EXCELERATE Deliverable D2.4

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:	**Final report on the impact of ELIXIR infrastructure for hosting scientific benchmarking and technical monitoring results**	
WP No.	2	
Lead Beneficiary:	12 - BSC	
WP Title	Benchmarking	
Contractual delivery date:	31 July 2019	
Actual delivery date:	7 August 2019	
WP leader:	Alfonso Valencia, Søren Brunak	12 - BSC, 38 - DTU
Partner(s) contributing to this deliverable:	12 - BSC; 26 - SIB.	

Authors and Contributors:

Salvador Capella-Gutierrez; Juergen Haas; Josep Ll. Gelpí; José Mª. Fernández.

Reviewers:

N/A

Table of contents

Table of contents 2

1. Executive Summary 2

2. Impact 3

3. Project objectives 4

4. Delivery and schedule 4

5. Adjustments made 5

6. Background information 5

7. Appendix 1: Final report on the impact of ELIXIR infrastructure for hosting scientific benchmarking and technical monitoring results. 8

7.1. Introduction. 8

7.2. OpenEBench. The ELIXIR benchmarking platform in the context of the tools platform ecosystem. 8

7.3. Standardization: Data Model (across communities). 11

7.4. CAMEO refactor to modern architecture. 13

7.5. OpenEBench General Architecture for integrating Scientific Communities. 14

7.6. OpenEBench Technical monitoring as a proxy for Life Sciences Research Software Quality observatory. 15

7.7. OpenEBench. Outlook. 18

1. Executive Summary

The objective of Deliverable 2.4 is to report on the current status of OpenEBench as the ELIXIR Benchmarking platform considering all work carried out in the H2020 ELIXIR-EXCELERATE project. This report will constitute a fundamental piece to establish the roadmap for OpenEBench further developments. During EXCELERATE, the platform has become an integral service for hosting both scientific and technical benchmarking data for bioinformatics tools, workflows and web-services. The scientific benchmarking component has focused on existing and newly created community-led efforts as community agreed datasets and metrics act as proxies of the current challenges in a given research area.

This final report builds on previously deliverables e.g. D2.1: Creation of a database warehouse infrastructure for storing and organizing data for online performance assessment experiments[1], D2.2: A report on the coordination with WP1 on the incorporation of monitoring statistics and benchmarking results in registry releases[2], and D2.3: A report on the features and nature of novel

[1] https://drive.google.com/file/d/0B9VTm8JIJMx4SG44RlFuNFB5Ukk/view?usp=sharing
[2] https://docs.google.com/document/d/1sMKdt2FokMuupcvxAKKdY8Mb77u_CHns7Y7cLkLuAtg/

data which are needed within online benchmarking experiments in different subareas[3]; and updates them in order to review the current infrastructure.

Two important aspects for the infrastructure are understanding the true nature of the data sets used by communities for their scientific benchmarking efforts as well as organizing the interactions with those communities. We have recently released the first stable version (1.0.1) of the OpenEBench scientific data model after extensive testing and adoption based on gathered feedback using real world data. Moreover, we have extensively documented and provide real examples to scientific communities for helping them to get organized as well as to facilitate their incorporation into OpenEBench. We have also consolidated our three-level models to facilitate the incorporation of scientific communities depending on their needs and maturity level. Level 1 and 2 are already implemented and are being used by different communities while level 3 has been a prototype but we depend on other efforts to fully implement it e.g. EOSC-Life.

In the technical monitoring side, we have consolidated the internal data management infrastructure, and extended the available metrics. Besides, we have been working as part of a community effort to translate and re-interpret, when needed, the FAIR data principles to FAIR principles applied to research software. This effort is important to us because it will offer a reference framework to position our set of technical and quality metrics for bioinformatics research software, which include tools, workflows and web-applications.

2. Impact

We have engaged to different degrees and kept interactions with the following scientific communities:

- CAMEO. Continuous Automated Model EvaluatiOn[4]
- QfO. Quest for Orthologs[5]
- GMI. Global Microbial Identification Initiative - Benchmarking Group
- CAID. Continuous Assessment of Intrinsic protein Disorder
- CAMI. Critical Assessment of Metagenome Interpretation
- CoCoBench. Community-based Continuous Benchmarking for Core Facilities
- TCGA. The working group for cancer driver genes and mutations from The Cancer Genome Atlas.

Along the lifetime of the project, we have interacted with community managers and members following different formats from periodic teleconferences (CAMEO, QfO, GMI, CAMI, TCGA), to participating in their regular meetings (QfO, CoCoBench, CAID, CASP), to organizing workshops to explain how to use OpenEBench to organize their scientific (GMI), as well as to organize

[3]https://docs.google.com/document/d/192JxUXBto1CZrfyXUd1eIGEudfDDa4K1aDDOchS202s/
[4]https://cameo3d.org
[5]https://questfororthologs.org/

Horizon 2020 grant n. 676559

workshops dedicated to exchange ideas on scientific benchmarking activities ([BC]2 workshop on Benchmarking).

