EXCELERATE Work Package 3 report:

Plan for collation of metrics and quality data at the ELIXIR Hub

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

Authors and Contributors:
Heinz Stockinger[1], Mary Barlow[2], Chuck Cook[2], Rachel Drysdale[3], Elisabeth Gasteiger[1], Jee-Hyub Kim[2], Rodrigo Lopez[2], Nicole Redaschi[1], Daniel Teixeira[1], Christine Durinx[1], Johanna McEntyre[2]

1 ELIXIR-Switzerland, SIB Swiss Institute of Bioinformatics
2 EMBL-EBI, UK
3 ELIXIR-Hub, UK

Status: 10th August 2017

Context of this report:

This report presents a **plan for the collation of indicators and quality data of ELIXIR's Core Data Resources and candidates** as defined in the following milestone:

Milestone M3.3 "Plan for collation of metrics and quality data at the ELIXIR Hub" (Due Month 24 - August 2017)

This work is done in the context of ELIXIR-EXCELERATE WP3, Task 3.2:

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**"

This report also prepares the work for the following deliverable and milestone:

Deliverable D3.3 "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" (Due Month 48 - August 2019)

Milestone M3.4 "Computer-assisted service-monitoring tools that deliver metrics data to the ELIXIR Hub" (Due Month 48 - August 2019)

Contents

1 Introduction

In July 2017, ELIXIR selected the initial set of Core Data Resources (CDRs)[R1, R2]: these are deposition databases and knowledgebases that are of fundamental importance for the life sciences community in Europe and worldwide. 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 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. The current set of Core Data Resources was chosen as the result of careful review by an independent expert panel and the ELIXIR Heads of Nodes (HoN). The set is not static: current and candidate Core Data Resources will be evaluated regularly, and more resources may be added, while others may be removed, as the landscape of biological data evolves.

Qualitative and quantitative information is required to support the life cycle management of the Core Data Resources and to shape future policy proposals. This information will be gathered by a defined and iterative process such that trends can be observed over time. A detailed list of the indicators is provided in [R1] (Box 1: https://f1000research.com/articles/5-2422/#B1), and a comprehensive summary of 23 indicators is available in Appendix 1. In this document, we provide a plan and some concrete examples of how data for these indicators can be collected and kept up-to-date. 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 carefully and - depending on the dataset and institution involved - might only be accessible on a granular level by a restricted set of people and for a clearly defined purpose: **the selection of Core Data Resources and their monitoring as part of their life-cycle management.** However, there will also be a need to use the data collectively in communications that describe the impact of the ELIXIR CDRs as a whole.

In summary, each ELIXIR Core Data Resource will be asked to provide information for each of the 23 indicators (cf. Appendix 1) to the ELIXIR Hub, who will then store and make appropriate use of the data. Since some of the indicator data change over time (e.g. usage or citation statistics), updates are necessary to allow for monitoring of trends for individual resources. We suggest that annual updates would provide a sufficiently granular profile. We provide a preliminary plan for how the indicator data can be exchanged, updated and made accessible. It is important to note that **managers of CDRs will be consulted in the second half of 2017 to adjust this plan and to agree on details on data collection and analysis.** This plan presented here represents a draft for discussion and agreement in 2018.

2 Objectives

2.1 *Why* is data collected?

The reasons for collecting indicator data are to inform the i**dentification and selection processes of ELIXIR Core Data Resources, and to monitor the usage, quality and impact of the ELIXIR Core Data Resources portfolio** over time, as part of the life-cycle management of those resources. In particular, the ELIXIR Hub will receive and store information about individual resources and make it available to a restricted group of relevant bodies (e.g. Head of Nodes committee, Scientific Advisory Board, funding agencies, etc.).

The entire network of ELIXIR Core Data Resources is a comprehensive data infrastructure and a peer of other large data infrastructures, such a the Square Kilometer Array (skatelescope.org). From a strategic perspective it should have a sustained funding model similar to these large central facilities. The collection of indicators from the Core Data Resources allows ELIXIR to demonstrate the need of sustained infrastructure funding to funders and stakeholders and monitor progress, trends and usage over time.

2.2 How will the data be used?

The primary aim is to monitor usage, quality and impact **trends over time** for each resource. The data are **not meant to compare individual resources to each other**, as each resource has its own specificities (cf. Section 3 Challenges). They will, however, provide indications about the trends in absolute number of users or volume of downloads for each resource.

Additionally, **data can be aggregated over all CDRs** and presented as a comprehensive data source representing the ELIXIR Core Data Resource portfolio. This aggregated data will be used to demonstrate the need for the resources and garner support. This may include overall usage, deposition rates, infrastructure size and growth or technology trends. These figures could also be used to calculate more abstract measures, like the overall research productivity gained through Core Data Resource use, or the economic returns the resource bring to the investments already made. Methods for how that data is collected and identifying data suitable for aggregation and public consumption will be key.

Details on how individual resource data or aggregated data will be used (i.e., for what exact purpose) and by whom (ELIXIR Hub, resource providers, etc.) will evolve, and are outside the scope of this document.

2.3 Collecting the data

Currently, resource providers have the full responsibility for the data provided, and guarantee the correctness of these data. It is the responsibility of each resource provider to make the necessary information available to the ELIXIR Hub. As the set of ELIXIR Core Data Resources may be eligible for international funding in the future, the accuracy of the indicator data is of utmost importance.

In addition, the data collection process needs to be optimised so that it becomes easy for resource providers to provide new or updated information in a consistent way: this process should be automated and standardised wherever possible and feasible. For example, external monitoring services might be used to check service availability for all Core Data Resources in a homogeneous way. Some information and corresponding data will be relatively static (scientific purpose or staff numbers) and others may change considerably every year. The update frequency will also be agreed with resource providers.

Details need to be defined how data should be transferred to the ELIXIR Hub (potentially stored in a central repository) and made accessible. **This work will be done in collaboration with resources providers.** A first outline is provided in Section 4.

History/background: approach taken up to now to select the first set of CDRs

In the first year of the ELIXIR-EXCELERATE project, a list of indicators was defined and a “case document template” (Box 2 in [R1]) was created which resource providers could use to provide the necessary information to the ELIXIR Heads of Nodes. The template also contained a minimal set of guidance and example indicators. In practice, resource providers who proposed a candidate Core Data Resource filled in a document describing each indicator. This was a manual process that allowed collection of a first complete set of data. It has not been necessary, and may not be technically feasible, for each resource provider to collect information using the same methods, nor have the individual submissions been analysed with respect to uniformity and standard usage.

The differences between resources, and in the technical facilities available at each Node, suggest that it is unlikely that the methods used to collect quantitative indicators, such as web usage statistics, can be made uniform for all resources across all Nodes. Nevertheless, **examples and guidelines in Appendix 2** as well as follow-up work to compare the different methodologies used for different resources and at different Nodes will **allow Nodes to share best practice.**

Currently, indicator data collected by the ELIXIR Hub for each Core Data Resource are static, and stored as text or PDF documents: no updates are scheduled.

