5845116).

Its purpose is to detail the information that will be required from those biodata resources invited to submit a Full Application for selection as a Global Core Biodata Resource.

This document is not an application form: applications will be made using an online system.
Details are available from the Global Biodata Coalition website:
https://globalbiodata.org/scientific-activities/gcbr-selection

Data required, for each of the five indicator categories:

1. Scientific focus and quality

1a. Deposition database and/or Knowledge base

Please indicate whether or not the biodata resource is a:

Deposition database (accepting deposition of in-scope experimental data from the wider international community)	YES/NO
Knowledge base (with added value, e.g. by expert curation)	YES/NO

1b. Scope statement

Please describe 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), species included, experimental methods represented and characteristics that distinguish the resource from other biodata resources of related focus. If relevant include the introductory documentation or "about" page url.

[Maximum 250 words]

1. Scientific focus and quality (*continued*)

1c. Global dimension

Please describe the global characteristics of the biodata resource with respect to the operation of the resource e.g. is it run via an international consortium? Is it funded by agencies in different countries?

[Maximum 250 words]

1d. Staff effort

Please indicate the number of Full Time Equivalents working on the biodata resource for each of the years stated. As this may fluctuate throughout the year, please state this in terms of a nominal mean for that year.

	2019	**2020**	**2021**
Total Staff			

2. Community

2a. Biodata resource usage - quantitative data

This section concerns the usage of the data resource over the last 3 years.

Please complete all Tables that apply for the data resource - it is expected that most data resources will supply figures for Table 2a.n (web analytics) or Table 2a.n (log analytics) but not both.

2a.i: Biodata Resource Usage
Access via a web browser determined using web analytics

Average monthly web traffic in…	**2019**	**2020**	**2021**
Visits per month (sessions)			
Unique IP addresses per month			
Page views per month			
Please state the technology used to determine these counts (e.g. Google Analytics, Open Web Analytics, Matomo, AWstats, GoAccess):			

2. Community (*continued*)

2a.ii: Biodata Resource Usage

Access via a web browser determined using log analytics

Average monthly web traffic in...	**2019**	**2020**	**2021**
Unique IP addresses per month			
Hits per month			
Sessions and/or pages per month (please specify)			
Please state the technology used to determine these counts (e.g. Matomo, Splunk):			

2a.iii: Biodata resource usage

Geographic distribution of users - as indicated by Unique IP addresses

For a representative recent 12 month period:

2a.iii.1 With respect to the web browser data presented in item 2a.i /2a.ii above, what percentage of the unique IP addresses accessing the biodata resource originated within the host country/countries, and what percentage originated in countries outside the host country/countries? If there are unique IP addresses where it is not possible to make that identification, please include that % in the third column.

Unique IP addresses accessing the biodata resource		
% from within host country/countries	**% from outside host country/countries**	**% for which it is not possible to identify country/countries**

2a.iii.2 With respect to the web browser data presented in item 2a.i /2a.ii above, please list the 10 countries from which the highest number of unique IP addresses access the biodata resource, and their percentage share of the total number of unique IP addresses.

Country	**% Usage**
1. 2. . . 9 10	1. 2. . . 9 10

2. Community (*continued*)

2a.iii: Biodata resource usage (*continued*)

Geographic distribution of users - as indicated by Unique IP addresses

2a.iii.3 Including the top 10 countries listed above in 2a.iii.2, how many countries (by unique IP address) are represented in the user base? Please describe how the user base is distributed between the countries.

[Maximum 100 words]

Please describe the basis on which the 2a.iii statements for this geographic distribution indicator are made including the technology/methodology used e.g. Google Analytics, or resource-specific customised software.

