EXCELERATE Deliverable 3.1

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:	List of metrics/quality criteria measuring the scientific impact, service usage, service delivery, and eligibility for “ELIXIR Named Resource” and “ELIXIR Core Resource” labels, to allow construction and extension of the ELIXIR Resource portfolio	
WP No.	3	
Lead Beneficiary:	EMBL (EBI)	
WP Title	Data Resources and Services	
Contractual delivery date:	31 August 2016	
Actual delivery date:	31 August 2016	
WP leader:	Jo McEntyre (EBI), Christine Durinx (SIB)	1: EMBL
Partner(s) contributing to this deliverable:	n/a	

Authors and contributors: Jo McEntyre, Christine Durinx, ELIXIR partners

Table of contents

1. Executive Summary

The core mission of ELIXIR is to build a stable and sustainable infrastructure for biological information across Europe. At the heart of this are the data resources, tools and services that ELIXIR Nodes offer to the life science community, providing stable and sustainable access to biological data. These resources vary from archives or deposition databases that contain research data outputs such as DNA sequences, to highly dynamic knowledge bases that aggregate, process and visualize research data, often adding layers of value through manual curation by highly qualified personnel. 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.

There are over 1000 data resources in Europe. Only a small fraction of these have institutional support and long-term funding commitments. The fact that the survival of many crucial bioinformatics resources in the mid- and long-term is not guaranteed is threatening the foundations of academic and industrial life science activities, and risks the loss of an immense wealth of biological and medical information as well as the associated investments.

Identifying ways to assess the quality and impact of these crucial data resources will (a) promote excellence in resource development and operation to support capacity building through spreading best practice, and (b) provide a basis for technical and science policy actions required to support the long-term sustainability of the data resources that are the backbone of bioinformatics infrastructure.

The ELIXIR Core Data Resources are defined as a set of European data resources that are of fundamental importance to the broad life science community and the long-term preservation of biological data. They provide complete collections of generic value to life science, are considered an authority in their field with respect to one or more characteristics, and show high levels of scientific quality and service. Thus, ELIXIR Core Data Resources are of wide applicability and usage.

The ELIXIR Core Data Resources will form the focal point of technical and science policy actions to drive long-term sustainability. Transparent indicators for the ELIXIR Core Data Resources will also provide strategic intelligence on resource quality and impact, notably to policy makers and funders.

The identification of the ELIXIR Core Data Resources involves a careful evaluation of the multiple facets of the data resources.

The indicators are grouped in five 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

The indicators recognise the heterogeneous nature of biological data, and the diversity of the supporting data resources, use cases and communities served. Indicators can be used to measure technical and/or scientific readiness of a resource compared to desirable levels of quality standards.

2. Project objectives

This deliverable fulfils one of the Ethics requirements of EXCELERATE.

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

No.	Objective	Yes	No
1	Build a framework to inform and drive the sustainable development of Europe's core life-science data resources	x	
2	Promote excellence in resource development and operation through providing a unified framework for the identification and monitoring of key bioinformatics resources across Europe	x	
3	Increase the sustainability of manually curated resources	x	

3. Delivery and schedule

The delivery is delayed: ☐Yes ☑ No

4. Adjustments made

No adjustments were made.

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

Participant number and person months per participant

1 - EMBL (85), 7 - CNIO (12), 9 - CIPF (3.6), 14 - UPF (14.1), 15 - IMIM (5.5), 25 - SIB (66)

Objectives

The overall objective of this Work Package (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)

Description of work and role of partners

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 (24PM)

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 (76.1PM)

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 (86.1PM)

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: EMBL-EBI, CH, ES

Annex 1: Identifying ELIXIR Core Data Resources and ELIXIR Services

Annex 2: Quantitative and qualitative indicators for the ELIXIR Core Data Resources

Identifying ELIXIR Core Data Resources and ELIXIR Services

1. Situating ELIXIR Core Data Resources and ELIXIR Services within the ELIXIR Mission

The core mission of ELIXIR is to **build a stable and sustainable infrastructure for biological information across Europe**. At the heart of this are the data resources, tools and services that ELIXIR Nodes offer to the life science community, providing stable and sustainable access to biological data. These resources vary from archives or deposition databases that contain research data outputs such as DNA sequences, to highly dynamic knowledge bases that aggregate, process and visualize research data, often adding layers of value through manual curation by highly qualified personnel. 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.

There are over 1000 data resources in Europe. Only a small fraction of these have institutional support and long-term funding commitments. The fact that the **survival of many crucial bioinformatics resources in the mid- and long-term is not guaranteed** is threatening the

foundations of academic and industrial life science activities, and risks the loss of an immense wealth of biological and medical information as well as the associated investments.

Identifying ways to assess the quality and impact of these crucial data resources will (a) **promote excellence in resource development and operation to support capacity building** through spreading best practice, and (b) provide a basis for **technical and science policy actions required to support the long-term sustainability of the data resources that are the backbone of bioinformatics infrastructure** (Figure 1).