3 Challenges

Distributed production, collection and usage of indicator data raises a few challenges that make certain comparisons difficult or impossible, and need to be considered:

- **Data confidentiality:** indicator data are not generally publicly accessible and are frequently considered to be confidential. This limits both collection of, and access to, these data. Access restrictions and roles need to be applied to specific user groups. Follow-up work is required once data privacy and access rights are clarified. Currently, only a restricted set of people has access to this information (incl. ELIXIR Hub, HoN, SAB and a small number of contributors to ELIXIR-EXCELERATE WP3). The confidentiality required will be reviewed with the General Data Protection Regulation (GDPR, www.eugdpr.org) in mind to ensure compliance.
- **Methods used to collect indicator data:** in Appendix 2, example methods and guidelines are given about how to measure usage and performance indicators. However, not all resource providers apply the same methods which makes aggregation difficult or impossible. It might not be feasible to apply the same methods across all resources but it is advisable to have the same methods where possible and feasible. The absolute minimal requirement is the following one: **each resource should aim to apply the same methods consistently over time** so trends can be measured. Past experience has shown that this is not always possible or feasible (e.g. in 2011 Google Analytics decided to change the way it counts visits; some resource providers changed from AWstats to Hadoop/PIG). Therefore, if changes are made, they must be well justified and documented.
- **Correctness of data:** it is of paramount importance that data provided for indicators is correct and up to date. A specific quality control and monitoring process needs to check that data are not faked nor artificially increased (as it might be the case for web or other usage statistics)

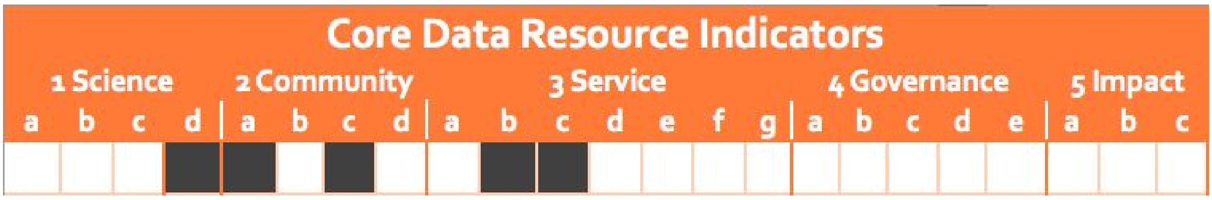
4 Basic methodology: outline of a preliminary approach

We will now look at the individual indicators in more detail and consider what information needs to be obtained. This is followed by a short discussion about a possible data collection and storage system.

4.1 Indicators and their update frequencies

In [R1] 23 main indicators are defined and grouped into 5 categories: **1 Science** (scientific focus and quality of science), **2 Community** (community served by the resource), **3 Service** (quality of service), **4 Governance** (legal and funding infrastructure, and governance), and **5 Impact**. Most of these indicators only need to be obtained once and require monitoring but not annual updating.

Figure 1. Overview of 23 main indicators (cf. Table 1 and Appendix 1 for details) and their update frequency. A dark box shows that an indicator needs to be updated regularly (usually, once a year). Indicators with a white box usually do not require regular updates, i.e., need to be provided at least once and may be updated "as appropriate" (cf. column "Update Frequency" in Table 1) meaning that changes should be reported when considered appropriate for a particular resource.



Core Data Resource Indicators

1 Science				2 Community				3 Service							4 Governance					5 Impact		
a	b	c	d	a	b	c	d	a	b	c	d	e	f	g	a	b	c	d	e	a	b	c
□	□	□	■	■	□	■	□	□	■	■	□	□	□	□	□	□	□	□	□	□	□	□

For example, indicator 1a (archives vs knowledgebases) states if a resource is an archive (i.e. deposition database) or a knowledge base. Usually, this feature does not change and does not need to be updated once the information has been initially collected e.g., via a web form. However, in the categories "Community" and "Service" four indicators **(2a overall usage, 2c usage measured through citation in literature, 3b data throughput and 3c technical performance)** can change and need to be updated to show trends over time (cf. Figure 1). Typically, these indicators require 1 data point per year in order to calculate yearly averages wherever possible and feasible as indicated in Table 1. Additionally, indicator **1d staff effort** changes over time but might not necessarily need to have yearly updates.

Table 1. Complete list of indicators. Most of this data need to be provided by the resource providers directly (decentralised data provision/collection) but some indicators (such as 2c and 3c) could be obtained in an automatic way without intervention by resource providers.

Indicator	**Update Frequency**	**Unit of measurement**	**How is data received (Automation)**	**Proposed Circulation - Visualisation**
1 Scientific focus and quality				
1a Archives vs knowledge bases	as appropriate	tick box	manually	public - ELIXIR web site
1b Scope statement	as appropriate	free text	manually	public - ELIXIR web site
1c International dimension	as appropriate	free text	manually	aggregation
1d Staff effort	yearly	FTE numbers	manually	aggregation - presentation, funding bids
2 Community				
2a Overall usage:				
Visits/sessions	yearly	numbers cf. Appendix 2	automation possible[1]	aggregation - or with written consent
Page views	yearly	numbers cf. Appendix 2	automation possible[1]	aggregation
Unique users/IP addr.[2]	yearly	numbers cf. Appendix 2	automation possible[1]	aggregation
Hits/requests	yearly	numbers cf. Appendix 2	manually	aggregation
Download	yearly	Gigabytes/year cf. Appendix 2	manually	aggregation
2b Potential usage	as appropriate	numbers	manually	aggregation
2c Usage in literature:				
Name mentioned	yearly	numbers cf. Appendix 2	automatically via EPMC	aggregation - or with written consent
Accessions mentioned	yearly	numbers cf. Appendix 2	automatically via EPMC	aggregation - or with written consent
Publications & #citations	yearly	list: text & numbers cf. Appendix 2	automatically via EPMC	aggregation - or with written consent

3d Dependency of other resources	as appropriate	free text	manually	map of network
3 Quality of service				
3a Identifier use	as appropriate	free text	manually	aggregation - or with written consent
3b Data throughput:				
Data entries	yearly	numbers	manually	aggregation
Data size	yearly	Gigabytes/year	manually	aggregation - or with written consent
3c Technical performance:				
Uptime	yearly	percentage/year cf. Appendix 2	automatically (e.g. via Monitis)	aggregation
Response time[3]	yearly	number in milliseconds cf. Appendix 2	automatically (e.g. via Monitis)	private or trends only
3d Use of standards	as appropriate	free text	manually	private or trends only
3e Links to doc. of provenance	as appropriate	free text	manually	aggregation
3f Data availability	as appropriate	free text	manually	public
3g Customer service	as appropriate	free text	manually	public
4 Legal and funding infrastructure, and governance				
4a Scientific Advisory Board	as appropriate	free text	manually	aggregation
4b Open Science	as appropriate	free text	manually	public - website
4c Privacy policy	as appropriate	free text	manually	public or internal
4e Ethics policy	as appropriate	free text	manually	public or internal
4f Sustainable support and funding	as appropriate	free text	manually	aggregation
5 Impact and translational stories				
5a Counterfactual	as appropriate	free text	manually	aggregation - or with written consent
5b Accelerating science	as appropriate	free text	manually	aggregation - or with written consent
5c Translational data	as appropriate	free text	manually	aggregation - or with written consent

[1] If a resource provider gives, for example, read access to Google Analytics data, information can be obtained automatically.
[2] Can not be accumulated over time (in contrast to other indicators mentioned here).
[3] Relative to a specific geographic location.

Table 1 includes a few **quantitative indicators** (2a, 2c, 3b, 3c) that can be measured in several ways. Details and **examples of how these quantitative indicators can be measured,** are provided in Appendix 2: Details for Community and Service Indicators and act as guidelines for resource providers. Furthermore, for some of the indicators, **example cases and practical implementations** can be found in Sections 5.1 and 5.2.

For mainly **qualitative indicators** like those found in Section 5 “Impact and translational stories”, it might be beneficial to adopt visualisation templates, which show resource owners how the data provided will be displayed. A current example being the ELIXIR impact infographic https://www.elixir-europe.org/system/files/elixir_infographic_30112016_office.pdf. Case studies which present information from all five indicator sections together will act as a powerful tool in raising the understanding of the Core Date Resources and in funding proposals. Examples (visualisation templates) for how information may be presented and circulated, could provide an incentive to resource managers to submit up-to-date information, support ELIXIR branding and make it clear how the data will be circulated/updated. Information gathered should highlight elements like the intended audience (public, funders, scientist), the area of science impacted (human health, agri-science etc.) and main outcome it is intending to highlight (new discovery, new method, a challenge overcome).