[Maximum 100 words]

Please state the 12 month period used to answer questions in this section 2a.iii:

MM/YYYY - MM/YYYY

2a.iv: Biodata Resource Usage

Data downloads

Average monthly downloads in ...	**2019**	**2020**	**2021**
Hits / Requests per month			
Unique IP addresses / Hosts per month			
Data transfer per month (GB)			
Please state technology used to provide and log these downloads (e.g. FTP, APIs):			

2. Community (*continued*)

2b. Usage in research as measured through biodata resource citation in the scientific literature

Table 2b: Biodata Resource Citation

Annual totals:	**2019**	**2020**	**2021**
Resource name mentioned in PubMed/EuropePMC/Google Scholar (citation of biodata resource by name)			
Resource-specific accession numbers or unique identifiers mentioned in PubMed/EuropePMC/Google Scholar (citation of specific entities from the biodata resource)			
Please indicate which literature database you used to generate this data. *If none of PubMed, EuropePMC or Google Scholar is a relevant literature database for this application, please make your assessment of these citation data points using an alternative source, and explain your choice/methodology here:*			

2c. Citation of key publications describing the biodata resource

Please list key articles that describe the biodata resource. Choose up to three articles, and include the url for the publication as well and the number of times they have been cited, with the source of the citation count (for example as stated in PubMed/EuropePMC).

Table 2c: Biodata Resource Key Article Citation

Article	**Year of publication**	**Link**	**Citation count (method)**
		<link>	
		<link>	
		<link>	

2. Community (*continued*)

2d. Connections to other data resources

This section describes the position the resource occupies in the ecosystem of biodata resources, as defined in terms of inward and outward data exchange relationships and/or dependencies with other data resources.

As part of your response, describe links and dependencies (where known) between this and other biodata resources for the service they provide, an estimate of the number, and the direction and nature of the dependence.

Examples might include manually curated hyperlinks that relate entities across data resources, and/or processes that exchange, update or rationalise specific data types or metadata between corresponding entities in the biodata resources, via automated updates/comparisons.

Please include a url link to pre-existing documentation on this topic, if relevant.

[Maximum 200 words]

3. Quality of service

3a. Identifier use

Does the biodata resource provide persistent and unique identifiers for the entities it stores? Please describe the format of the identifiers used.

Does the biodata resource participate in any identifier resolution mechanisms/services?

[Maximum 100 words]

3b. Data volume

Please state the cumulative total number of entries, records processed, depositions, assays, as appropriate, and data volume in gigabytes or other standard units. for each year indicated.

Table 3b: Data volume

Cumulative Data in…	**2019**	**2020**	**2021**
Total number of entries/records/deposition/assays...			
Please state the corresponding entity chosen for the counts given in the row above:			
Total size in GB			

3. Quality of service (*continued*)

3c. Technical performance:

3c i. Uptime

Please state this in terms of percentage availability per month for a sample of indicative web pages and/or search functions over the past 12 months (e.g. search results, homepage, data record pages).

[Maximum 100 words]

3c ii. Response times of key web pages

Please give response times for web pages that represent the typical web-base use case.

[Maximum 100 words]

3c iii. Back-up and disaster recovery

Please describe the strategy for ensuring adequate back-up for the data housed within the biodata resource. Is a disaster recovery plan in place and adequately disseminated to multiple relevant staff? Is a Failover system in place?

[Maximum 100 words]

3d. Use of standards

Please list community interoperability standards, ontologies and/or controlled vocabularies used for metadata and data housed in the biodata resource and/or requested as part of a data submission protocol.

Examples might include MIAPPE, INSDC features, GA4GH standards, Darwin Core, JATS, OBO foundry ontologies, Taxon, Geographic information - Metadata, EML - Ecological Metadata Language, or similar.

Provide a link to documentation describing the implementation of these standards in the biodata resource.

[Maximum 200 words]

3. Quality of service (*continued*)

3e. Documentation

3e i. Data Curation
Does the resource provide documentation of the data curation process/deposition workflow?

Please describe, with a url link to public documentation, where relevant.
[Maximum 100 words]

3e ii. Provenance and Evidence
Does the resource link to the primary scientific literature for provenance of/evidence for data statements or biological context?

