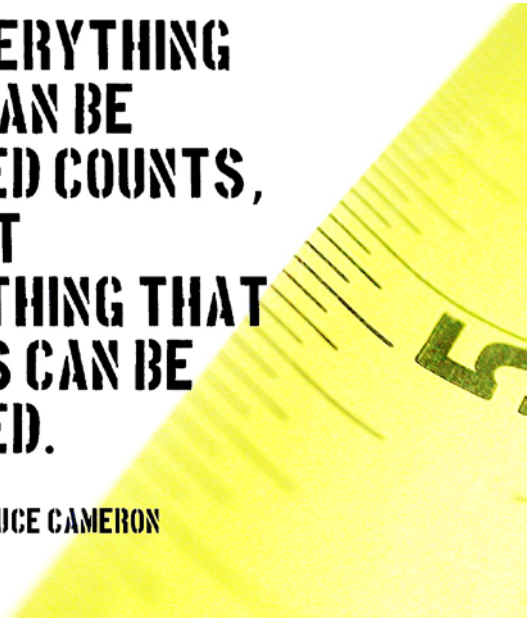
ELIXIR KPI framework and alignment of KPIs between e-Infrastructures and research infrastructures

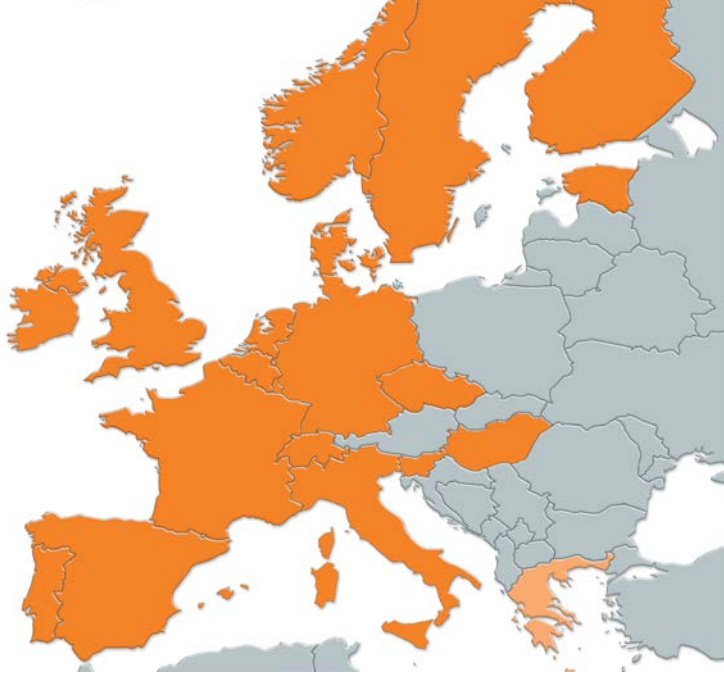
Rafael C Jimenez, ELIXIR CTO

June 8, 2017

e-IRG workshop June 2017



ELIXIR

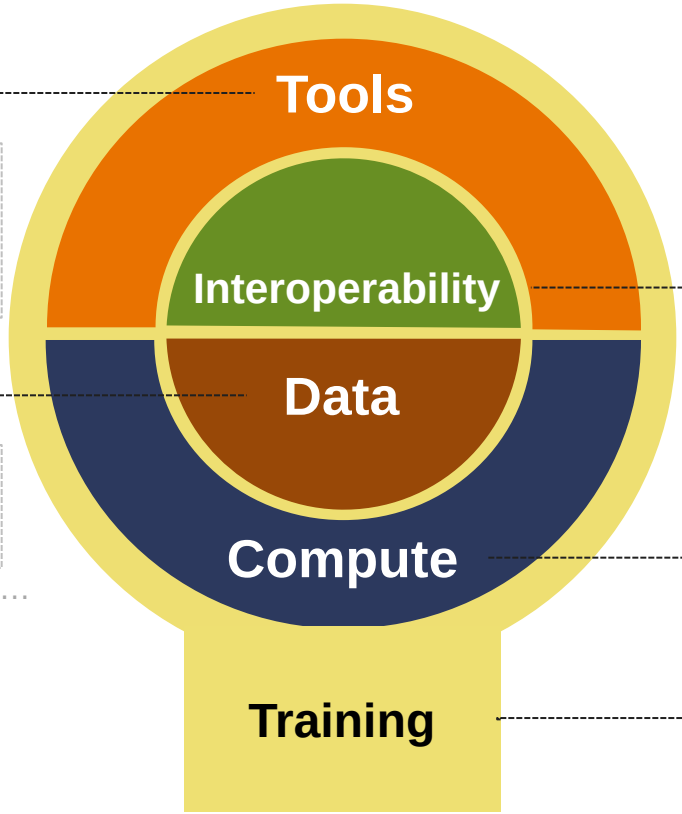


European **distributed** Research Infrastructure for **biological research**

Participated by major European bioinformatics **service providers (~180)** and supported by **EU member states (21)** & **EMBL-EBI**

Provide **data services** essential to **enable, sustain, or enhance** biological science

Infrastructure for Life Sciences



Services & connectors to drive access and exploitation

Access, Search, Analysis …

Sustain core data resources

Integration, Optimization, Privacy, …

Formats, Ontologies, Guidelines, …

Integration of data and services

Storage, Network & Computing

Access, Exchange & Compute on sensitive data

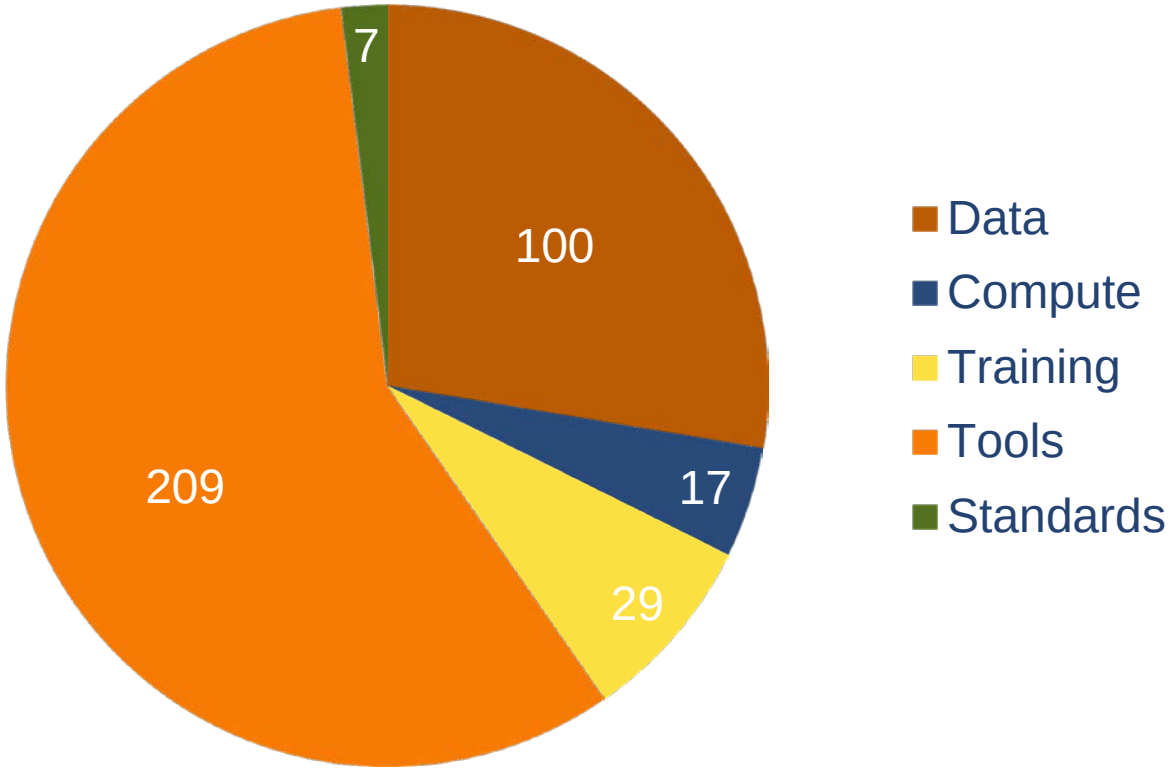
Scientific & technical

Professional skills for managing and exploiting data

ELIXIR programme

http://www.elixir-europe.org/sites/default/files/documents/elixir_scientific_programme_final.pdf

