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

Rachel Drysdale[1*], Charles E. Cook[2], Robert Petryszak[2], Vivienne Baillie-Gerritsen[3], Mary Barlow[2], Elisabeth Gasteiger[3], Franziska Gruhl[4], Jürgen Haas[5], Jerry Lanfear[1], Rodrigo Lopez[2], Nicole Redaschi[3], Heinz Stockinger[4], Daniel Teixeira[4], Aravind Venkatesan[2], ELIXIR Core Data Resource Forum[6], Niklas Blomberg[1], Christine Durinx[4], Johanna McEntyre[2].

1: ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
2: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
3: SIB Swiss Institute of Bioinformatics, CMU, rue Michel Servet 1, 1211 Geneva, Switzerland
4: SIB Swiss Institute of Bioinformatics, Quartier Sorge - Bâtiment Amphipôle, 1015 Lausanne, Switzerland
5: SIB Swiss Institute of Bioinformatics & Biozentrum, University of Basel, Klingelbergstr 50/70, 4056 Basel, Switzerland
6: The ELIXIR Core Data Resource Forum members: Alex Bateman[2], Alan Bridge[3], Guy Cochrane[2], Rob Finn[2], Frank Oliver Glöckner[7], Marc Hanauer[8], Thomas Keane[2], Andrew Leach[2], Luana Licata[9], Per Oksvold[10], Sandra Orchard[2], Christine Orengo[11], Helen Parkinson[2], Bengt Persson[12], Pablo Porras[2], Jordi Rambla[13], Ana Rath[8], Charlotte Rodwell[8], Ugis Sarkans[2], Dietmar Schomburg[14], Ian Sillitoe[11], Dylan Spalding[2], Mathias Uhlén[10], Sameer Velankar[2], Juan Antonio Vizcaíno[2], Kalle von Feilitzen[10], Christian von Mering[15], Andrew Yates[2].
7: Alfred Wegener Institut, Helmholtz Zentrum für Polar- und Meeresforschung and Jacobs University Bremen, Am Handelshafen 12, 27570 Bremerhaven, Germany
8: INSERM, US14 - Orphanet, Plateforme Maladies Rares, 96 rue Didot, 75014 Paris, France
9: University of Rome Tor Vergata ,Department of Biology, Via della Ricerca scientifica, 00133, Rome, Italy
10: KTH Royal Institute of Technology, CBH, Department of Protein Science, Science for Life Laboratory, Box 1031, SE-171 21 Solna, Sweden
11: University College London, Gower Street, London, WC1E 6BT, UK
12: Elixir-Sweden, Science for Life Laboratory, Dept of Cell and Molecular Biology, Uppsala University, BMC, Box 596, S-751 24 Uppsala, Sweden
13: Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology and UPF, Dr. Aiguader 88, Barcelona 08003, Spain
14: Technische Universität Carolo Wilhelmina zu Braunschweig, Braunschweig, Germany
15: University of Zurich and Swiss Institute of Bioinformatics, Winterthurerstrasse 190, 8057 Zurich, Switzerland

*Corresponding author: rachel.drysdale@elixir-europe.org

Authorship Note: The named authors contributed to the content and writing of this article. The listed members of the ELIXIR Core Data Resource Forum are the current representatives of the Core Data Resources working with ELIXIR in this context, who acted together as consultants in the framing of this article.

Abstract

Motivation

Life science research in academia, industry, agriculture, and the health sector is critically dependent on free and open data resources. ELIXIR, the European Research Infrastructure for life sciences data, has undertaken the task of identifying the set of Core Data Resources within Europe that are of most fundamental importance to the life science community for the long-term preservation of biological data. Having defined the Core Data Resources, we explored characteristics of the usage, impact and sustainability of the set as a whole to assess the value and importance of these resources as an infrastructure, to understand sustainability to the infrastructure, and to demonstrate a model for assessing Core Data Resources worldwide.

Results

The nineteen resources designated as Core Data Resources by ELIXIR together form a data infrastructure in Europe that is a subset of the wider worldwide open life sciences data infrastructure. These resources are of crucial importance to research throughout the world. We show that, from 2013 to 2017, data managed by the Core Data Resources tripled and usage doubled while staff numbers increased by only a sixth. Additionally, support for the Core Data Resources is precarious, with all resources together having assured funding for less than a third of current staff after only three years.

Our findings demonstrate the importance of the ELIXIR Core Data Resources as repositories for research data and the knowledge generated from those data,while also demonstrating the precarious nature of the funding environment for this infrastructure. The ELIXIR Core Data Resources are part of a larger worldwide life sciences data resources ecosystem. Both within Europe and as part of the Global Biodata Coalition, ELIXIR will work for longer-term support for the worldwide life sciences data resource infrastructure and for the subset of that infrastructure that is the ELIXIR Core Data Resources.

Introduction

Life science data resources have been used extensively in academia and industry for well over two decades, and are increasingly used in clinical settings. These resources are critical for ensuring the reproducibility and integrity of the entire life sciences research enterprise (Bourne *et al.,* 2015*)*. Despite their importance, many are supported in whole or in part by short-term grants and there is little coordination of funding across these resources (Berman, 2008; Gabella *et al.*, 2017; https://www.biorxiv.org/content/10.1101/110825v3).

ELIXIR (www.elixir-europe.org) brings together life sciences resources from across Europe. More than 20 countries contribute to ELIXIR's infrastructure with scientific tools and databases, as well as compute infrastructure, standards for interoperability, and training. Here, we focus on existing, well-established data resources. One of ELIXIR's goals is to support the most valuable, used and useful resources, i.e., those with a very high scientific impact. To fulfill this goal ELIXIR has created a formal process to identify the most critical life sciences data resources in Europe, designated ELIXIR Core Data Resources (https://www.elixir-europe.org/platforms/data/core-data-resources; Durinx *et al.*, 2016). There are currently 19 Core Data Resources (CDRs, Table 1), spanning a broad range of life sciences data types including genes and genomes, proteins, chemistry, molecular structures and interactions, and the research literature. The process to identify these resources (Durinx *et al.*, 2016) uses a set of qualitative and quantitative indicators of scientific and technical quality and impact. The indicators fall into five categories: Scientific focus and quality of science; Community served by the resource; Quality of service; Legal and funding infrastructure, and governance; Impact and translational stories. The resources identified in this way are of fundamental importance to the wider life sciences community and the long-term preservation of biological data: they are comprehensive, are considered an authority in their fields, are of high scientific quality and provide a high level of service delivery. It is of critical importance that these resources are sustained for the benefit of all researchers.