OpenEBench contributes to the scientific communities' activities in different ways: standardizing the scientific benchmarking activities, technical and scientific monitoring based on community inputs. Recently, the OpenEBench Data Model v1.0.1 supporting scientific benchmarking activities has been released. In the midterm time range, this will enable the life sciences for the first time to exchange benchmarking results. This is crucial for benchmarking workflows compiled of individual tools that combine diverse scientific tasks e.g. metagenomics assembly and annotation. OpenEBench already exposes benchmarking information via available APIs to other platforms like the ELIXIR Tools Registry bio.tools and the Galaxy ToolShed aiming to inform the potential users of the tools. OpenEBench currently monitors technical aspects of 15,000+ bioinformatics tools, servers and workflows. Some of the metrics computed for each tool are: the availability of schema.org annotations in the associated web pages; the impact of the manuscripts associated to that tool, in terms of citations; identification of features, like availability of documentation, an issues system, the existence of a license at the source level, and similar, when the source code is in GitHub. Most of these computed metrics come from specialized tool data enrichers, which have been developed along the project lifetime. But, tool data enrichers are only useful on those tools whose entries contain the annotations used as seed by them.

3. Project objectives

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

No.	Objective	Yes	No
1	Establish an ELIXIR discovery portal that provides a transparent route to tools and services for data access and exploitation by users.	X	
2	Stimulate innovation by supporting industry uptake of ELIXIR resources, particularly in SMEs.	X	

4. Delivery and schedule

The delivery is delayed: Yes • No ☑

Horizon 2020 grant n. 676559

5. Adjustments made

The scope of this deliverable has been extended to include a review of previously reported deliverables, especially deliverables D2.1 and D2.3.

6. Background information

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

Work package number	2	Start date or starting event:	month 1
Work package title	**Benchmarking**		
Lead	Alfonso Valencia (ES) and Søren Brunak (DK)		
Participant number and person months per participant 7 - CNIO 2.00; 8 - CRG 20.60; 10 - IRB 12.00; 12 - BSC 28.00; 25 - SIB 24.50; 38 - DTU 6.00			
Objectives The concept of assessing bioinformatics methods in terms of quantitative performance and user friendliness is crucial to the development of the infrastructure in the general field of bioinformatics. Accordingly, WP2 will focus on the following objectives: • Systematically organize the relations to communities already running benchmarking exercises within biology and medicine. (Task 2.1) • Development and maintenance of a generic infrastructure to support benchmarking exercises in different subareas. (Task 2.2) • Develop the technology to perform online, uninterrupted methods assessment in key areas of bioinformatics. (Task 2.3) • Development and implementation of data warehouse infrastructures to store benchmarking results and to make them accessible to benchmark participants and method developers for subsequent transfer to the ELIXIR registry. (Task 2.4) • Development of the procedures to create standards in the different fields subject to benchmarking. (Task 2.5) • Establish workshops, hackathons and jamborees for different user communities. (Task 2.6) Work Package Leads: Alfonso Valencia (ES) and Søren Brunak (DK)			
Description of work and role of partners **WP2 - Benchmarking [Months: 1-48]** **BSC, CNIO, CRG, IRB, SIB, DTU**			

Horizon 2020 grant n. 676559

World-wide, bioinformaticians already engage significantly with evaluation exercises in the form of open challenges. The role model for this type of effort is the still on-going "Critical Assessment of protein Structure Prediction, or CASP, which is a community-wide, world-wide experiment for protein structure prediction taking place every two years since 1994. This effort, as well as others, provide research groups with an opportunity to objectively test their prediction methods and delivers an independent assessment of the state of the art to the research community and software users. CASP has inspired many other similar experiments, including analysis of text mining methods (BioCreative), docking (Capri): force-field evaluation for atomistic simulations and benchmarking of small molecule docking, evaluation of multiple alignments, NGS sequencing variation analysis, gene finding and others. All these community efforts have a similar organization and similar basic infrastructure needs. A further challenge is to make these challenges not only static annual or bi-annual competitions, but to evaluate the systems in an online fashion, which would make them more sustainable. A few experiments were organized in the past (e.g. the EVA effort organized by Burkhard Rost and co-workers), but abandoned for technical reasons. The WP will reintroduce these concepts such that methods can be benchmarked based on data, which are novel to all, including the methods developers in more sustainable frameworks. It is an essential part of the European infrastructure since:

• It provides a strong connection between the ELIXIR infrastructure and the communities carrying out benchmarking exercises within their expert knowledge domains.

• It is directly linked to the information to be disseminated in the ELIXIR tools and services registry.

• Provides direct access to information on methods and performance measures for end-users.

• Provides the benchmarking data needed for training of new methods making progress in the different subareas of field.

• Furthermore, the benchmarking activities will provide a great vehicle for developing novel standards for data and methods thus also providing useful input to other WPs.

Task 2.1: Organize the relations with communities already running benchmarking exercises (9.6PM)

Obtain agreement with existing communities on the conditions of challenges, organizes, formats, goals and other organizational issues that can lead to harmonization of efforts world-wide in addition to division of labour decisions.

Partners: ES, DK

Task 2.2: Development and maintenance of a generic infrastructure to support benchmarking in different areas(16.5PM)

The emerging ELIXIR registry will be a reference for the research community. The methods to be benchmarked will be described in the registry with the proper version control and automatic access procedures. At the same time a generic infrastructure is needed in order to organize data for new and existing benchmarking efforts. WP2 will be responsible for implementing the guidelines and standards for data organization and submission of the different methods subsequently to be incorporated in the registry. We will also collect qualitative and quantitative data about the usage of these services, and different indicators about the service itself (i.e. data grow rate, uptime, etc.). These data will be stored in the data warehouse infrastructure (Task 2.4). Opinion leaders in the field will be surveyed about how useful they consider the resources are and the results will be included in the registry.

