EXCELERATE Deliverable 2.3

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	A report on the features and nature of novel data which are needed within online benchmarking experiments in different subareas	
WP No.	2	
Lead Beneficiary:	12 - BSC	
WP Title	Benchmarking	
Contractual delivery date:	31st August 2018	
Actual delivery date:	31st August 2018	
WP leader:	Alfonso Valencia, Søren Brunak	12 - BSC, 38 - DTU
Partner(s) contributing to this deliverable:	12 - BSC; 26 - SIB.	

Authors and Contributors:

Authors: Salvador Capella-Gutierrez; Juergen Haas; Josep Ll. Gelpí; José Mª. Fernández; Javier Garrayo;

Table of contents

1. Executive Summary

The objective of Deliverable 2.3 is to report on the different types of data sets involved in the management of scientific benchmarking and their main features. Understanding the nature of these data sets is crucial to properly include them into the OpenEBench's data model. Importantly, understanding those data sets will contribute to interact with the different scientific communities facilitating their migration into OpenEBench. Deliverable 2.3 updates the initial report presented at (D2.1: Creation of a database warehouse infrastructure for storing and organizing data for online performance assessment experiments) and offers detailed real world examples on how the different proposed data sets types map to a broad set of benchmarking initiatives.

As OpenEBench becomes an ELIXIR integrated platform for both scientific and technical benchmarking, it is important to identify commonalities across scientific communities which can be used at the platform level, but also specific divergences that should be considered to give a better support to the communities. The effort made to identify novel data types and map existing ones into OpenEBench is crucial to offer specific actions for each data type e.g. accessibility modes, data ownership, etc. Moreover, the nature of data sets dictate how workflows can be articulated at the platform and beyond; e.g. communities can choose to run

their evaluation workflows at their own platforms, for privacy reasons, and deposit only the results into OpenEBench using the APIs developed for that end.

As data is the key component of the infrastructure, this report directly touches on most of the objectives associated to the ELIXIR-EXCELERATE WP2 (1 to 5). The remaining one, will be addressed indirectly through specific hackathons that will be organized to identify, link and/or deposit different data types for the benchmarking communities in OpenEBench.

2. Impact

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

- CAMEO. Continuous Automated Model EvaluatiOn.
- QfO. Quest for Orthologs.
- 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

OpenEBench currently monitors technical aspects of 15,002 bioinformatics tools, servers and workflows. Moreover, it exposes that information via available APIs to other platforms like the ELIXIR Tools Registry bio.tools and the Galaxy Shed.

Data Model v1.0 supporting scientific benchmarking activities has been released.

3. Project objectives

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

No.	Objective	Yes	No
1	Systematically organize the relations to communities already running benchmarking exercises within biology and medicine. (Task 2.1)	X	
2	Development and maintenance of a generic infrastructure to support benchmarking exercises in different sub-areas. (Task 2.2)	X	
3	Develop the technology to perform online, uninterrupted methods assessment in key areas of bioinformatics. (Task 2.3)	X	

4	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)	X	
5	Development of the procedures to create standards in the different fields subject to benchmarking. (Task 2.5)	X	
6	Establish workshops, hackathons and jamborees for different user communities. (Task 2.6)		X

4. Delivery and schedule

The delivery is delayed: Yes ☑ No

5. Adjustments made

The scope of this deliverable has been extended to include an updated version of deliverable D2.1, which presented the data model used to capture, store and expose benchmarking data provided and/or consumed by scientific communities.

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 (BSC) and Søren Brunak (DTU)		
Participant number and person months per participant 7 – CNIO 2.00; 8 - CRG 20.6; 10 - IRB 12; 12 - BSC 28; 25 - SIB 23; 38 - DTU 6. *CNIO and BSC activities are reported together since CNIO partners moved to BSC within the Grant Agreement 2nd Amendment.*			

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:

1. Systematically organize the relations to communities already running benchmarking exercises within biology and medicine. (Task 2.1)
2. Development and maintenance of a generic infrastructure to support benchmarking exercises in different subareas. (Task 2.2)
3. Develop the technology to perform online, uninterrupted methods assessment in key areas of bioinformatics. (Task 2.3)
4. 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)
5. Development of the procedures to create standards in the different fields subject to benchmarking. (Task 2.5)
6. 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

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. In a close collaboration with the Continuous Automated Model Evaluation (CAMEO) platform, which is running continuously since 2012, the WP will build on 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 sub-areas 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 (7.6PM).

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 (12PM).

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 of 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

Task 2.3: Develop the technology to perform online, uninterrupted methods assessment in key areas of bioinformatics (18PM).

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 (18PM).

In this task we will develop with the 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 (7PM).

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 (7PM).

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

Appendix 1: A report on the features and nature of novel data which are needed within online benchmarking experiments in different subareas

A1.1. 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 from large scale genomics projects has put an extra concern in the possibility of properly analyzing such data, and hence assuring the outcomes of such projects. Bioinformatics as a science has become a need at all levels of biology. It is no longer a private space where some 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 of the biological community. This puts an extra challenge in the development of research software[1]. 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 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[2].

With the advent of different personalized medicine initiatives, there is an emerging need to guarantee, and to certain extent to certify, that analytical workflows used routinely in the clinical practice 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.