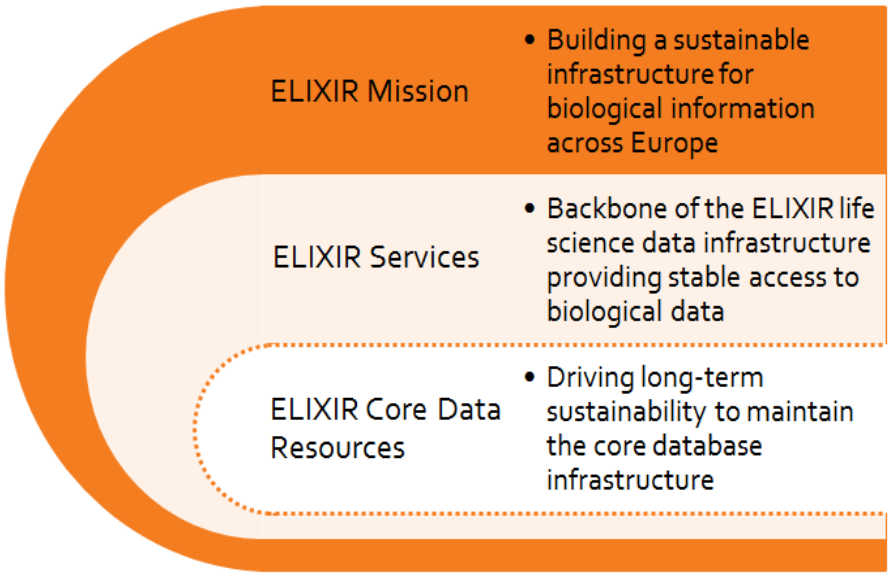

Figure 1. The place of ELIXIR Services and ELIXIR Core Data Resources within ELIXIR's mission.

The proposal for establishing ELIXIR Services and ELIXIR Core Data Resources was put to the ELIXIR Scientific Advisory Board (SAB) in December 2014 [1]. This document describes how to take the proposal into practice and provides the guidelines for the implementation of life-cycle management.

ELIXIR Nodes define, through their Node applications and Service Delivery Plans (in the case of EMBL-EBI, the 'Work Programme'), a set of services and data resources that are offered to the research community, the **ELIXIR Services**. These resources form **the backbone of the life science data infrastructure**.

The **ELIXIR Core Data Resources** are defined as a set of European *data* resources that are of fundamental importance to the broad life science community and the long-term preservation of biological data. They provide complete collections of generic value to life science, are considered an authority in their field with respect to one or more characteristics, and show high levels of scientific quality and service. Thus, ELIXIR Core Data Resources are of wide applicability and usage.

Core data resources tend to be well known within the life science community and also reach through to key stakeholders such as funders and journals. ELIXIR Core Data Resources are well maintained with a professional service delivery plan based on well-established life-cycle management processes and well-understood dependencies with related data resources. The ELIXIR Core Data Resources coexist with a broader range of databases with diverse motivations, often specialising in a particular scientific topic.

The ELIXIR Core Data Resources will form the **focal point of technical and science policy actions to drive long-term sustainability**. Transparent indicators for the ELIXIR Core Data Resources will also provide **strategic intelligence on resource quality and impact, notably to policy makers and funders**.

Through the ELIXIR Scientific Programme and ELIXIR-EXCELERATE grant, the infrastructure will deliver and enable a range of initiatives that aim to support and strengthen the ELIXIR Services and ELIXIR Core Data Resources. ELIXIR Services and ELIXIR Core Data Resources will be the most widely used and outwardly visible part of ELIXIR. Establishing the portfolio of these data resources and services is the key priority for ELIXIR and publicly marks the transition towards a cohesive infrastructure. Through the establishment of the ELIXIR Services portfolio, ELIXIR also aims to support and implement best practice in resource management and bring European bioinformatics resources to the next level, building confidence among users.

2. The life-cycle of ELIXIR Services

This section outlines the framework for life-cycle management of the ELIXIR Services, defining the stages within their technical life cycle (Table 1). Through the ELIXIR-EXCELERATE Node Capacity Building and Communities of Practice (WP10) and Training Programme (WP11) work packages, this framework will be put into practice, thus strengthening the ELIXIR infrastructure by creating a stairway to excellence.

Table 1. Stages of the technical life cycle of ELIXIR Services.

Stage	Definition	Status
Emerging	A resource in active development towards maturity. Emerging Services may have lower reliability compared to Mature Services and go through more changes in their presentation and APIs. 'Emerging' status should not exceed 2 years. If an Emerging Service does not become Mature, it is good practice to display an "end of service" date prominently for at least 6 months before being withdrawn.	ELIXIR Emerging Services
Mature	An ELIXIR Service that has passed the development stage. It is active, i.e. new data are currently integrated, and reliable. If feasible, major changes to its API and/or user interface that may break existing functionality and/or are not fully backwards compatible, are notified at least 6 months on beforehand. A Mature Service relies only on other Mature or Legacy Services. In exceptional cases, a Mature Service might rely on an Emerging Service that is close to becoming mature. It is good practice that withdrawal of the Service is notified at least 1 year on beforehand, a period during which the Service has Legacy status.	ELIXIR Services
Legacy	A previously Mature Service scheduled for archiving or decommissioning. A Service must spend at least 1 year in the Legacy state before final withdrawal. Reliability should be at the	ELIXIR Services - Legacy

	same level as Mature Services, but compromises on content (e.g. data not updated, no new content is added) are allowed.	

