

The Global Biodata Coalition

Chuck Cook
GBC Program Manager



GLOBAL
BIODATA
COALITION

AgBioData.org webinar 3 Feb 2021

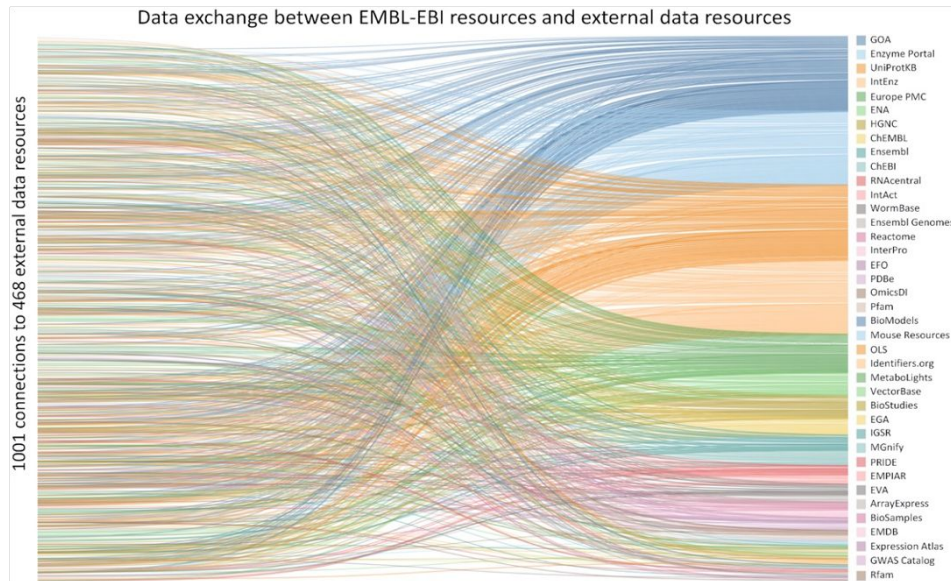
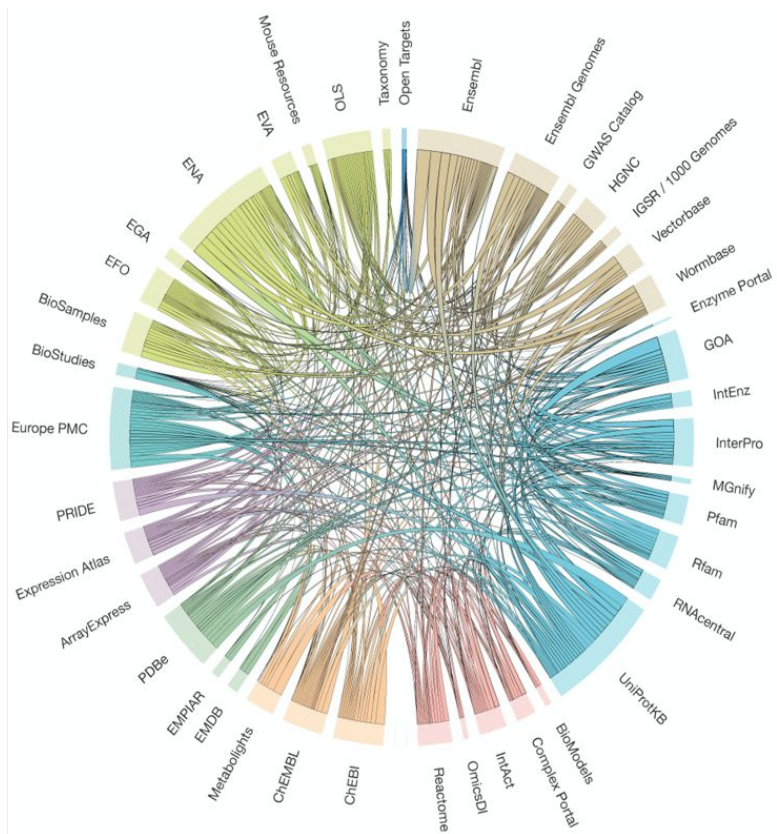
Science Infrastructures