Many of the Core Data Resource indicators, particularly qualitative indicators such as those concerned with governance or the provision of user support, were collected as part of the initial selection process but tend to be stable and are therefore not useful for describing evolutionary changes to the infrastructure as a whole. In this paper we characterise the Core Data Resources using a subset of the quantitative indicators helpful for portraying aspects of the utility and value of the resources to the research community over time.

Rather than considering data resources individually, ELIXIR views the Core Data Resources as a collective entity, together forming an integrated life sciences data infrastructure. As previously described (Durinx *et al.*, 2016), managers of the Core Data Resources supply Indicator data as part of the selection process, with updates provided on an annual basis. Here, we have for the first time used data collected from the Core Data Resources covering the years 2013-2017, to characterise this emerging infrastructure as a whole.

Name	Overview	References
ArrayExpress	Functional Genomics Data from high-throughput functional genomics experiments	Athar *et al.*, 2019
BRENDA	Database of enzyme and enzyme-ligand information	Jeske *et al.*, 2019
CATH	Hierarchical domain classification of protein structures PDB	Sillitoe *et al,*. 2019
ChEBI	Dictionary of molecular entities focused on 'small' chemical compounds	Hastings *et al.*, 2016
ChEMBL	Database of bioactive drug-like small molecules	Mendez *et al.*, 2019
EGA	Personally identifiable genetic and phenotypic data	Lappalainen *et al.*, 2015
ENA	Nucleotide sequencing information	Harrison, 2019
Ensembl	Genome browser for vertebrate genomes	Cunningham, *et al.*, 2019
Ensembl Genomes	Genome browser for non-vertebrate genomes, with sites for bacteria, protists, fungi, plants, and invertebrate Metazoa	Kersey *et al.* 2018
Europe PMC	Repository to life sciences articles, books, patents and clinical guidelines	Levchenko *et al.*, 2018
Human Protein Atlas	Information on human protein-coding genes	Uhlén *et al.*, 2015
IMEx Consortium (IntAct and MINT)	IntAct: experimentally-verified molecular interactions MINT: experimentally verified protein-protein interactions	Orchard *et al.*, 2012
InterPro	Functional analysis of protein sequences	Mitchel *et al.*, 2019
Orphadata[1]	Comprehensive, high-quality datasets related to rare diseases	Rath *et al.*, 2012
PDBe	Biological macromolecular structures	Mir *et al.*, 2018
PRIDE	Mass spectrometry-based proteomics data	Perez-Riverol *et al.*, 2019

[1] Orphadata was only recently introduced to the Core Data Resource list, in the second round of selection (concluded late in 2018) and is not yet fully integrated into the indicator update cycle, so is not included in the graphics presented here.

SILVA	Resource for quality checked and aligned ribosomal RNA sequence data	Glöcnker *et al.*, 2017
STRING	Known and predicted protein-protein interactions.	Szklarczyk *et al.*, 2019
UniProt	Comprehensive resource for protein sequence and annotation data	UniProt Consortium 2019

Table 1: List of ELIXIR's Core Data Resources

Methods

Qualitative and quantitative information to support the life cycle management of the Core Data Resources is gathered by a defined and iterative process that has been described elsewhere (https://zenodo.org/record/1194123#.XG_anC10eL5). This work depends on close collaboration between the managers of the ELIXIR Core Data Resources, the ELIXIR team, and tools and infrastructure providers who facilitate access to the necessary information.

Data were collected in two phases. For the first round of Core Data Resource selection (https://f1000research.com/documents/7-1711) a Case Document was prepared by the applicant resource managers, providing information about 23 indicators (Durinx *et al.*, 2016) for the calendar years 2013-2015. Annual updates were subsequently requested for 2016 and 2017 from the selected Core Data Resources. For the second round of selection (https://f1000research.com/documents/7-1712) the applicants provided indicator data for the calendar years 2014-2016, later updated with 2017 data.

In the following section, the methods used to generate each Figure are described in turn. The data from which the Figures were generated and additional specific descriptions of methodology and techniques can be found in the accompanying Supplementary Data.

Figure 1:
Data entries: This indicator corresponds to Indicator 3b "Data entries - Total, cumulative" from Durinx *et al.* (2016). Each CDR decides which data entity is its primary entry type and provides counts on an annual basis. Data types include nucleic acid and protein sequences, genomes and metagenomes, macromolecular structures, molecular complexes, publications, complex assemblies, and articles from the scientific literature. The items that constitute "Data Entries" therefore vary between the resources, but the counts down the years are of the same entity for each CDR.

Users: This indicator corresponds to Indicator 2a "Overall usage: visitors" from Durinx *et al.*, 2016. The CDRs are, by virtue of the selection criteria, open to all users with no requirement to register for an account. Because usage is unrestricted, determining the number of users poses a challenge. One way to measure the user community is to count the average monthly web access for each year in terms of unique IP addresses. This is necessarily a proxy for

user numbers and both under- and over-reporting is possible, e.g., users may access resources from multiple devices and thus have multiple IP addresses, and users may also be connected using systems with dynamic IP address assignment: both situations generate more IP addresses than individuals. Conversely, some institutions representing hundreds or thousands of users may appear as a single IP address, leading to underreporting. Additionally, a single IP address that accesses different Core Data Resources will be counted separately for each resource. On balance, the number of unique IP addresses is almost certainly an overestimate of the number of users.

Web access can be measured with web analytics or log analytics. Web analytics ("web page tagging") is based on tags that are embedded in web pages and cookies stored on a user's device, and are typically collected through services such as Google Analytics. Log analytics are based on the analysis of IP address data collected on the server hosting the resource. Although web analytics are generally easier to set up, they do not track 100 percent of requests because JavaScript may not be executed on the client side, for example when cookies or image downloading are blocked, as is typical on mobile devices. Log analytics, on the other hand, are more complicated to set up, requiring dedicated hardware and infrastructure. The system used depends on the technology that is preferred by the hosting institution of the respective CDRs. For 13 CDRs, the estimation of the usage was based on log analytics, and for five resources on Google Analytics. When both measures were reported, log analytics figures were chosen for this analysis.

