EXCELERATE Deliverable D3.3

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	Report describing ELIXIR-wide systems for the computer-assisted collection and delivery of harmonised metrics and quality criteria from multiple ELIXIR resources and collation of these at the ELIXIR Hub	
Contractual delivery date:	31 August 2019	
Actual delivery date:	7 August 2019	
WP no.:	WP3	
Lead Beneficiary:	1 - EMBL	
WP Title:	Data Resources and Services	
WP leader:	Johanna McEntyre (EMBL-EBI) and Christine Durinx (SIB)	1-EMBL-EBI and 26-SIB
WP partners:	1-EMBL-EBI , 7-CNIO, 9-FVIB, 14-UPF, 16-IMIM, 26-SIB	

Authors:

Christine Durinx (SIB), Rachel Drysdale (EMBL-EBI), Jo McEntyre (EMBL-EBI), Heinz Stockinger (SIB), Juergen Haas (SIB), Nicole Redaschi (SIB), Aravind Venkatesan (EMBL-EBI), Franziska Gruhl (SIB)

Contributors:

Chuck Cook (EMBL-EBI)

1. Table of contents

2. Executive Summary

The mission of the ELIXIR Data Platform is to deliver a sustainably funded portfolio of Core Data Resources that exemplify excellence within a coordinated and vibrant ecosystem of Node Data Resources, to meet the needs of the European life science research community.

The overall objective of Work Package 3 is to build a framework to inform and drive the sustainable development of Europe's Core life-science Data Resources. ELIXIR aims to ensure that these resources are available long-term and that the life cycles of these resources are managed such that they support the scientific needs of the life sciences and biological research. In addition, Work Package 3 promotes excellence in resource development and operation through providing a unified framework for the identification and monitoring of key bioinformatics resources across Europe.

Running two iterations of the process to select ELIXIR Core Data Resources has informed us how best to collect all the required indicators. While some indicators never or rarely need updating, others can only be collected manually. As discussed in detail in this report, some indicators are automatically collected on a local level, while for others, collection can be centralised. The adopted approach therefore maintains a pragmatic balance between automated collection (local and centralised) and manually collated updates. The Core Data Resource selection and update processes that are currently in place only require modest human resources at the ELIXIR Hub and little specialized technical knowledge, making these processes easily sustainable over time.

Having defined the Core Data Resources, and implemented annual collection of Indicator data, Work Package 3 has produced an article "The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences" and posted it as a preprint on bioRxiv as well as submitting it to the Bioinformatics journal. This is the first such analysis to be conducted on such a collection, and will be a powerful tool in the ongoing work towards long term sustainability of life science data resources.

3. Impact

We have demonstrated that the ELIXIR Core Data Resource selection process is a practical and workable process that requires modest resources to maintain in the long term. Running the process has resulted in the selection of 19 Core Data Resources to date in Europe and the paper describing this work has been viewed almost 20,000 times. This experience has informed aligned processes within ELIXIR to identify key resources, such as the ELIXIR UK Node Service selection process and that for the selection of the ELIXIR Recommended Interoperability Resources. Most importantly, it is proving exemplary in the formative discussions within the Global Biodata Coalition regarding how to identify the most critical resources for life sciences research globally.

4. Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Implement the coordination of data archives and literature sources within ELIXIR	x	
2	Stimulate innovation by supporting industry uptake of ELIXIR resources, particularly in SMEs		x

5. Delivery and schedule

The delivery is delayed: Yes • No☑

6. Adjustments made

No adjustments have been made.

7. Background information

Background information on this WP as originally indicated in the description of action (DoA) is included here for reference.

Work package number	3	Start date or starting event:	month 1
Work package title	**Data Resources and Services**		
Lead	Jo McEntyre (EMBL-EBI) and Christine Durinx (SIB)		

Participant number and person months per participant

1 - EMBL 83.00; 9 - CIPF 1.33; 12 - BSC 12.00; 14 - UPF 4.34; 15 - IMIM 4.70; 25 SIB 68.50

Objectives

The overall objective of this WorkPackage (WP) is to build a framework to inform and drive the sustainable development of Europe's core life-science data resources. The goals of WP3 are to:

- Promote excellence in resource development and operation through providing a unified framework for the identification and monitoring of key bioinformatics resources across Europe.
- Increase the sustainability of manually curated resources, which, while of high value and essential to the life- science community, are very labour-intensive to operate. This will be done by integrating the literature with data, with particular emphasis on maximizing value added by curation.

Work Package Leads: Jo McEntyre (EMBL-EBI) and Christine Durinx (SIB)

The core mission of ELIXIR is to build a sustainable infrastructure for biological information across Europe. Data resources and services (hereinafter referred to as "resources") are a key part of this infrastructure and can vary; from submission databases that contain research data outputs such as DNA sequences (e.g. European Nucleotide Archive), to highly dynamic resources that aggregate, process and visualise research data, often adding layers of value through manual curation by highly qualified personnel. (e.g. UniProtKB/Swiss-Prot).

Task 3.1. Promote and implement good practice in data resource and service management through the formalization of metrics and quality criteria enabling the identification of ELIXIR Named and Core Resources, and informing their life-cycle management (22.67PM)

The first requirement for the development of a unified framework for the management of key bioinformatics resources across Europe is to identify which resources (a) meet a variety of quality criteria with respect to scientific impact and level of service, and (b) which of these are of fundamental importance to the life-sciences community. Therefore, ELIXIR resources will be identified and classified into two categories:

- ELIXIR Named Resource will be attributed to ELIXIR Resources from the project partners (ELIXIR Nodes) that are compliant with a set of metrics/criteria that guarantee their quality.
- ELIXIR Core Resources will be the subset of ELIXIR Named Resources that, based on metrics/quality criteria, are of fundamental importance to the life-science community and that are considered to be an authority in their field with respect to one or more characteristics.

Definition of clear metrics/quality criteria that measure current and projected use of ELIXIR resources as well as their scientific impact, and the reliability of the service, will underpin the identification of ELIXIR named and core resources and provide data to inform life-cycle management on an ongoing basis.

The initial set of metrics and quality criteria for ELIXIR resources will be identified based on prior resource management experience of WP partners, on the work completed by the ELIXIR technical coordinators group, and experiences from other disciplines such as the Data Seal of Approval project56. Formal opportunities for ELIXIR-wide review of the proposed criteria will be conducted through presentations and workshops aligned with project management meetings.

Metrics and quality criteria will evaluate both the scientific impact of the resources on the life-science community and the reliability of the service. They include, but are not limited to: uptime and download speeds, usage statistics (IPs, page views, downloads), citations in the literature, data submission rates, international collaborations, programmatic access, and curational effort.

In defining measures of quality it is important to recognise the context in which the service is being provided and to base categorization on a range of criteria. For example, a resource that serves a small community may not have as many page views as a large resource, yet reach 90% of the community it supports. Other may play a foundational role to derived services. It will be important to differentiate between submission databases and "added-value" databases that organize, curate, or otherwise represent submitted data, as the profile of use of these types of resource may be very different.

Equipped with an agreed set of criteria, it will be possible to effect a number of actions:

- Identify new resources for inclusion in the ELIXIR set.
- Set quality standards for emerging resources and inform their development.
- Build confidence among users through the identification of ELIXIR resources directly (such as a "badge" on the resource itself) and through a variety of portals such as the Tools and Data Services Discovery Portal (WP1).