The agreed set of indicators for the ELIXIR Core Data Resources sets quality standards that can guide and inform the managers of Emerging Services in the development of their Resource towards an "ELIXIR Service" status.

The monitoring of usage trends and the scientific impact of the ELIXIR Services further provides information to support their life cycle management, contributing to the maintenance of the ELIXIR Service status, or – where appropriate - leading a resource towards the Legacy stage.

3. Measuring the quality and impact of data resources

3.1 Key indicators for data resources

As pointed out by Wilsdon et al. (2015) in their report on the role of metrics in research assessment and management in the United Kingdom [2], the term 'metric' can be open to misunderstanding. For example, the number of citations received by a publication is a citation metric, not an impact metric, as it does not directly measure the impact of that researcher's work. In other words, it suggests "measurement" of a quantity or quality which has not in fact been measured.

Wilsdon et al. (2015) propose to use the term "**indicator**" in contexts in which there is the potential for confusion. The term "indicator" is defined as a measurable quantity that 'stands in' or substitutes for something less readily measurable and is presumed to associate with it without directly measuring it. Citation counts could be used as indicators for the scientific impact of journal articles even though scientific impacts can occur in ways that do not generate citations. **We will therefore use the term "indicators" throughout this document.**

3.2 Five categories of indicators, reflecting the multiple facets of data resources

The identification of the ELIXIR Core Data Resources involves a careful evaluation of the multiple facets of the data resources.

The indicators are grouped in five 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

Figure 2. A carefully chosen basket of qualitative and quantitative indicators for bioinformatics resources.

When collecting and interpreting the indicators, it is important to articulate the methods used and standardise terminology where possible. This facilitates the understanding of the indicators and avoids misinterpretation across different nodes.

(1) Scientific focus and quality of science

This includes the inherent scientific quality of the resource, its uniqueness and comprehensiveness. Other indicators included here are the positioning of the resource with other similar resources, if applicable, and whether the resource is an authority in its field.

A differentiation should be made between *archival or deposition databases* that receive and archive *de novo* data sets and well-structured metadata deposited by scientists, and added-value databases or *knowledge bases*, which are based on the archival data and add substantial value through expert curation, annotation of metadata, sophisticated data processing and/or data integration. The curation effort and curation outputs linked to a certain resource is an important measure of its quality.

(2) Community

This category reflects the size and the measured demand of the communities that are served by the resource: web statistics, user reach, and international use. The community that is served can be the depositors, as some resources are vital for deposition and/or the end-users. There are different ways to identify and measure the community, such as by measuring access to URLs, to download servers, and through APIs, and also through the citation of data and data resources in publications. In addition, certain resources play a foundational role to derived resources and/or services that rely on their existence.

The scientific context in which the resource operates should be taken into account. A resource that serves a small scientific community may not have as many users as a resource serving a broader interest, and yet it may reach 90% of the community it supports (coverage) and be crucial for the scientific work of that community.

(3) Quality of service

Certain service levels and reliability can be quantified with specific technical indicators such as the uptime of the resource, response times in general, periodic application of meaningful and automated tests, user support and related training, use of community-recognised standards, diversity of data retrieval mechanisms, and so on. Usually, this requires a quality assurance process during service development and operation. This training aspect is related to the ELIXIR-EXCELERATE Work Package 11 (Accelerating the ELIXIR Training Programme) and the ELIXIR Training Platform that help resources delivering training and provide evaluation systems and good practice guidelines for training.

(4) Legal and funding infrastructure, and governance

As stable research infrastructures, Core Data Resources can demonstrate that they have a sound legal, funding and governance structure.

A viable resource has a suitable legal framework (clear terms of use, licensing, data security etc.). Open data is a critical driver for life sciences research and therefore for ELIXIR, but the policy of the access to data must be considered in the light of the existence of funding for the resource. A resource's commitment to longevity is shown by its institutional support, by its funding schemes and long(er) term financial stability. Core Data Resources have demonstrated that they can manage transitions through different funding sources. A strong governance structure includes an international, independent Scientific Advisory Board (SAB), which allows community input and provides a permanent oversight framework.

(5) Impact and translational stories

Impact evaluation attempts to provide a definite answer to the question of whether the Resource is meeting its objective of fulfilling a specific need of the scientific community. In the UK, the HM Treasury's Magenta Book provides guidelines for policy makers and analysts, on how policies and projects should be assessed and reviewed. According to this guidance, the key characteristic of a good impact evaluation is that it recognises that most needs can be met by a range of elements, not just the policy (or, in ELIXIR's case, resource) [3]. To test the extent to which the Resource is

responsible for meeting the need, it is necessary to estimate – usually on the basis of an (often quite technical) statistical analysis of quantitative data – what would have happened in the absence of the Resource. This is known as the counterfactual. Establishing the counterfactual is not easy, since by definition it cannot be observed – it is what would have happened if the Resource did not exist. A strong evaluation is successful in isolating the effect of the Resource from all other potential influences, thereby producing a good estimate of the counterfactual.