Staff effort in Full Time Equivalents (FTEs): This corresponds to Indicator 1d "Staff effort: number of FTEs per year for the past 2–3 years" from Durinx *et al.* (2016) and includes curators, bioinformaticians and technical staff representative of each calendar year as reported by each resource manager. This reflects the staff required to develop and maintain a data resource. The distribution of types of staff varies between the CDRs. In Deposition Databases, such as ArrayExpress or ENA, the focus is on technical staff and bioinformaticians. By contrast, knowledgebases, for example the Human Protein Atlas or UniProt, add layers of value through teams of highly qualified curators who manually analyse and standardise research data. Each resource uses its own method to settle on an FTE count to provide in its annual update, then uses that same method for each year. This consolidates part-time and full-time contributors to the equivalent number of full-time positions, so it does not necessarily reflect the actual number of people involved in the resource. It is likely that the FTE count recorded for CDRs housed within large bioinformatics institutes underestimates the actual staff effort required to support such resources, due to economies of scale and institutional support provided within those large institutes. By contrast, a resource operating in a smaller institute may be the only hosted service, and must manage all core IT tasks itself.

Figure 2:
Literature citations: This corresponds to Indicator 2c "Usage in research as measured through citation in the literature" from Durinx *et al.* (2016). This indicator aims to evaluate how the CDRs contribute to specific research projects. For each CDR three different types of citation indicators in Europe PMC have been counted on a yearly basis: a) mentions of the name of the CDR, through mining of the patterns of the resource name, b) citation of

individual records within the CDR, identified through mining of the patterns of their unique identifiers, and c) citations of selected Key Articles describing the individual resources in other publications (see Supplementary Data for further details).

These citation indicators conservatively estimate usage of CDRs in research projects as the estimates are constrained by the number of full text papers available in Europe PMC, *de facto* excluding the non-open access literature. Mining resource-name mentions was carried out for 16 of the 19 CDRs: BRENDA, SILVA and Orphadata were not included in the initial list of CDRs, and have not yet been folded into the "Resource Name Mentions" text mining pipeline. Mining of entry identifiers was carried out for 13 of those 16 resources: three resources do not assign their own unique identifiers to individual data sets (see Supplementary Data for further details). A caveat to this methodology is that use of certain resources has become so "core" to everyday research practice that they are rarely cited. This is the case for literature repositories such as Europe PMC, which are heavily used but rarely explicitly acknowledged. Additionally, while initiatives to encourage data citation are gaining traction (https://doi.org/10.25490/a97f-egyk), these are relatively recent and not yet comprehensively adopted. These factors contribute to significant, but difficult-to-quantify, undercounting of literature citations to the CDRs.

Figure 3:
Categories of the top 20 CDR-citing journals: Three citation indicators of CDRs were collected: a) mentions of the name of the CDR, through mining of the patterns of the resource name, b) citation of individual records within the CDR, through mining of the patterns of their unique identifiers, and c) citations of selected Key Articles describing the individual resources in other publications (see Supplementary Data for details). For each unique PMID across the three citation indicators, the journal title and citation count were retrieved from Europe PMC. The top 20 CDR-citing journals were identified and mapped to a set of categories, based on the category model used in the Scimago Journal & Country Rank (https://www.scimagojr.com/journalrank.php). Finally, the number of citations to CDRs in all three indicators in journals within each category were tallied and plotted against the categories.

Figure 4:
Core Data Resource interconnectivity: Lists of the data resources to which each CDR directly link were requested from the CDR managers. For Figure 4 the interrelationships between the CDRs were plotted. The relationships are expressed in a chord diagram, with the arc width weighted according to the number of links from each CDR to the other CDRs.

Figure 5:
Heat map of Core Data Resource co-citation: The citations of CDRs were collected as for Figure 3. For each unique PMID across the three citation indicators, Cited-by counts were retrieved from Europe PMC. For each pair of resources, the number of common unique PMIDs were counted and displayed graphically as the log of the co-citation count for those two resources. Although co-citations do occur across the full set of CDRs, for legibility only the 12 CDRs that are most co-cited are displayed in Figure 5.

Figure 6:
Horizon of assured funding: This is related to Indicator 1d "Staff effort" from Durinx *et al.* (2016). CDR managers were asked "As of January 2019, for how many Full Time Employees (FTEs) do you have committed funding, on 1 January in the following years?" Data were requested for 2019 to 2024. The figures reported do not imply that the baseline (January 2019) count reflects optimal staffing; resources may have been sub-optimally funded at the time of the survey. Nor do the figures imply that the resources anticipate that support will necessarily decline as shown — efforts to secure future funding are foremost in the minds of the resource managers, and ongoing. The survey question was intentionally specific, aiming to capture the assured security of staff funding for the infrastructure, projected forwards.

Results

Scale of the Core Data Resources

Figure 1 shows the cumulative number of data entries across the Core Data Resources, including all deposited, curated and computed records. The total number of data entries more than tripled, from 766 million to 2.72 billion, between 2013 and 2017.

The number of unique IP addresses accessing the data resources almost doubled in the same time. As noted in the Methods, IP address figures are proxies for the number of individuals who use the CDRs. However, even with very conservative modelling (see discussion in https://beagrie.com/static/resource/EBI-impact-report.pdf) the number of scientists using the CDRs per month, given almost three million unique IP addresses, is in the hundreds of thousands. Additionally, we are confident that the increase in unique IP addresses is a indicator of real growth in users: this figure almost doubled from 2013 to 2017.