Partners: ES, DK

Horizon 2020 grant n. 676559

Task 2.3: Develop the technology to perform online, uninterrupted methods assessment in key areas of bioinformatics (24.5PM)

In order to make online methods performance assessment several infrastructure elements need to be in place in order to support the various challenges. These include:

• Organization of a collection of training data (validated by experts),

• Identification, collection and organization of a collection of testing data which are kept secret,

• Community agreements on the data standards, submission formats and evaluation methods (quality assessment),

• Hosting or accessing methods (e.g. by programmatic access) to obtain results from them automatically without human intervention,

• Parsing, organization and display of the results with proper statistics and comparison facilities.

Partners: ES, DK, CH

Task 2.4: Development and implementation of data warehouse infrastructures to store benchmarking results and to make them accessible to users and method developers (24.5PM)

In this task we will develop with each one of the communities the necessary data framework and method standards, based on the community recommendations and the experience acquired in each challenge. The standards will be essential for the operation of the benchmarking infrastructure. The standards will also facilitate the end- users interpretation of the results, and we will develop tools for the conversion of the data from different formats into the most frequent standards in collaboration with WP1. We will also develop tools to diagnose and rate the ELIXIR resources according to the level of agreement with those standards.

Partners: ES, DK, CH

Task 2.5: Development of the procedures to create standards in the different fields subject to benchmarking (9PM)

Data warehouses are key to storing and analysing the very large collection of data that will be generated by the prediction methods.Part of the WP2's mission is to store these data in a way such that they can be used for the continuous evaluation of the methods and for training of new methods. With time the ambition is that this infrastructure will be the main infrastructure of the different communities in subareas from protein structure and feature prediction to genomics and chemoinformatics.

Partners: ES, DK

Task 2.6: Establish workshops and jamborees for the different user communities (9PM)

The final goal of the infrastructure is to provide users with a continuous evaluation of bioinformatics methods and to have a positive influence on tools development. The effort requires a robust system for the provision of testing data, running methods and evaluating results. The design of the most adequate representation system for each of the areas will require additional software development efforts. In the training workshops and jamborees representatives of the scientific communities involved in the project will participate alongside new communities interested in adapting their challenges to the use of the infrastructure. The training aspects will be coordinated with the other training efforts in the project.

Partners: ES, DK

Horizon 2020 grant n. 676559

7. Appendix 1: Final report on the impact of ELIXIR infrastructure for hosting scientific benchmarking and technical monitoring results.

7.1. Introduction.

The dependence of the scientific advance on research software is increasing in all science fields. Notably in biology, where the availability of growing amounts of data coming e.g. from large scale omics and non-omics projects, has put an extra concern in the possibility of properly analyzing such data, and hence assuring the outcomes of such projects, as well as in the possibility of reproducing performed analyses for further interpretation and/or integration with others. Bioinformatics as a science has become a need at all levels of biology. Indeed, it is no longer a private space where specialized researchers develop and test new methodologies for the sake of their own scientific objectives. Bioinformatics methods and tools have now to be consumed by the whole biological community. This puts an extra challenge in the development of research software[6]. Bioinformaticians should prepare software for the use of non experts, and have to compete in a continuously evolving market of alternative options, proving with objective metrics that the software is usable, efficient, and gives the adequate research answers. Benchmarking has been a traditional activity in bioinformatics, although it has been mostly conducted by scientific communities, for internal consumption and seldom considered by final users of the software[7].

With the advent of grand initiatives including different personalized medicine ones, there is an emerging need to guarantee, and to a certain extent certify, that analytical workflows used routinely are complaint with the highest standards, implement state-of-the-art technologies, and consistently process input data as expected. Thus, there is a clear need of establishing standards, relevant scientific challenges and meaningful metrics by knowledgeable scientific communities. However, those efforts should be complemented by a stable platform which can support these activities, provide a reference place for different stakeholders and give a general overview on how tools and workflows, scientific challenges, metrics and data sets evolve over time.

7.2. OpenEBench. The ELIXIR benchmarking platform in the context of the tools platform ecosystem.

In this context, the need for an open platform around benchmarking has become evident. **OpenEBench**[8], the main outcome of ELIXIR-EXCELERATE WP2, seeks to fill in this gap and three

[6]Silva, L. B., Jimenez, R. C., Blomberg, N., & Luis Oliveira, J. (2017). General guidelines for biomedical software development. F1000Research, 6, 273. https://doi.org/10.12688/f1000research.10750.2

[7]Capella-Gutierrez, S., de la Iglesia, D., Haas, J., Lourenco, A., Fernandez Gonzalez, J. M., Repchevsky, D., ... Valencia, A. (2017, August 31). Lessons Learned: Recommendations for Establishing Critical Periodic Scientific Benchmarking. Cold Spring Harbor Laboratory. https://doi.org/10.1101/181677