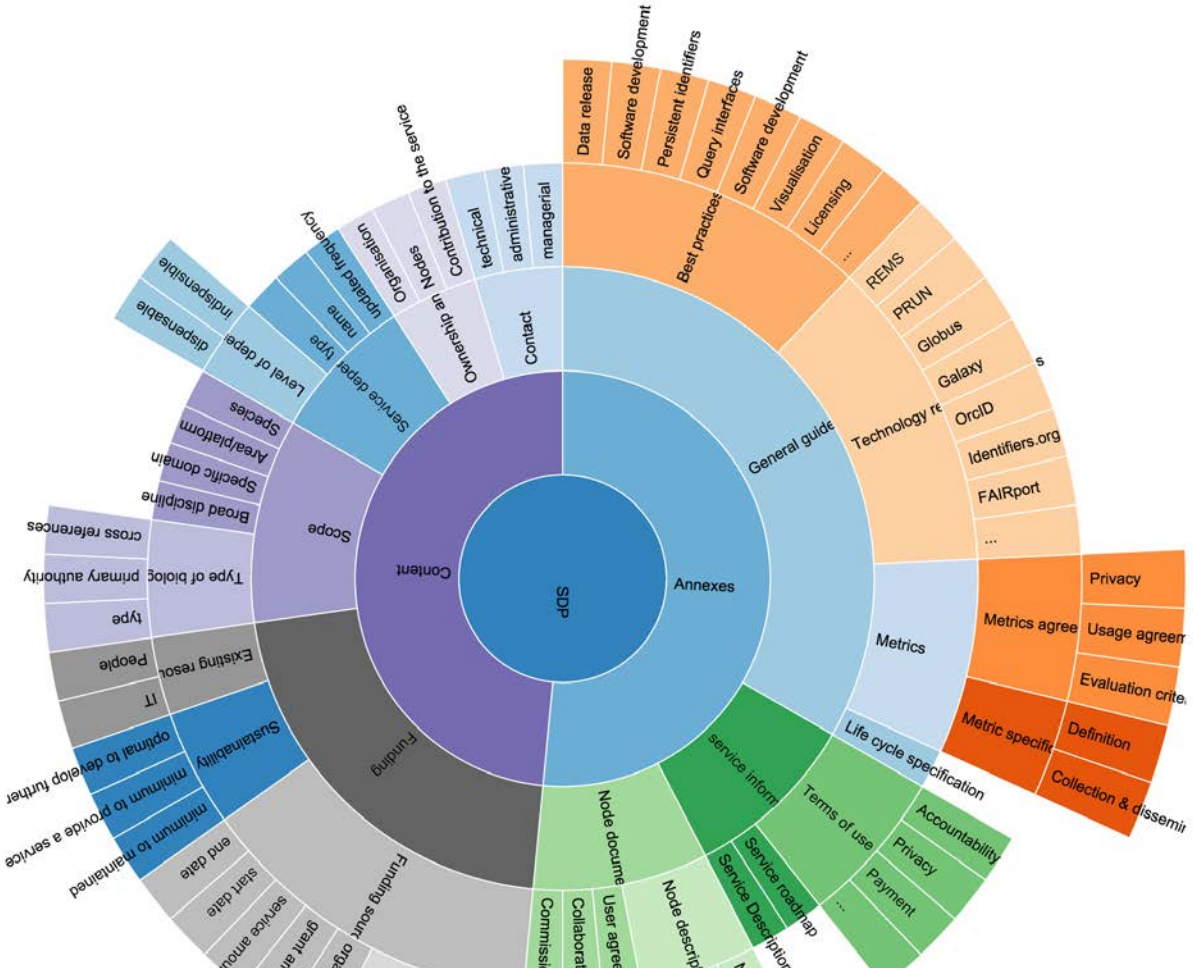
ELIXIR Nodes Service Inventory



https://www.elixir-europe.org/services

Service Delivery Plans

- Description of services provided by ELIXIR nodes
- Commitment from Nodes to provide ELIXIR services



Evaluation and assessment

- Infrastructure
  - **Services**
  - Projects
  - Groups
    - Nodes, Platforms, Teams, …
- Not just performance (efficiency) but impact, quality, adoption of best practices, …

Indicators for

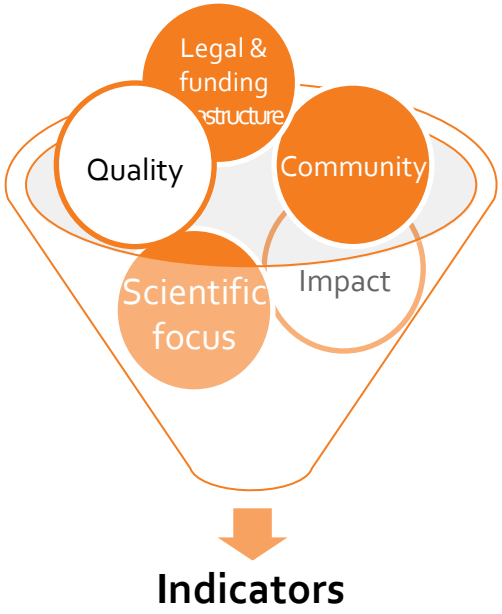
Progress	Platform	Target	Assessment
+	Data	**Data resources**	Relevancy, Usage, Reliability, Sustainability, Impact, …
	Tools	**Software development**, functional benchmarking	Discoverability, Openness, Reusability, Transparency, Best practices, …
	Training	**Courses**	Participation, Impact *(Geographical, Career, Research, …),* Demand, Applicability, …
	Interoperability	FAIR principles, Data management plans	Findability, Accessibility, Interoperability, Reusability, …
-	Compute	Compute services	Usage, reliability, …

Indicators for data
& interoperability

Objective

- **Identification** and **evaluation** of **data resources** fundamental for global life sciences research
  - Support long term **sustainability**
  - Improve **management**, **operation** and **development**

A carefully chosen basket of indicators, reflecting the multiple facets of bioinformatics resources



1. **Scientific focus** and quality of science
2. **Community** served
3. **Quality** of service
4. **Legal and funding infrastructure** and governance
5. **Impact** and translational stories

Christine Durinx (SIB) & Jo McEntyre (EMBL-EBI)

1. Scientific focus and quality of science

Measuring what? Inherent scientific quality, of the resource, its uniqueness and comprehensiveness. Relevance of the resource.

Indicators:

a. ***Archives vs knowledge bases***

b. ***Scope*** *statement*: scientific coverage and comprehensiveness

c. ***International*** *dimension*

d. *Staff* ***effort***: including curation effort

Christine Durinx (SIB) & Jo McEntyre (EMBL-EBI)

2. Community served

Measuring what? Usage of the resource

Indicators:

a. *Overall **usage**:* access via web browser and other methods

b. ***Potential** usage*

c. *Usage in research as measured through **citation** in the literature:* the resource name, data of a resource

d. ***Dependency** of other resources*

Christine Durinx (SIB) & Jo McEntyre (EMBL-EBI)

3. Quality of service

Quality

Measuring what? Service levels and reliability

Indicators:

a. *Use of persistent and unique **identifiers***
b. *Data **throughput**:* number of entries, depositions
c. *Technical **performance**:* uptime, response time
d. *Use of community-recognized **standards** for (meta)data*
e. *Links to documentation of **provenance***
f. *Data **availability*** - access services and formats
g. ***Customer* service**: helpdesk, user feedback, training activities

4. Legal and funding infrastructure, and governance

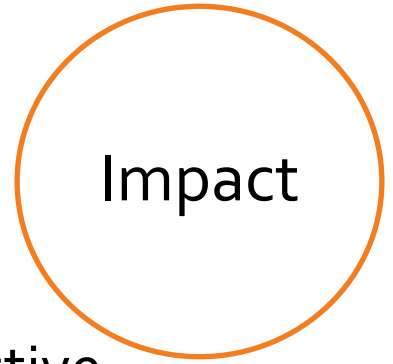
Measuring what? Soundness of the legal, funding and governance structure guaranteeing its long-term stability

Indicators:

a. ***Scientific Advisory Board***
b. *Legal framework supporting* ***Open Science***
c. ***Privacy policy***
d. ***Ethics policy***
e. ***Sustainable*** *support and funding*

Christine Durinx (SIB) & Jo McEntyre (EMBL-EBI)

5. Impact and translational stories



Measuring what? Is the resource meeting its objective of fulfilling a specific need of the scientific community

Indicators:

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

How FAIR are our indicators?

FAIR Principles		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
Findable:	F1									■														
	F2				■								■											
	F3														■									
	F4									■			■											
Accessible:	A1									■					■									
	A1.1														■			■						
	A1.2																	■	■					
	A2																				■			
Interoperable:	I1								■				■		■									
	I2												■											
	I3								■					■										
Re-usable:	R1				■								■											
	R1.1																	■						
	R1.2								■				■	■										
	R1.3												■											

For more details

https://f1000research.com/articles/5-2422/#B2

Check for updates

METHOD ARTICLE

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

Christine Durinx[1], Jo McEntyre[2], Ron Appel[1], Rolf Apweiler[2], Mary Barlow[2], Niklas Blomberg[3], Chuck Cook[2], Elisabeth Gasteiger[4], Jee-Hyub Kim[2], Rodrigo Lopez[2], Nicole Redaschi[4], Heinz Stockinger[1], Daniel Teixeira[1], Alfonso Valencia[5]

Author details

Grant information

This article is included in the ELIXIR gateway.

Abstract

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 offers to the life-sciences community, providing stable and sustainable access to biological data. 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, including biological research.

ELIXIR Core Data Resources are defined as a set of European data resources that are of fundamental importance to the

Open Peer Review

Referee Status: ✓ ✓

Version(s)	Invited Referees 1	2
REVISED Version 2 published 09 Mar 2017		
Version 1 published 30 Sep 2016	✓ read report	✓ read report

1 **Helen Berman** , Rutgers, The State University of New Jersey , USA

2 **Maryann E. Martone** , University of California, San Diego, USA Hypothes.is, USA SciCrunch.com, USA

All reports (2)

Comments on this article

All comments (0)

Add a Comment

Indicators for tools

Objectives

- Raise **Quality** and **Sustainability** in software **development** in the life sciences
  - Promote best practices
  - Assess adoption of recommendations
- Benchmark **functional performance** of software (bioinformatics **methods**)

netherlands eScience center

Software Sustainability Institute

elixir

Version 1. F1000Res. 2016; 5: ELIXIR-2000.
PMCID: PMC5007752
Published online 2016 Aug 16. doi: 10.12688/f1000research.9206.1

Top 10 metrics for life science software good practices

Haydee Artaza,[#1] Neil Chue Hong,[#2] Manuel Corpas,[#a,1] Angel Corpuz,[#3] Rob Hooft,[#4] Rafael C. Jimenez,[#5] Brane Leskošek,[#6] Brett G. Olivier,[#7] Jan Stourac,[#8] Radka Svobodová Vařeková,[#b,9] Thomas Van Parys,[#10] and Daniel Vaughan[#11]

Author information ► Article notes ► Copyright and License information ►

Peer Review Summary

Go to:

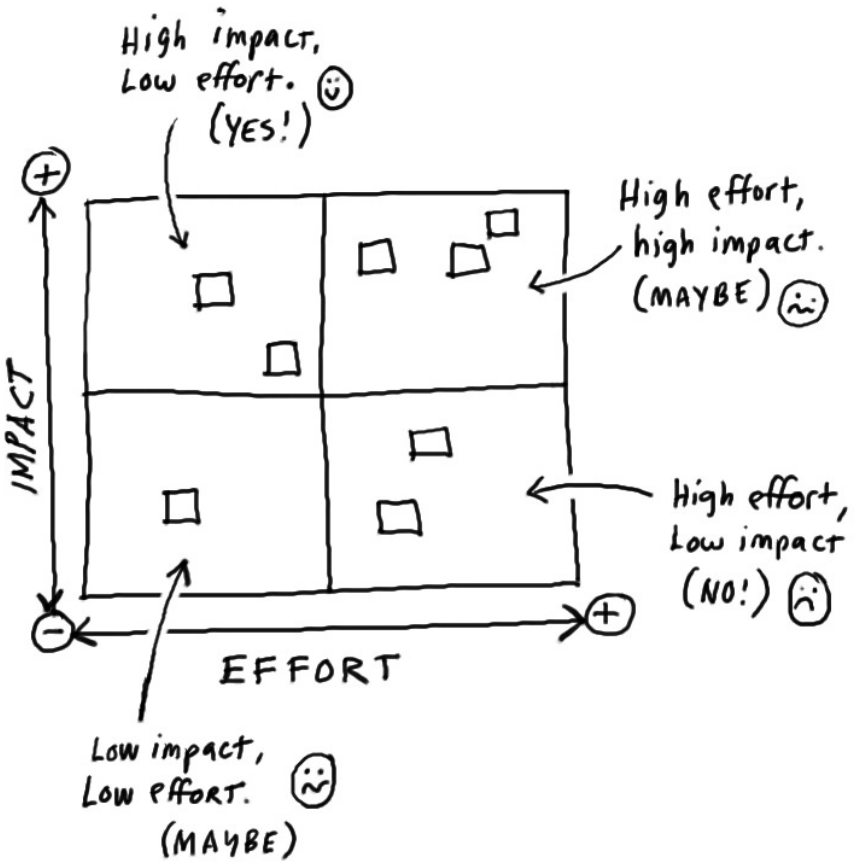
Mapping metrics to good enough practices