When communicating the impact of ELIXIR's resources and their role in accelerating science to funders and the general public, the indicators should be made relevant to the audience. This can be done by "translating" them to information the audience is familiar with e.g. *x* visits daily to a database is comparable with the number of people accessing the BBC web page on a daily basis.

3.3 Detailed description of the indicators and related methodology

For each of the five categories, the indicators and the related information are described in Appendix 1 of this document. In Appendix 2, a "Case Document" template is provided that can be used to describe a data resource through these indicators.

3.4 Indicators to support expert judgment

Taking into account that "**Not everything that can be counted counts, and not everything that counts can be counted**" (William Bruce Cameron), the different indicators will be used to **inform a peer-review process** that is described further in this document.

A carefully chosen set of qualitative and quantitative indicators, tailored to the specific situation of bioinformatics resources, will inform the identification process of the ELIXIR Core Data Resources. The indicators will support, but not supplant, expert judgment.

ELIXIR Core Data Resources should have an international independent Scientific Advisory Board (SAB) in place. These SABs consist of distinguished academic and industry researchers and professionals who conduct the scientific and/or technological review, ensuring scientific quality and providing strategic advice to the resources' management. The identification of ELIXIR Core Data Resources does not encroach on these governance structures. The establishment of Core Resource and Node SABs is part of the best practices that will be disseminated by the Node Capacity Building and Communities of Practice work package (WP10).

Indicators can only meet their potential if they are underpinned by an open, transparent and coherent indicator collection infrastructure, therefore it is important that the methods of collection and processing are clear.

4. Examples of ELIXIR Core Data Resources

Starting from the definition of "ELIXIR Core Data Resources" above, we identified a "seed list" of candidate core resources (Table 2) that we used to inform the portfolio of Core Data Resource indicators.

Table 2. Examples of European data resources that are unequivocally considered as core for the life science community.

Resource name	Institutions	Type

UniProt	EMBL-EBI (European Molecular Biology Laboratory – European Bioinformatics Institute); SIB Swiss Institute of Bioinformatics; Protein Information Resource (PIR) - Georgetown University Medical Centre	Knowledgebase of proteins
European Nucleotide Archive (ENA)	EMBL-EBI, in the framework of the International Nucleotide Sequence Database Collaboration (INSDC)	Comprehensive archive of nucleotide sequences, annotations and associated data
PRIDE (Proteomics identifications database)	EMBL-EBI	Archive of mass spectrometry based proteomics data
Europe PubMed Central (Europe PMC)	EMBL-EBI; Jisc; University of Manchester (Mimas and NaCTeM); British Library	Archive for full-text biomedical and life sciences journal articles
InterPro	Consortium of databases based at EMBL-EBI, EMBL Heidelberg; SIB Swiss Institute of Bioinformatics; WTSI; University of Manchester; PRABI; J. Craig Venter Institute, Rockville ; PIR ; University of Bristol; University College London; University of Southern California	Knowledgebase of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)
Protein Data Bank in Europe (PDBe)	EMBL-EBI, in collaboration with the Worldwide Protein Data Bank (wwPDB) and EMDataBank partners	Protein structure knowledgebase
Human Protein Atlas	AlbaNova and SciLifeLab, KTH - Royal Institute of Technology, Stockholm, Sweden, the Rudbeck Laboratory, Uppsala University, Uppsala, Sweden and Lab Surgpath, Mumbai, India	Knowledgebase with high-resolution images showing the spatial distribution of proteins in normal human tissues and cancer types, as well as human cell lines

5. Establishing the portfolio of ELIXIR Core Data Resources

5.1 Identifying the ELIXIR Core Data Resources

The identification of the ELIXIR Core Data Resources involves a careful evaluation of the multiple facets of the data resources (cf. Section 3.2). The indicators and the related information that support this evaluation are described in Appendix 1 of this document. The ELIXIR Node(s) that are home to the candidate ELIXIR Core Data Resource submit the completed "Case Document" (Appendix 2) to the ELIXIR Hub.

Only data resources that are part of an ELIXIR Node Application and/or Service Delivery Plan (in the case of EMBL-EBI, the 'Work Programme') can be candidate ELIXIR Core Data Resources.

The initial evaluation of the ELIXIR Core Data Resources takes place on an annual basis.

The ELIXIR Hub checks the Case Document for completeness and verifies whether the proposed Resource is included in the Node's Service Delivery Plan (or Work Programme). The ELIXIR Hub has an advisory role in the selection of the ELIXIR Core Data Resources, as a service to the Nodes that

are submitting candidate Core Data Resources. The Hub has no decision-making power and does not evaluate the proposals.

The ELIXIR Director informs the Heads of Nodes Committee of the candidate ELIXIR Core Data Resources. The Heads of Nodes Committee can request additional information about a candidate Resource from the Head of Node who has submitted the respective application.

During a face-to-face meeting, the Heads of Nodes Committee reviews and discusses the submitted Case Documents, and establishes the list of ELIXIR Core Data Resources. It is expected that this will be a limited list in the first selection that can grow over time, based on an annual review of candidate Core Data Resources.