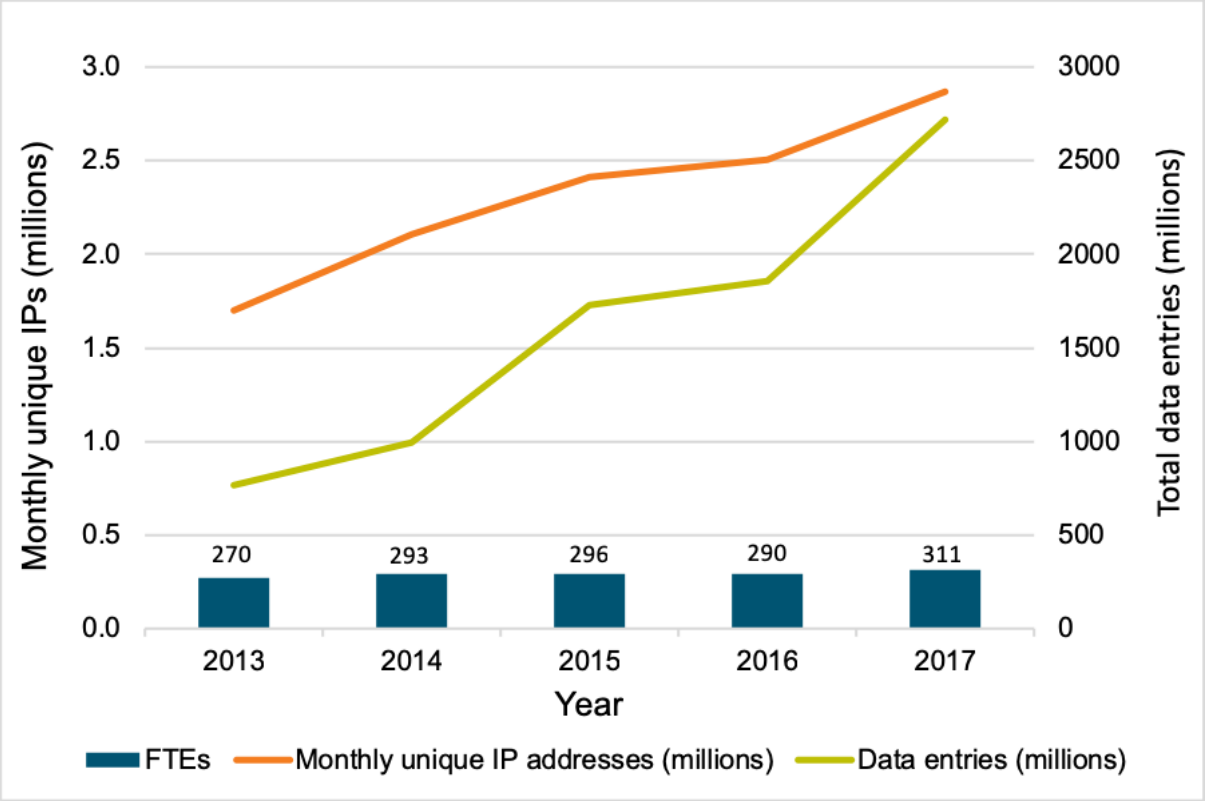

Figure 1. Scale of the Core Data Resources.

Cumulative number of data entries in all Core Data Resources, plotted in conjunction with usage (as measured via the number of unique IP addresses accessing the CDRs per month), and the number of staff at the CDRs (as measured by Full Time Equivalents), per year.

How many people are needed to maintain, curate and serve these data to all these users? The number of FTEs employed in the Core Data Resources grew from 270 to 311, or just 15 percent, over the observed five-year period (Figure 1). Staff numbers are thus growing only slowly despite substantial increases in usage of the Core Data Resources and in their size as measured by the number of records and bytes (their “storage footprint”). This reflects the scalability of the technical solutions that have been adopted, the highly skilled workforce, and the value for money these resources offer. For each FTE employed, requests from at least 10,000 unique IP addresses per month are recorded.

Science evolves continually; developing data services such as metadata schemas, ontologies and user interfaces to support those evolving needs, while also maintaining backward compatibility to older data, is a distinctly human effort. Retaining and finding talented and knowledgeable staff to maintain the scientific relevance of CDRs and support their continued growth in usage requires continual investment.

Open Data and FAIR Data Leadership

Wide usage of Core Data Resources depends critically on the legal right to reuse data. All ELIXIR Core Data Resources are open access[2], with either Terms of Use statements (12 of the resources) or specific licences (7 of the resources) that allow reuse, corresponding to Indicator 4b "Open science" in the selection process (Durinx *et al.*, 2016). Indeed, during the process of identifying Core Data Resources, six resources changed their licences to be more permissive to fulfill this openness criterion.

The Core Data Resources exemplify FAIR data (Wilkinson et al., 2016), maximizing findability, accessibility, interoperability and reusability. For example, Core Data Resources use persistent identifiers, standard vocabularies, and ontologies as the norm in their metadata (included in the "Quality of service" Indicator 3a and 3d category). Data exchange is enabled via standard protocols such as HTTP(S) (websites and APIs) and FTP ("Quality of service" Indicator 3f). The Core Data Resources provide user support and customer service via helpdesks, user feedback mechanisms, and outreach and training activities ("Quality of service" Indicator 3g).

Core Data Resource Citations in the Scientific Literature

Citation in the scientific literature is an established indicator of the value and significance of data resources (Bousfield *et al.*, 2016). We investigated the impact of Core Data Resources by mining the full text open access publications available in Europe PMC for mentions of Core Data Resources by their name and by their specific data entry identifiers, with open citations of Key Articles describing each specific resource being also included in the analysis. Figure 2 shows the growth in the number of publications in Europe PMC on the basis of these three citation indicators.

Given the total of 51,434 name or data identifier mention citations in 2017, a year in which around 305,000 open access articles were published, 17 percent of the open access articles in Europe PMC refer to a Core Data Resource by mentioning the resource name or an entry identifier. This is a significant proportion. As shown in Figure 2, the combined citation indicators for the Core Data Resources grew by a third over the four-year period analysed.

[2] ELIXIR is committed to Open Access as a core principle for publicly funded research. ELIXIR Core Data Resources reflect this commitment and have terms of use or a licence that enables the reuse and remixing of data. The Creative Commons licenses CC0, CC-BY or CC-BY-SA are all conformant with the Open Definition (http://opendefinition.org/licenses/), as are equivalent open terms of use.