4.2 Data collection: towards a repository for indicator data

It is envisaged that ELIXIR will maintain a **single repository for indicator data** for the various Core Data Resources. The technical details still need to be defined but the assumption is that such a data repository will be managed by, or on behalf of, the ELIXIR Hub. The repository needs to provide a way to achieve the following:

- For a new Core Data Resource, information about all 23 indicators can be submitted via the submission Case Document. Exceptions can be made for indicators that can be obtained automatically.
- For a Core Data Resource already registered with the repository, annual updates can be submitted and historic values (time series) are stored.
- Only authorised personnel is allowed to change or view data. This is true for both personnel of the CDR as well as personnel at the Hub or other ELIXIR bodies.
- A reporting and visualisation feature allows for creating individual or aggregated reports for one or more resources.

Depending on the indicator, the mode of data submission to such a repository needs to be defined, including what actions a resource provider needs to do explicitly (manually) and what can be automated or done centrally for a set of resources. Depending on the indicator, different data collection/update options are possible with various complexity and automation effort:

1. A resource provider sends an **email** to the ELIXIR Hub (attaching a **single document** in doc or PDF format) or uploads a document to a specified site. The Hub can then use a simple document system to manage these documents and apply version control on document level.
2. A resource provider **updates data in a central service addressing each indicator individually**. Examples:
   a) web-accessible spreadsheet (such as Google docs)
   b) specialised on-line web application (can either be via a graphical user interface and/or a programmatic interface).
3. Mainly for quantitative indicators that need updates:
   a. A resource provider **publishes and populates a local service**: Subsequently, a specific ELIXIR service pulls the data in an automatic way to populate the repository for indicator data.
   b. A resource provider **uses an existing, central service** (e.g. Google Analytics) to monitor/store indicator data, and a specific ELIXIR service pulls data in an automatic way to populate the repository for indicator data.
   c. Indicators are monitored via external services (e.g. EPMC for indicator 2c Usage in literature) without any intervention of a resource provider.

Option 1 is the simplest and requires little technical effort. In fact, it was already partially used for the first round of CDRs. However, it does not allow for easy data analysis or reporting. Therefore, option 2 is more appropriate as a long term solution. Current practise and experience of ELIXIR Nodes (cf. examples in Section 5.2) and their respective Core Data Resources will be taken into account for the selection of a technical solution.

5 Examples and Case Studies

The process of defining and selecting a first set of ELIXIR Core Data Resources has taken almost two years since the beginning of the ELIXIR-EXCELERATE project. Each candidate ELIXIR Core Data Resource has used a template to provide the data related to the indicators (Box 2, [R1]). As a result, experience with the case template has been gained for 26 candidate resources. However, currently there is no indicator data repository in place where indicator information can easily be retrieved in an automatic way except by browsing through a large number of documents. In order to gain experience with potential solutions, we present two cases that can be used to obtain, store and query indicator information. The two cases are currently not connected to each other but each addresses a specific set of indicators listed in Table 1:

- Case 1: indicator 2c literature usage
- Case 2: indicators 2a overall usage and 3c technical performance

5.1 Case 1: Literature usage

To analyse the usage of CDRs in the research literature, a text-mining module has been developed as a part of the core Europe PMC text mining pipeline to extract two types of usages:

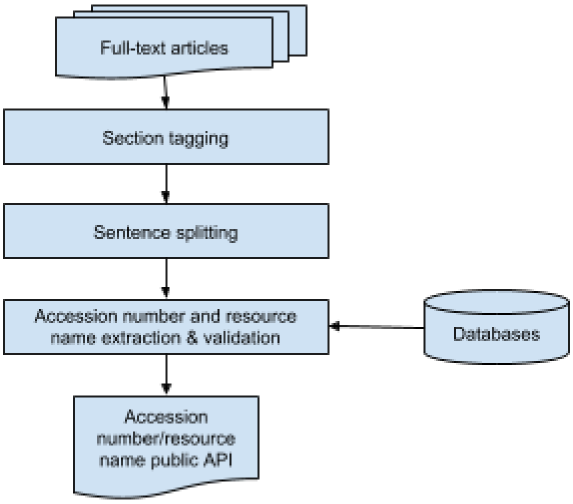
- Mentions of the **names of resources,**
- Mentions of specific datasets through the **citation of accession numbers** in text (expressed via an explicit “accession type” and respective patterns).

The system to extract resource names and accession numbers consists of three modules (shown in **Figure 2**):

1. Identification of regions of interest (ROI)
2. Sentence splitting (identification of sentences)
3. Identification of data citation statements, extraction and validation.

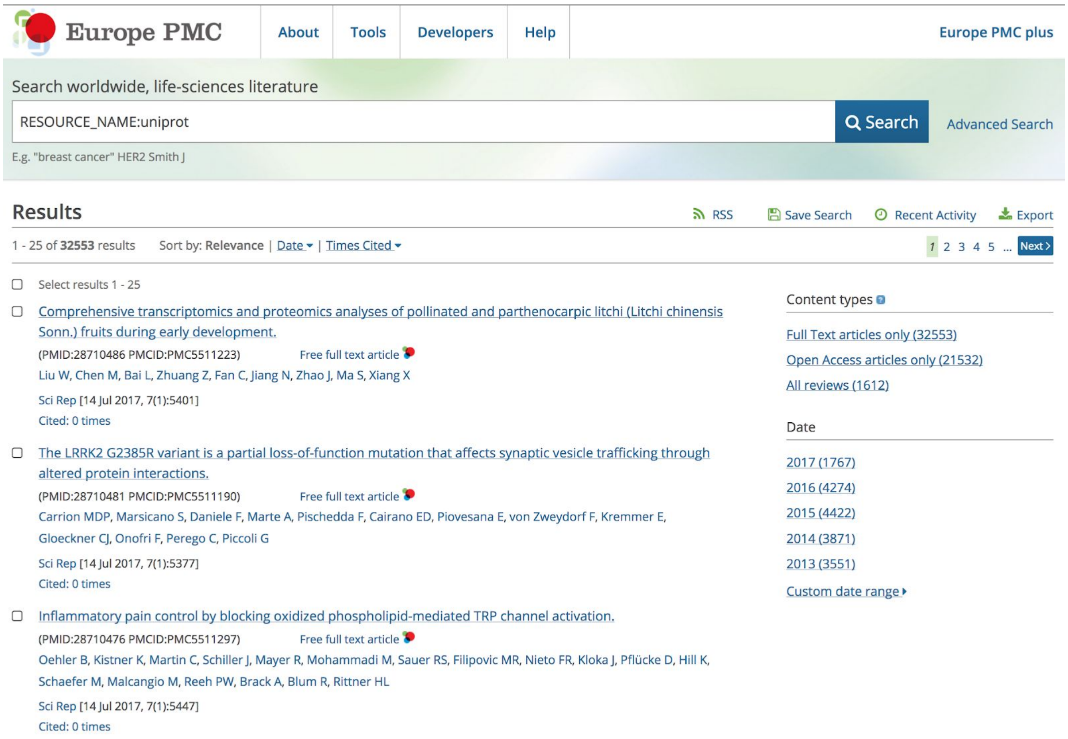
The first module tags sections of articles (e.g. introduction, methods, results, etc) as ROIs, to which the second module is applied in order to split paragraphs in sections into a list of sentences. Then, the last module - the core of the system - text mines data citation attributions and validates these putative attributions against known attributions in the reference databases.