Global Life Science Data Resources Infrastructure

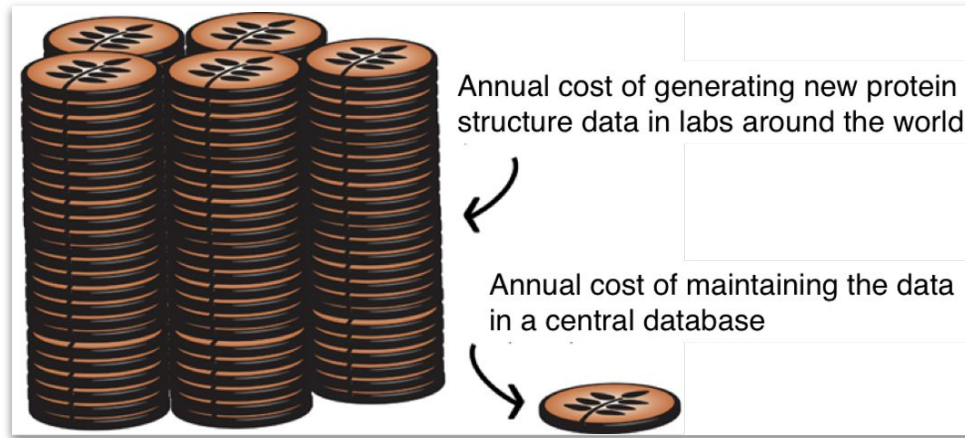
- Crucial **distributed** global infrastructure for biology
- Repositories for research data, added-value curation of those data and analytical platforms
- Essential—use is ubiquitous in research in biology, life sciences, biotech, pharma, and agriculture, and increasingly in clinical settings
- Biological, 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



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

OUTLOOK BIG DATA IN BIOMEDICINE

PERSPECTIVE

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say Philip E. Bourne, Jon R. Lorsch and Eric D. Green.



Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-adjusted terms in many countries (including the United States), and data are being generated at unprecedented rates, so the research community must find more efficient models for storing, organizing and accessing biomedical data. Simply putting more and more money into the current systems is unlikely to work in the long term. To better understand this situation, we are examining the current and projected costs of managing biomedical data at the US National Institutes of Health (NIH). Our initial analyses indicate that even if we rely on the National Center for Biotechnology Information, which is a special case, the 20 largest NIH-funded data resources have a collective annual budget of US\$1.0 billion. And this figure represents just the tip of the iceberg for future needs.

UNDERSTANDING USAGE

Today biomedical data resources typically treat all items in their collections equally. This does not always make sense, given that the usage patterns of the data vary. But how do we decide which data get more attention? As larger and larger data sets are generated more easily, and the cost of maintaining and annotating these data continues to rise, this question is becoming increasingly important.

Answering it requires a better understanding of how research data are used. This has rarely been thoroughly explored. Historically, funders have been interested primarily in knowing how the data resources that they support are used and by whom. They tended not to look closely at the details of how and why individual items and types of data within a collection are used.

Analysis of these details can be revealing. Preliminary studies suggest that typically a small subset of the data is used frequently, whereas most of the data are rarely accessed. However, the exact subset of data that is used heavily may change over time, and most of the data access may be performed after the data are downloaded, so this is not

recorded. All of this means that absolute numbers are hard to interpret. These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

FAIR AND EFFICIENT

When we have a better understanding of data usage, we can develop business models that consider supply and demand, and develop sustainable practices. In addition, finding economies of scale and harnessing market forces will be essential.

For a typical biomedical data resource, the cost of simply keeping the data is only a small fraction of the total cost of data management. The remainder is largely the cost needed to support the finding, accessing, interpreting and reusing (the FAIR principles; see go.nature.com/1xvdy) of the data—a cost that is widely underappreciated.

Is the FAIR fraction of the cost justified? Are services from different data resources redundant? Are resources subject to 'feature creep'—the addition of costly 'bells and whistles' that are of limited value? Do our funding mechanisms contribute to these problems? And most importantly, in the way we currently maintain biomedical data optimal for the science that needs to be done both today and in the future?

Current practices typically use many disparate sources of data to conduct a study. These data are located in a variety of repositories, often with different modes of access. The lack of centralization and commonality may hinder their ease of use and reduce productivity. We need a better understanding of usage patterns across multiple data resources to use as a basis for organizing such resources to preserve valuable expertise and curation, and for improving how the data are found, accessed, integrated and reused.

The nature of curation and the quality assurance for biomedical data must also change. Complete and accurate automated or semi-automated extraction of metadata and annotation is needed to provide metadata and annotation. We should consider crowdsourcing curation, with appropriate validation and incentives. Additionally, the role of professional curators must be better appreciated by data users by the institutions where the curators work, and by the funder.