- Monitor usage trends and manage resource life cycles effectively using objective criteria.

- Build understanding of the impact of ELIXIR resources both within the ERA and within global research infrastructures.

Resource development based on Metrics and Quality Criteria

Alongside the definition of the metrics and quality criteria, coordinated management processes will be required to review candidate resources, encourage use of ELIXIR-approved badges (or similar), and monitor resource life cycles.

We expect the organizations running the resources to actively contribute to this process, and that this in itself will provide feedback mechanisms to improve and refine the criteria. This coordinated feedback model will have the added benefit of providing opportunities for peer-peer capacity building (WP10) in the areas of life-cycle management and sustainability, and metrics/quality criteria implementation as we share expertise between ELIXIR Nodes.

Partners: EMBL-EBI, CH

Task 3.2. Inform ELIXIR Resources life-cycle management and improve the ELIXIR Resource portfolio through the implementation of an active and computer-assisted infrastructure for the monitoring of ELIXIR Named and Core Resources based on the metrics and quality criteria formalized in Task 3.1 (71.1PM)

In the interest of transparency and to build excellence across resources, metrics and quality criteria for ELIXIR named and core resources will be held centrally at the ELIXIR Hub (see also WP12.3). Access to this collated data will be made available to all Nodes and resources involved, and potentially more widely as aggregated data.

In this task, technical processes will be developed to generate and collate the metrics and quality criteria agreed in Task 3.1. Operating in active mode over a period of time, the emerging trends will inform ELIXIR Resources life- cycle management and improve the ELIXIR Resource portfolio overall.

The processes developed will gather, report and upload metrics and other quality criteria in agreed formats and to an agreed timescale to the ELIXIR Hub.

The need to collate metrics/quality criteria centrally for analytical and comparative purposes raises questions regarding the technical implementation of such a system. There are a number of challenges in doing this, not least the willingness of the resource providers to share detailed metrics and quality criteria regarding their resource. Subsequent to this will be the need to provide confidence, particularly in the case of metrics, that what is being measured/reported from different resources is comparable in a fair manner; this will require sharing of methodological approaches (such as how robot traffic to websites is treated) through a shared understanding of what is considered a page view across different resources. Finally, agreement on a timetable and format for quality and metrics information will be required so that it can be easily collated in one place. These challenges may give rise to a need for technical effort in the participating resources and such requirements will be supported through the ELIXIR Hub core budget if required.

Partners: EMBL-EBI, EMBL-ELIXIR, CH

Task 3.3. Increase the sustainability of curated resources through literature-data integration and resource crosslinking (80.1PM)

The integration of the literature with data is critical for understanding the biological context of new results, for showing clear provenance of scientific assertions, and for discovering new information. While these are important activities for all of the scientific community using online resources, the requirement is most intense within scientific curation processes. The excellent quality of many European bioinformatics resources relies on manual curation, a process in which trained experts review experimental data reported in publications and extract relevant information for inclusion in data resources. This requires searching, reading, filtering, verifying and recording information; labour-intensive, and therefore costly, processes. However, curation saves time and adds significant value for researchers, obviating the need for potential users to individually seek out and synthesise threads of scientific information. Technological advancements in the past few years provide new opportunities to expedite the work of curators and also provide novel approaches to integrating the literature with data for the wider scientific community. For example, when a curator adds a new piece of information to a data record, the source article is cited in the record. However, it would be useful to link from that specific annotation directly to the precise point in the article that was extracted by the curator, for example, a figure legend. This will allow researchers and curators alike to understand exactly where that piece of information came from when viewing the data record, or conversely, to follow a link to see more data when reading the article - a connection that is currently not possible to traverse. Such developments will provide efficiency savings in resource and tool interfaces, reduce repetition, and when published, will provide granular deep links between the literature and data for users.

In this task, a roadmap for infrastructure that integrates the literature with data through a variety of novel approaches, including text mining, will be developed. Elements of this roadmap will be demonstrated by a collection of pilot developments that provide deep links between the literature and established or emerging ELIXIR data resources. Automated approaches, such as text mining, that identify and extract useful biological concepts will be a necessary part of this activity, from generating granular links to suggesting articles to curate in the longer term. Harnessing the expertise of the text and data mining community as a whole would maximize the impact of this aspect. This task aims to engage with existing database providers and novel Use Cases (WP6 to 9) to develop a roadmap that combines the above elements to develop an infrastructure for literature-data integration and enrichment, and furthermore to demonstrate this through a collection of pilot developments. To do this we will use known high-quality annotations such as GeneRifs (sentences extracted from articles that have been included in gene database records) and the Europe PMC database of life science research articles.

Partners: EBI, CH, ES

8. Appendix 1: Report describing ELIXIR-wide systems for the computer-assisted collection and delivery of harmonised metrics and quality criteria from multiple ELIXIR resources and collation of these at the ELIXIR Hub

8.1 Introduction

In July 2017, ELIXIR selected the initial set of **Core Data Resources** (**CDR**s) [1]. CDRs are deposition databases and knowledgebases that are of fundamental importance for the life sciences community in Europe and worldwide [2]. Candidate Core Data Resources were evaluated using a set of five high-level indicator categories:

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

Identification of the Core Data Resources is a key step in the collective endeavour to ensure that funders, contributors (i.e., researchers generating data) and users are aware of the impact of these resources. This in turn highlights the need for sustained long-term funding to secure the data and knowledge they contain.

As part of the selection process, each candidate ELIXIR Core Data Resource is asked to provide information for each of the 23 indicators [1] to the ELIXIR Hub, who store and make appropriate use of the data. The set of Core Data Resources [2] is the result of careful review by an independent expert panel and the ELIXIR Heads of Nodes (HoN).

Data have been aggregated over all CDRs and presented as a comprehensive infrastructure representing the ELIXIR Core Data Resource portfolio [3]. The results can be used to calculate more abstract measures such as the overall research productivity gained through Core Data Resource use, or the economic returns the resources bring to the investments already made.

Qualitative and quantitative information is required to support the life cycle management of the Core Data Resources. This information is gathered by a defined and iterative process such that trends can be observed over time. Indeed, for the indicator data that change over time, such as usage or citation statistics, annual updates are subsequently requested from the selected Core Data Resources to allow for monitoring of progress, impact trends and usage for individual resources over time. The data are not intended to be used to compare individual resources to each other, as each resource has its own specificities (see Section 3). They provide, however, indications about the trends in absolute number of users or volume of downloads for each resource. Based on these monitoring data, the set of Core Data Resources will be evaluated every 2-3 years. More resources may be added, while others may be removed, as the landscape of biological data evolves.

In this document, we detail how data for the 23 indicators are collected, kept up-to-date, and made available. We also explain how this collection process has evolved, as well as the associated challenges. This report builds on Milestone M3.3 “Plan for collation of metrics and quality data at the ELIXIR Hub” (Due Month 24 - August 2017), which details why the indicator information is gathered and how the data are used [4].

8.2 Methods of data collection

8.2.1 Data Collection during the selection of Core Data Resources

The initial Core Data Resource list was defined in July 2017. As part of this first round of selection [5], a Case Document was prepared by the resource managers, which provided information about the 23 indicators described in the original publication [2]. The collected data covered calendar years 2013-2015.