Figure 2. Overall text-mining system architecture


Full-text articles
Section tagging
Sentence splitting
Accession number and resource name extraction & validation
Databases
Accession number/resource name public API

The text-mined results are indexed and available both in public APIs and via the Europe PMC website (http://europepmc.org/). The system is running and updating daily. Figure 3 and Figure 4 respectively show examples of mentions of resource names and accession numbers.

Figure 3. Screenshot of search results for UniProt (www.uniprot.org) as a resource name, mentioned in full-text articles. The query URL can be used in a web browser as follows where “%3A” HTML-encodes the character “:” to pass the following key:value pair to the query “**RESOURCE_NAME**:**uniprot**”:
http://europepmc.org/search?query=**RESOURCE_NAME**%3A**uniprot**



Search worldwide, life-sciences literature

RESOURCE_NAME:uniprot

E.g. "breast cancer" HER2 Smith J

Results

1 - 25 of **32553** results   Sort by: **Relevance** | Date | Times Cited   1 2 3 4 5 ...

□ Select results 1 - 25

□ Comprehensive transcriptomics and proteomics analyses of pollinated and parthenocarpic litchi (Litchi chinensis Sonn.) fruits during early development.
(PMID:28710486 PMCID:PMC5511223) Free full text article
Liu W, Chen M, Bai L, Zhuang Z, Fan C, Jiang N, Zhao J, Ma S, Xiang X
Sci Rep [14 Jul 2017, 7(1):5401]
Cited: 0 times

□ The LRRK2 G2385R variant is a partial loss-of-function mutation that affects synaptic vesicle trafficking through altered protein interactions.
(PMID:28710481 PMCID:PMC5511190) Free full text article
Carrion MDP, Marsicano S, Daniele F, Marte A, Pischedda F, Cairano ED, Piovesana E, von Zweydorf F, Kremmer E, Gloeckner CJ, Onofri F, Perego C, Piccoli G
Sci Rep [14 Jul 2017, 7(1):5377]
Cited: 0 times

□ Inflammatory pain control by blocking oxidized phospholipid-mediated TRP channel activation.
(PMID:28710476 PMCID:PMC5511297) Free full text article
Oehler B, Kistner K, Martin C, Schiller J, Mayer R, Mohammadi M, Sauer RS, Filipovic MR, Nieto FR, Kloka J, Pflücke D, Hill K, Schaefer M, Malcangio M, Reeh PW, Brack A, Blum R, Rittner HL
Sci Rep [14 Jul 2017, 7(1):5447]
Cited: 0 times

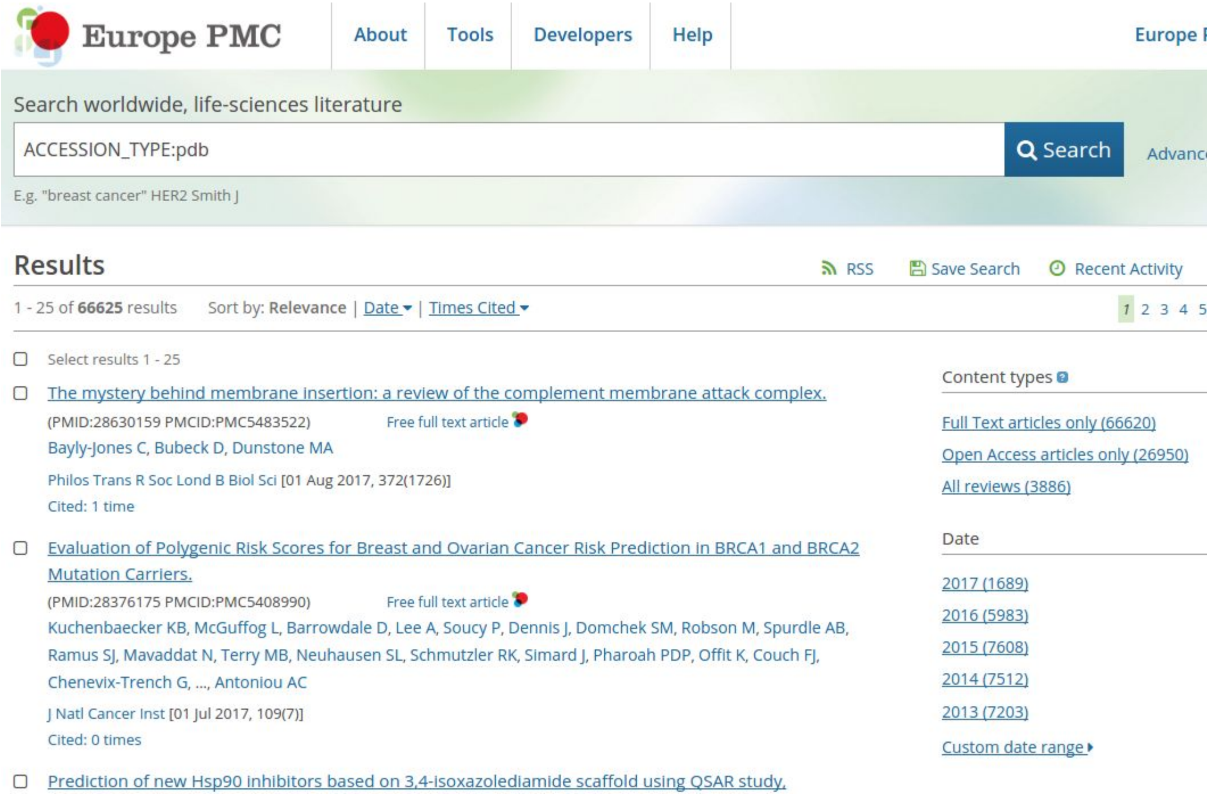
Content types

Full Text articles only (32553)
Open Access articles only (21532)
All reviews (1612)

Date

2017 (1767)
2016 (4274)
2015 (4422)
2014 (3871)
2013 (3551)
Custom date range

Figure 4. Screenshot of a search result for papers that mention PDB accession numbers in full-text articles (http://europepmc.org/search?query=**ACCESSION_TYPE**%3A**pdb**). Note that for each database - in this case PDB (www.wwpdb.org) - patterns for accession numbers are used to identify the respective datasets (cf. Appendix 2, Section A2.2.2).



Search worldwide, life-sciences literature

ACCESSION_TYPE:pdb

E.g. "breast cancer" HER2 Smith J

Results

1 - 25 of **66625** results Sort by: **Relevance** | Date ▾ | Times Cited ▾ 1 2 3 4 5

□ Select results 1 - 25

□ The mystery behind membrane insertion: a review of the complement membrane attack complex.
(PMID:28630159 PMCID:PMC5483522) Free full text article
Bayly-Jones C, Bubeck D, Dunstone MA
Philos Trans R Soc Lond B Biol Sci [01 Aug 2017, 372(1726)]
Cited: 1 time

□ Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers.
(PMID:28376175 PMCID:PMC5408990) Free full text article
Kuchenbaecker KB, McGuffog L, Barrowdale D, Lee A, Soucy P, Dennis J, Domchek SM, Robson M, Spurdle AB, Ramus SJ, Mavaddat N, Terry MB, Neuhausen SL, Schmutzler RK, Simard J, Pharoah PDP, Offit K, Couch FJ, Chenevix-Trench G, ..., Antoniou AC
J Natl Cancer Inst [01 Jul 2017, 109(7)]
Cited: 0 times

□ Prediction of new Hsp90 inhibitors based on 3,4-isoxazolediamide scaffold using QSAR study.

