The Global Biodata Coalition

Chuck Cook
GBC Program Manager

GLOBAL BIODATA COALITION

AgBioData.org webinar 3 Feb 2021
Science Infrastructures
Global Life Science Data Resources Infrastructure

- Crucial **distributed** global infrastructure for biology
- Repositories for research data, added-value curation of those data and analytical platforms
- Essential—use is ubiquitous in research in biology, life sciences, biotech, pharma, and agriculture, and increasingly in clinical settings
- Biological, life sciences, and biomedical research depend on a healthy ecosystem of data resources
- Reliant on a small number of specific funding programmes from an even smaller number of funders
Biodata resources are interconnected

Data exchange between EMBL-EBI resources and external data resources

https://doi.org/10.1093/nar/gkt21033
Scale of the infrastructure

• $>$3000 data resources have been created
• $>$ 100 new resources per year
• Globally, public spending to support resources is perhaps US$500 million per year
• This is perhaps 0.1% of the total global spending on life sciences research
Biodata resources are vulnerable

- Exponential data generation increases demand
- Open access policies increase demand
- New technologies require new data resources
- As infrastructure, biodata resources need **long-term** funding
  - Yet funding is fragmentary, fragile, and haphazard
- Little international coordination among funders of resources
- Curtailment of funding risks loss of important data resources, hindering research
- Risk of retreat behind paywalls
  - Resources would become inaccessible to many researchers and scientists
Funders recognize the challenges to sustaining the infrastructure

Nature 2015 (https://doi.org/10.1038/s41559-016-0210)
Local support: global usage

UNITING FUNDERS
The medical research community has too little money to start new data resources or to support the growth of more mature databases and services. Moreover, current funding schemes do little to foster the development of best practices; for example, each data resource is usually reviewed in isolation.

Changes to funding practices need to extend across both agency and international borders. Data generation and maintenance are typically funded nationally, but the data are used internationally. As a result, we need to develop more equitable funding models. The first step is for funding agencies to communicate more effectively about data science problems and to seek collaborative solutions. Working from the bottom up, scientists have been doing this for a long time.
Global Biodata Coalition

Towards Coordinated International Support of Core Data Resources for the Life Sciences


doi: https://doi.org/10.1101/110825

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract  Full Text  Info/History  Metrics  Preview PDF

A global coalition to sustain core data

As members of an international working group to support the rapidly growing core-data resources in the life sciences, we aim to create a sustainable and accessible data infrastructure that will benefit scientists worldwide.

Although researchers have relied on international resources such as the Protein Data Bank and Flybase for decades, the current system is unsustainable because it is largely funded by short-term grants (P.E. Bourne et al., Nature 527, S16–S17, 2015). A global coalition of data resources would provide much-needed governance structure, active service management and community-driven scientific development, which together are currently well beyond the scope of an individual investigator's typical research programme.

Nature 2017 543: 379

Science funders globally should support these data resources on the basis of their value to the research community. The coalition would define indicators to establish the core-data resources that are eligible for international support, develop models for free global access and help to assess the fraction of total research funding needed. It would also compile a set of metrics to estimate the impact, costs and benefits of each resource, including the consequences of curtailing support.

The set of data resources designated as 'core' for the life sciences would reflect a dynamic, reliable and managed portfolio that could adapt to changing scientific needs. The Global Life Sciences Data Resources Coalition will follow the lead of other international coalitions, such as those in health and physical sciences, in setting priorities and evaluating effort.

(For details, see W. Anderson et al. Preprint at bioRxiv http://doi.org/b2g4; 2017).