For the selected CDRs, the data collected through these Case Documents have been collated into a master “Core Data Resource Statistics” spreadsheet in which each Core Data Resource is represented in a single row while the columns correspond to indicators described in the publication [2]. This master spreadsheet is maintained at the ELIXIR Hub, with new data collected via the annual updates (see Section 2.2). Access to these data, and the spreadsheet, is restricted because many of the indicators correspond to confidential information not generally communicated publicly, such as numbers that relate to user access, or rate of growth of content, or changes in staff FTEs (Full Time Equivalents).

As part of the CDR selection process, each resource that applies, lists scientific publications about their resource. These are chosen by the applicant, with some of them opting to limit the choice to a small number of highly significant articles, and others opting to comprehensively list all reference articles. When a data resource joins the CDR list, it is asked to select up to five significant reference articles for use in citation and impact analysis. These are entered into a "CDR Key Article Citations" spreadsheet, maintained at the ELIXIR Hub.

As the collaboration with, and the buy-in of, the managers of CDRs is essential for the continued success of this work, they were consulted in the second half of 2017, once the initial selection process had been completed. Consequently, both the data collection form and analysis have been adjusted to introduce clarifications based on their feedback in light of them having worked through the process. For example, Indicator 3b "Quality of Service: Data Throughput" was interpreted by some data resources as requesting statements of increments of record numbers and/or data volume for each year, and others as requesting statements of cumulative totals. In this case, cumulative totals were chosen as the preferred method of expression for this Indicator. Such clarifications have been carried through to the annual update procedure (see Section 2.2) and subsequent rounds of Core Data Resource selection.

A second round of Core Data Resource selection was held in 2017/2018. The process is described elsewhere [6]. The data collected as part of the selection process covered calendar years 2014-2016.

News articles announcing Core Data Resource additions were published by ELIXIR on 25th July 2017 [7], 25th June 2018 [8] and 31st January 2019 [9].

The ELIXIR Deposition Databases

For the purposes of data resource description, a distinction is made between archival resources, such as a deposition database for primary experimental data, and knowledgebases, which add value to primary data through curation and analysis. During the course of the initial round of CDR selection, a set of deposition databases were identified that fulfill all requirements for consideration as Core Data Resources. However, others were found to be of high quality with respect to all selection criteria except their fundamental importance to the "wider life-science community". They might be the depository of record for a specific type of data pivotal to a specific community of researchers, for example BioModels is a database of computational models of biological processes. Additionally, while Core Data Resources are required to be well established, with a long track record of service, important deposition databases may legitimately have arisen quite recently in response to the emergence of new scientific techniques, and the resultant new types of data. This circumstance is illustrated by the MetaboLights database, established in 2012 to store experimental data from new technologies that lead to the development of the discipline of metabolomics.

Given that there is an increasingly pressing need to define a set of deposition databases suitable for funders, publishers and institutions to reference as they craft open data policies and data management plans, the need for a list of recommended ELIXIR Deposition Databases, distinct from the Core Data Resource list, was recognised. Data resources may thus be listed on the ELIXIR Core Data Resource list, the recommended Deposition Database list, or both.

8.2.2 Annual Data Collection following initial Core Data Resource selection

While some of the indicators are qualitative, or of a type that does not change - such as license type, or provision of a help desk - several are quantitative. These are FTE count, Web access (e.g. number of IP addresses, or page views), Data downloads, Citations, Total number of records and Total size (e.g. in Gigabytes) of entries. A full list can be found in Table 1 of [4]. For the indicators that are quantitative, annual updates are collected, as part of our efforts to monitor the ongoing utility of the CDR set.

Starting in 2018, annual updates were requested from each Core Data Resources, using an update form that corresponds closely to the original Case Document (see Annex 1). This method is currently in use for BRENDA, CATH, EGA, HPA, the IMEX consortium, Orphadata, SILVA, String-DB, and UniProtKB.

For those resources based entirely at the EMBL-EBI, where the collection of web access statistics is centralised as part of the EBI's operations, much of the update information is retrieved from a single point of contact. In addition to this, any necessary follow-up queries were sent to individual Resource Managers, for example, to establish the number of employees, and database Entries (cumulative totals, in terms of number of records, and total size in GB). This method is currently in use for ArrayExpress, ChEBI, ChEMBL, ENA, Ensembl, Ensembl Genomes, EuropePMC, InterPro, PDBe and PRIDE.

During the initial round of CDR selection, it became obvious that not all CDRs had supplied consistently derived information regarding the citation of the resource in the research literature. Therefore, the collection of this data has been automated and centralised across all CDRs, based on the Europe PMC publications database and its routine text mining workflows. Each CDR supplies to Europe PMC:

(1) Patterns for accession numbers

(2) Resource names, abbreviations, and qualifying terms

The text mining patterns used to mine both the data resource accession numbers and the resources names are available on GitHub. While the text mining approach is not perfect, it provides an automated and consistent approach to gathering citation metrics in the research literature, allowing for straightforward comparisons longitudinally through the years. All full text open access incoming papers are mined daily for this information, and the results made available via Europe PMC's public APIs. (The source code for the text mining pipeline element and the Annotation API is also available on GitHub, for transparency.)

Citation data in terms of citations of Key Articles about the CDRs are generated on an "as needed" basis. As mentioned above, Resource Managers nominate up to five articles that they regard as authoritative references for their resource, with an opportunity to review the list annually. Citations for these articles are also based on Europe PMC's open citations.

8.3 Examples of usage of the collected indicator data

The collected indicator data are used by different entities under different formats.

Original case documents

The original case documents generated by the candidate Core Data Resources are provided by email to the Heads of Nodes Committee for review and to inform the selection process.

Core Data Resource Statistics master spreadsheet

Access to the "Core Data Resource Statistics" master spreadsheet maintained by the ELIXIR Hub is strictly limited to the Data Platform Coordinator, with back up from the ELIXIR CTO, and two additional persons directly involved in the EXCELERATE Task 3.2 work. Editorial access is restricted to the Data Platform Coordinator. The data in the spreadsheet provide the Data Platform Coordinator with a detailed view on the trends in usage of the Core Data Resources.

Aggregated graphical representations

As mentioned earlier, the underlying purpose for identification of the Core Data Resources is to build awareness among funders, submitters, and users regarding their impact. In order to describe the Core Data Resources as a whole infrastructure, it is therefore essential to create high-quality graphic material for communication, outreach and advocacy purposes.

This is first of all done via the creation of graphics from the "Core Data Resource Statistics" master spreadsheet, such as the one in Figure 1.