THE RESEARCH COMMUNITY MUST FIND MORE EFFICIENT MODELS FOR STORING, ORGANIZING AND ACCESSING BIOMEDICAL DATA.

BIG DATA IN BIOMEDICINE OUTLOOK

In the longer term, we need models that are better aligned with the research life cycle. There is an unnecessary cost in a researcher interpreting data and putting that interpretation into a research paper, only to have a bioeditor extract that information from the paper and associate it back with the data. We need tools and rewards that incentivize researchers to submit their data to data resources in ways that maximize both quality and ease of access.

BUSINESS MODELS

One business model worth exploring is the 'freemium model'. Here, the primary data are available free of charge, but services that add value to these data have an associated charge that generates funds that are used to maintain the primary data. This approach is used in other disciplines, notably chemistry. But there are two critical questions: Should for-profit institutions be charged the same as non-profits? And who should own the intellectual property associated with value-added content?

Another potential business model is the 'subscription model', which is used to access the genetic and molecular databases that are provided by the Arabidopsis Information Resource (TAIR), for example. This option delivers support for a data resource from its active users, but it restricts access, which may be problematic for public-access data policies.

Taking the business-model idea further, what happens if data resources are merged, acquired or go out of business? Would existing resources be more useful and cost-effective if they were merged in some way? Should some services be dropped owing to lack of demand to make way for new services? Would reducing funding for particular data resources over time promote increased efficiency? To answer such questions, we would benefit from advice and help from the private sector and from other scientific communities.

COMMON GROUND

Cloud computing creates an element of data virtualization, takes computing to the data, and may help to solve some of the problems facing biomedical big data. At the NIH, we propose to exploit these opportunities by creating a 'commons' as one possible sustainable model.

Physically speaking, the commons will be collections of public and private resources (including cloud resources) for storing data and computing with those data. To be commons-compliant, such resources must abide by two simple rules. First, each research object in the commons—for example, data, software, narratives or papers—must be uniquely identified, sharable (taking into account privacy issues), and responsibly to its source by using a common identifier. Second, each research object must be defined by a minimal amount of metadata, as defined by the community.

The NIH Big Data to Knowledge (BD2K) programme (bd2k.nih.gov) aims to bring about the creation of the commons. The 12 new BD2K centres are encouraged to share research objects within the commons, and a BD2K consortium is prototyping an index that makes it easy to find commons content.

We are also studying the notion of computing credits, in which a grant recipient is given credits instead of funding to pay for computational time. A principal investigator would be able to spend these credits at any commons-compliant resource. Researchers whose work involves extensive computation on small amounts of data may spend their credits at different commons-compliant resources to investigators who do minimal computing on large amounts of data.

This model is very different from the situation today. It shifts the initial burden of hardware and software maintenance from awardees and their institutions to third parties, notably cloud service providers. The funding model also has the effect of paying only for services used, and aims to create competition in the marketplace, so this approach could result in more data science per dollar.

If the pilot studies at the NIH are successful, it will be important



Research organizations such as the Broad Institute are rapidly evolving their practices for storing and accessing biomedical big data.

to consider the longer-term implications of a commons model. One outcome is that data and software usage will be tracked both during an initial period and after it has expired. Such tracking will yield

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.

Sustaining the biomedical big-data ecosystem is the responsibility of all stakeholders, and will require coordinated efforts among data generators, data maintainers, data users, funders, publishers and others in the private sector. The NIH BD2K programme, in collaboration

Philip E. Bourne is associate director for data science at the US National Institutes of Health. He was previously associate vice-chancellor for innovation and industry affairs at the Office of Research Affairs at the University of California, San Diego. Jon R. Lorsch is director of the National Institute of General Medical Sciences. He was previously professor of biophysics and biophysical chemistry at Johns Hopkins University in Baltimore, Maryland. Eric D. Green is director of the National Human Genome Research Institute. He was previously its scientific director, chief of its genome technology branch and director of the NIH Intramural Sequencing Center. e-mail: phil.bourne@nih.gov

5 NOVEMBER 2015 | VOL 527 | NATURE | 517

Nature 2015 (<https://doi.org/10.1038/527516a>)