[8]https://openebench.bsc.es

different but yet complementary levels of benchmarking: i) scientific benchmarking related to the scientific quality of bioinformatics tools and workflows; ii) technical monitoring related to software quality; and iii) performance benchmarking regarding the usability and efficiency of the technical deployment of bioinformatics tools, servers and/or workflows. Indeed, benchmarking (WP2) is central to distinguish the effort of the ELIXIR Tools Platform from popular web search sites such as google, bing, ask, duckduckgo, or yahoo. Overall, OpenEBench provide information for i) end-users, deciding which resource is the most appropriate for their problem at hand, ii) software developers, seeking for accepted best practices in research software, testing their own tools against accepted and/or possibly competing alternatives using relevant datasets and metrics established by scientific communities, iii) infrastructure providers, seeking to design an adequate provision of tools, servers and/or workflows, iv) funders, requiring an overview of a given field, and checking the outcome of funded activities, v) policy-makers, requiring to understand the current practices in a given field for establishing minimum complaint mechanisms, and vi) research journals, willing to understand the performance of methods in the context of a neutral evaluation platform. A number of other initiatives do exist within and outside ELIXIR that clearly intersect OpenEBench aims. In particular, tools registries, mainly bio.tools registry[9] (from ELIXIR-EXCELERATE WP1), aggregated tools platforms like BioConda[10] or Galaxy tool-shed[11], or software deployment platforms like BioContainers[12]. Interactions with other systems and platforms from the ELIXIR platforms, particularly from the tools one, are highly relevant for OpenEBench. For instance, all individual tools and web servers being technically evaluated and/or taking part of community-led scientific benchmarking efforts should be registered at least in bio.tools. In those cases, where a software container is available for those tools in biocontainers and/or BioConda, an appropriate link is also included in OpenEBench. Moreover, all OpenEBench components have been designed and implemented following the recommendations made by the ELIXIR tools platform e.g. making code available in public repositories from day 1; are available as software containers, and use workflow managers promoted by ELIXIR. Figure 1 illustrates the interconnection of OpenEBench to other ELIXIR tools platforms systems and platforms and beyond.

[9] https://bio.tools
[10] https://bioconda.github.io
[11] https://toolshed.g2.bx.psu.edu
[12] http://biocontainers.pro

Horizon 2020 grant n. 676559

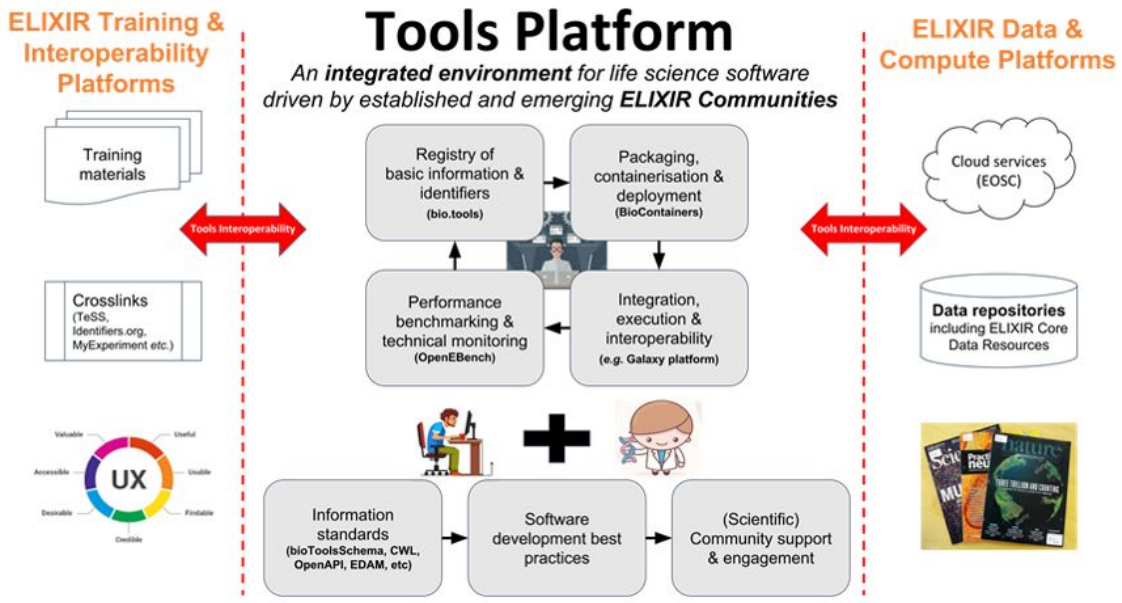

Figure 1. OpenEBench as part of the ELIXIR tools platform.

OpenEBench has been designed as an information Hub (Figure 2), where data is being collected from different sources, and others, processed, and redistributed back for the use of those platforms and also to the already mentioned group of users via a web-interface and/or APIs.

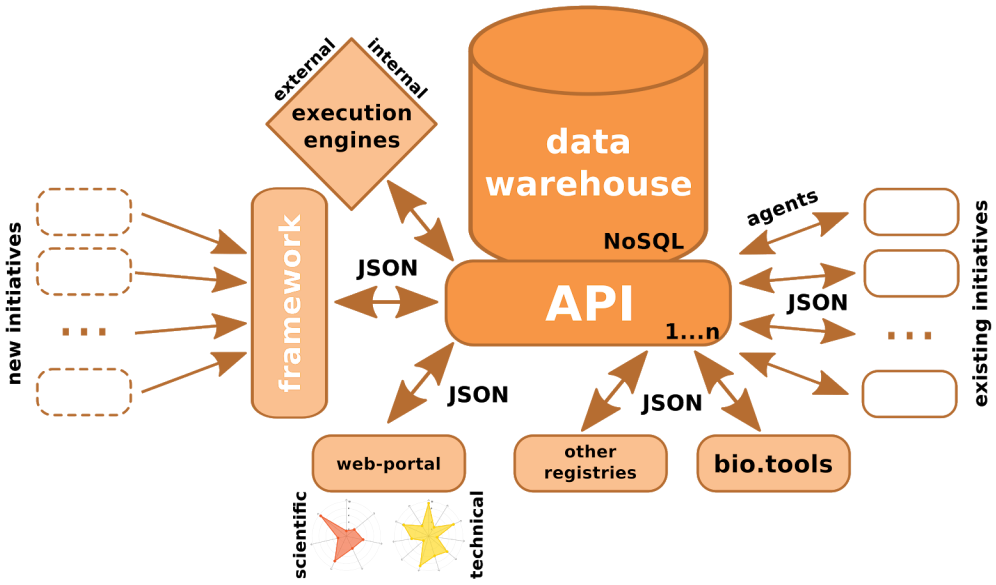

Figure 2. OpenEBench philosophy as information hub.