After this face-to-face meeting of the Heads of Nodes Committee, the ELIXIR SAB will review the process before the first initial selection of ELIXIR Core Data Resources is confirmed.

In addition, the ELIXIR Scientific Advisory Board (SAB) comments on the overall portfolio of ELIXIR Core Data Resources and advises on the process for the identification of the ELIXIR Core Data Resources. Since the individual ELIXIR Core Data Resources already have a governance structure that includes an independent, international SAB, this individual review is not to be duplicated by the ELIXIR SAB. The outcome will be presented to the ELIXIR (governance) Board for information and for review of a correct application of the process.

The work that will be done in the context of Task 3.2 of WP3, as well as more generally at the ELIXIR Hub, will provide the technical means and a standardised methodology for collecting and monitoring certain of the indicator-related data. In collaboration with the nodes, these monitoring data will be automatically collected at the ELIXIR Hub on an ongoing basis and will be regularly transmitted to the Heads of Nodes. The nodes need to provide the necessary data to the specification defined through the work of Task 3.2.

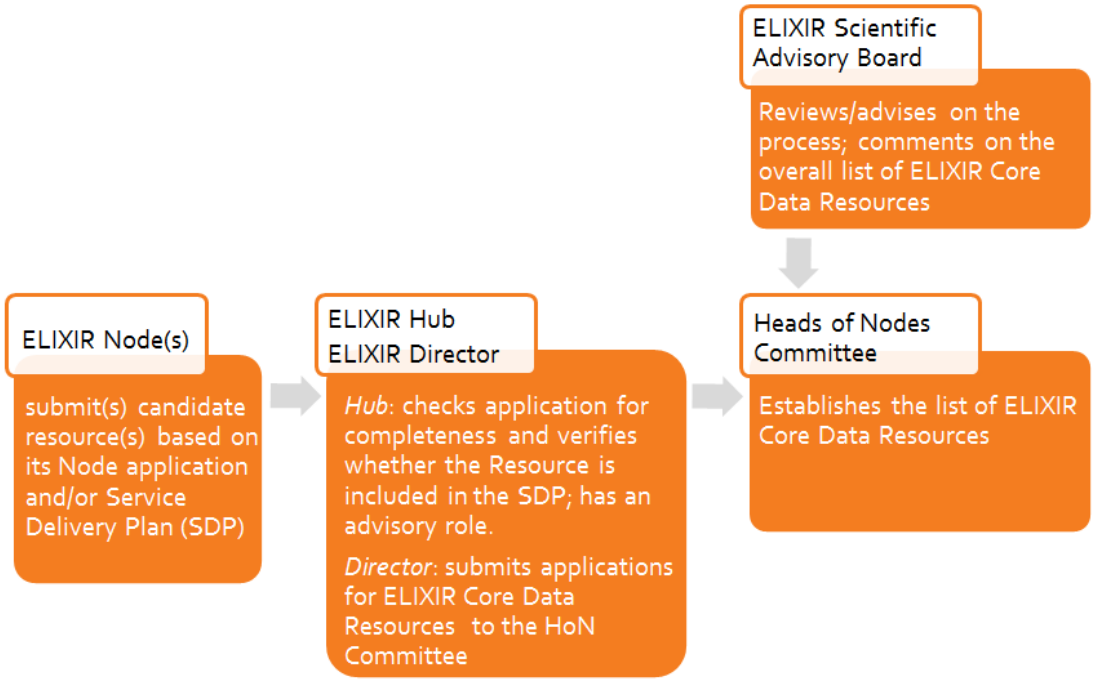

Figure 3. Process for the identification of the ELIXIR Core Data Resources.

5.2 Reviewing the ELIXIR Core Data Resources

ELIXIR Core Data Resources may be requested to report the data for certain indicators on a regular basis, as well as updates on any major changes to the Resource.

A review of all ELIXIR Core Data Resources will be carried out as a live meeting every two to three years, unless three Heads of Nodes request an extraordinary evaluation of an individual resource, notably based on the monitoring data. If the review shows that there is an issue with an ELIXIR Core Data Resource, it is the Heads of Nodes Committee's responsibility to decide what action to take.

6. Practical implementation: driving the long-term sustainability of the ELIXIR Core Data Resources

6.1 A basis for science policy actions to support their long-term sustainability

ELIXIR Core Data Resources form the centre of ELIXIR's sustainability strategy and science policy actions. The collected key indicators for these bioinformatics resources, and more specifically the impact and translational stories, will be used to make a case to funders. This information in turn will help them to translate the impact that Core Data Resources make for their treasury stakeholders.

In addition, the ELIXIR Core Data Resources could contribute to impact and econometric analysis of life science data within ELIXIR, as well as European Commission-focused events on the value of infrastructure for open sustainable data.

ELIXIR will assume limited responsibility for ELIXIR Core Data Resources due to their importance for the life science community. In particular, should an ELIXIR Core Data Resource become endangered, ELIXIR would commit to intervene in order to ensure the continuous availability of the ELIXIR Core Data Resource by actively supporting fundraising. The ultimate aim here is to ensure long-term funding for the ELIXIR Core Data Resources.