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

W. Anderson, R. Apweiler, A. Bateman, G.A. Bauer, H. Berman, J.A. Blake, N. Blomberg, S.K. Burley, G. Cochrane, V. Di Francesco, T. Donohue, C. Durinx, A. Game, E.D. Green, T. Gojobori, P. Goodhand, A. Hamosh, H. Hermjakob, M. Kanehisa, R. Kiley, J. McEntyre, R. McKibbin, S. Miyano, B. Pauly, N. Perrimon, M.A. Ragan, G. Richards, Y-Y. Teo, M. Westerfield, E. Westhof, P.F. Lasko

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: 179

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

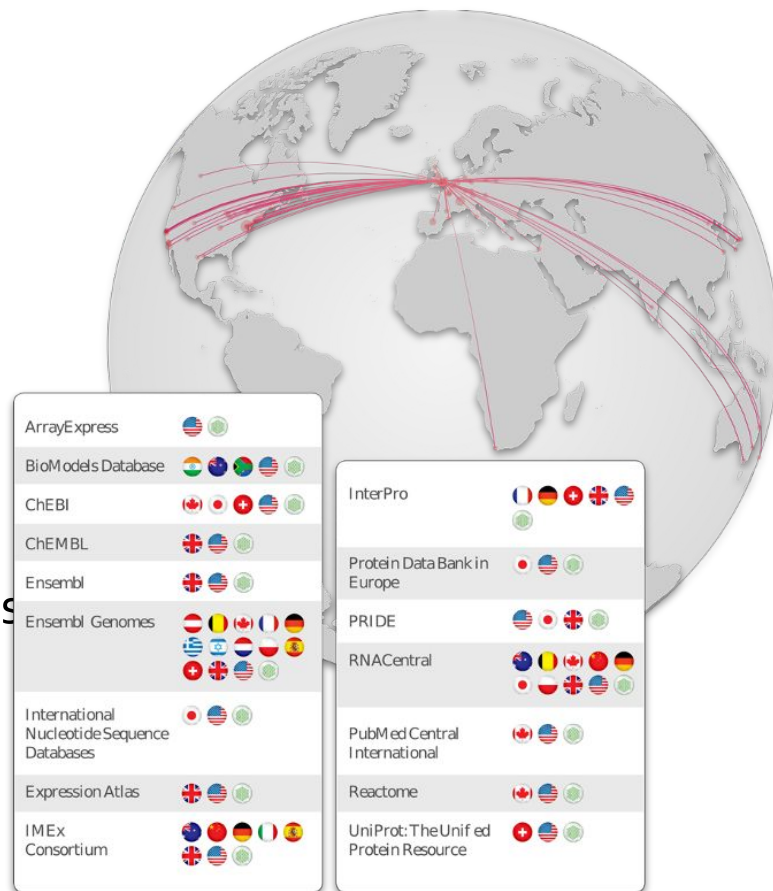
- How can the biodata resource infrastructure be strategically managed for the benefit of all?
- What is the infrastructure?
- How do we define the components?
- What is the value?

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

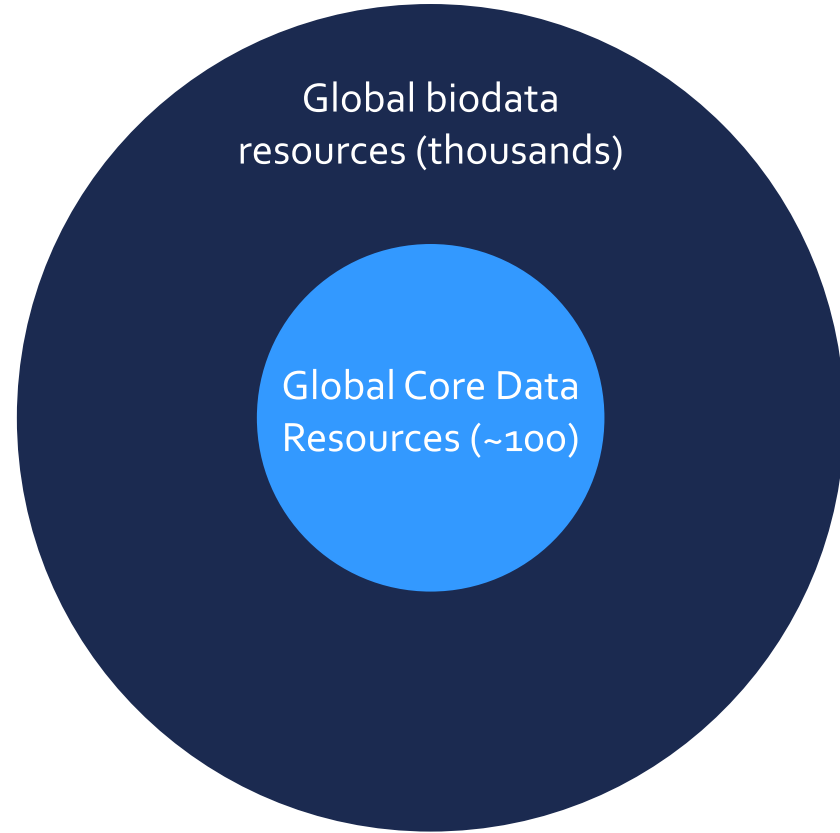
- Pioneered by ELIXIR in Europe
- Data resources that are of fundamental importance to the broad life science community and the long-term preservation of biological data
- Provide complete collections of generic value to life science, and show high levels of usage, scientific quality and service