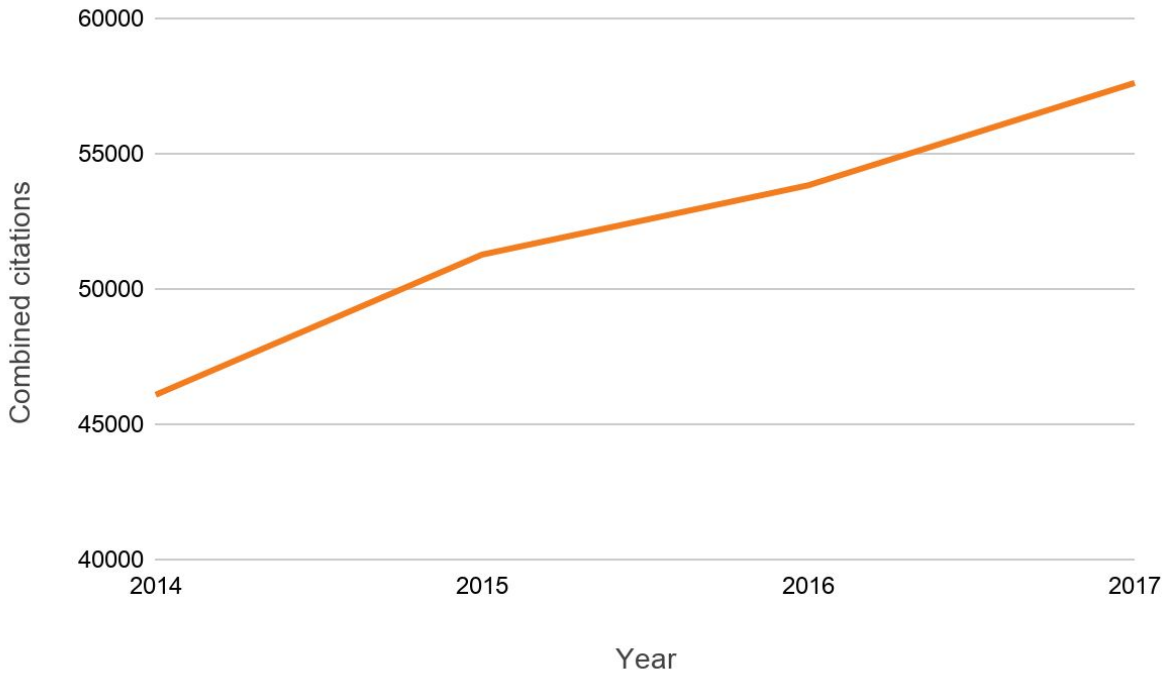

Figure 2. Usage of Core Data Resources in research.

Number of citations in the open access literature per year (citations of the name of the resources (16 CDRs), of resource entry identifiers (12 CDRs), and of pre-identified Key Articles describing the respective resources (18 CDRs)).

Having established that the Core Data Resources are widely cited in the literature we assessed the scientific fields of the citing journals. As shown in Figure 3, the impact of the Core Data Resources beyond the immediate basic research domain from which they originated is clearly evident. The Core Data Resources are used more widely than within bioinformatics and molecular biology, ranging from primary research into applied and health sciences, food security and the environment.

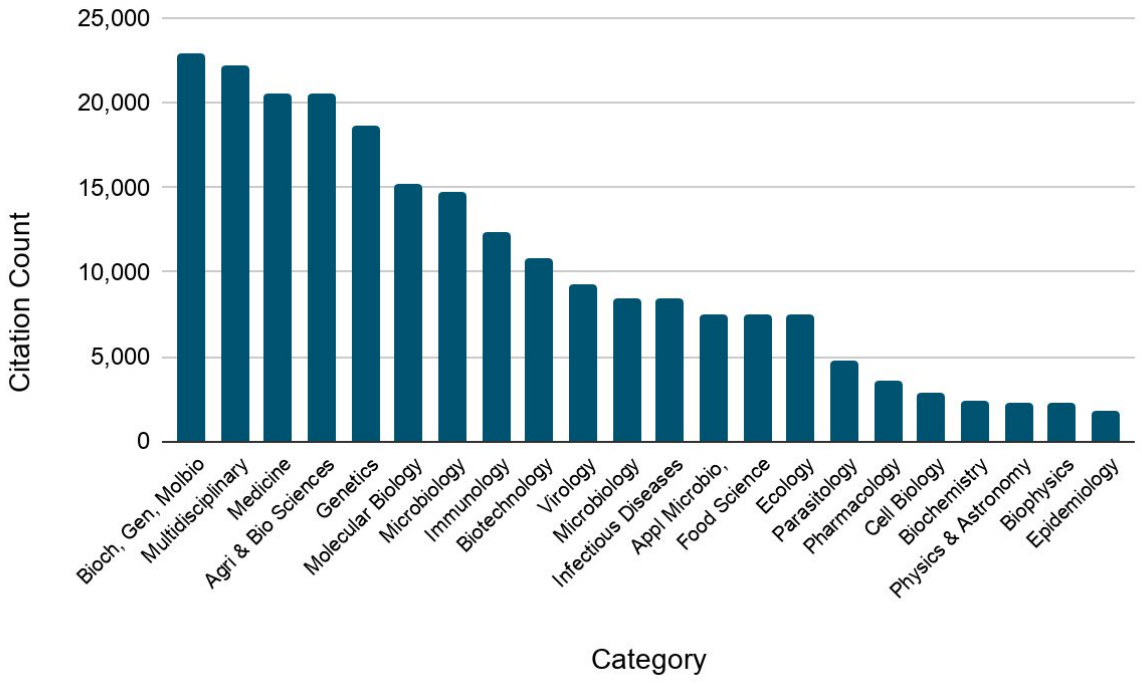

Figure 3. Citation counts for the categories of scientific fields in which the 20 journals that most frequently cite the Core Data Resources are active.

Integration, Dependency and Ecosystem

The Core Data Resources exhibit high connectivity and interdependencies, reflecting the biological relationships between the different data types. The use of persistent identifiers is the primary method of cross-referencing between different resources, alongside the use of standard shared vocabularies such as the Gene Ontology (The Gene Ontology Consortium 2019). For example, UniProt protein sequences are translated from ENA sequences and Ensembl, and linked to corresponding PDBe structures. Records for compounds in ChEMBL link to IntAct interactions in which they are involved. The InterPro consortium builds on UniProt sequences to generate protein family signatures, which in turn are used to annotate uncharacterised UniProt sequence data. All resources link to publications (Europe PMC) for biological context, which in turn cite identifiers to link back to the data. Figure 4 shows a representation of the interconnectivity between 18 of the CDRs. As new Core Data Resources are identified, it is expected that they will contribute to and extend the ecosystem.

While the CDRs support each other with the interconnections illustrated in Figure 4, they also interact with multiple resources outside this set. For example, ChEBI is used in UniProt enzyme annotations in the form of Rhea chemical reactions (https://www.rhea-db.org/), and UniProt enzymes are annotated using the IUBMB enzyme classification (https://iubmb.org/) as represented by the ENZYME database (https://enzyme.expasy.org/). While SILVA links to the ENA Core Data Resource, it also cross-references to RNACentral

(https://rnacentral.org/), and the prokaryotic standard name resource LPSN (http://www.bacterio.net/) among others.

Between them, the Core Data Resources link out to more than 350 external resources, listed in Table S6 in the Supplementary Data. The diversity of these resources illustrates the foundational role of the Core Data Resources in the wider bioinformatics landscape. Worldwide, the life science data resource ecosystem is an interlinked network, and the Core Data Resources are important nodes in that they integrate and make findable the data from hundreds of other resources, many of which are smaller, or domain-specific. In this way the Core Data Resources enhance the value of the other resources to which they are linked by multiplying re-use of their data.