- 17 good enough practices
- 43 metrics
- Quantitative as well as qualitative
- 10 metrics Prioritized by impact/effort matrix

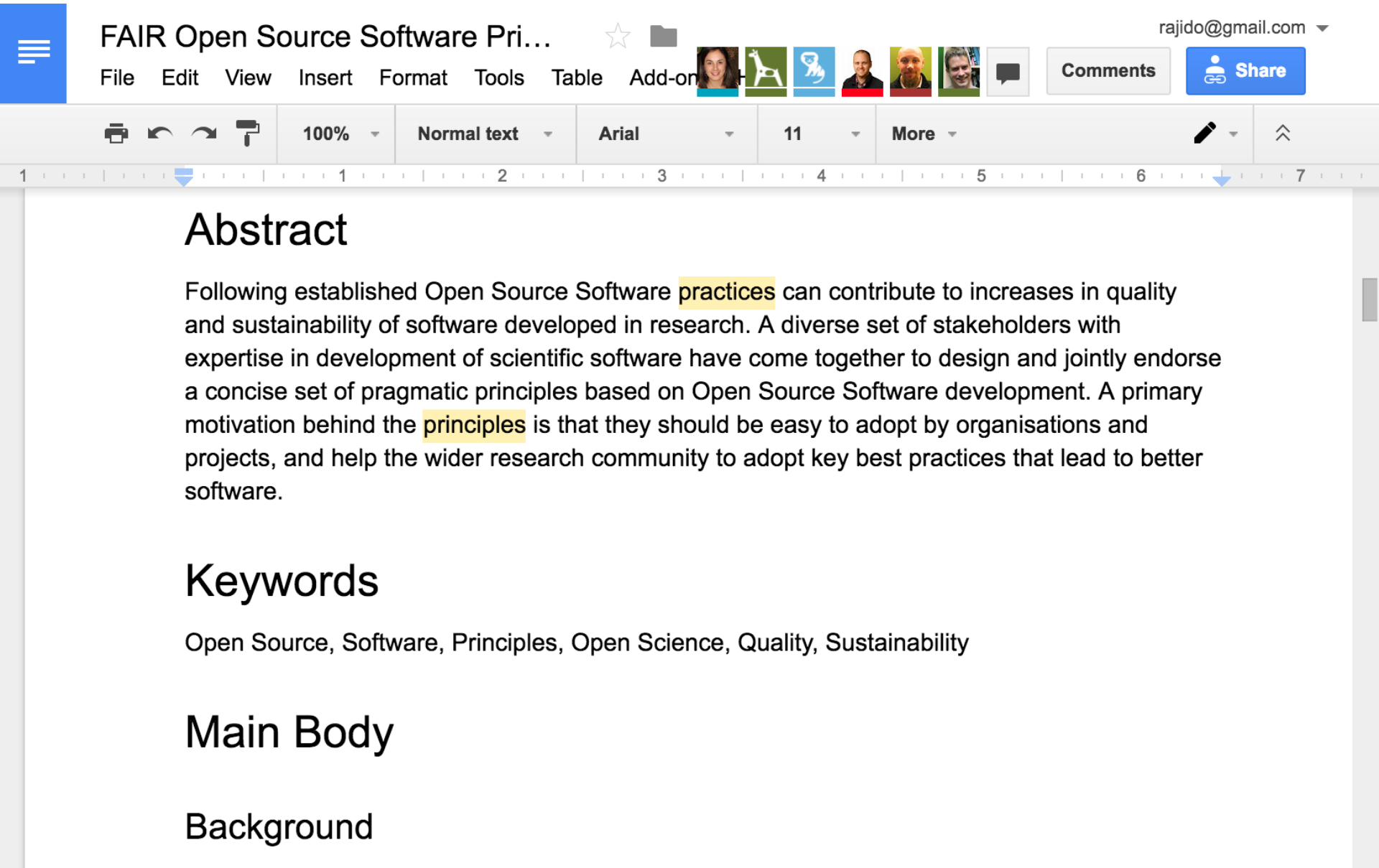
1. **Version control:**
    a. Yes/no?
    b. How many committers?
    c. When was the version control started?
    d. When was the last commit?
2. **Code reviews:**
    a. Yes/no?
    b. Star rating based on code description
3. **Automated testing:**
    a. Yes/no?
    b. Coverage for unit tests
    c. Yes/no for individual tests:
        i. Unit tests

Top 10 metrics (from 43)

1. Is version control used?
2. Is the software discoverable?
3. Is an automated build system used?
4. Are test data available?
5. Does software contain parts that reimplement existing technology?
6. Is the software compliant with community standards?
7. Are code reviews performed?
8. Is automated testing performed?
9. Is the code documented?
10. How high is the code complexity?



Open Source Software Recommendations



Open Source Software Recommendations

Publicly accessible open source code from day one

Start your project in the open from the very first day in a publicly accessible version controlled repository

Source code easily discoverable

Register your software, source code repository, license and contributors in a public registry. i.e. bio.tools or biojs.io

Source code that can be used and reused by other software

Include a license within your publicly accessible repository, and also ensure your software complies with third party software licenses

Clear and transparent contribution, governance and communication processes

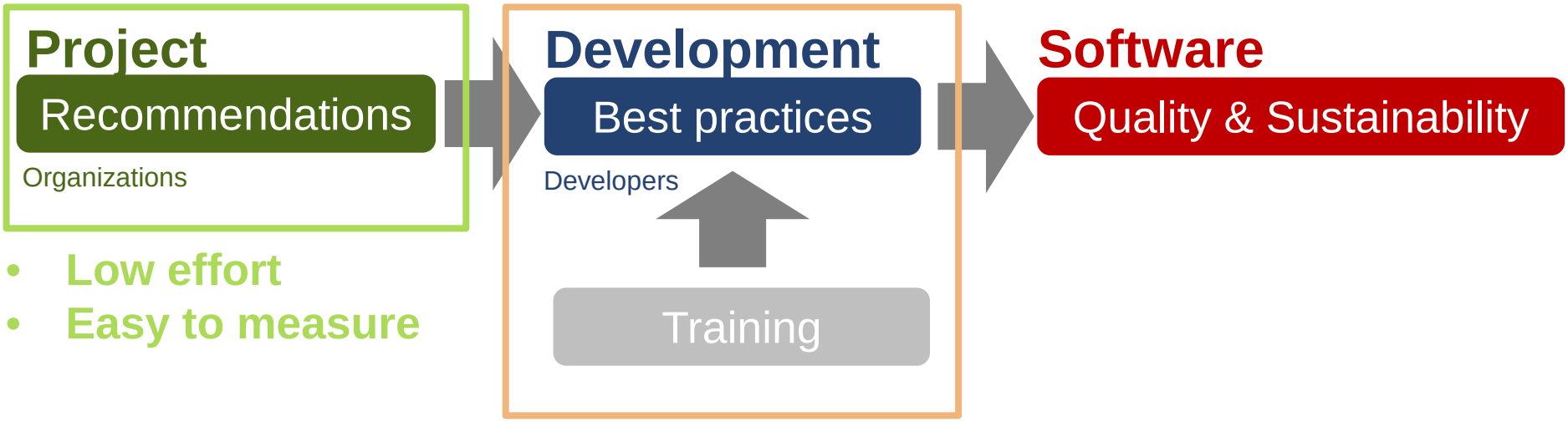
Projects should be clear about how contributions can be made and incorporated by having transparent governance model and communication channels

https://softdev4research.github.io/recommendations

Maximize a chain of cause-effect events

Application of metrics

- Starting to measure adoption of recommendations



- **Low effort**
- **Easy to measure**

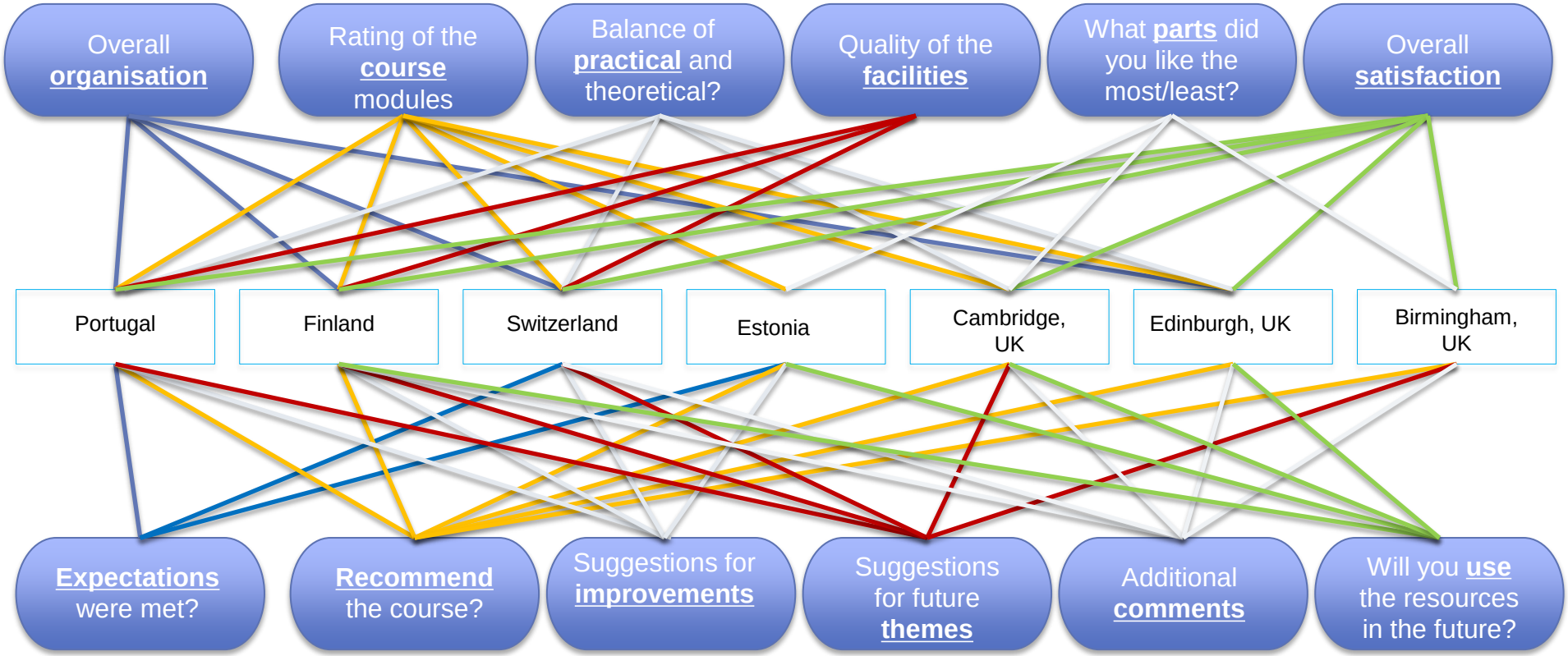
- **High effort**
- **Difficult to measure**