In the context of ELIXIR-EXCELERATE WP2, we have established collaborations within the project with WP1 for increasing the available information for tools, web-services and workflows; WP3 for analysing the gathered data for the ELIXIR Core Data Resources and Deposition Databases; WP4 for developing initial prototypes for software containers-based workflows, WP5

Horizon 2020 grant n. 676559

for implementing recommendations on how to FAIRify data, WP6 for understanding key issues with meta-genomics pipelines, which lead to an exchange with the CAMI (Critical Assessment of Metagenome Interpretation) effort; and WP8 for identifying an initial set of data for the further benchmarking of variant calling workflows. We have also established collaborations with a number of communities e.g. the cancer driver genes and mutations benchmarking group from the TCGA. The insights gathered from these interactions is now culminating in this report. One key aspect is a general data-model that allows efficient and transparent exchange of benchmarking data across communities, e.g. to gather data in OpenEBench. In the following we will elaborate on the different aspects and data types involved in benchmarking and how we manage to unite these in the OpenEBench framework.

7.3. Standardization: Data Model (across communities).

In an effort to standardize the benchmarking process *per se* across scientific communities, we have developed a refined data-model to reflect the process itself and allow scientists to refer to a particular step and/or data set in a defined way. Participants represent systems e.g. individual tools, analytical workflows, web-servers, taking part of a specific benchmark event. The detail of OpenEBench data model is available on GitHub[13], recently a stable version (1.0.1) has been released as part of the WP2 activities. All activities in the context of OpenEBench, especially the interactions with the communities, are periodically updated and revisited to ensure the full alignment with different stakeholders. Below there is a summary of the different OpenEBench working model data sets generated after a number of iterations with representatives of different communities e.g. QfO, CAMEO, GMI, TCGA, among others.

- ***Public Reference data sets***. These are widespread, publicly available and well characterized data sets, which can be used by developers and/or interested users to gather performance data of their systems in a controlled set-up. When public reference data sets are not publicly available through recognized archives and databases e.g. EGA, ENA, EVA, UniProt; OpenEBench could contribute to their dissemination by depositing them in archives like Zenodo and then sharing the pointers to these data sets.

- ***Input data sets***. Represent the data sets to be processed as input by participants in the benchmarking activities. Those data sets can be publicly available for download at specific repositories e.g. UniProtKB; and/or can be submitted automatically by benchmarking platforms e.g. CAMEO, to participants web-servers. *Input data sets* should provide enough metadata describing the data sets to facilitate reproducibility, data provenance and, potentially, the evolution of participants across different benchmarking challenges editions with different input data sets of varying degrees of complexity.

- ***Participant data sets***. These data sets represent the data e.g. predictions, produced by participants given a specific *Input data set* associated to specific benchmarking activities.

[13] https://github.com/inab/benchmarking-data-model

Horizon 2020 grant n. 676559

Unless previously agreed, participant data sets are often kept private to participants and/or communities.

- ***Metrics Reference data sets***. These data sets contain data used to evaluate the benchmarking process, i.e. the "true" responses to the challenges. These data sets are often kept private by benchmarking events organizers while a challenge is active. This standard practice prevents participants from adjusting their systems to have the best performance for very specific data sets, which is often referred to as overfitting. In continuous efforts there is the additional benefit of a guaranteed fully blind assessment, where both, the automated platform and the participants do not know the reference structure reducing *Metrics Reference data sets* bias even further. Often *Metrics Reference data sets* become public e.g. *Public Reference data sets*, once a given challenge has concluded because of its intrinsic value to address valuable scientific challenges.

- ***Assessment data sets***. These data sets are produced after applying specific metrics to *participants data sets* while considering metrics *reference data sets*. *Assessment data sets* establishes how close or far are participants from the expected results. Often preliminary assessment data sets tend to be private to each participant e.g. understanding the initial characteristics of the platforms and/or metrics reference data sets nature; while final assessment data sets tend to be shared among benchmarking participants before the challenge ends, and made public once the events end.

- ***Challenge data sets***. These data sets are considered metadata sets grouping either i) assessment data sets from different participants for the same reference metrics data set and applied metrics, ii) assessment data sets from the same participant but for different reference metrics data sets and/or applied metrics in the same benchmarking event, or iii) the grouping of the assessment data sets from the same participant and the same applied metrics across different benchmarking events. *Challenge data sets* are the foundations of the community-led scientific benchmarking activities as they offer a unified framework to compare participants performance among themselves for a specific scientific challenge and/or the evolution of individual participants along time.

Importantly, it is recommended that participant data sets, which are part of scientific benchmarking publications should be made available for reproducibility purposes, data reuse in downstream analysis and/or further meta-analysis.

Among the different data sets proposed, the *challenge data sets* are highly relevant for many end-users of the platform as they are the ones consumed by experts and non-experts for taking decisions on what systems to use for their own scientific problems. *Challenge data sets* can be directly offered at OpenEBench using available views e.g. experts and non-experts data views; and/or using available APIs. Those data sets, due to their own nature, would be mostly public although they might remain private to scientific communities and/or benchmarking participants while challenges remain open.