Please provide examples, with a url link to public documentation, where relevant.
[Maximum 100 words]

3e iii. Quality Assurance
Does the resource provide versioning and/or evidence trails for modifications to datasets or data/metadata statements?

Please provide examples, with a url link to public documentation, where relevant.
[Maximum 100 words]

3f. Data availability

3f i. Data sharing services
Please list services through which data is shared (e.g. APIs, FTP/lftp/sftp, TripleStore, Globus, Aspera).
[Maximum 100 words]

3f ii. Data sharing formats
Please list formats in which data is made available for download (e.g. plain text (txt), FASTA, XML, HTML, SMBL, Dublin Core, xlsx, csv, tsv, JSON, docx, pdf, image formats, RDF, Hierarchical Data Format, MP3, MP4).
[Maximum 100 words]

3. Quality of service (*continued*)

3g. User support

3g i. Helpdesk

Does the resource operate a Helpdesk via forms, email or social media Q&A? If so, how can the users find out how to lodge a query with the Helpdesk?

[Maximum 100 words]

3g ii. User feedback

How does the resource seek and incorporate user input into service design decisions?

[Maximum 100 words]

3g iii. Training

By what means does the resource undertake and publicise training activities or provide training materials?

[Maximum 100 words]

3g iv. Communications

By what means are updates and announcements communicated with the biodata resource users?

[Maximum 100 words]

3g v. Language

Please list the language(s) in which the biodata resource web interface is available.

[Maximum 100 words]

4. Funding, governance and legal infrastructure

4a. Funding

Please list the funding secured for the resource by the host institution or other entities over the past five years, including current as well as any future funding commitments.

[Maximum 300 words]

4b. Scientific Advisory Board

Please describe the composition, function and activities of the Scientific Advisory Board, or other equivalent body. Please include a url for the SAB, if available, or brief description, if not.

[Maximum 200 words]

4c. Data preservation

Please describe the long-term data preservation plan adopted by the biodata resource. In particular, what would happen to the data held in the resource if the resource were to lose funding or otherwise cease operating?

Please stipulate whether the data you refer to is user-contributed datasets (for deposition databases) or data, metadata and data/metadata relationships contributed either by depositors or by in-house curators (for deposition databases and knowledge bases).

Please provide url for public-facing documentation of the long term data preservation plan, where relevant.

[Maximum 200 words]

4d. Open Science

Please describe the licensing arrangements that support Open Science, including which particular open licence or Terms of Use apply, and the url for the documentation or public notice of that status.

[Maximum 200 words]

4e. Privacy policy

Does the biodata resource have a privacy policy/notice where the personal data collected and employed in the provision of the resource services to the user is described? Are the measures put in place to manage the security issues around that personal data documented in that policy/notice?

Please include the url for relevant privacy and security statement(s).

[Maximum 100 words]

4. Funding, governance and legal infrastructure (*continued*)

4f. Ethics policy

Does the resource have an ethics policy that complies with relevant data standards and best practices?

An ethics policy might, for example, include requirements regarding consent for sample collection for deposited human DNA sequence data, or responsible use of deposited geolocation data for threatened species. If relevant, does the resource require ethical approval for the research that generated the data?

Please include the relevant url for the statement of that policy.

[Maximum 100 words]

5. Impact stories

5a. Accelerating science

How does the resource accelerate science? Has the biodata resource made a specific contribution that has potentiated scientific progress or discovery, or facilitated scientific methodologies? For example, does the resource set standards; promote reuse of data or software; promote research efficiencies; extend technical products in other areas?

[Maximum 500 words]

5b. Counterfactual

If the resource were to disappear from the biodata ecosystem and its data, services, and functions not be replaced, what would be the impact on ...

the scientific community?

other biodata resources?

primary scientific research?

Is the biodata resource globally unique?

What would no longer be possible in the absence of this data resource?

[Maximum 500 words]