Content types
Full Text articles only (66620)
Open Access articles only (26950)
All reviews (3886)

Date
2017 (1689)
2016 (5983)
2015 (7608)
2014 (7512)
2013 (7203)
Custom date range ▸

In addition to the web interface, a REST API (cf. documentation) is also available and can be used in the following way, taking the example of Interpro (www.ebi.ac.uk/interpro): http://www.ebi.ac.uk/europepmc/webservices/rest/search?query=RESOURCE_NAME%3A**interpro**

http://www.ebi.ac.uk/europepmc/webservices/rest/search?query=ACCESSION_TYPE%3A**interpro**

One of the limitations of the text-mining system is that the ambiguity of some terms (e.g. short abbreviations such as ENA, EGA, etc.) may cause false positives to be annotated as resource names. To guard against this, strict rules are applied that require additional co-occurring terms to be present close to the putative resource name. In this way, the text-mining algorithms deployed favour precision over recall, i.e. it is more likely that resource names are undercounted than over counted. Another limitation is that the system currently uses only Europe PMC full-text articles, therefore any accession numbers or resource names mentioned outside of this collection are missed. The service is fully operational, and the results available in Europe PMC APIs and website. Further scientific resources may be added on demand via the Europe PMC helpdesk.

5.2 Case 2: Resource usage and technical performance

5.2.1 Method used at ELIXIR-Switzerland

ELIXIR-Switzerland (SIB Swiss Institute of Bioinformatics) has implemented a service called “SIB Insights” (https://insights.expasy.org) that focuses on indicators that are related to usage (indicator 2a overall usage) and technical performance (indicator 3c). SIB Insights allows the collection and analysis of usage and performance statistics for SIB resources in a coherent and automated manner via a unified dashboard. In this way, resource providers get better insights on how their resources are used, informing life cycle management decisions. Access is also granted to the SIB Scientific Advisory Board in order to inform their advice.

SIB Insights can collect and visualise data from various sources such as Google Analytics, HTTP (web) and FTP servers. Each resource provider can control who has access to these data. Currently, the following indicators are analysed:

- Web access, e.g. via Google Analytics (currently the main source),
- Programmatic access and data download via HTTP server logs.

The following features are planned to be added:

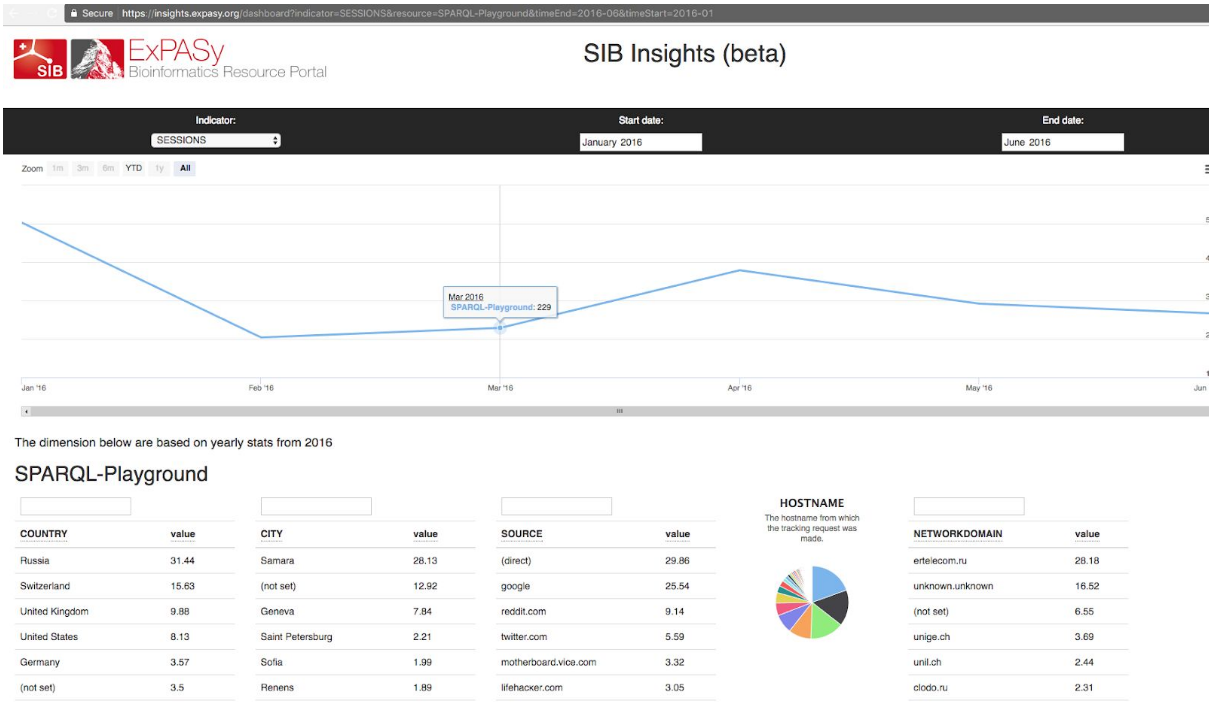
- Analysis of FTP server logs,
- Integration of uptime and response time data, e.g. via monitis.com,
- Integration of literature indicators from EuropePMC (similar as Section 5.1).

Indicator data can be collected in two different ways:

1. automatically via external services such as Google Analytics or Monitis, or
2. manually for **off-line analysis** of HTTP and FTP server logs.

Note that current access statistics include monthly averages where possible, i.e. 12 data points per year are collected rather than only 1 data point as discussed in Section 4.

Figure 5. **Screenshot of SIB Insights** related to a resource that is used in workshops to learn SPARQL http://sparql-playground.sib.swiss/ (SPARQL playground).



Secure https://insights.expasy.org/dashboard?indicator=SESSIONS&resource=SPARQL-Playground&timeEnd=2016-06&timeStart=2016-01
SIB
ExPASy
Bioinformatics Resource Portal
Zoom 1m 3m 6m YTD 1y All
Mar 2016
SPARQL-Playground: 229
Jan '16
Feb '16
Mar '16
Apr '16
May '16
Jun
HOSTNAME
The hostname from which the tracking request was made.

SIB Insights (beta)

Indicator: SESSIONS

Start date: January 2016

End date: June 2016

The dimension below are based on yearly stats from 2016

SPARQL-Playground

COUNTRY	value
Russia	31.44
Switzerland	15.63
United Kingdom	9.88
United States	8.13
Germany	3.57
(not set)	3.5

CITY	value
Samara	28.13
(not set)	12.92
Geneva	7.84
Saint Petersburg	2.21
Sofia	1.99
Renens	1.89