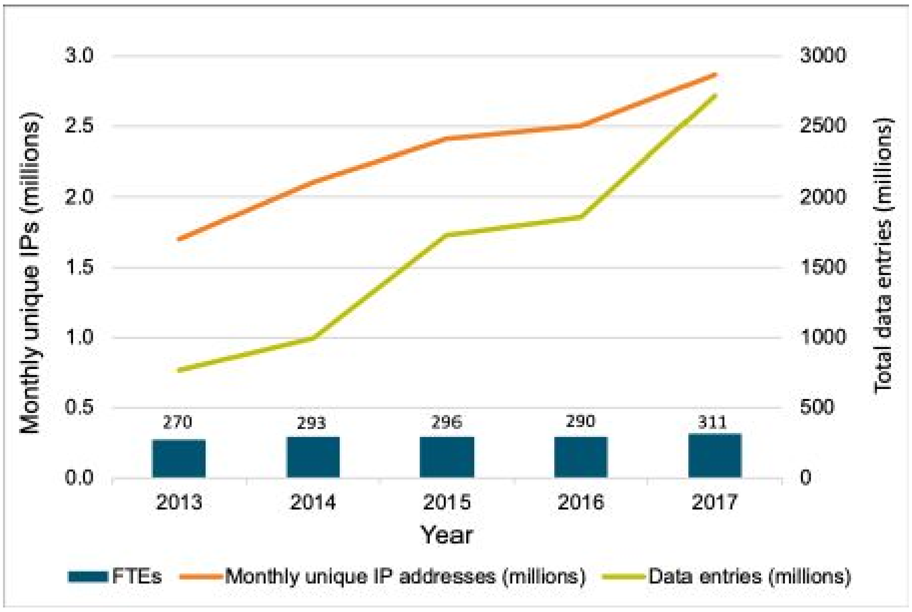

Figure 1. Scale of the Core Data Resources: Cumulative number of data entries in all Core Data Resources, plotted in conjunction with usage (as measured via the number of unique IP addresses accessing the CDRs per month), and the number of staff at the CDRs (as measured by Full Time Equivalents), per year.

In addition, a password-protected data visualisation application was developed on top of the “Core Data Resource Statistics” master spreadsheet. This application allows creation of aggregated graphics as in Figure 2 below.

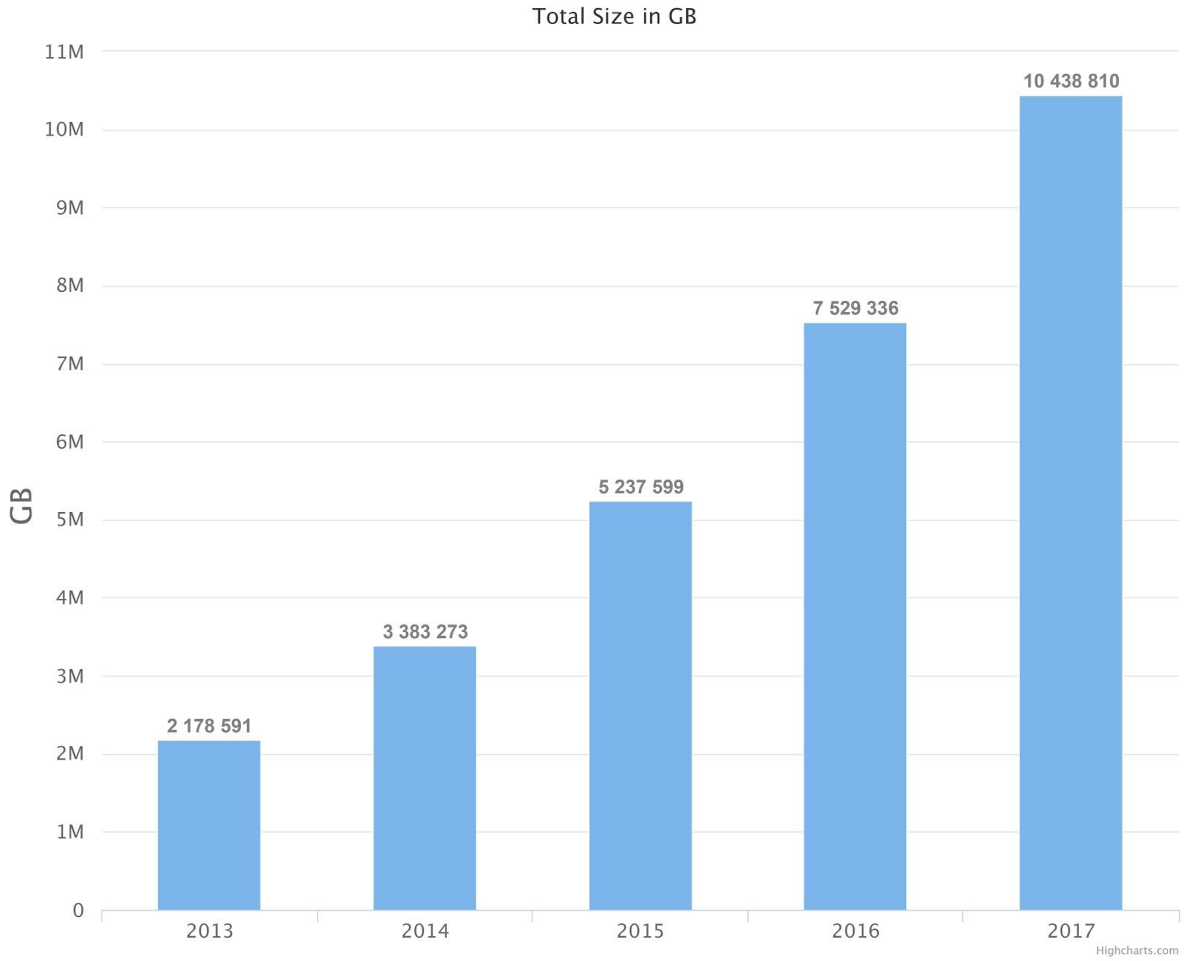

Figure 2. Scale of the Core Data Resources: Cumulative number of data entries in all Core Data Resources.

Periodic reviews

The HoN Committee will review ELIXIR Core Data Resources and recommended ELIXIR Deposition Databases every two to three years. This activity will start in Q3 2020, i.e., three years after the publication of the first set of ELIXIR CDRs and EDDs.

While the details of this process are in development, it will depend heavily on the collected indicator data for each CDR resource. These will be examined for unexpected departures from anticipated trends, which would then be highlighted for consideration by the Heads of Nodes committee working in partnership with the Resource Managers. This will form part of the life cycle management of each resource and the maintenance of the relevance of the Core Data Resource list as a whole: in the longer term we can expect that some resources will become less relevant as technologies evolve in the life sciences, and new resources will come to the fore to reflect the emergence of new technologies and domains of knowledge.

8.4 Discussion

Computer-assisted elements of indicator collection and collation

The system described here includes several examples of computer-assisted collection and collation of metric data:

- Each resource compiles its own usage statistics, or the institute does it on its behalf, based on Web page tagging ("web analytics") and/or log analysis ("log analytics"); the collection at the CDR host institutes is highly automated.
- Each resource compiles its own data content statistics (number of records, submissions, etc). Resources generally automate this within their release processes and publish these statistics in their regular release announcements and/or websites.
- For the Resource name mention and Resource accession number mentions, the text mining technology in place in Europe PMC provides a comparable methodology for each CDR and across the set, removing the burden of manually collating this information from each Resource Manager (or their staff) or the ELIXIR Hub while ensuring that a standard methodology is used for each resource

Challenges

Confidentiality considerations

This work depends on a trusted collaboration between the managers of the ELIXIR Core Data Resources, the ELIXIR Hub, and tools and infrastructure providers who facilitate access to the necessary information. Data must be treated with care and - depending on the dataset and institution involved - might only be accessible on a granular level by a restricted set of people and the clearly defined purpose of selecting the Core Data Resources and monitoring them as part of their life-cycle management. Only aggregated data that do not allow singling out a specific resource, are published externally.