Indicators for training

Objective

- Measure training **quality** and **impact** collecting feedback from **training events** via **surveys** and face-to-face interviews
  - Improve training activities

Analysis of surveys across training providers



Gabriella Rustici, Sarah Morgan and Louisa Bellis

Stakeholder analysis

Metrics to be recorded per training event	Stakeholder(s)	Why?
No. of **participants**, no. of **courses** and overall **days** of training	Heads of Node (HoN), SAB, Industry Advisory Committee (IAC), Elixir Hub, general public, taxpayers	Has there been appropriate investment in funding? Show users what's been carried out?
Geographical breakdown of home country of employment	EU funders, Elixir users	What is the geographical reach and **geographical impact**?
No. people on the **waiting list**	EU funders	Are we meeting the **demand**?
Training resulted in **scientific papers** / **collaborations** / **grant proposal**	General public, taxpayers, national funders, RCs, policy makers, ministries, observers, potential new members	Has there been a positive effect on research/practice/collaborations, **applicability** and **impact** on research at a national level?
Type of **industry/sector**	National funders, RCs, policy makers, ministries	Is there a transfer of skills to different industries?
Career level	HoN, SAB, IAC industry advisory committee, Hub	**Impact** on career on a longer term
Rating of **confidence after training**	Elixir trainers, trainees, Elixir governance, Elixir users	Positive **impact** on the research/work

Data collection – short term

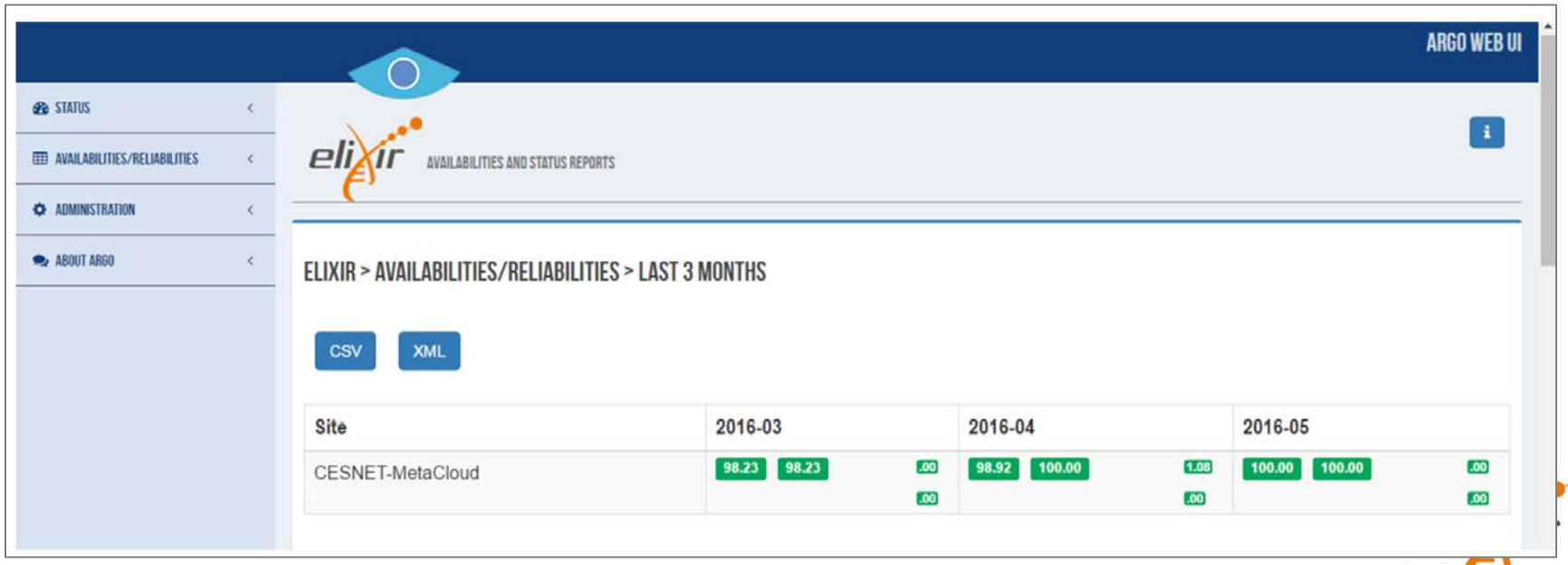
- Same/similar questions across multiple nodes
- Common sets of options for the answers
- Minimum information to capture
  - **Participant** numbers
  - **Gender**
  - **Career level**
  - **Employment sector**
  - **Country** of employment
  - Where did the participant see the event being **advertised**?
  - Overall **satisfaction**
  - Would they be happy to be **contact**ed in the future?

Gabriella Rustici, Sarah Morgan and Louisa Bellis

Indicators for compute

Objective

- ELIXIR Metrics Portal for compute resources
  - Service level monitoring and **reliability** via "EGI ARGO"
  - **Usage** based reports via "EGI APEL"



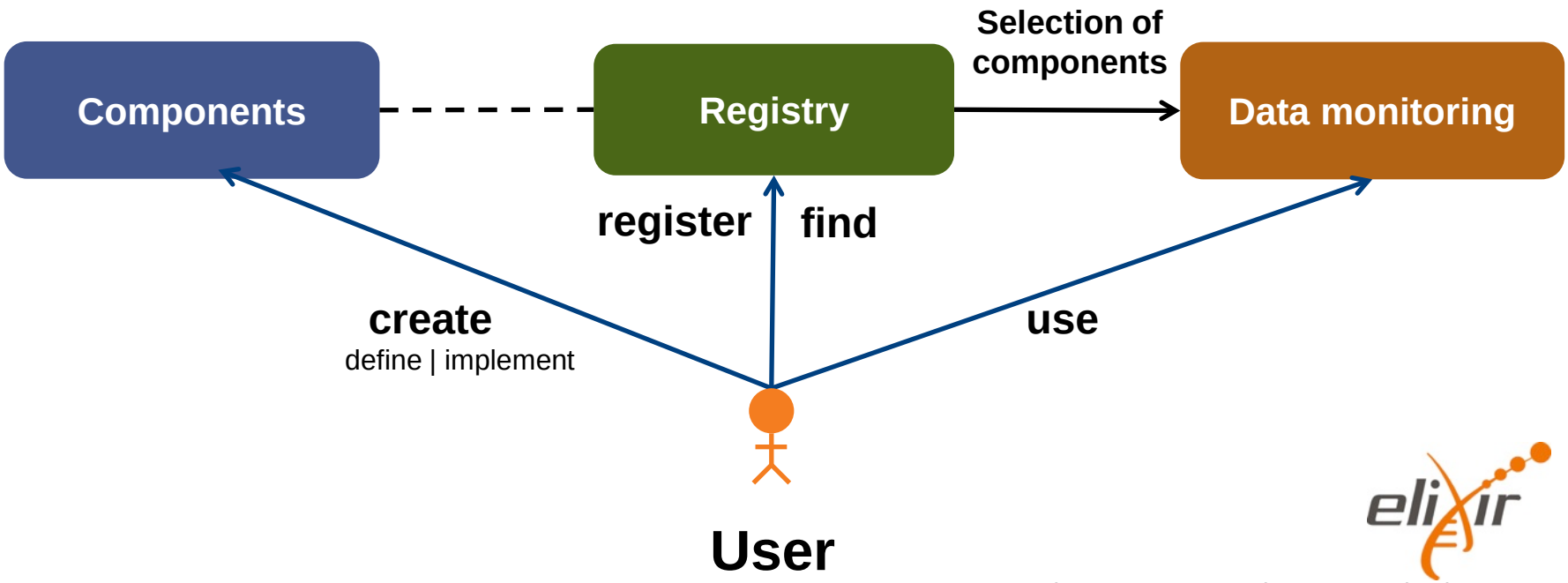
Pisco - metrics framework:

Define and register and monitor

https://github.com/BioPisCO/pisco-metrics-framework

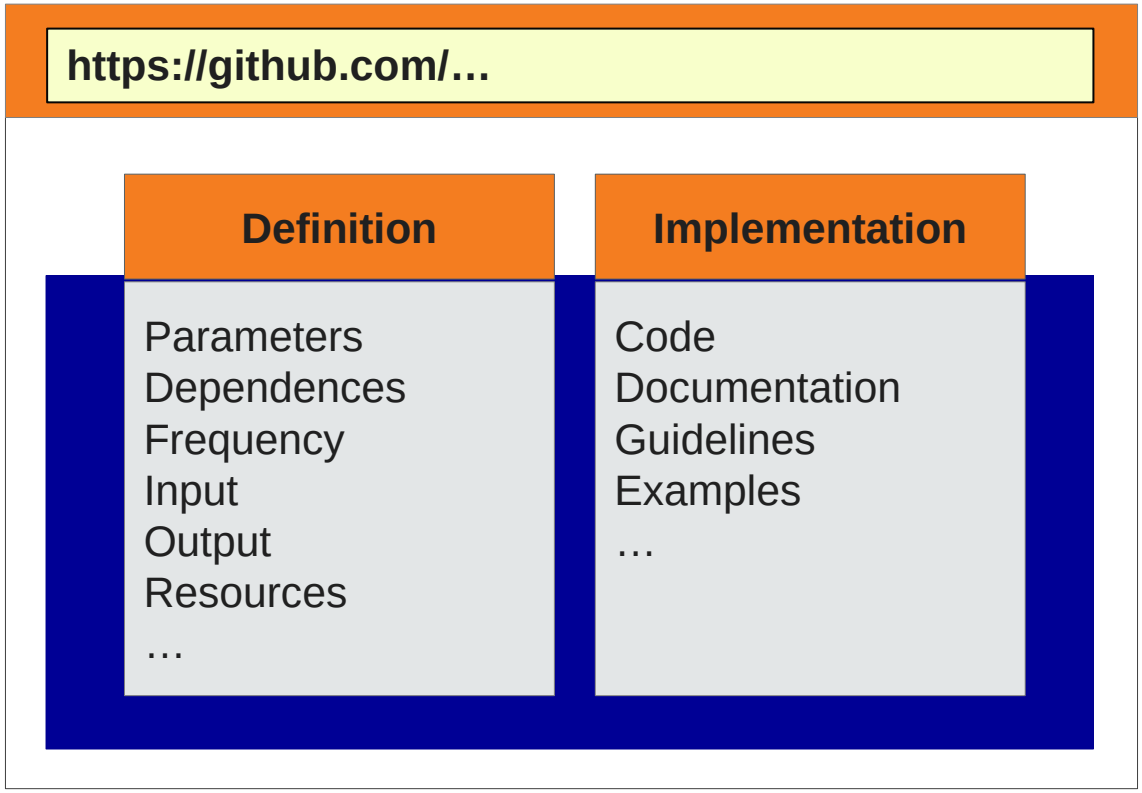
Elements of the framework

- Metrics **components**
- Metrics components **registry**
- Metrics **data** and **monitoring** repository



Metrics components

- Common schema for metrics **definition**
- Common structure for metrics **implementation**
- Defined and implemented **by experts**
- The framework just provide the **guidelines**



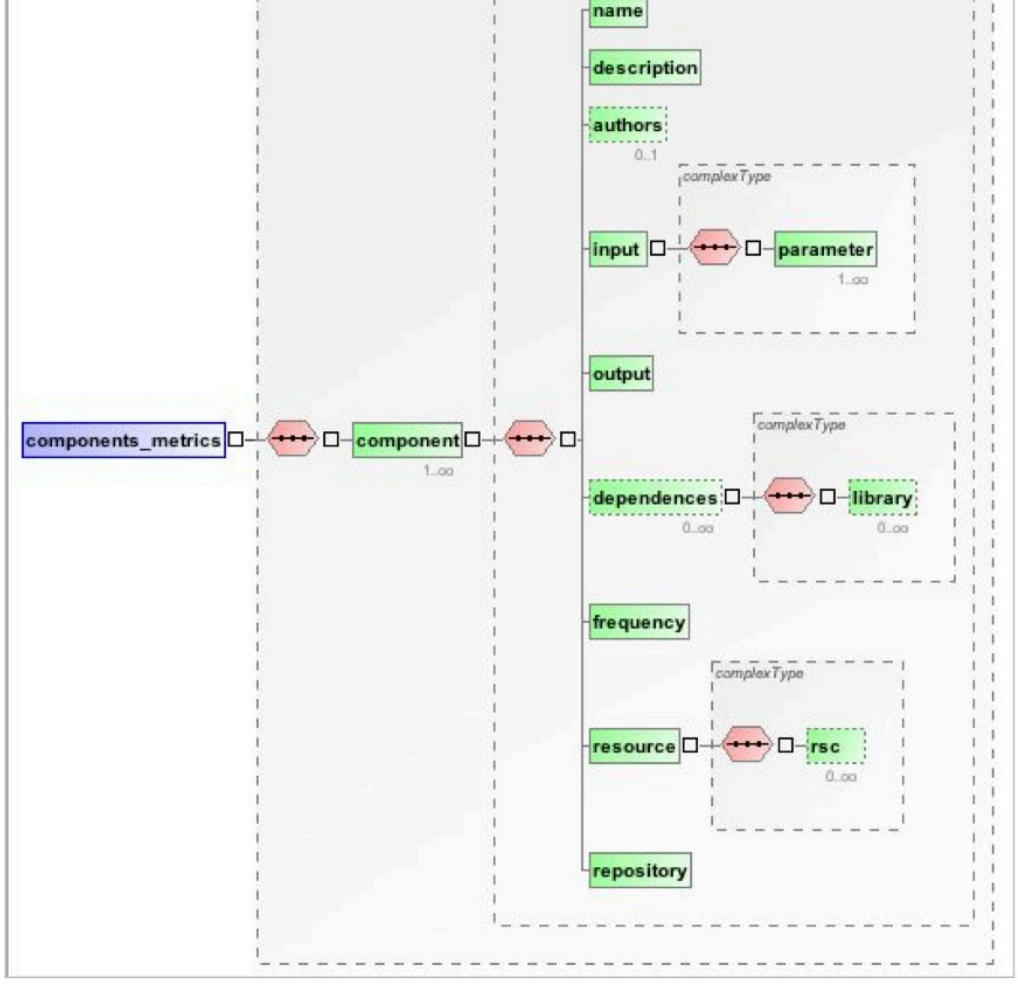
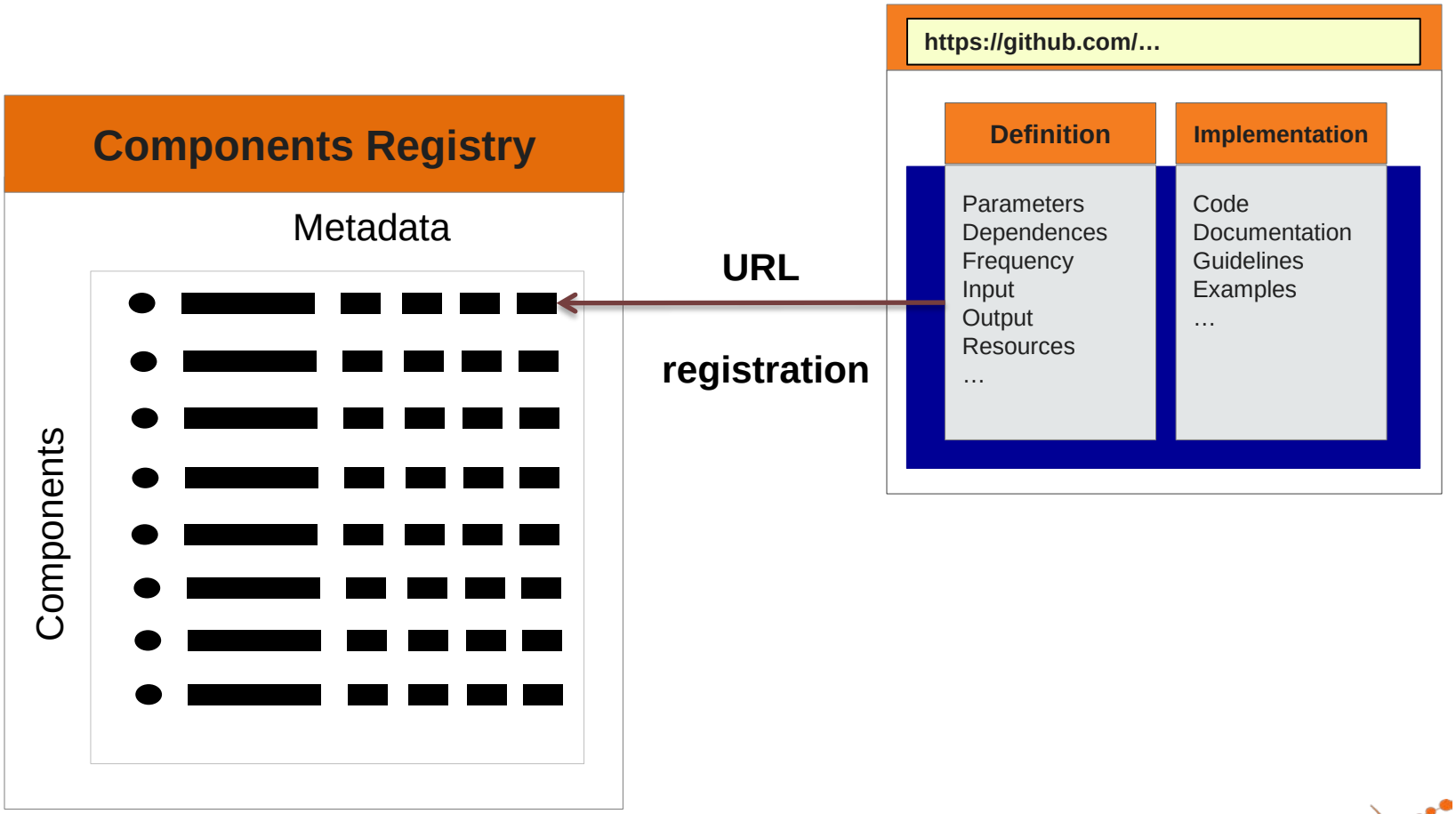


Figure 2. Graphical component schema. This graphic shows the component schema hierarchy.