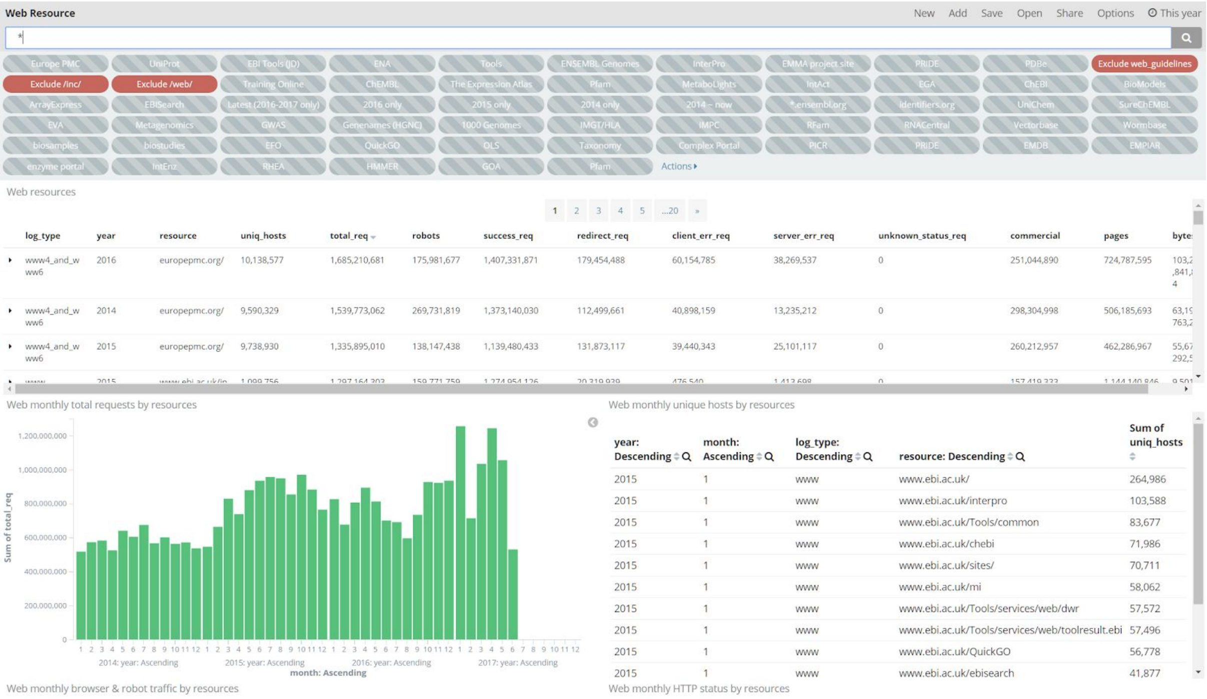
SOURCE	value
(direct)	29.86
google	25.54
reddit.com	9.14
twitter.com	5.59
motherboard.vice.com	3.32
lifehacker.com	3.05

NETWORKDOMAIN	value
ertelecom.ru	28.18
unknown.unknown	16.52
(not set)	6.55
unige.ch	3.69
unil.ch	2.44
clodo.ru	2.31

5.2.2 Method used at EMBL-EBI

Similarly, the **EMBL-EBI has a uniform method of collecting and aggregating web and download usage data**. This relies on Elastic Stack technologies (www.elastic.co/products) and provides a presentation layer and a statistical analysis framework. In summary, EMBL-EBI keeps the summary usage data on a single spreadsheet and a summary sheet for each resource. These are updated once a year. In reference to Section 4.2, the EMBL-EBI "data usage repository" interface basically consists of a few customisable documents and a (many-tabbed) spreadsheet. Figure 6 shows one of the dashboards developed to provide access to this data to the data resource owners.

Figure 6. EBI’s dashboard providing access to its baseline statistics.



Web resources

log_type	year	resource	uniq_hosts	total_req	robots	success_req	redirect_req	client_err_req	server_err_req	unknown_status_req	commercial	pages	byte:
www4_and_www6	2016	europepmc.org/	10,138,577	1,685,210,681	175,981,677	1,407,331,871	179,454,488	60,154,785	38,269,537	0	251,044,890	724,787,595	103,2 ,841, 4
www4_and_www6	2014	europepmc.org/	9,590,329	1,539,773,062	269,731,819	1,373,140,030	112,499,661	40,898,159	13,235,212	0	298,304,998	506,185,693	63,1 763,2
www4_and_www6	2015	europepmc.org/	9,738,930	1,335,895,010	138,147,438	1,139,480,433	131,873,117	39,440,343	25,101,117	0	260,212,957	462,286,967	55,67 292,5
www	2015	www.ebi.ac.uk/in	1,099,756	1,297,164,303	159,771,759	1,274,954,126	20,319,929	476,540	1,413,698	0	157,419,333	1,144,140,846	9,501

Web monthly total requests by resources

Web monthly browser & robot traffic by resources

Web monthly unique hosts by resources

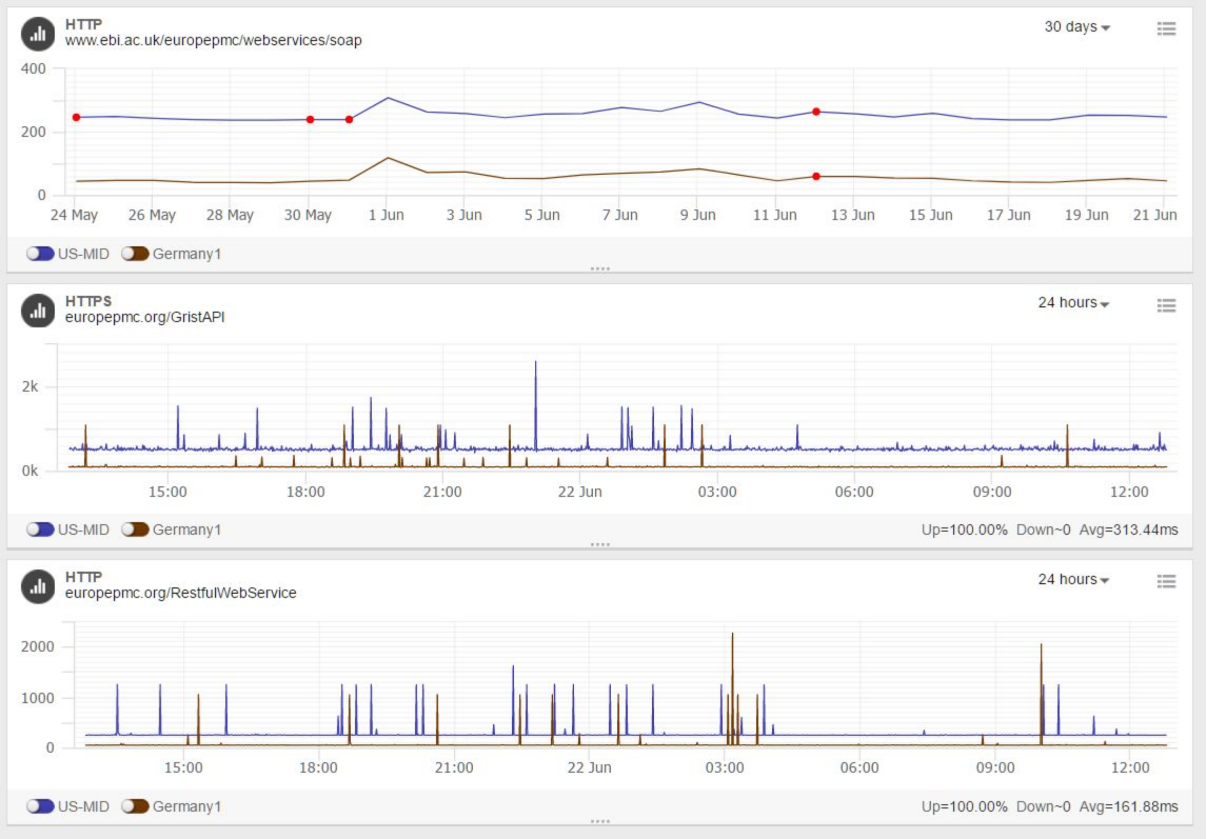
year: Descending	month: Ascending	log_type: Descending	resource: Descending	Sum of uniq_hosts
2015	1	www	www.ebi.ac.uk/	264,986
2015	1	www	www.ebi.ac.uk/interpro	103,588
2015	1	www	www.ebi.ac.uk/Tools/common	83,677
2015	1	www	www.ebi.ac.uk/chebi	71,986
2015	1	www	www.ebi.ac.uk/sites/	70,711
2015	1	www	www.ebi.ac.uk/mi	58,062
2015	1	www	www.ebi.ac.uk/Tools/services/web/dwr	57,572
2015	1	www	www.ebi.ac.uk/Tools/services/web/toolresult.ebi	57,496
2015	1	www	www.ebi.ac.uk/QuickGO	56,778
2015	1	www	www.ebi.ac.uk/ebisearch	41,877

Web monthly HTTP status by resources

For monitoring of uptime and customisable Service Level Agreements EMBL-EBI has looked at and in some cases tested a number of services, including interseek, netnanny, akamai aleksa, monitis.com, and of course Google Analytics (which is not suitable for EMBL-EBI due to its high traffic). Of these, EMBL-EBI has chosen to use monitis.com, which offers flexible and multiple checks including: HTTP, TCP and UDP ports, FTP, IMAP, and SOAP, and which, over time, has proven to be quite reliable.