<https://doi.org/10.1093/nar/gkx1154>

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

METHOD ARTICLE

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

Christine Durinx ¹, 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 Haas⁵, Jerry Lanfear¹, Rodrigo Lopez², Nicole Redaschi³, Heinz Stockinger⁴, Daniel Teixeira^{4,6}, Aravind Venkatesan², Elixir Core Data Resource Forum¹, Niklas Blomberg ¹, Christine Durinx^{4,*} and Johanna McEntyre²

2020: <https://doi.org/10.1093/bioinformatics/btz959>

ELIXIR Core Data Resource list

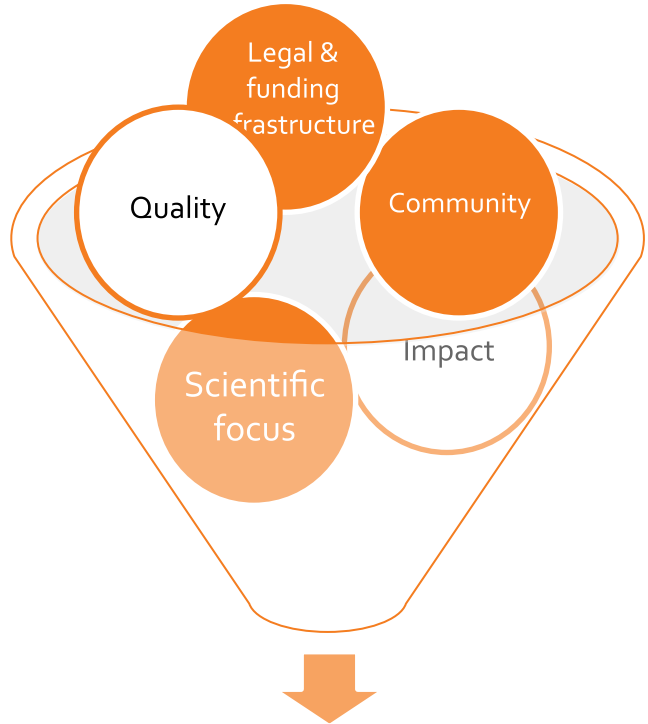
Core Data Resource	Data type
ArrayExpress	Functional Genomics Data from high-throughput functional genomics experiments.
BRENDA	Database of enzyme and enzyme-ligand information, across all taxonomic groups, manually extracted from primary literature and extended by text mining procedures, integration of external data and prediction algorithms.
CATH	A hierarchical domain classification of protein structures in the Protein Data Bank.
ChEBI	Dictionary of molecular entities focused on 'small' chemical compounds.
ChEMBL	Database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties and abstracted bioactivities.
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.
ENA	Nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
Ensembl	Genome browser for vertebrate genomes that supports research in genomics, evolution, sequence variation and transcriptomics.
Ensembl Genomes	Comparative analysis, data integration and visualization for vertebrate genomes.



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



Indicators (Mandatory and optional)

- **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

Durinx C, McEntyre J, Appel R *et al.* Identifying ELIXIR Core Data Resources *F1000Research* 2016, **5**(ELIXIR):2422



Support

Current funding



Biotechnology and
Biological Sciences
Research Council



Agency for
Science, Technology
and Research

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.





GLOBAL
BIODATA
COALITION