Computer-assisted collection of metrics - limitations

As described in "Identifying ELIXIR Core Data Resources" [2] and elaborated upon in "Plan for collation of metrics and quality data at the ELIXIR Hub" [4], several of the 23 indicators are qualitative rather than quantitative. Examples are the description of the scientific scope of a data resource or of the international dimension of a resource. As they require creative human input, these qualitative descriptions cannot be collected in a computer-assisted way. Even some of the quantitative indicators cannot be obtained in a computer-assisted manner.

For example, to compute the FTE count, it is necessary to average the number of full time equivalents who have worked on the resource over a given year and classify them into curators, bioinformaticians and technical staff. This is not suited for automated collection.

Those metrics that are purely linked to parameters that can be monitored automatically have been found to be technically heterogenous between the different resources. As an example, 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. Whereas Google Analytics is typically easy to activate and use, these web analytics 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. However, sending data to Google, a commercial provider, often triggers privacy concerns, and furthermore, more than 10 million ‘hits’ per month are associated with significant Google Analytics fees. Log analytics, on the other hand, are more typically handled locally by the resource or host organization, requiring dedicated hardware and infrastructure. In addition, the General Data Protection Regulation or GDPR, which regulates data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA) and addresses the export of personal data outside the EU and EEA areas, constrains the central collection and analysis of log files, as IP addresses are considered personal data. Sharing and combining such data from multiple servers distributed across different hosting sites within the GDPR framework would require a significantly different and coordinated approach to gaining consent from the millions of users of the Core Data Resources than the “local collection only” model.

The system used by a specific database therefore depends on the technology in place at its hosting institution(s) and legal constraints. We therefore accept whatever data the resources already provide, so as not to burden those teams with significant technical development purely for the purpose of harmonizing the collection of these data.

Within individual institutes, the collection and collation of such statistics is often a highly automated process. Setting up and maintaining such systems, such as “SIB Insights” at the SIB Swiss Institute of Bioinformatics or the central weblog services at EMBL-EBI requires considerable effort from highly skilled staff. Given that the Core Data Resources are part of a distributed infrastructure, the technological and legal challenges for setting up an automated process to operate across all resources would be significant.

Considerations regarding extension of the automation

Two additional computer-assisted methods of CDR data collection and archiving were considered.

(1) *Replace the "Core Data Resource Statistics" master spreadsheet* with a customised database solution. However, as the size of the dataset fits easily into a spreadsheet, the spreadsheet approach was considered sufficient, and has the advantage of being easy to maintain over time.

(2) *Develop a more deeply automated approach to data collection and storage.* Specifically, streaming usage data direct from CDRs to a centralised repository. However, this would require a common approach to data collection, which is unlikely to happen at present for several reasons. The ELIXIR Core Data Resources are a distributed infrastructure for life science research. The resources are hosted in different institutions and jurisdictions with their own technical infrastructures and methodologies. In order to automate the collection of certain parameters, technical interoperability of the collection methods would need to be established. As described above, different institutes have adopted different monitoring methods and technical interoperability would therefore come at a significant cost.

Overall conclusion

Running two iterations of the process to select ELIXIR Core Data Resources has informed us how best to collect all the required indicators (see Table 1, [4]). While some indicators never or rarely need updating (i.e., whether the resource is a knowledgebase or a deposition database, the existence of a Scientific Advisory Board), others can only be collected manually (e.g., staffing levels and their roles). As discussed above, some indicators are automatically collected on a local level (e.g., usage statistics), while for others, collection can be centralised (e.g., mentions of accession numbers in publications). The adopted approach therefore maintains a pragmatic balance between automated collection (local and centralised) and manually collated updates. The Core Data Resource selection and update processes that are currently in place only require modest human resources at the ELIXIR Hub (a matter of a few days per year) and little specialized technical knowledge, making these processes easily sustainable over time.

In summary, we have demonstrated that the ELIXIR Core Data Resource selection process is a practical and workable process that requires modest resources to maintain in the long term. Running the process has resulted in the selection of 19 Core Data Resources to date in Europe and the paper describing this work has been viewed almost 20,000 times. This experience has informed aligned processes within ELIXIR to identify key resources, such as the ELIXIR UK Node Service selection process [15] and that for the selection of the ELIXIR Recommended Interoperability Resources [16]. Most importantly, it is proving exemplary in the formative discussions within the Global Biodata Coalition regarding how to identify the most critical resources for life sciences research globally [17].

8.5 References

1. Durinx C, McEntyre J, Appel R, Apweiler R, Barlow M, Blomberg N, et al. Identifying ELIXIR Core Data Resources. F1000Res. 2016;5. doi:10.12688/f1000research.9656.2
2. ELIXIR Core Data Resources | ELIXIR [Internet]. [cited 24 May 2019]. Available: https://elixir-europe.org/platforms/data/core-data-resources
3. Drysdale R, Cook CE, Petryszak R, Baillie-Gerritsen V, Barlow M, Gasteiger E, et al. The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences [Internet]. bioRxiv. 2019. p.598318. doi:10.1101/598318
4. Stockinger H, Barlow M, Cook C, Drysdale R, Gasteiger E, Kim J-H, et al. Plan for collation of metrics and quality data at the ELIXIR Hub [Internet]. 2018. doi:10.5281/zenodo.1194123
5. Drysdale R, Repo S, Garcia PR, McEntyre J, Durinx C, Blomberg N. Implementing a Process for the Selection of Core Data Resources. F1000Res. 2018;7. doi:10.7490/f1000research.1116247.1
6. Drysdale R, McEntyre J, Durinx C, Blomberg N. The Process for the Selection of ELIXIR Core Data Resources. F1000Res. 2018;7. doi:10.7490/f1000research.1116248.1
7. ELIXIR announces initial list of Core Data Resources and Deposition Databases | ELIXIR [Internet]. [cited 24 May 2019]. Available: https://elixir-europe.org/news/core-data-resources
8. BRENDA and SILVA named ELIXIR Core Data Resources | ELIXIR [Internet]. [cited 24 May 2019]. Available: https://elixir-europe.org/news/brenda-and-silva-elixir-core-data
9. Orphadata selected as ELIXIR Core Data Resource | ELIXIR [Internet]. [cited 24 May 2019]. Available: https://elixir-europe.org/news/orphadata-selected-elixir-core-data-resource
10. Text mining patterns used to mine the CDR data resource accession numbers are available on Github: https://github.com/EuropePMC/EuropePMC-Identifier-Extractor/blob/master/automata/acc181210.mwt
11. The text mining patterns used to mine the CDR resources names are available on GitHub: https://github.com/EuropePMC/EuropePMC-Identifier-Extractor/blob/master/automata/resources180904.mwt
12. The source code for the text mining pipeline element is available on GitHub: https://github.com/EuropePMC/EuropePMC-Identifier-Extractor
13. The source code for the Annotation API is available on GitHub:

https://github.com/EuropePMC/Annotations-API

14. The password-protected data visualisation application developed on top of the “Core Data Resource Statistics” master spreadsheet is available here: https://devmaster.vital-it.ch/elixir-data-visualisations/