Of interest in the context of core services, the HTTP checks from at least two locations provide a reliable measurement and have been in constant use for the last four years. The monitoring service EMBL-EBI uses is not free: however the free version of the service does offer the monitors described above for a limited number of hosts/services.

Monitoring reports can be configured in various ways. EMBL-EBI receives reports quantifying uptime on a daily, weekly, and monthly basis, and on Service Level Agreements monthly. Figure 7 below shows a screenshot of one of the Monitis dashboards, in this case showing the Europe PMC service (europepmc.org).



6 Outlook

Up until now, indicators have been only received once, via “static” documents, as part of the initial ELIXIR Core Data Resource selection procedure, and no follow up about updates has been communicated to resource providers. In the next period, more details about the data collection process, workflow and periodicity and a possible technical implementation need to be defined. The case examples in Sections 5.1 and 5.2 can be used as building blocks for a future information system. Any process and solution needs to be defined in cooperation with resource providers, the ELIXIR Hub and various other stakeholders in ELIXIR (such as ELIXIR-EXCELERATE WP3, etc.).

The next step is to **invite the managers of the ELIXIR Core Data Resources to discuss the basic methods defined here** and to **agree on a complete data exchange process and a technical solution for a data indicator repository**. A first meeting is foreseen for Q4 2017.

References

[R1] Durinx C, McEntyre J, Appel R *et al.* Identifying ELIXIR Core Data Resources [version 2; referees: 2 approved]. *F1000Research* 2017, **5**(ELIXIR):2422 (doi: 10.12688/f1000research.9656.2)

[R2] ELIXIR Core Data Resources https://www.elixir-europe.org/platforms/data/core-data-resources, 25 July 2017

Appendix 1: List of 23 Indicators

In [R1], a set of indicators is listed. Here, a comprehensive list is provided adding details to the main indicators listed in Figure 1 and Table 1:

1. Scientific focus and quality (4 main indicators)

- a. *Archives vs knowledge bases*
- b. *Scope statement*
- c. *International dimension*
- d. *Staff effort*
  - i. Curators
  - ii. Bioinformaticians
  - iii. Technical staff

2. Community (4 main indicators)

- a. *Overall usage*
  - i. Access via a web browser: number of visits, unique visitors, hits, and page views
  - ii. Access via additional access methods: visits, unique visitors, hits, and downloads
- b. *Potential usage*
- c. *Usage in research as measured through citation in the literature*
  - i. Citation of a resource name
  - ii. Citation of data of a resource
  - iii. Key publications describing the resource list
- d. *Dependency of other resources*

3. Quality of service (7 main indicators)

- a. *Identifier use*
- b. *Data throughput*
- c. *Technical performance*
  - i. Uptime
  - ii. Response times of key web pages.
- d. *Use of standards*
- e. *Links to documentation of provenance*
- f. *Data availability* - access services and formats
  - i. Data sharing services
  - ii. Data sharing formats
- g. *Customer service*
  - i. Helpdesk
  - ii. User feedback
  - iii. Training

4. Legal and funding infrastructure, and governance (5 indicators)

- a. *Scientific Advisory Board*
- b. *Open Science*
- c. *Privacy policy*
- d. *Ethics policy*
- e. *Sustainable support and funding:*

5. Impact and translational stories (3 indicators)

- a. *Counterfactual*
- b. *Accelerating science*
- c. *Translational data*

Appendix 2: Methods for Community and Service Indicators

We here describe **example methods and suggestions** to obtain certain indicators in a consistent, standardized way where possible and practical. The indicators we will consider are:

A2.1: Community indicators *Overall usage (**indicator 2a**)*
A2.2: Community indicators *Usage as measured through citation in the literature (**indicator 2c**)*
A2.3: Quality of service indicators *Technical performance (**indicator 3c**)*

A2.1 Community indicators - Overall Usage

A2.1.1 Technology

There are two technologies to collect overall usage indicators:

- **Web analytics,** based on a JavaScript library that runs on the **client** browser (e.g. Google Analytics), allow obtaining information about *sessions*, *unique visitors,* and *pageviews.*
- **Log analytics**, based on the analysis of log files on the **server** side, allow obtaining information about *sessions*, *unique visitor*, *hits* and *downloads. Page views* can be estimated. Note that *sessions* may be difficult to compute if log files are distributed.

Web analytics is generally easier to setup but does not track 100% of requests because JavaScript may not be executed on the client side (blocking add-on, bot, etc.). On the other hand, log analytics is in general more complicated to setup and often requires dedicated hardware and infrastructure.

A possible setup which takes advantage of both technologies is to use *web analytics technology* to compute access via a web browser, and *log analytics technology* to compute additional access methods.

A2.1.2 Overall usage indicators - definitions

(1) Visits/Sessions
A session, also referred to as a visit, is a set of requests/interactions done by the same uniquely identified client, 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 closer together. 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). A unique client is commonly identified by an IP address or a unique ID placed in the browser.

(2) Pageviews (also referred to as “pages”, “impressions” or “URLs”)
Pageviews or impressions correspond 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 pageviews. This indicator varies considerably based on the implementation and technology.

This indicator can be computed using web analytics. Alternatively, it can be estimated by using log analytics by filtering HTML files only, and this works for traditional websites (one single HTML page). Note that with the emerging technology of the web (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.

(3) Unique users (visitors) or IP addresses
This indicator 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 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).
- unique visitors can also be determined by a user cookie in case of web technology.

Note that the concept of “unique IP addresses” is a proxy for the number of users, but is not an exact count. Almost all users have multiple IP addresses due to use of multiple devices and/or dynamic IP addressing. Conversely, many users may appear to have the same IP address if their institution configures its system to show only one or a few addresses to the outside world.

(4) Hits/Requests
Hits or requests 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. If there are 12 such files on a given web page, each time the page is viewed this will correspond to 12 hits. A hit includes the results of a single request made to a server via HTTP/HTML, FTP, REST API or other. It includes the following files: .html, .css, .js, .png, .jpg, .xml, .json, .txt etc. When a web page is rendered in a web browser, the number of "hits" or "page hits" or “requests” is equal to the number of files requested (for example the website www.bbc.com needs more than 150 objects to render the home webpage and therefore generates more than 150 hits for one single web page). This value may differ considerably depending on the technology and server configuration caches. This indicator can be used to analyse trends of a specific resource.

(5) Downloads
This indicator measures the size of the data downloaded from resource in terms of volume / bandwidth (commonly measured in GB). This indicator can only be computed using log analytics: the value can be taken from server logs (HTTP, FTP, etc. assuming that the respective servers are explicitly configured to allow for this) and summed up to compute this indicator.

A2.1.3 Views and representation of the data

Each of the indicators mentioned above can be segmented (represented) in different *dimensions* to give reports an even more valuable insight. Some examples are given below to illustrate how data can be reorganised/aggregated and visualised.