As a visible central part of the ELIXIR architecture, the ELIXIR Core Data Resources naturally form part of the ELIXIR offering in grant applications, discussions and integrations, for example in the context of the INFRAIA programmes. As such, the Core Data Resources will be the backbone of ELIXIR's application to funding opportunities.

6.2 Capacity building

Key indicators for Core Data Resources, in particular those around user policies and procedures, will be useful as flagships of excellence and best practice to support capacity building within the ELIXIR Community. This may be extended to interoperability best practices on concept naming, identifier resolution, identifier mappings, and data identity provision.

For example, the ELIXIR Core Data Resources, especially the knowledge bases, can function as "concept authorities" within and beyond ELIXIR, having a clear role in standardizing what the community understanding is of a given biological concept.

Certain indicators could be used outside of ELIXIR (e.g. uptime) to consolidate confidence across a wide range of stakeholders, providing that there is full transparency as to how the indicators are produced, in order to avoid misunderstanding or misuse.

6.3 Life-cycle management

The key indicators will inform life-cycle management, identifying trends and supporting decision-making around a given resource. This is important not only for the teams managing the resources, but also for the identification of Emerging Services that may evolve into ELIXIR services. As new

resources are listed on the ELIXIR node Service Delivery Plans, the indicators and capacity building around the Core Data Resources will support the growth of Emerging Services as they mature.

6.4 A basis for technical actions to support their long-term sustainability and integration with ELIXIR Services

ELIXIR Core Data Resources will be prioritised for technical actions, as well as training on their use. ELIXIR Core Data Resources become the primary resources for ELIXIR Cloud, storage and data distribution efforts within the ELIXIR Nodes network. Where appropriate, ELIXIR Commissioned Services and Implementation studies may be used to establish distributed components that contribute to the Core Data Resources from different nodes, for example, the development of remote deposition tools for core archives in collaboration with national data management efforts. These actions will have important implications for supporting the evolution of Emerging Services associated with Core Resources.

ELIXIR will devote efforts to add value to all ELIXIR resources, including the ELIXIR Services, by supporting interactions of the Core Data Resources with each other and with ELIXIR Services and Emerging Services for the benefits of the user community at large. Examples of this are the use-case driven enhancement of the interoperability of the ELIXIR Core Data Resources with each other and with other ELIXIR Services, supporting helpdesks to scale national operations, as well as implementation studies to explore links to national infrastructures and data services.

7. References

1. ELIXIR Scientific Advisory Board paper 2014 (ELIXIRSAB/2014/4).
2. Wilsdon J et al. (2015) The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. Higher Education Funding Council for England (HEFCE). DOI: 10.13140/RG.2.1.4929.1363
3. HM Treasury. The Magenta Book – Guidance for evaluation. April 2011

Appendix 1: Quantitative and qualitative indicators for the ELIXIR Core Data Resources

The ELIXIR Core Data Resources are defined as a set of European *data* resources that are of fundamental importance to the broad life science community and the long-term preservation of biological data (see main document for an overview of Core Data Resources and the processes used to identify and manage their life-cycles).

This document lists the key indicators that may be used to make a case for a Core Data Resource. The indicators aim to reflect the essence of the definition of an ELIXIR Core Data Resource and support the promotion of excellence in resource development and operation.

The indicators have been grouped in five 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.

The indicators recognise the heterogeneous nature of biological data, and the diversity of the supporting data resources, use cases and communities served. Indicators can be used to measure technical and/or scientific *readiness* of a resource compared to desirable levels of quality standards.

One of the challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, and access to, scientific data. **FAIR** is a set of guiding principles to make data **Findable, Accessible, Interoperable, and Reusable** (Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. DOI: 10.1038/sdata.2016.18). Through the indicators listed below, ELIXIR Core Data Resources should be demonstrated to be compatible with the FAIR data principles. The table below provides a mapping of the indicators to the corresponding FAIR criteria.

Table. The FAIR criteria mapped to the corresponding Core Data Resource indicators.

FAIR Principles	Core Data Resource Indicator(s)
To be Findable: F1 (Meta)data are assigned a globally unique and eternally persistent identifier. F2 Data are described with rich metadata. F3 (Meta)data are registered or indexed in a searchable resource. F4 Metadata specify the data identifier.	3a 1d, 3d 3f(i) 3a, 3d
To be Accessible: A1 (Meta)data are retrievable by their identifier using a standardized communications protocol. A1.1 The protocol is open, free, and universally implementable. A1.2 The protocol allows for an authentication and authorization procedure, where necessary. A2 Metadata are accessible, even when the data are no longer available.	3a, 3f(i,) 3f(ii) 3f(i), 4b 4b, 4c 4e

To be Interoperable: I1 (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2 (Meta)data use vocabularies that follow FAIR principles. I3 (Meta)data include qualified references to other (meta)data.	2d, 3d, 3f(ii) 3d 2d, 3e
To be Re-usable: R1 (Meta)data have a plurality of accurate and relevant attributes. R1.1 (Meta)data are released with a clear and accessible data usage license. R1.2 (Meta)data are associated with their provenance. R1.3 (Meta)data meet domain-relevant community standards.	1d, 3d 4b 2d, 3d, 3e 3d

The context of a core resource is critical to understanding its importance, which is why indicators alone are not sufficient, and qualitative evidence will be required to review cases throughout the resource's lifecycle through the expert judgment of the ELIXIR Heads of Nodes and Scientific Advisory Board.