15. Hancock JM, Game A, Ponting CP and Goble CA. An open and transparent process to select ELIXIR Node Services as implemented by ELIXIR-UK. F1000Res. 2017;5(ELIXIR):2894. doi.org/10.12688/f1000research.10473.2

16. ELIXIR Recommended Interoperability Resource (RIR) Selection [Internet]. [cited 11 July 2019]. Available: https://elixir-europe.org/platforms/interoperability/rir-selection

17. W. Anderson, R. Apweiler, A. Bateman, G.A. Bauer, H. Berman, J.A. Blake, et al. Towards Coordinated International Support of Core Data Resources for the Life Sciences. bioRxiv 110825; doi: https://doi.org/10.1101/110825

8.6 Annex 1 - CDR Indicators Annual Update Form

CDR Indicators Annual Update Form: [Resource Name]

Document owner: [Name] [Email Address]

Please complete this Update Form by adding information for your data resource to the 2016 and 2017 columns. We have stocked the 2013, 2014, 2015 columns with the values you provided as part of the initial selection (possibly with adjustments deemed necessary as we worked with you to drive out ambiguities from the first round of collection) to assist you.

Please address any questions to Rachel Drysdale rachel.drysdale@elixir-europe.org

Reference:
Identifying ELIXIR Core Data Resources: Durinx C, McEntyre J, Appel R et al. F1000Research 2017, 5(ELIXIR):2422 (doi: 10.12688/f1000research.9656.2)

Indicator: Full Time Equivalent

Table 1

Number of FTE	2013	2014	2015	2016	2017
Total Staff	xxx	xxx	xxx		
Curators ❑ support for submission adherence to metadata requirements ❑ support for extraction of information from the scientific literature					

Bioinformaticians *Technical staff*					

Indicator: Web Access - Unique visitors/Users

Table 2

Access via a web browser: Please indicate methodology

Either

- ○ Web page tagging (e.g. Google Analytics, Matomo/Piwik PRO) YES/NO

or

- ○ Web log analysis (e.g. Matomo/Piwik PRO, AWStats) YES/NO

Average monthly web traffic ...	in 2013	in 2014	in 2015	in 2016	in 2017
Unique visitors per month (users)	xxx	xxx	xxx		

Indicator: Web Access - Visits/Sessions

Table 3

Access via a web browser: Web page tagging (e.g. Google Analytics, Matomo/Piwik PRO)

Average monthly web traffic in...	in 2013	in 2014	in 2015	in 2016	in 2017
Visits/Sessions per month	xxx	xxx	xxx		

Indicator: Web Access - Page Views

Table 4

Access via a web browser: Web page tagging (e.g. Google Analytics, Matomo/Piwik PRO)

Average monthly web traffic in...	in 2013	in 2014	in 2015	in 2016	in 2017

Page views per month	xxx	xxx	xxx		

Indicator: Web Access - Requests/Hits

Table 5

Access via a web browser: Web log analysis (e.g. Matomo/Piwik PRO, AWStats)

Average monthly web traffic ...	in 2013	in 2014	in 2015	in 2016	in 2017
Hits per month	xxx	xxx	xxx		

Indicator: Downloads - Unique IPs/Hosts

Table 6

FTP, APIs etc: Log analysis (e.g. Matomo/Piwik PRO, AWStats)

Average monthly downloads ...	in 2013	in 2014	in 2015	in 2016	in 2017
Unique IP addresses / Hosts per month	xxx	xxx	xxx		

Indicator: Downloads - Number of Items/Entries/Files

Table 7

FTP, APIs etc: Log analysis (e.g. Matomo/Piwik PRO, AWStats)

Average monthly downloads ...	in 2013	in 2014	in 2015	in 2016	in 2017
Hits / Requests per month	xxx	xxx	xxx		

Indicator: Downloads - Data Transfer Volume

Table 8

FTP, APIs etc: Log analysis (e.g. Matomo/Piwik PRO, AWStats)

Average monthly downloads …	in 2013	in 2014	in 2015	in 2016	in 2017
Data transfer per month (GB)	xxx	xxx	xxx		

Indicator name: Entries - Number

Table 9

Cumulative data …	in 2013	in 2014	in 2015	in 2016	in 2017
Total number entries/ records/ depositions/ assays	xxx	xxx	xxx		

Indicator name: Entries - Total Size

Table 10

Cumulative data …	in 2013	in 2014	in 2015	in 2016	in 2017
Total Size in GB	xxx	xxx	xxx		

Indicator: Citations - Resource name mentions in the literature

We generate reports from EuropePMC for this Indicator, using patterns shown in the "Core data Resource names and Accession number mining" spreadsheet[1]. Please review the pattern that corresponds to your data resource, and if updates are necessary please contact rachel.drysdale@elixir-europe.org.

Indicator: Citations - Resource accession number mentions in the literature

We generate reports from EuropePMC for this Indicator, using patterns shown in the "Core data Resource names and Accession number mining" spreadsheet. Please review the pattern that corresponds to your data resource, and if updates are necessary please contact rachel.drysdale@elixir-europe.org.

Indicator: Citations - of key data resource articles

We generate reports from EuropePMC for this Indicator, based on a list of up to five key publications about your resource, such as Nucleic Acid Research Database issue publications, or similar. The Table shows the key publications we have listed for your resource - please amend if you would like to update this list:

PMID	Title	Year	NAR DB issue?	Please amend if you make a change
nnnnnnnn	xxx	xxx	Y/N	Retain
nnnnnnnn	xxx	xxx	Y/N	Retain
nnnnnnnn	xxx	xxx	Y/N	Retain
nnnnnnnn	xxx	xxx	Y/N	Retain
nnnnnnnn	xxx	xxx	Y/N	Retain

[1] https://docs.google.com/spreadsheets/d/1x-nhzUIiS_Ff0WUG2bZP6y9fKUVdgQA_GUdaRdJOfx4/edit?usp=sharing

8.7 Supplementary data - Technical Specification: Core Data Resource Indicators for Annual Collection (internal document)

Technical Specification:
Core Data Resource Indicators for Annual Collection

All hyperlinks have been removed from this document as they gave access to sensitive data.

Purpose of this document

This document lists the Indicators that we will collect or generate on an annual, or on-demand, basis. For each Indicator it states their nature, measuring methodology and associated caveats, and includes relationships to the CDR Usage Stats Spreadsheet and CDR Indicators Annual Update Form - for Internal EXCELERATE WP3 Task 3.2 use. This document was prepared by Rachel Drysdale, Nicole Redaschi, Heinz Stockinger, and Rodrigo Lopez, with input from the wider Task 3.2 group.

Sections from the CDR Usage Stats Spreadsheet that are not included in this list are:

- **Archive and/or Knowledgebase (Case Document Indicator 1a)** - this is a characteristic of the resource that is not generally speaking dynamic, and there is no value to its staying the same year on year. If it were to change, that would be interesting, but of no value to track in a longitudinal manner.
- **Web access (monthly average): Sessions and pages - Log analytics (Case Document Indicator 2a)** - this indicator was reported in Case Documents for only two of the 16 Core Data Resources selected in the first round of selection, and one of those two also reported a Google Analytics version also reported by the majority of the Core Data Resources. As such, this metric is of little value as an Indicator.
- **Persistent Identifiers (Case Document Indicator 3a)** - this is a characteristic of the resource that is not generally speaking dynamic, and of no value to track in a longitudinal manner.