(1) Demographics
This dimension tells in which **continent, country** or even **city** most of the community is located. This representation is easily achievable using web analytics technology like Google Analytics. For log analytics solutions, one may require an external database like: ip2location (commercial), maxmind (commercial), geolitecity (free but not comprehensive). Note that some countries may pass all their international traffic through a single city, or small number of cities, and therefore for some countries the geographic access data will be of low granularity. Depending on the database used, some countries may not specifically "exist" (e.g. Scotland and Northern Ireland are considered part of the UK): this is the case for maxmind. However, ip2location does provide this regional information.

Other aspects of usage that are often important when reporting usage is the type of use. The list below illustrates what kind of usage types are available in ip2location DB24:

> (COM) Commercial, (ORG) Organization, (GOV) Government, (MIL) Military, (EDU) University/College/School, (LIB) Library, (CDN) Content Delivery Network, (ISP) Fixed Line ISP, (MOB) Mobile ISP, (DCH) Data Center/Web Hosting/Transit, (SES) Search Engine Spider, (RSV) Reserved

(2) Traffic type
There are 3 different types of traffic:

A. **Real users** (humans) who access via a web browser like Google Chrome, Firefox, Internet Explorer, Edge, Safari, etc.
B. **Scripts and programs** that may download some data (for example "wget" on an XML file)
C. **Spiders, robots, crawlers** that index content on the website (examples: Google and Bing bots, Baidu and Saigu robots, Yahoo spiders and Yandex crawler).

The traffic referred to in B "Scripts and programs" is in general not possible to capture with web analytics technology. The C traffic "Spiders, robots, crawlers" is also difficult to measure with a web analytic solution. (Piwik image tracker is able to track some of that traffic (see Section A2.1.4 below)). For ELIXIR, traffic measured by A and B is most relevant.

(3) Content Type / File format
This dimension tells what **file formats** are used such as JSON, XML, TSV, RDF etc. This information is only available using log analytics. For Apache web servers: if "mode extended" is defined, this can be captured using the "Apache Log Extended" format, looking at the MIME type. Another possibility is to analyse the extension of a hit (i.e., analyse URLs).

(4) Errors
Errors are captured using log analytics. For web access logs the errors correspond to **HTTP status codes** that are 400 or higher. For example, error code 404 indicates that a web page is not found. This information is very useful to help troubleshooting of web applications, i.e., to help improve quality of service.

(5) Top pages (Equivalent: Top URLs)
This dimension tells which is the **page or URL** that is accessed the most.

(6) Top users (Equivalent: Top IP addresses)
This dimension tells which is the **user or IP address** that has made most requests. Note that within an institution a single IP address may represent many different users.

A2.1.4 Web analytics solutions / products

This section presents a few concrete solutions that can be used to measure the usage indicators defined in Section A2.1.2 using **web analytics**.

(1) Google Analytics
Google Analytics (http://google.com/analytics) is one of the most popular software systems for web analytics. It has the advantage of being very easy to set up and has a user-friendly dashboard that gives instantaneous insight. This solution is free up to 10 million hits per month. After this value either the data is sampled or one must purchase the 360Insight solution. Google Analytics is biased towards e-commerce web sites. Additionally, one must consent with the terms of data privacy since the data is hosted at Google. Google Analytics has made efforts itself to be consistent along the years, the only major impact which has been seen was in 2011, where they have changed the way sessions were computed.

(2) Piwik
Piwik (https://piwik.org/) is another very popular web analytics software. It can be installed locally and is therefore preserving the data privacy. It also supports *image*

tracker which will allow to have a more representative information about bot tracking and overcome limitations like browsers blocking JavaScript.

(3) Others
A more complete list of available solutions (e.g. Adobe Analytics, Kissmetrics, Open Web Analytics, etc.) can be found here https://en.wikipedia.org/wiki/List_of_web_analytics_software).

A2.1.5 Log Analytics solutions / products

There is also a wide range of server side analytics products for **log analytics** that complement the above-mentioned web analytics products.

(1) Piwik
Piwik can be used both as a web analytics solution and as a log analytics solution: http://piwik.org/log-analytics.

(2) Splunk
Splunk (splunk.com) is relevant if the index volume does not go higher than 500 MB/day. One can benefit from a free license under those conditions.

(3) Others
There are several additional solutions available (e.g. graylog, analytics.AngelFishStats.com, GoAccess.io., etc.). If log data is so big that it does not fit into a single machine, a solution is to use Big Data technology to create log analytics. There are multiple technologies that are suitable to analyse big data logs: ELK Elastic Stack. Apache Spark, Hadoop with PIG, etc.

A2.2 Community indicators - Usage as measured through citation in the literature

A2.2.1 Background

There are several different methods to measure citations, e.g. via Google scholar, ISI Web of Science, PubMed etc. If one wants to obtain more detailed information about a scientific resource, e.g. the number of times the resource name is mentioned in articles or the number of times data (e.g. accession numbers) are cited, one needs to have access to full text articles. Since Europe PubMed Central (Europe PMC, https://europepmc.org/) provides such features, Europe PMC is the recommended method to measure indicators related to literature and usage in research.

For measuring citations of resource names and accession numbers, text mining (TM) has been applied to Europe PMC Open Access articles (mentioned in Section 5.1).

A2.2.2 Indicators

(1) Citations of resource name in Europe PMC
This indicator lists the number of times a **resource name** is mentioned in various scientific articles in the literature, i.e., a full-text search of published literature is done in order to obtain the number of times a resource is mentioned for a defined period, e.g. over the last 5 years.

In order to obtain this number from Europe PMC, one needs to specify the *resource name*, *its alias and abbreviations*. Then, a namespace will be assigned for the use of search APIs on Europe PMC. Please contact: helpdesk@europepmc.org

(2) Citations of accession numbers in Europe PMC
This indicator measures how often **data** of a resource are cited. In particular, data items of a resource are identified by an accession number or other means. A full-text search of available literature is done in order to retrieve this number of data citations.

In order to add new accession numbers for a new resource not yet known to Europe PMC, the pattern of accession numbers needs to be defined and provided to Europe PMC. Then, the search algorithm will be adapted to search for the new accession numbers. Please contact: helpdesk@europepmc.org

(3) Citations of articles of resource
This is the “classical” citation count for published **articles** of a resource. In detail, for each article describing the specific resource the number of citations is counted, i.e., the number of times this article is used in the reference list of another scientific article.

A2.3 Quality of the service (Performance)

Performance of a web resource can be measured using two indicators:

- Uptime
- Response time

A2.3.1 Indicators

(1) Uptime
The **uptime** quantifies the amount of time a website is up and running. For example, if a site has an uptime of 99.6%, this means that during one month the site was down for a bit less than 3 hours in total. Example calculation of a month with 30 days: ((100-99.6) /100) * 30 * 24) = 2.88 hours, i.e., ~ 3 hours.

(2) Response time

Response time measures how long a site takes to respond to a request. Essentially, this depends on the geolocation of the resource compared to the point of monitoring, but also depends on the technology and tuning of a website.

A2.3.2 Possible solution

In order to measure those indicators, one can use **Monitis** (monitis.com) which is a reliable service and relatively cheap (80 cents per resource, as of July 2016). Monitis has the following features that are interesting for ELIXIR resources:

- Allows to deploy monitoring points around the world (for example USA, Switzerland, etc.) and compare the different response times around the globe.
- Has a robust system of alerts where one can configure groups and send alerts (email, SMS, phone calls) to a certain group of people.
- Has a live support (chat) which may be very handy for troubleshooting and initial setup.

In order to avoid false results it is recommended to tune the different monitors: for example, it may be useful to set up alerts after 2 consecutive fails instead of 1 fail which is the value by default.