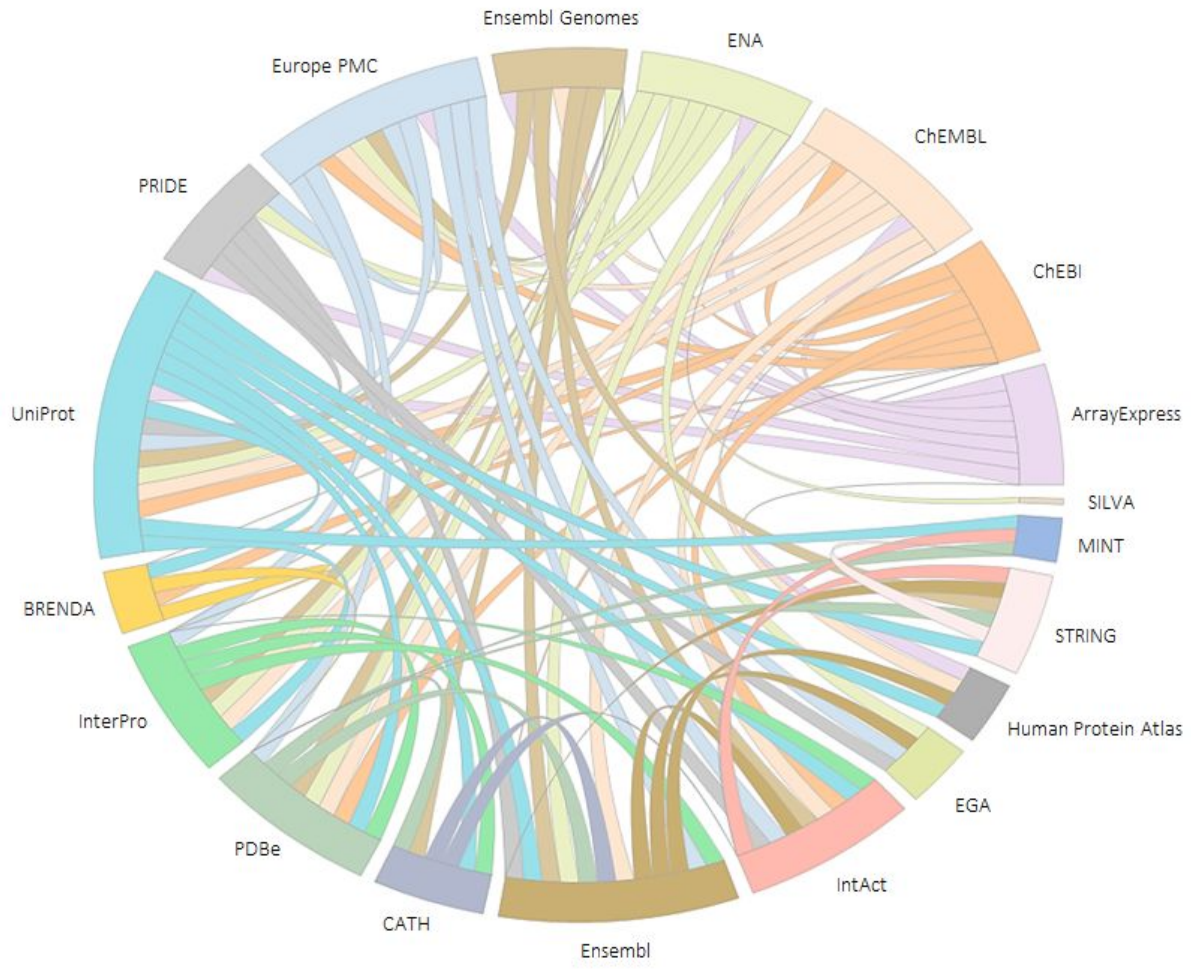

Figure 4. Core Data Resource interconnectivity.

The Core Data Resources are placed on the circumference of the circle, with each resource represented by an arc proportional to the total number of interactions. The width of each internal arc, which transects the circle and connects two different resources, is proportional to the number of different data types that are exchanged between the two resources at the ends of the arc.

Another way to represent the integrated nature of Core Data Resources is to analyse the co-citation of different data resources in full text publications. That is, to count the number of

times two or more resources (name or entry identifiers) are cited in the same publication. Figure 5 depicts the co-citation distribution for the 12 Core Data Resources that show the most co-citation. Notable co-citation hotspots include UniProt, PDBe and ENA, attesting to their frequent use in conjunction with each other and with other Core Data Resources.