Format of this document

Sections for each Indicator include:

- **Indicator name**
- **Corresponding columns in CDR Usage Stats Spreadsheet** used to house the data
- **Corresponding table in "CDR Indicators Annual Update Form"**
- **Description of Concept that Indicator intends to express**
- **Method used to measure the indicator**
- **Caveats** - what is actually being measured and what it might not represent, and how (more or less directly) they relate to the Concept

Table of Contents:

Indicator name: Full Time Equivalent

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - FTE (Case Document Indicator 1d)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - Table number 1
- **Description of Concept that Indicator intends to express**
  - Scale of human effort required to maintain the resource.
- **Method used to measure the indicator**
  - Staff measured as Full Time Equivalents.
- **Caveats**
  - FTE count varies over the year - until such time as a specific protocol for reporting FTE is developed that takes this into account, the values reported by different resources are likely not to be strictly comparable.
  - FTE does not list the *number of people* but the equivalent number of full time positions, i.e., two people working 50% each count as 1 FTE.

Indicator name: Web Access - Unique visitors/Users

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - ○ Web access (monthly average): Unique visitors/Users - Google Analytics (Case Document Indicator 2a)
  - ○ Web access (monthly average): Unique visitors/Users - Log analytics (Case Document Indicator 2a)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - ○ Table number 2
- **Description of Concept that Indicator intends to express**

  This indicator is used to measure how many distinct individuals (users) access a website over a specified period of time, regardless of how often they visit.
- **Method used to measure the indicator**[1]

Unique visitors/Users can be determined in different ways:

- ○ Web **page tagging** ("web analytics"): uses tags that are embedded in web pages and cookies that are stored on a person's device to identify users. Specific examples include Google Analytics, Matomo[2] (formerly Piwik) and PIWIK PRO[3].
- ○ Web **log analysis** ("log analytics"): uses the IP address of a person's device (optionally combined with the user agent used, such as the type of web browser) that is recorded in a Web server's log file to identify users. Specific examples include Matomo (formerly Piwik), PIWIK PRO and AWstats[4].

- **Caveats**

Both Web page tagging and Web log analysis use approximations to estimate the number of real users:

- Page tagging counts only interactive users that look at a web page with a browser. Because it relies on cookies that are stored on a person's device, it may count the same person as several users: once for each device that is used by that person, and as a new user each time the cookies are deleted.
- Log file analysis uses the information that a Web server logs for each request. This covers interactive and programmatic users, as well as robots like search engines. The separation of these three classes of users is difficult and requires heuristic methods that may vary between different analysis softwares. Like with page tagging, the same person is counted as multiple users when using multiple devices with different IP addresses or a device with dynamic IP addressing. Conversely, many persons may appear to have the same IP address, if their institutional network shows only one or a few IP addresses to the outside world.
  - ○ For the above mentioned reasons, metrics determined with different

methods, or the same method but different analysis software[1], must be compared with great caution.

Indicator name: Web Access - Visits/Sessions

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - Web access (monthly average): Visits / Sessions - Google Analytics (Case Document Indicator 2a)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - Table number 3
- **Description of Concept that Indicator intends to express**

  A session, also referred to as a visit, is a set of requests/interactions by the same uniquely identified visitor/user (commonly identified by an IP address or a unique ID placed in the browser), who has not visited the site recently (typically, within the past 30 minutes). The number of sessions is a measure of how much traffic a website gets. A visit is considered a visit as long as the events (individual page requests for example) are 30 minutes or less apart. If a user visits a site at noon and then again at 15:00, that counts as two visits. A visit can consist of one page view or many (practically, there is no limit).
- **Method used to measure the indicator**
  - Web page tagging (e.g. Google Analytics, Matomo (formerly Piwik), PIWIK PRO)
- **Caveats**
  - The number of visits is under-counted if users return within 30 mins on the same device to conduct two entirely different "tasks" (which should in fact be counted two different sessions).
  - The number of visits is over-counted if the user switches between devices to continue the same task.

Indicator name: Web Access - Page Views

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - Web access (monthly average): Pageviews - Google Analytics (Case Document Indicator 2a)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - Table number 4
- **Description of Concept that Indicator intends to express**

  Pageviews (also known as impressions) correspond to a request to load a single HTML file (web page) of a website, identified by the URL in a browser. During a visit or session, a person can access several different pages of a website, which results in several pageviews.
- **Method used to measure the indicator**
  - Web page tagging (e.g. Google Analytics, Matomo (formerly Piwik), PIWIK PRO)
- **Caveats**
  - **Note about pageviews and log analysis:** Pageviews can be estimated by using log analytics by filtering HTML files only, and this works for traditional websites (one single HTML page). However, with the emerging web technology of HTML5, one single page may request several HTML files (partials) and therefore this indicator should not be computed via log analytics for websites using technology such as AngularJS, web components, etc. For this reason - because we generally do not have that much technical detail from the CDRs - we do not use log analytics to measure Page views.
  - This indicator varies considerably based on a website's design (e.g. a complex page with many sections vs separate pages for each section) and technology (e.g. a "Single Page Application" provides a single addressable URL even if many tasks can be achieved on the same page; otherwise, each task could be implemented as a respective page with a distinct URL).

Indicator name: Web Access - Requests/Hits

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - ○ Web access (monthly average): Hits - Log analytics (Case Document Indicator 2a)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - ○ Table number 5
- **Description of Concept that Indicator intends to express**

  Hits or requests refer to the number of files requested and downloaded when a web page is viewed. A web page is typically made up of a number of individual files such as HTML documents, images, JavaScript files (i.e., .html, .css, .js, .png, .jpg, .xml, .json, .txt etc.). When a web page is viewed, each of these files is requested from the web server, adding up to the hit-count for the website. For example, the website www.bbc.com needs more than 150 files to render the home page and therefore generates more than 150 hits for one single pageview.
- **Method used to measure the indicator**
  - ○ Web log analysis (e.g. Matomo/PIWIK PRO, AWStats)
- **Caveats**
  - ○ This value may differ considerably depending on a website's design and technology.
  - ○ Because this measure is so dependent on the specifics of the website design, this indicator can be used to analyse trends of a specific resource over time but it is not adequate to compare between various resources whose designs might be significantly divergent.
  - ○ Another issue is caching: web browsers can cache visited pages in order to avoid requesting the same page again. Web log analysis will miss such requests/hits.

Indicator name: Downloads - Unique IPs/Hosts

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - Data Downloads (monthly average) - FTP, etc: Unique IPs/Hosts (Case Document Indicator 2a)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - Table number 6
- **Description of Concept that Indicator intends to express**

  This indicator measures the number of Unique IP addresses/Hosts via which Downloads are made.
- **Method used to measure the indicator**
  - Log analysis (e.g. Matomo/PIWIK PRO, AWStats)[1]
- **Caveats**
  - "Downloads" data is generally associated with file servers (e.g. FTP, Aspera, rsync), but can also refer to items/entries downloaded via RESTful and SOAP API calls over HTTP.
  - For websites that offer both an interactive user interface and a programmatic interface (RESTful and SOAP APIs), it can be difficult to determine the proportion of API interface usage that should be reported for this indicator.
  - Server type is relevant to the calculation of total values e.g. Aspera traffic relates to/includes a high percentage of data exchange between service providers/partners (e.g. NCBI/EMBL/DDBJ, EBI/SIB/PIR) and does not only represent data downloaded by individual users.