Warwick P. Anderson* Human Frontier Science Program, Strasbourg, France.
Managing Biodata Resources as a global infrastructure

Goal
- Ensure sustainability of biodata infrastructure and continued open access to data

Scope
- The GBC will focus initially on only those biodata resources that are available with unrestricted access

Benefits
- Prevent data loss
- Coordinate a fragmented ecosystem
- Reduce redundancy
- Strengthen international coordination
- Provide opportunities for additional funders to support biodata resources
- Reduce dependency on small number of funders
Global coordination and knowledge exchange

- Monitor and share funder/national data resource-related strategies and policies and expertise
- Improve coordination among funders to reduce duplication of funding
- Identify new data resource funding opportunities for interested funders
- Link funder/national research data strategies across borders
- Share international expertise and best practice in managing data
- Leverage national expertise and capabilities to build expertise worldwide
- Promote national participation in development and adoption of standards for data management
GBC Vision

Before establishing the GBC

After establishing the GBC
Current status

• (Virtual) Secretariat established
• Human Frontier Science Program (HFSP) as initial host
• Supported by research funders
  • Lead by Board of Funders
• Ongoing outreach to stakeholders
• GBC programmatic activities
  • Coordination and knowledge exchange
  • Scientific program
Stakeholder engagement

- The Global Biodata Coalition is supported and managed by research funders
- Stakeholders in the global biodata infrastructure include
  - Data producers
  - Data users
  - Data resource managers and staff
  - Other research funders
- Outreach activities
  - Meeting to share information and approaches among funders and stakeholders
  - Publication of scientific program activities
  - Presentations at scientific meetings
Scientific program: basis

• How can the biodata resource infrastructure be strategically managed for the benefit of all?
• What is the infrastructure?
• How do we define the components?
• What is the value?
Scientific program overview

• Describing the biodata infrastructure
  • What is total global spending on life sciences research and biodata resources?
  • How many data resources are there?
  • What are funders’ strategies for supporting data resources?
• Global Core Data Resources
Global Core Data Resources

- Pioneered by ELIXIR in Europe
- Data resources that are of fundamental importance to the broad life science community and the long-term preservation of biological data
- Provide complete collections of generic value to life science, and show high levels of usage, scientific quality and service
Why focus on Core Data Resources

- Core Data Resources are fundamental for the entire global infrastructure
  - Analogous to keystone species in an ecosystem
  - Focus on Core Data Resources will also help protect the entire global infrastructure
- Impractical to focus on all data resources
Established approach to selecting Core Data Resources

Identifying ELIXIR Core Data Resources [version 2; peer review: 2 approved]

Christine Durinx1, Jo McEntyre2, Ron Appel1, Rolf Apweiler2, Mary Barlow2, Niklas Blomberg3, Chuck Cook2, Elisabeth Gasteiger4, Jee-Hyub Kim2, Rodrigo Lopez2, Nicole Redaschi4, Heinz Stockinger1, Daniel Teixeira1, Alfonso Valencia5

2017: https://f1000research.com/articles/5-2422

Databases and ontologies

The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences

Rachel Drysdale1, Charles E. Cook2, Robert Petryszak2, Vivienne Baillie-Gerritsen3, Mary Barlow2, Elisabeth Gasteiger3, Franziska Gruhl4, Jürgen Haas6, Jerry Lanfear1, Rodrigo Lopez2, Nicole Redaschi3, Heinz Stockinger4, Daniel Teixeira4,6, Aravind Venkatesan2, Elixir Core Data Resource Forum1, Niklas Blomberg1, Christine Durinx4,* and Johanna McEntyre2

2020: https://doi.org/10.1093/bioinformatics/btz959
Global Core Data Resources selection process

1. Expressions of Interest
   a. Eligibility criteria
      i. In development
   b. Very short EoI form
   c. Shortlisting

2. Selection from shortlist
   a. Longer application
   b. Independent review panel
   c. Multiple indicators
Core Data Resources – a data infrastructure

- **Scientific focus** and quality of science
  - Curation level, benchmarking
- **Community** served by the resource
  - Web statistics
- **Quality of service**
  - Uptime, user support and training
- **Legal and funding infrastructure**
  - Institutional support, use policy
- **Impact** and translational stories
  - Foundational role

**Indicators** (Mandatory and optional)

Support

Current funding

In-kind support
GLOBAL BIODATA COALITION

The Global Biodata Coalition (GBC) is a forum for research funders to better coordinate and share approaches for the efficient management and growth of biodata resources worldwide. The GBC aims to stabilize and ensure sustainable financial support for the global biodata infrastructure and in particular to identify for prioritized long-term support a set of Global Core Data Resources that are crucial for sustaining the broader biodata infrastructure.