Indicators and Related Information

All elements in sections 1-4 require a response.

Quantitative indicators are underlined.

1. Scientific focus and quality

a. *Archives vs knowledge bases*: is the resource archival (taking submissions) or a knowledge base (added-value)?

b. *Scope statement*: describe the scientific coverage and comprehensiveness of the resource. For example, all species or a subset of species, families, outputs from a particular experimental method? How is the resource positioned with respect to other similar data resources?

c. *International dimension*: does the resource have a global footprint? (Demonstrated through, for example, an international consortium delivering the resource, geographical diversity in the submissions, global literature curated, international diversity of delivery partners and/or funders)

d. *Staff effort:* number of FTEs per year for the past two or three years
   i. Curators
      - support for submission adherence to metadata requirements? (see also 3d)
      - support for extraction of information from the scientific literature?
   ii. Bioinformaticians
   iii. Technical staff

2. Community

e. *Overall usage:* what is the usage of the resource for the past two or three years?*
   i. Access via a web browser: number of visits, unique visitors, hits, and page views[1]

[1] **Visits (also referred to as sessions):** a session (or visit) is a set of requests/interactions done by the same uniquely identified client within a specific time window (typically, within 30 minutes). The number of sessions is a measure of how much traffic a website gets.
Users (also referred to as unique IP addresses, unique visitors, or visitors): is used to measure how many distinct individuals access a web site over a specified period of time, regardless of how often they visit. It can be

ii. Access via additional access methods: visits, unique visitors, hits, and downloads (includes FTP downloads and programmatic access)

f. *Potential usage:* what is the estimated size of the global potential user community?

g. *Usage in research as measured through citation in the literature:**
   i. Citation of a resource name: the number of times the resource name is mentioned in scientific articles per year (in Europe PMC)
   ii. Citation of data of a resource: the number of times accession numbers from the resource are mentioned or cited in research articles (in Europe PMC)
   iii. Key publications describing the resource list (e.g. publications in NAR Database issue) and the number of citations (in Europe PMC).

h. *Dependency of other resources:* do other resources have a dependency on the resource described here to provide that service (i.e. what is the reach through)? Please list.

* The method used to derive these indicators needs to be supplied.

3. Quality of service

a. *Identifier use*: does the resource provide persistent and unique identifiers?

b. *Data throughput:* number of entries, depositions (records or bytes ingested per year), records processed, genomes assembled, etc. per year, for past 2 or 3 years.

c. *Technical performance*:
   i. Uptime: percentage availability per month for a sample of key web pages (or similar) over the past 12 months (e.g. search results, homepage, data record pages).
   ii. Response times of key web pages.

d. *Use of standards*: which community-recognised standards are used for metadata and data (e.g. MIAME, JATS, INSDC features, ontologies)? Provide a link to documentation.

e. *Links to documentation of provenance*: does the resource link to the scientific literature for provenance of facts or biological context?

f. *Data availability* - access services and formats
   i. Data sharing services: list services through which data is shared (e.g. website, APIs, FTP, TripleStore)
   ii. Data sharing formats: list formats data is available in (e.g. plain text, FASTA, XML, RDF, Dublin Core, tsv, JSON)

g. *Customer service*
   i. Helpdesk: does the resource run a helpdesk?
   ii. User feedback: does the resource seek and incorporate user input into service design decisions?

---

determined in different ways: number of unique IP addresses, number of unique IP addresses + user agent (a "user agent" refers to the client that is used to access a web site), by a user cookie in case of web technology.
Hits: can be used to analyse trends of a specific web resource. Hits refer to the number of files 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. 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.
Downloads: measures the size of the data downloaded from resource in terms of volume / bandwidth (commonly measured in GB).
Pages (also referred to as page views, impressions or URLs): corresponds to a request to load a *single* HTML file (web page) of a web site, identified by the URL in a browser. During a visit or session, a person can access several different pages of a web application, which results in several impressions or page views.

iii. Training: does the resource undertake training activities?

4. Legal and funding infrastructure, and governance

a. *Scientific Advisory Board:* does the resource have an international, independent Scientific Advisory Board?

b. *Open Science:* does the resource have a legal framework that supports Open Science? E.g. open licenses or public statement of open terms of use.

c. *Privacy policy:* does the resource have a public privacy policy in which security around personal data and cookies are described? Does the resource have an ethics policy?

d. *Ethics policy*

e. *Sustainable support and funding:* demonstrate the past and future funding commitments secured by the resource by the host institution or otherwise.