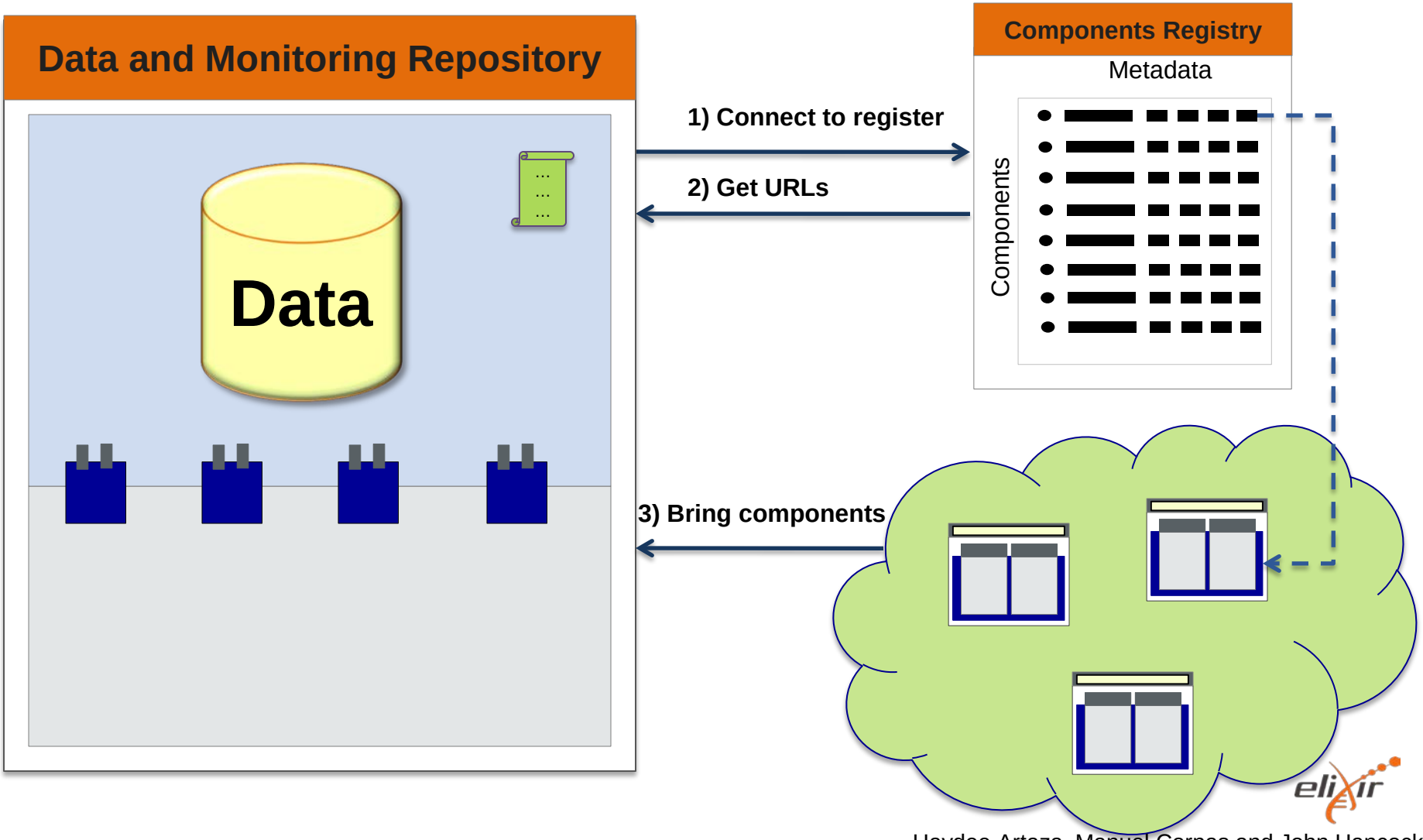
Parameters	Description	Cardinality	Type
Name	Component name	1	String
Description	Short component explanation	1	String
Authors	Authors names separated by commas	0..1	String
Input	Input parameters list used in the component execution.	1..*	String
Output	Type execution result: number, string or list.	1	String
Dependencies	Libraries list necessaries for this component.	0..*	String List
Frequency	Time interval to execute a component:	1	String

Registry

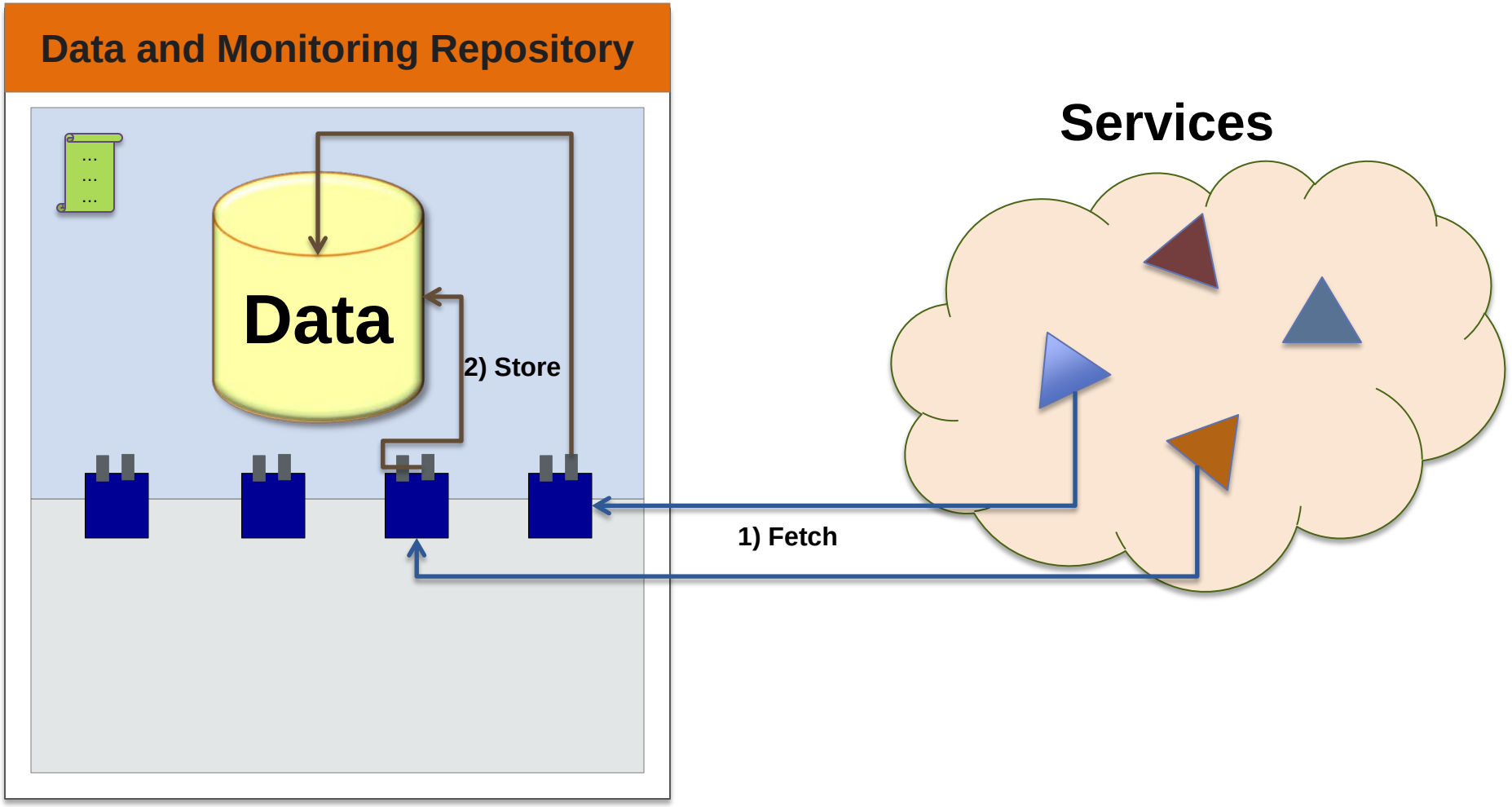
- Metadata registry to **discover** and **register** Components.



Automated installation



Collect data



```
[15:06:23][admin]~/DataMonitoringRepository$ node installcomponents.js
https://nodeload.github.com/elixirhub/metrics-module-citation/zip/master
https://github.com/elixirhub/metrics-module-citation.git download into
components/metrics-module-citation-4ywN_j5H completed
installing library xml2js
xml2js@0.4.9 ../node_modules/xml2js
├── sax@0.6.1
└── xmlbuilder@2.6.4 (lodash@3.9.3)
Library xml2js installed
Component metrics-module-citation-4ywN_j5H has been updated in
schema/selectedcomponents.xml
```

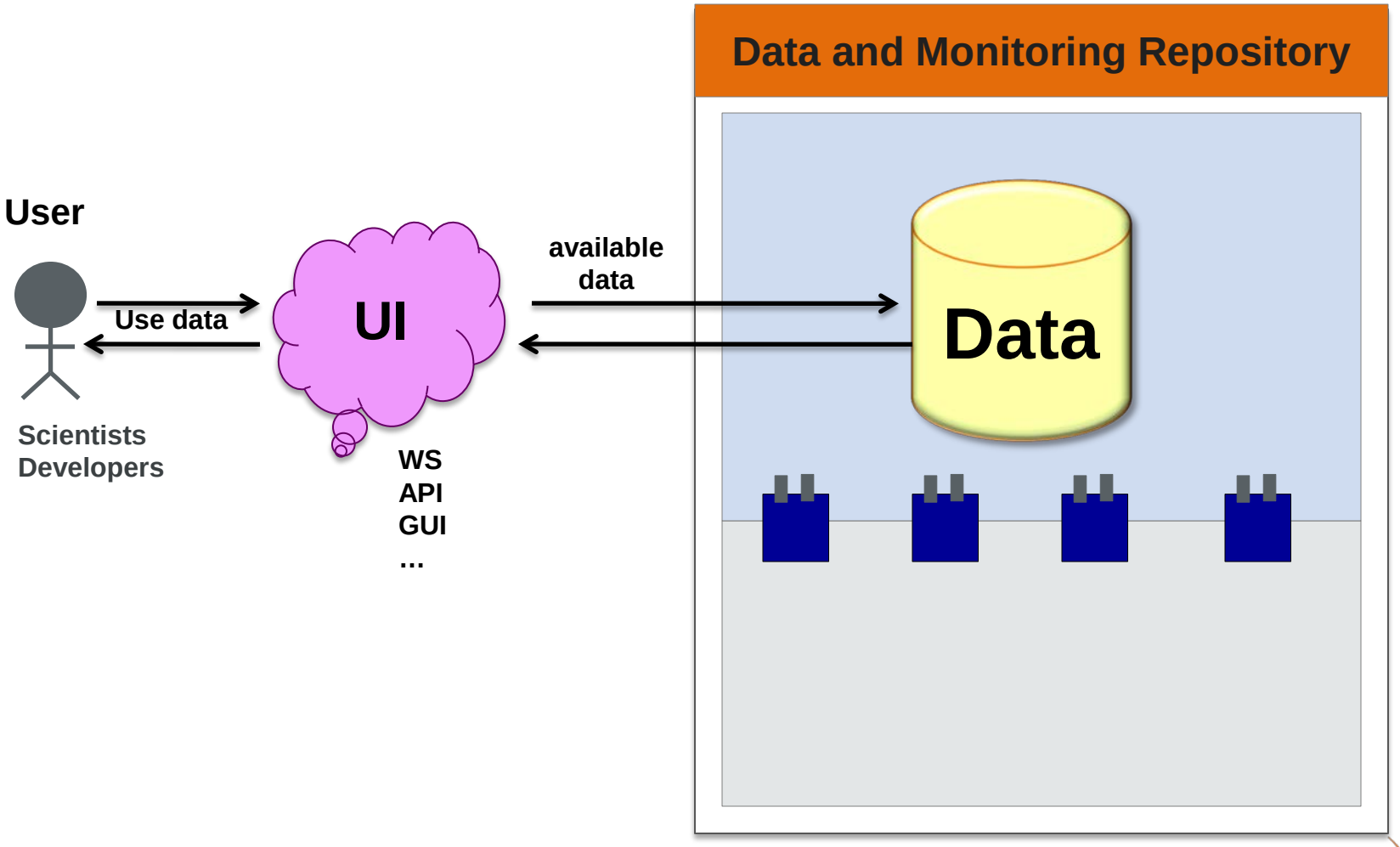
Monitoring execution: example of console log for the component monitoring. The parameters of execution are defined in XML schema:

- Component ID: metrics-module-citation-4ywN_j5H
- Frequency: minute
- Installed: true (upgrade in the installation execution).
- Parameter: swissprot,swiss prot,swiss-prot,UniProtKB,uniprot
- Executable: test/testcitations.js

```
[15:24:33][admin]~/DataMonitoringRepository$ node runcomponents.js
components/metrics-module-citation-4ywN_j5H/test/testcitations.js
Frequency: minute (cron format 0 * * * * *)
Resource: "swissprot,swiss prot,swiss-prot,UniProtKB,uniprot"
30411 citations for swissprot,swiss prot,swiss-prot,UniProtKB,uniprot
data save into > citation.txt

component metrics-module-citation-4ywN_j5H was executed
```

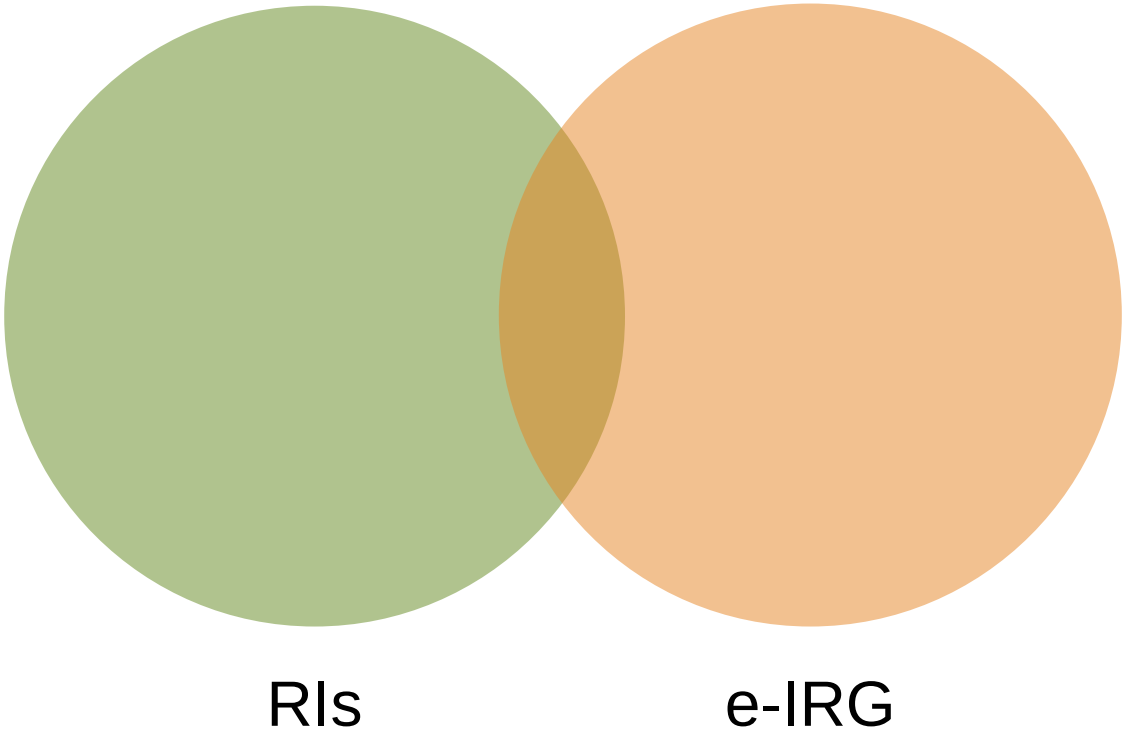
Access Data


Data and Monitoring Repository
User
Use data
UI
available
data
Data
Scientists
Developers
WS
API
GUI
...

Suggestions

Suggestions

- More involvement of RIs
- More reuse
- Agreement of core service metrics



Thanks for your attention!

Describe metrics with Schema.org markup

Data repositories

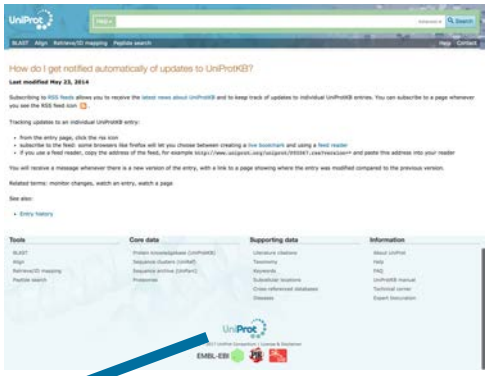
...

Description

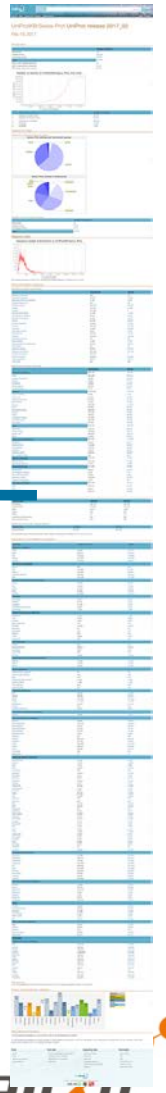
Download

APIs

Release



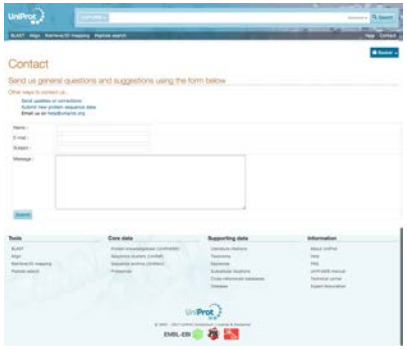
Metrics



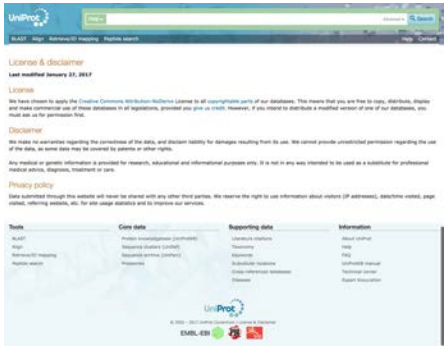
...