Each *Benchmarking event* can be represented by a data flow composed by these six different data types, as illustrated in figure 3. In the case of continuous benchmarking systems, the red arrow at figure 2 indicates the start of the subsequent cycles which often tend to adhere to community established metrics and change the *Reference Metrics data sets* e.g. CAMEO.

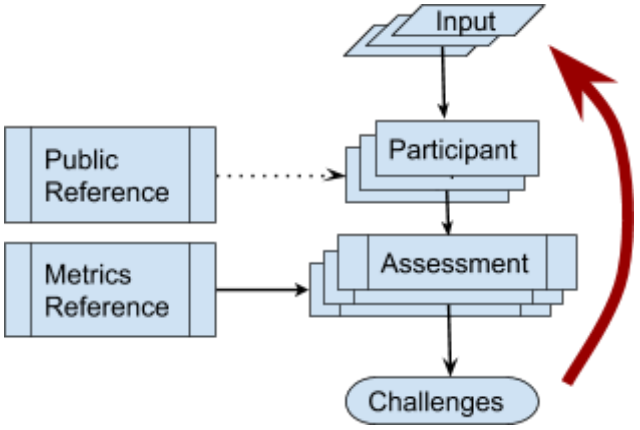

Figure 3. OpenEBench definition of datasets and how they relate to each other.

Despite the nature of each data set, it is crucial that all data sets which are part of community-led scientific benchmarking efforts become public during their data life cycle. This effort will incite open discussions and decisions within the community around which scientific challenges are relevant. Moreover, those efforts can be re-used by other communities maximizing the added value of data and/or metadata. For some communities e.g. health, biotechnology, it is accepted that (some) reference data sets are private, and therefore they cannot be made publicly available for ethical and/or intelligent competitive reasons. Here, only assessment data sets can be published along with the assessment workflow, making sure that the original data cannot be reconstructed, e.g. for very small datasets. As a general rule, data should follow the FAIR data principles[14] [Wilkinson et al. 2016], which states how to make data Findable, Accessible, Interoperable and Re-usable.

7.4. CAMEO refactor to modern architecture.

CAMEO[15] was launched in early 2012 based on a community decision at the CASP9 conference in Asilomar. It first featured the "protein structure prediction - 3D" category and was fully automated through a selection of bash scripts handling the overall workflow and its own basic workflow manager written in Python based on OpenStructure[16]. Adding further categories such as "ligand-binding residue prediction" in 2012 (deceased in 2016), "quality estimation prediction" in 2013 and "residue-residue contact prediction" in 2016 increased complexity drastically. The

[14]Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[15]Haas, J., Barbato, A., Behringer, D., *et al.* (2018). Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins 86:387-398. https://doi.org/10.1002/prot.25431

[16]Biasini, M., Schmidt, T., Bienert, S., *et al.* (2013). OpenStructure: an integrated software framework for computational structural biology. Acta Crystallographica Section D. 69, 5, 701-709.(doi: 10.1107/S0907444913007051)

original in-house workflow manager was to be refactored to allow parallel processing of the categories, because the CAMEO team had to invest considerable time in ensuring smooth operations. This altogether led to major rewrite of the code base, with the first stage of the rewrite having been concluded just recently. It features a Jinja2[17] based automated configuration for installation with all configuration centralized accessible to all CAMEO components. It allows parallel processing of the three categories and is based on Nextflow[18]. Nextflow immediately gave access to several major scheduling systems and resembles a big step towards cloud readiness. We are currently preparing to employ CAMEO within OpenEBench, which would at minimum allow an important fallback ensuring even more robust operations. This includes the adoption of the OpenEBench data model in order to unify the various categories at the technical level and enable CAMEO to seamlessly share data with OpenEBench.

7.5. OpenEBench General Architecture for integrating Scientific Communities.

In the context of WP2, we adopted a development strategy based on the regular release of Minimum Viable Products (MVPs) to capture early in the process the feedback provided by different end-users. Despite the potential delay introduced by this strategy, we make sure that OpenEBench covers a broad audience. One of the main aims for adopting such strategy is the early adoption of the platform by different type of end-users and the possibility for those users and communities to actively contribute in the development of the platform either by using the offered functionality and/or participating in the development process itself.

After a number of iterations, we proposed a three level architecture to support benchmarking activities across Life Sciences by scientific communities at different maturity stages (figure 4). Level 1 is used for the long-term storage of benchmarking events and challenges aiming at reproducibility and provenance; Level 2 allows community to assess participants' performance given one or more reference datasets, and one or more metrics; Level 3 goes further by getting workflows specifications from participants, and then evaluating them in terms of technical and scientific performance. Importantly, each level makes use of the architecture defined in the previous level e.g. participants' data generated by workflows at Level 3 are evaluated using the metrics and reference datasets in Level 2, and the resulting data is stored following the data model in Level 1 for private and/or public consumption.

Levels 1 and 2 are already in production while level 3 has been prototyped to demonstrate its feasibility. EOSC-Life will provide the technology needed e.g. workflows repository, to implement level 3. Those workflows will be annotated with all necessary metadata to facilitate their use and re-use including references to software containers. In this way, users will be able to start using the workflow of their choice for solving their research questions. By leveraging developments at EOSC-Life, OpenEBench will make sure to use community agreed standards in the Life Sciences and beyond.