5. Impact and translational stories

a. *Counterfactual*: what would the impact on the scientific community be if the resource had not existed or were to disappear and not be replaced? Is the resource globally unique? What would the impact on other dependent resources be?

b. *Accelerating science*: how does the resource accelerate science? For example, does the resource set standards; promote reuse of data or software; promote research efficiencies; extend technical products in other areas?

c. *Translational data*: are there any "translational" figures that are familiar to the audience that will help them grasp the core nature of the resource? E.g. citation in patent documents, x visits daily to database y is comparable with the number of people accessing the BBC web page on a daily basis.

Appendix 2: Case Document Template

A "Case Document" describes a (candidate) Core Data Resource and is based on the all indicators introduced in Appendix 1.

Case Document: [Resource Name] v1.0

Document owner: [Insert Name] [email address]

1. Scientific focus and quality

a. **Archival vs knowledge base:** is the resource

- ❑ archival (taking submissions)
- ❑ knowledge base (added-value)

b. **Scope statement:** describe the scientific coverage and comprehensiveness of the resource. For example, all species or a subset of species, families, outputs from a particular experimental method? How is the resource positioned with respect to other similar data resources?

c. **International dimension:** does the resource have a global footprint? (E.g. demonstrated through an international consortium delivering the resource, geographical diversity in the submissions, global literature curated, international diversity of delivery partners and/or funders)

d. **Staff effort:**

Number of FTE	**[Year 1]**	**[Year 2]**	**[Year 3]**
Curators **support for submission adherence to metadata requirements** **support for extraction of information from the scientific literature** ***Bioinformaticians*** ***Technical staff***			

2. Community

a. **Overall usage - quantitative:** what is the usage of the resource for the past 2/3 years?

Please indicate the method used to derive these indicators.

Access via a web browser (using web analytics, example: Google Analytics)

Average monthly web traffic	[Year 1]	[Year 2]	[Year 3]
Visits (sessions)			
Unique visitors (users)			
Page views			

Access via a web browser (using log analytics)

Average monthly web traffic	[Year 1]	[Year 2]	[Year 3]
Unique visitors (users)			
Hits			
Sessions and pages (if possible)			

Data downloads (FTP, APIs, etc.)

Average monthly downloads	[Year 1]	[Year 2]	[Year 3]
Hits / Requests			
Unique IP addresses / Hosts			
Data transfer (GB)			

b. **Potential usage:** what is the estimated size of the global potential user community?

c. **Usage in research as measured through citation in the literature**:

Please indicate the method used to derive these indicators.

Annual totals:	[Year 1]	[Year 2]	[Year 3]

Resource name mentioned in Europe PMC (citation of resource name)			
Accession numbers mentioned in Europe PMC (citation of data of the resource)			

Key publications describing the resource list (e.g. publications in NAR Database issue) and the number of citations (in Europe PMC):

d. **Dependency of other resources:** do other resources have a dependency on the resource described here to provide that service (i.e. what is the reach through)? Please list.

3. Quality of service

a. **Identifier use**: does the resource provide persistent and unique identifiers?

b. **Data throughput**: number of entries, depositions (records or bytes ingested per year), records processed, genomes assembled, etc. per year, for past 2 or 3 years.

	[Year 1]	[Year 2]	[Year 3]
Total number of entries/depositions			
Size in GB			
Size (other)			

c. **Technical performance**:
   i. **Uptime:** percentage availability per month for a sample of key web pages (or similar) over the past 12 months (e.g. search results, homepage, data record pages).
   ii. **Response times of key web pages**.

d. **Use of standards:** which community-recognised standards are used for metadata and data (e.g. MIAME, JATS, INSDC features, ontologies)? Provide a link to documentation.

e. **Links to documentation of provenance:** does the resource link to the scientific literature for provenance of facts or biological context?

f. **Data availability – access services and formats:**
   i. **Data sharing services:** list services through which data is shared (e.g. website, APIs, FTP, TripleStore)
   ii. **Data sharing formats:** list formats data is available in (e.g. text, FASTA, XML, Dublin Core, tsv, JSON)

g. **Customer service**:
   i. **Helpdesk:** does the resource run a helpdesk?

ii. **User feedback:** does the resource seek and incorporate user input into service design decisions?
iii. **Training:** does the resource undertake training activities?

4. Legal and funding infrastructure, governance

a. **Scientific Advisory Board**: does the resource have an international, independent Scientific Advisory Board?
b. **Open Science**: does the resource have a legal framework that supports Open Science? E.g. open licenses or public statement of open terms of use.
c. **Privacy policy**: does the resource have a public privacy policy in which security around personal data and cookies are described? Does the resource have an ethics policy?
d. **Ethics policy**
e. **Sustainable support and funding**: demonstrate the past and future funding commitments secured by the resource by the host institution or otherwise.

5. Impact and translational stories

a. **Counterfactual:** what would the impact on the scientific community be if the resource had not existed or was to disappear and not be replaced? Is the resource globally unique? What would the impact on other dependent resources be?
b. **Accelerating science:** how does the resource accelerate science? For example, does the resource set standards; promote reuse of data or software; promote research efficiencies; extend technical products in other areas?
c. **Translational data:** are there any "translational" figures that are familiar to the audience that will help them grasp the core nature of the resource? E.g. citation in patent documents, x visits daily to database y is comparable with the number of people accessing the BBC web page on a daily basis.