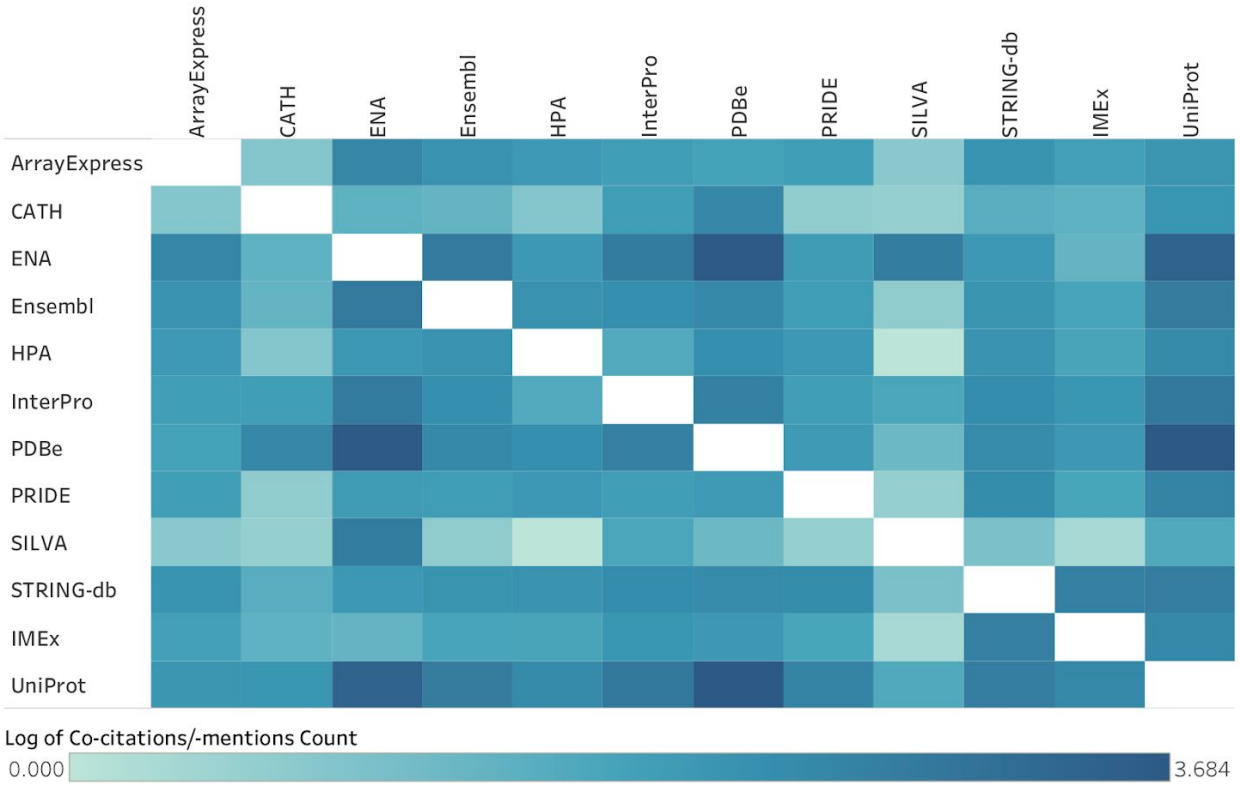

Figure 5. Heat map of the pairwise co-citation of the 12 ELIXIR Core Data Resources that are most frequently co-cited.

The intensity of shading correlates with the frequency of co-citation.

Funding Horizon

ELIXIR Core Data Resources are the repository of record for a number of data types. Funders, journals, and submitters treat the Core Data Resources as stably funded infrastructure, but funding is in fact not assured past a very short horizon for many resources.

To assess the magnitude of this problem we asked managers of each Core Data Resource to report the funding for their staff that is currently confirmed. Figure 6 shows that as of early 2019, the resources have assured funding for on average 88 percent of the staff within a one-year horizon, and for on average 31 percent of the staff over a three-year horizon. Only four of the 18 resources (22 percent) have the assurance that, one year from January 2019, they would have funds to support the same level of staffing as on that date.

These results show that beyond 2020 the assured funding levels are not sustained, implying a clear risk to the continued existence of this essential research infrastructure. The lack of assured long-term support for these resources demonstrates the fragility of data infrastructure upon which the research ecosystem depends and upon which funders rely for storing research data generated from public monies.

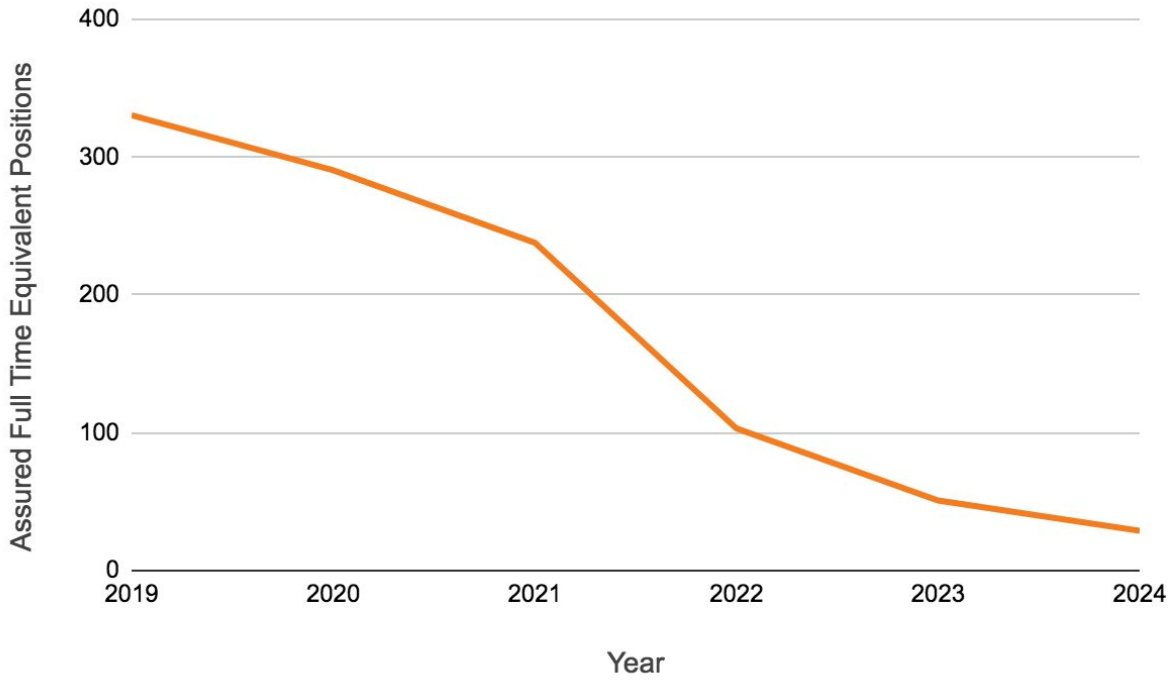

Figure 6. Horizon of assured funding: number of Full Time Equivalent positions for which the CDRs have assured funding, by year, as of January 2019.

It is unlikely, of course, that staffing for the infrastructure will collapse as shown in Figure 6. Funding for much of the infrastructure is currently awarded on the basis of short-term grants or contracts whose terms are often suited more to research projects than to funding infrastructure. Consequently, resource managers spend a significant proportion of their time demonstrating the value of their resources to funders and preparing applications for funding renewal. It is entirely appropriate for funders to exercise mechanisms that continually assess the fit of the infrastructure with the scientific need, but Figure 6 suggests that the frequency of this assessment is faster than might be warranted for an infrastructure, which by definition must be established, of proven utility, and stable over time.

Discussion

During the past four decades, the massive growth of data in life sciences research, and the demonstration by researchers and funders that these data are more valuable if shared and re-used, have led to the creation of thousands of data resources to store, curate, and share

these data (Imker, 2018; http://bigd.big.ac.cn/databasecommons/). Together, these data resources represent a new type of research infrastructure, which, unlike traditional "bricks and mortar" research infrastructures, is both virtual and distributed. The resources that make up this infrastructure are developed and maintained through the expertise of highly qualified personnel. The physical components of the infrastructure are these staff and the computational resources within which the data are stored and through which they are distributed to users.