Indicator name: Downloads - Number of Items/Entries/Files

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - ○ Data Downloads (monthly average) - FTP, etc: Number of Items/Entries/Files (Case Document Indicator 2a)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - ○ Table number 7
- **Description of Concept that Indicator intends to express**

  This indicator measures the number of items/entries/files downloaded.
- **Method used to measure the indicator**
  - ○ Log analysis (e.g. Matomo/PIWIK PRO, AWStats)
- **Caveats**
  - ○ "Downloads" data is generally associated with file servers (e.g. FTP, Aspera, rsync), but can also refer to items/entries downloaded via a RESTful and SOAP API calls over HTTP.
  - ○ For websites that offer both an interactive user interface and a programmatic interface (RESTful and SOAP APIs), it can be difficult to determine the proportion of API interface usage that should be reported for this indicator.

Indicator name: Downloads - Data Transfer Volume

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - ○ Data Downloads (monthly average) - FTP, etc: Data transfer (GB) (Case Document Indicator 2a)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - ○ Table number 8
- **Description of Concept that Indicator intends to express**

  This indicator measures the size of the data downloaded from resource in terms of volume / bandwidth (commonly measured in GB).
- **Method used to measure the indicator**
  - ○ Log analysis (e.g. Matomo/PIWIK PRO, AWStats)
- **Caveats**
  - ○ "Downloads" data is generally associated with file servers (e.g. FTP, Aspera, rsync), but can also refer to items/entries downloaded via a RESTful and SOAP API calls over HTTP.
  - ○ For websites that offer both an interactive user interface and a programmatic interface (RESTful and SOAP APIs), it can be difficult to determine the proportion of API interface usage that should be reported for this indicator.

Indicator name: Entries - Number

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - Total Entries (Case Document Indicator 3b)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - Table number 9
- **Description of Concept that Indicator intends to express**
  - Scale of resource as measured by total number of entries.
- **Method used to measure the indicator**
  - Counts provided by the data resource.
- **Caveats**
  - Each data resource has its own definition of "entry" - some resources have several types of entries. In comparing year on year the same item must be counted each time, for comparability.
  - The distinction between curated knowledgebases and deposition archives is pertinent to assessments of comparisons based on this indicator. For example:
    - UniProtKB: growth for TrEMBL is much bigger (automatic annotation) than for Swiss-Prot entries (manual curation) year by year. Each increase has a different value, and should be regarded in that context.

Indicator name: Entries - Total Size

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - Size (GB) (Case Document Indicator 3b)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - Table number 10
- **Description of Concept that Indicator intends to express**
  - Scale of resource as measured by total size of data, in Gigabytes.
- **Method used to measure the indicator**
  - Size in GB provided by the data resource.
- **Caveats**
  - In comparing year on year the same protocol for determining “size” must be employed, for comparability.
  - Many resources distribute their data in several data formats that may vary greatly in size (also depending on whether and how the data is compressed). This indicator therefore does not reflect the importance of the resource, but the trend in amount of data it needs to manage, over time.

Indicator name: Citations - Resource name mentions in the literature

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - Resource Name Mentions in Europe PMC - Number of articles per year (Case Document Indicator 2c)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - Not applicable - To be generated by Europe PMC for all CDRs. The pipeline runs daily - reports can be generated as required.
- **Description of Concept that Indicator intends to express**

  Value of each resource in scientific research as indicated by direct mentions of the corresponding resource name in open access full-text publications.
- **Method used to measure the indicator**
  - The resource name is identified using a pattern based approach.
    - A dictionary with the patterns/variants that correspond to the resource name (e.g. to include stemming relationships or abbreviated forms) is applied to all sections in the full-text articles under examination (i.e., for a given date range) to recognise the resource name mentions in the text and tag them.
    - The total number of resource names in the corpus of articles is counted, and from that output the unique number of articles mentioning the resource name is counted.
    - The patterns/variants mined are reviewed periodically by the data resource managers, to ensure optimal retrieval.

- **Caveats**
  - Due to the different nature of the resource names, and the impact of that on precision/recall, it is unlikely that there will be parity in retrieval across the full set of data resources.

Indicator name: Citations - Resource accession number mentions in the literature

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - Resource Accession Numbers Mentions in Europe PMC - Number of articles per year (Case Document Indicator 2c)
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - Not applicable - To be generated by Europe PMC for all CDRs. The pipeline runs daily - reports can be generated as required.
- **Description of Concept that Indicator intends to express**

  Value of each resource in scientific research as indicated by direct mentions of the corresponding resource accession numbers in open access full-text articles.
- **Method used to measure the indicator**
  - The resource accession numbers are identified using a pattern based approach.
    - A dictionary with the regular expression that corresponds to the resource accession number pattern is applied to all sections of the full-text articles under examination (i.e., for a given date range).
    - The accession numbers occurring within the same sentence as a corresponding resource name mention are tagged, and validated as being current (i.e., non-deprecated).
    - The total number of validated accession numbers in the corpus of articles is counted, and from that output the unique number of articles mentioning the resource accession numbers are counted.
    - The regular expressions are reviewed periodically by the data resource managers, to ensure optimal retrieval

- **Caveats**
  - Due to the different nature of the resource accession numbers, and the effect on the power of the regular expression to identify them, it is unlikely that there will be parity in retrieval across the full set of data resources.
  - Not all accession number mentions will necessarily occur in the same sentence as the corresponding resource name, so the article counts may be an underestimate.
  - Some data resources share accession number space, so we cannot always uniquely assign accession numbers tagged to corresponding resource e.g. the European Variation Archive (EVA) reuses a subset of accession numbers from the European Nucleotide Archive (ENA).

Indicator name: Citations - of key data resource articles

- **Corresponding columns in CDR Usage Stats Spreadsheet**
  - ○ Not currently represented in Usage Stats Spreadsheet - this new spreadsheet was compiled for Global Coalition business case purposes
- **Corresponding table in "CDR Indicators Annual Update Form"**
  - ○ Not Applicable - To be generated at the Hub, annually.
- **Description of Concept that Indicator intends to express**
  - ○ Significance of data resource within research and service provision sectors, as indicated by direct citation of significant publications specifically about the data resource and its operations.
- **Method used to measure the indicator**
  - ○ Citation counts as retrieved from Europe PMC for articles nominated by the resource providers themselves as being (up to) the top five most significant articles about the resource.
- **Caveats**
  - ○ Not all resources providers contribute five significant articles to the list, so coverage of resources is differentially represented.
  - ○ Some resource providers will necessarily report the five most-cited - others may choose to include their five most significant. The five most significant will not necessarily be the same as the five most cited.

[1] Various free, open source and proprietary software and hosted services exist for these methods (see https://en.wikipedia.org/wiki/List_of_web_analytics_software for an overview). Most implement only one method, a few offer both. Some institutions build their own customized solutions (e.g. based on “Elastic Stack”).

[2] https://matomo.org/log-analytics/

[3] https://help.piwik.pro/category/web-log-analytics/

[4] https://www.awstats.org/