[17] http://jinja.pocoo.org

[18] Di Tommaso, P., Chatzou, M., Floden, E.W., *et al* (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316-319. https://doi.org/10.1038/nbt.3820

Horizon 2020 grant n. 676559

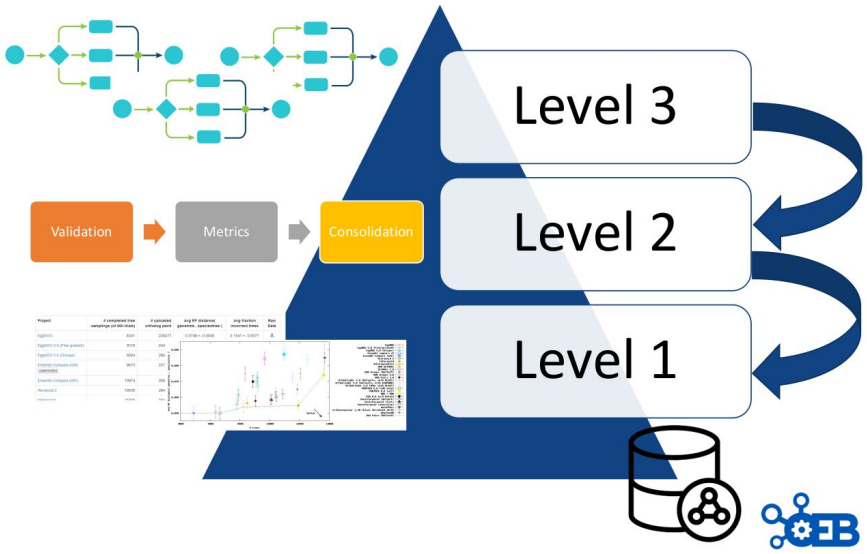

Figure 4. OpenEBench general architecture for facilitating the engagement with communities performing scientific benchmarking at different maturity stages.

7.6. OpenEBench Technical monitoring as a proxy for Life Sciences Research Software Quality observatory.

Software quality is a key issue in research[19], as the quality of scientific outcomes is clearly ligated to the quality of the tools used to deliver them. Bioinformatics as a whole has been largely accused of generating poor research software due to the prioritization of the scientific results over the optimization and standardization of the tools used. Due to the fast evolution of Bioinformatics itself, accepted algorithms become obsolete far before the software made out of them can reach the usual quality standards normal in other disciplines. While this is traditionally accepted as normal use by researchers, it puts strong questions in the reproducibility of research results and on the validity of processed data deposited in permanent archives.

OpenEBench, as indicated above, holds a specific infrastructure to monitor software quality. A series of quality metrics taken from a number of sources have been selected and implemented (Table 1 of D2.2). The source of such metrics includes documents by the Software Sustainability Institute[20], recommendations for open source software development[21], or for software quality metrics[9]. The main source of information corresponds to ELIXIR-EXCELERATE WP1's bio.tools registry[5], but primary data is also collected from BioConda repository[6] and Galaxy Tool-Shed[7].

The present contents of OpenEBench Tools monitoring repository contains 15,002 tools corresponding to over 22,000 deployments, all of them actively checked. Other metrics are less encouraging, although almost all show a clear description of the tools: only 2,800 (19%) have an

[19]Artaza H, Chue Hong N, Corpas M *et al.* Top 10 metrics for life science software good practices. *F1000Research* 2016, **5**(ELIXIR):2000. (doi: 10.12688/f1000research.9206.1)
[20]https://www.software.ac.uk/software-evaluation-guide
[21]Jiménez RC, Kuzak M, Alhamdoosh M et al. Four simple recommendations to encourage best practices in research software, F1000Research 2017, 6:876 (doi: 10.12688/f1000research.11407.1).

Horizon 2020 grant n. 676559

easily accessible documentation other than a simple description; 3,300 (23%) an open source licence or a terms of use document. Table 1 shows a summary of the analytic metrics at the time of writing.

Table 1. Summary of quality metrics at OpenEBench repository.

Metrics Name	Total	Percentage %
Tools Id/s		
Resource ID	15,002	100%
Documentation		
Description	14,850	99%
Help	670	5%
Manual	2,322	15%
Tutorial	222	2%
Publications	11,107	74%
Identity & Findability		
Website	15,002	100%
bioschemas	455	3%
Buildability & Installability		
Language	5,232	35%
Operating system	5,758	38%
Copyright		
Copyright statement	479	3%
Credits	2,089	14%
Licensing		
Project license	5,411	36%
Open source	2,801	19%
OSI	2,724	18%
Accessibility		
Binary distribution	4,346	29%
Source code	4,639	30%
Source code repositories	586	4%
Supportability & User Support		
e-mail	8,138	54%
Changeability		
Issues tracker	17	<1%

Horizon 2020 grant n. 676559

7.6.1. FAIR principles for software

As indicated, software plays a crucial role in contemporary scientific research. Computational tools are increasingly becoming constitutive parts of scientific research, from experimentation and data collection, to the dissemination and storage of results. This new paradigm regarding Bioinformatics is characterized by the use and reuse of massive amounts of data, usually unifying theory, experiments and simulation. The quality of software tools, their availability, and the reproducibility of their results are key to the required level of trust required to avoid the unnecessary repetition of large scale analysis. Unfortunately, software is not required to meet the requirements that are normally a must for other scientific methods: being peer-reviewed, being reproducible and allowing to build upon another's work. In the last years, the urgent need to trust and re-use data produced by large-scale projects, especially in the genomics field, has led to define a series of principles (the FAIR principles) that will allow, eventually to assess the quality and conditions of re-use of data. FAIR, standing for (F)indability, (A)ccessibility, (I)nteroperabilty, and (R)e-usability, principles are now understood and adopted by many research projects and have become a mandatory component of data management, no equivalent exists for the software used for such management.