The successful selection by ELIXIR of a set of Core Data Resources for Europe has shown that it is possible to develop a data-driven process to measure the impact of data resources and to use this process to identify a subset of those resources from within the larger data resource ecosystem that are most crucial to the larger infrastructure. The ELIXIR Core Data Resources define a cohort within the global life sciences infrastructure that funders and other stakeholders may use as a basis for structuring policies that support long-term sustainability, for both the Core Data Resources and the greater worldwide life sciences data infrastructure of which they are a part.

In addition to making the case for more sustainable funding support, the named Core Data Resources are models of good practice for managing data resources. They provide a focus for initiatives to integrate data and workflows from other, smaller data resources. Several of the Core Data Resources serve as the repository of record for archiving the data type they store: they are crucially important for the long term preservation of hard-won experimental data generated with public funding. The selection process itself provides a basis for selecting exemplars of good practice for other resource types, such as ELIXIR's Recommended Interoperability Resources (https://www.elixir-europe.org/platforms/interoperability/rirs), as part of building the European research infrastructure across all components necessary for life sciences research.

The Core Data Resources identified by ELIXIR are, by definition, of fundamental importance to the life sciences research infrastructure in Europe and the rest of the world, and, for the first time here, this assertion is quantitatively demonstrated across the set of Core Data Resources. We have shown that these Core Data Resources are accessed by hundreds of thousands of users per month (Figure 1); they are explicitly cited in 17 percent of open access publications in Europe PMC (Figure 2); and they are used extensively across all fields in academic life sciences, medical sciences, and in various life sciences-related commercial activities (Figure 3). It is clear from our analysis that the value of the Core Data Resources infrastructure for the scientific research effort is continually increasing over time as archived data and the use of those data grows.

Risk to this critical infrastructure. This infrastructure has become essential to life sciences research worldwide, as well as in more applied settings such as healthcare, environmental science, biotechnology and food science, and operates in the commercial sector such as the pharmaceutical industry and many small-to-medium-sized companies (https://f1000research.com/documents/7-590). In recognition of the underpinning nature of open data to both research and the science-driven economy, virtually all research funders, both public and charitable, now strongly recommend or require deposition of research data

into open access data resources (European Research Council: https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf; Science Europe: https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf). Leading scientific journals, addressing their concerns about research reproducibility, increasingly advocate and in some cases require deposition in open access data repositories of research data associated with the articles they publish (natureresearch: https://www.nature.com/sdata/policies/data-policies; PLOS: https://journals.plos.org/plosone/s/data-availability). Consequently, the core resources in this global enterprise should have more sustainable funding (Bourne *et al*., 2015; Anderson *et al.* 2017).

Worldwide data ecosystem. The European resources from which ELIXIR Core Data Resources are selected represent only a fraction of life sciences data resources worldwide. The rest of the world also develops and hosts data resources, and many of these are as important to the global life sciences data ecosystem as are the ELIXIR Core Data Resources. Indeed, several of the ELIXIR Core Data Resources are already members of international consortia, with ENA (INSDC; http://www.insdc.org/), PDBe (wwPDB; https://www.wwpdb.org/), and UniProt (https://www.uniprot.org/) being three prominent examples. Many of the global resources are also at risk from short-term and unstable funding cycles. The ELIXIR Core Data Resource selection process provides a model for identification of other crucial resources worldwide that will allow funders to more efficiently support the worldwide life sciences data resource ecosystem. The nascent Global Biodata Coalition (Anderson, 2017; Anderson *et al.* 2017), supported by funders and heads of international research organisations, will use this process as a model for a worldwide effort to identify and secure long-term funding for crucial data resources.

Acknowledgements

This work was supported in part by European Union's Horizon 2020 research and innovation program, ELIXIR- EXCELERATE, grant agreement 676559, and by EMBL and SIB.
The authors thank John Hancock for comments on the manuscript.

References

Anderson W.P., Global Life Science Data Resources Working Group. (2017) Data management: A global coalition to sustain core data. *Nature*, 543, 179.

Athar A. et al. (2019) ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res*; 47(D1), D711-D715.

Berman H.M. (2008) The Protein Data Bank: a historical perspective. *Acta Crystallogr A*, 64, 88-95.

Bourne P.E. et al. (2015) Perspective: Sustaining the big-data ecosystem. *Nature*, 527, S16-17.

Bousfield D. et al. (2016) Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources. *F1000Res*, 5(ELIXIR), 160.

Cunningham F. et al. (2019) Ensembl 2019. *Nucleic Acids Res,* 47(D1), D745-D751.

Durinx C. et al. (2016) Identifying ELIXIR Core Data Resources. *F1000Res*, 5(ELIXIR), 2422.

Gabella C. et al. (2017) Funding knowledgebases: Towards a sustainable funding model for the UniProt use case. *F1000Res*, 6(ELIXIR), 2051.

Glöckner F.O, et al. (2017) 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J Biotechnol,* 261, 169-176.

Harrison P.W. et al. (2019) The European Nucleotide Archive in 2018. *Nucleic Acids Res,* 47(D1), D84-D88.

Hastings J. et al. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res,* 44(D1), D1214-1219.

Imker, H.J. (2018) 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance. *Frontiers in Research Metrics and Analytics* 8, 3.

Jeske L. et al. (2019) BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res,* 47(D1), D542-D549.

Kersey P.J. et al. (2018) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res,* 46(D1), D802-D808.

Lappalainen I. et al. (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet,* 47(7), 692-695.

Levchenko M. et al. (2018) Europe PMC in 2017. *Nucleic Acids Res,* 46(D1), D1254-D1260.

Mendez D. et al. (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res,* 47(D1), D930-D940.

Mir S. et al. (2018) PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res,* 46(D1), D486-D492.

Mitchell A.L. et al. (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res,* 47(D1), D351-D360.

Orchard S. et al. (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods,* 9(4), 345-350.

Perez-Riverol Y. et al. (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res,* 47(D1), D442-D450.

Rath, A. et al. (2012) Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum Mutat, 33(5), 803-808.

Sillitoe I. et al. (2019) CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res,* 47(D1), D280-D284.

Szklarczyk D. et al. (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res,* 47(D1), D607-D613.

The Gene Ontology Consortium. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res,* 47(D1), D330-D338.

Uhlén M. et al. (2015) Tissue-based map of the human proteome. Science, 347(6220), 1260419.

UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res,* 47(D1), D506-D515.

Wilkinson M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data,* 3, 160018.