UniProt

Contact



License

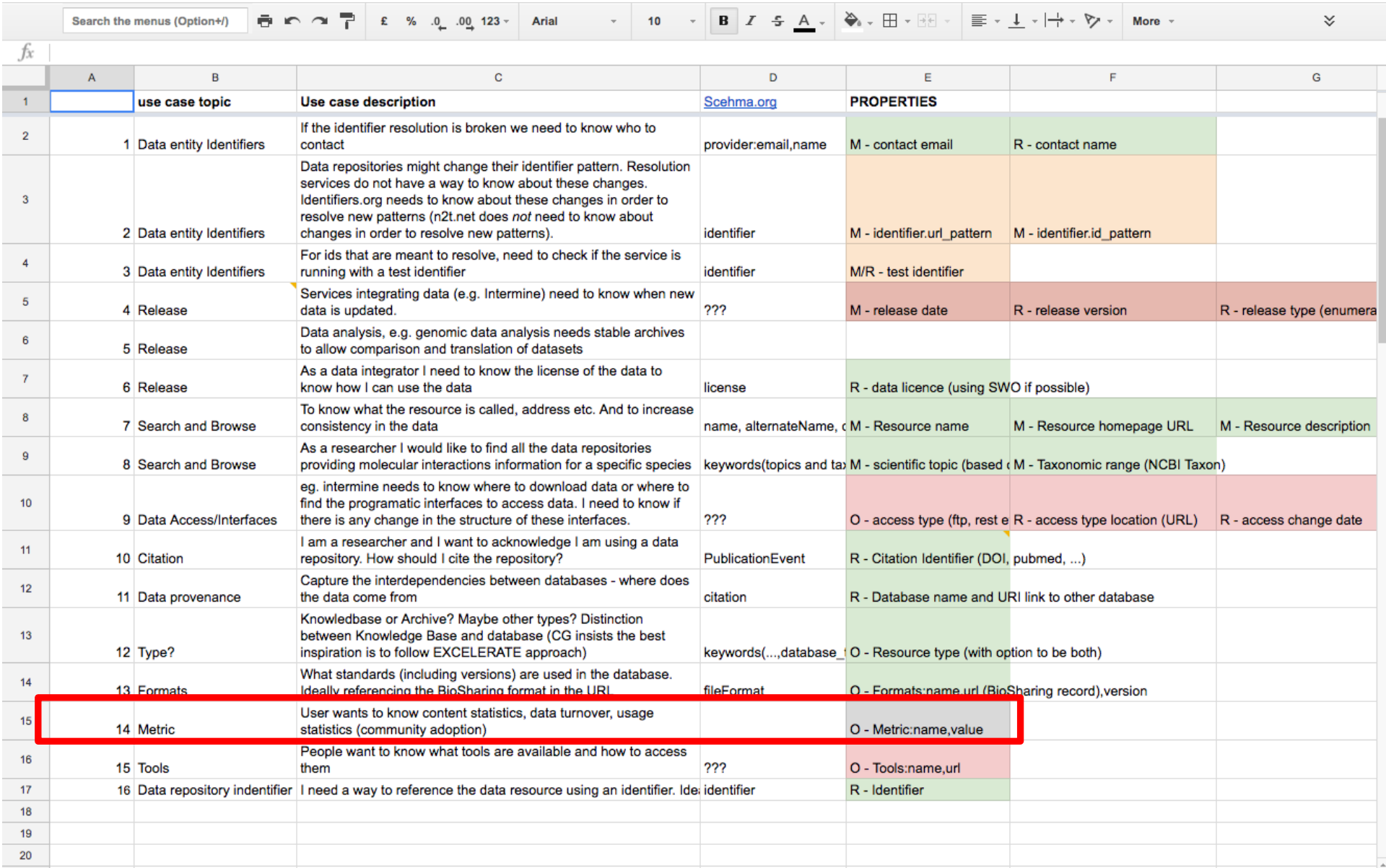


Funding

Citations

Authors

Finding use cases to describe data repositories



	A	B	C	D	E	F	G
1		use case topic	Use case description	Scehma.org	PROPERTIES		
2	1	Data entity Identifiers	If the identifier resolution is broken we need to know who to contact	provider:email,name	M - contact email	R - contact name	
3	2	Data entity Identifiers	Data repositories might change their identifier pattern. Resolution services do not have a way to know about these changes. Identifiers.org needs to know about these changes in order to resolve new patterns (n2t.net does *not* need to know about changes in order to resolve new patterns).	identifier	M - identifier.url_pattern	M - identifier.id_pattern	
4	3	Data entity Identifiers	For ids that are meant to resolve, need to check if the service is running with a test identifier	identifier	M/R - test identifier		
5	4	Release	Services integrating data (e.g. Intermine) need to know when new data is updated.	???	M - release date	R - release version	R - release type (enumera
6	5	Release	Data analysis, e.g. genomic data analysis needs stable archives to allow comparison and translation of datasets				
7	6	Release	As a data integrator I need to know the license of the data to know how I can use the data	license	R - data licence (using SWO if possible)		
8	7	Search and Browse	To know what the resource is called, address etc. And to increase consistency in the data	name, alternateName,	M - Resource name	M - Resource homepage URL	M - Resource description
9	8	Search and Browse	As a researcher I would like to find all the data repositories providing molecular interactions information for a specific species	keywords(topics and ta	M - scientific topic (based	M - Taxonomic range (NCBI Taxon)	
10	9	Data Access/Interfaces	eg. intermine needs to know where to download data or where to find the programatic interfaces to access data. I need to know if there is any change in the structure of these interfaces.	???	O - access type (ftp, rest e	R - access type location (URL)	R - access change date
11	10	Citation	I am a researcher and I want to acknowledge I am using a data repository. How should I cite the repository?	PublicationEvent	R - Citation Identifier (DOI, pubmed, ...)		
12	11	Data provenance	Capture the interdependencies between databases - where does the data come from	citation	R - Database name and URI link to other database		
13	12	Type?	Knowledbase or Archive? Maybe other types? Distinction between Knowledge Base and database (CG insists the best inspiration is to follow EXCELERATE approach)	keywords(...,database_	O - Resource type (with option to be both)		
14	13	Formats	What standards (including versions) are used in the database. Ideally referencing the BioSharing format in the URI	fileFormat	O - Formats:name,url (BioSharing record),version		
15	14	Metric	User wants to know content statistics, data turnover, usage statistics (community adoption)		O - Metric:name,value		
16	15	Tools	People want to know what tools are available and how to access them	???	O - Tools:name,url		
17	16	Data repository indentifier	I need a way to reference the data resource using an identifier. Ide	identifier	R - Identifier		
18							
19							
20							

Metrics type - proposal

https://docs.google.com/spreadsheets/d/19myommNwBTHVPmJ2R_0RM6UuK_1xWpiGiqgSYZjuQmU/edit?usp=sharing

schema.org			bioschemas		
Property	**Expected Type**	**Description**	**SubProperties**	**Minimum Fields**	**Cardinality**
description	Text	A description of the item.		Recommended	ONE
image	ImageObject or URL	An image of the item. This can be a URL or a fully described ImageObject.		Optional	MANY
name	Text	The name of the item.		Minimum	ONE
category	Text	Reflect the essence of the definition of this metric. E.g.: scientific, community, quality, legal, ...		Recommended	MANY
measurement	QuantitativeValue QualitativeValue variableMeasured	Value you are measuring. E.g.: number of visits, visitors, hits, page views, ...	QuantitativeValue: value, maxValue, minValue	Minimum	MANY
source	Organization	Who provided this metric. E.g.: Uniprot, Wikipedia, ELIXIR, etc.	Organization: name, email, address	Minimum	ONE
policy	CreativeWork	License for using the information provided for this metric (terms of use). E.g.: proprietary, free, open	CreativeWork: accessMode, author, license	Recommended	MANY
breakdown	URL	URL where we are describing more details about the content of this metric. E.g.: List of cited by, external link.		Optional	MANY

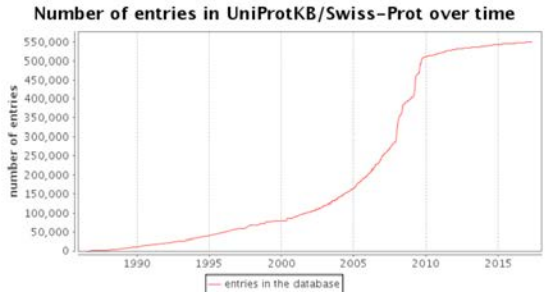
Examples

UniProtKB/Swiss-Prot UniProt release 2017_05

May-10, 2017

Introduction

	Number of entries
New entries	290
Updated entries	214,201
Unchanged entries	340,024
Total	554,515
Entries with updated sequences	136
With a fragmented AA sequence	9,133
With known alternative products	24,852



http://www.uniprot.org/statistics/Swiss-Prot

EMBL-EBI

Protein Data Bank in Europe

Bringing Structure to Biology

Latest archive statistics

As of 17 May 2017 the PDB contains 130102 entries (latest PDB entries, chemistry, biology) and EMDB contains 3633 entries (latest map releases, latest header releases, latest updates).

https://www.ebi.ac.uk/pdbe/

EMBL-EBI

Recent Pfam blog posts ☒Hide this

Pfam 31.0 is released (posted 8 March 2017)

Pfam 31.0 contains a total of 16712 families and 604 clans. Since the last release, we have built 415 new families killed 9 families and created 11 new clans. We have also been working on expanding our clan classification; in Pfam 31.0, over 36% of Pfam entries are placed within a clan. The new "stuff" [...]

http://pfam.xfam.org/

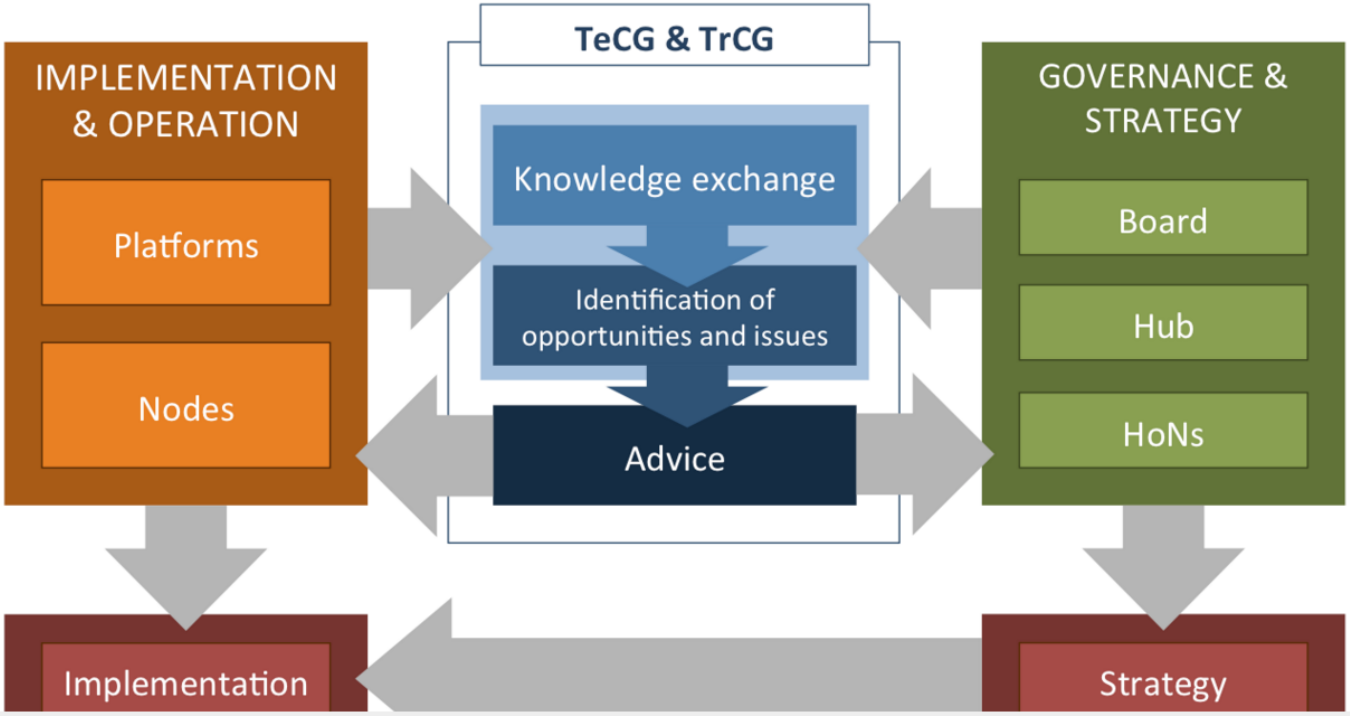
Milestone M12.9 - Define performance indicators of permanent working groups to assess effectiveness and impact

ELIXIR-EXCELERATE is funded by the European Commission within the Research Infrastructures programme of Horizon 2020, grant agreement number 676559.

www.elixir-europe.org/excelerate

Scope

Three sequential categories define the scope of the coordinators groups: "**knowledge exchange**", "**identification of opportunities and issues**", and "**advice**" (*Figure 1*). The coordinators groups will identify gaps and promote connections among Nodes and platforms. It will be the responsibility of the coordinators groups to provide advice and recommendations on topics prioritised by HoNs. Coordinators will be able to participate in the implementation of ELIXIR strategies as representatives of an operation group. HoNs will be able to commission work to coordinators groups to explore specific opportunities and issues aiming to get advice and recommendations.



Coordinators groups objectives

- Facilitate **knowledge exchange**
- Identify **opportunities and issues**
- Act as an **advisory body** to HoNs and ELIXIR Nodes
- **Lead and manage** assigned working groups

Coordinators groups metrics

- Facilitate inter-node and platform **knowledge exchange**
  - Involvement of the coordinator members in implementation groups
  - Knowledge exchange between nodes in coordinators meetings
- Identify **opportunities and issues** among Nodes and Platforms
  - Collection of issues, opportunities and actions
- Find the right domain experts to seek **advice** on identified opportunities and issues
  - Engagement of domain experts
- Act as an **advisory body** to HoNs and ELIXIR Nodes when broad representation across Nodes is needed.
  - Advice provided to HoN and the Hub
- Lead **working groups** assigned by HoNs
  - Working groups managed by coordinators permanent groups
  - How well working groups operate