Within OpenEBench we have started, FAIRsoft, an initial effort to assess the quality of research software using a FAIR-like framework. To be able to do that, we propose here a preliminary set of FAIR principles and metrics for research software. This effort is based on the adaptation of the FAIR Data Principles to evaluate research software. We have used a set of 2,000 tools (the test set selected for the Tools Platform Ecosystem) to automatically evaluate the initial proposed FAIRsoft principles. This effort serves a double purpose. First, it allows to draw a landscape of software quality-related features we can use to assess the applicability of our proposed metrics, making our metrics refinement loop partly evidence-based. Second, it allows us to evaluate the feasibility of the FAIRness automated monitoring we aim to implement in OpenEBench. Table 2, summarizes the initial proposal for the FAIRsoft principles.

Table 2. The FAIRsoft principles and high-level metrics. Each high-level metric has several associated low-level metrics. Preliminar aggregated scores for high-level metrics are shown in parentheses.

F	**To be Findable: a software can be found and unequivocally identified.**
F1	Software has a proper, unique and persistent identifier (i.e. unique program name) (0.4
F2	Software is described with rich metadata including scientific applicability (0.2)
F3	Software is included in public software registries (0.4)
A	**To be Accessible: it is possible to access a usable form of a software.**
A1	A working version of the software can be accessed/downloaded/built (0.7)
A2	Software history is traceable (Not measured)
A3	Software can be used without restriction (0.3)
I	**To be Interoperable: a software can be integrated with other tools in the users computational workflow.**

Horizon 2020 grant n. 676559

I1	Software Input and output data types and formats are documented (0.6)
I2	Software can be deployed in a format to be included in workflows (0.1)
I3	Software dependencies are documented and mechanisms to access them exist. (0.3)
R	**To be (Re)Usable: the software can be properly used and/or contributed to.**
R1	Software provides adequate usage documentation (0.3)
R2	Software provides a clear and accessible usage license (0.3)
R3	Software provides a contribution policy (0.2)
R4	Software provenance is available (0.2)

An initial set of scores have been defined to obtain a quantitative assessment of FAIRsoft metrics. Figure 5 summarizes the distribution on the 2,000 tools assayed. As expected Findability of such tools is high (they come from the bio.tools registry), while the other principles have a much lower coverage. The existence of "good behaving" tools in all categories reinforce the validity of the metrics despite the general low scoring.

FAIRsoft principles are being integrated with more general community efforts, including ELIXIR's software best practices group.

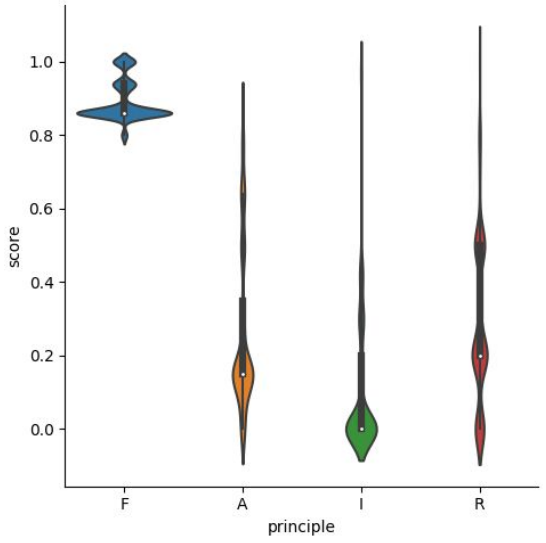

Figure 5. FAIRsoft scores for the initial testset

7.7. OpenEBench. Outlook.

OpenEBench has been designed and implemented during ELIXIR-EXCELERATE as a platform to support the community-led scientific benchmarking activities and the technical monitoring of bioinformatics tools, web servers and workflows. There are still a number of features to be incorporated into OpenEBench that will facilitate their use and adoption by different types of end-users.

Horizon 2020 grant n. 676559

One relevant aspect that will be implemented in the near future is the extended data accessibility model for scientific benchmarking. At the moment, benchmarking results can be either private e.g. only participants can have access to their results, and public e.g. once a given benchmarking event closes, the final submission by participants are made publicly available. The extended data accessibility model will include two additional access modes to facilitate sharing the results across communities prior to the publication of the final results as well as via specific links e.g. when submitting a manuscript to a peer-reviewed journal for its publication.

Another relevant aspect to include in future iterations of OpenEBench is the use of access controlled data sets. Depending on the scientific community focus, input and/or reference data sets may have some access limitations e.g. human omics data that can be used to identify individuals should be protected. Thus, OpenEBench will take into consideration those limitations by preventing direct access to those data sets. This will be possible using architecture level 3 where workflows will be run into a secure environment where participants will not have access to the data used as input[22].

Finally, the OpenEBench leads foresee a co-production model together with scientific communities and other stakeholders. The co-production model will contribute towards the sustainability of the platform by making other stakeholders part of it. Moreover, the co-production model will contribute to implement relevant functionality and gather feedback about it in order to refine it and maintain it. Indeed, the co-production model is already being used in different European projects where scientific benchmarking activities are foreseen as part of expected work to be done e.g. hosting DREAM challenges-like events in OpenEBench for communities working with single-cell transcriptomics.

[22]Hie, B., Cho, H., Berger, B. (2018). Realizing private and practical pharmacological collaboration. Science 362, 6412, 347-350. https://dx.doi.org/10.1126/science.aat4807

Horizon 2020 grant n. 676559