A1.2. OpenEBench. The ELIXIR benchmarking platform

In this context, the need for an open platform around benchmarking has become evident. **OpenEBench**[3], the main outcome of ELIXIR-EXCELERATE WP2 seeks to fill in this gap and three 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

[1] 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

[2] 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

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

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 gen. Overall, OpenEBench should 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, and testing their own tools against the accepted and/or possibly competing alternatives, iii) infrastructure providers, seeking to design an adequate provision of tools, servers and/or workflows, and iv) funders, requiring an overview of a given field, and checking the outcome of funded activities. A number of other initiatives do exist within and outside ELIXIR that clearly intersects of OpenEBench aims. In particular, tool's registries, mainly bio.tools registry[4] (from ELIXIR-EXCELERATE WP1), aggregated tools platforms like BioConda[5] or Galaxy tool-shed[6], or software deployment platforms like BioContainers[7]. OpenEBench is designed as an information Hub (Figure 1), where data is being collected from those sources, and others, processed, and redistributed back for the use of those platforms and also to the already mentioned group of users.

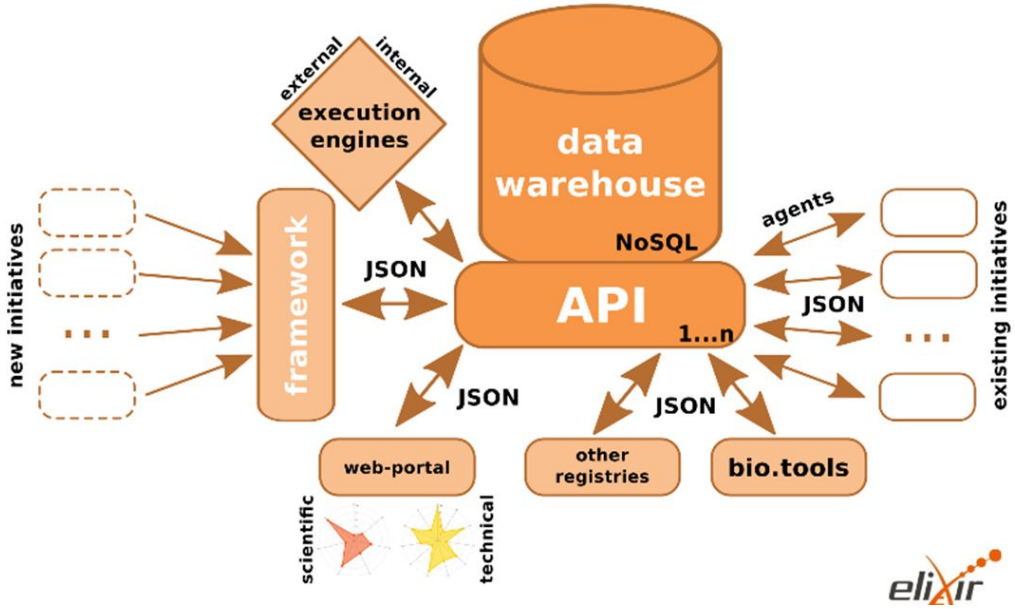

Figure 1. OpenEBench philosophy as information hub.

In the context of efforts at ELIXIR-EXCELERATE WP2, we have established collaborations within the project with WP6 for understanding key issues with meta-genomics pipelines, which lead to an exchange with the CAMI (Critical Assessment of Metagenome Interpretation) effort. 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 as the various communities have very different needs. Underlying is a general data-model that allows efficient and transparent exchange of benchmarking data

[4] https://bio.tools
[5] https://bioconda.github.io
[6] https://toolshed.g2.bx.psu.edu
[7] http://biocontainers.pro

Horizon 2020 grant n. 676559

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. OpenEBench. Roadmap

The roadmap for OpenEBench development has been organized as follows:

1. Design and development of a comprehensive data model to hold benchmarking data (Task 2.4, Completed and updated within this deliverable, BSC, SIB)
2. Select a series of scientific communities active and mature in benchmarking activities, and incorporate a digested summary of their already available benchmarking data into OpenEBench repository (Task 2.1, Completed, BSC, SIB, CRG, IRB)
3. Experiment and develop visualization alternatives to offer a quick overview of scientific benchmarking data suitable to be incorporated to software registries (T2.2, ongoing, BSC, SIB)
4. Define an extensive list of software quality metrics, and develop the necessary interfaces for gathering such information (completed, reported at deliverable D2.2, BSC, DTU)
5. Develop a simple visual element (HTML widget) to summarize software quality data that can be exported to registries like bio.tools (Completed, reported at D2.2, BSC, DTU)
6. Develop the necessary APIs to distribute benchmarking data to the community and to ELIXIR information ecosystem (ongoing, BSC, SIB, CRG).
7. Develop an automated platform to apply benchmarking metrics, that can evolve to host complete scientific benchmarking events (Task 2.3, prototype for benchmarking metrics, BSC, CRG).
8. Develop an automated platform to evaluate technical and scientific performance of bioinformatics tools under the same technical environment (ongoing, Task 2.3, BSC).

A1.4. Novel data types for organizing benchmarking activities.

In an effort to standardize the benchmarking process *per se*, 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. Figure 2 depicts the workflow for a single *Benchmarking Event*. Participants represent those systems e.g. individual tools, analytical workflows, web-servers, taking part of a specific benchmark event. The detail of OpenEBench data model is available at https://github.com/inab/benchmarking-data-model. A detailed explanation of created data sets types follows:

- ***Public Reference data sets***. They are a 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. Scientific communities tend to make available *Public Reference data* to facilitate the engagement of participants

within the challenges at hand. These data sets could comprise data from previous benchmarking editions but it is highly dependant on the community and the scientific problem at hand.

- ***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 specific reference proteome sets for the Quest for Orthologs participants; and/or can be submitted automatically by benchmarking platform e.g. CAMEO, to participants web-servers. *Input data sets* should follow at least the same data formats as the *Public Reference data sets*, and 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. Depending on the level of automation, participant data sets can be submitted manually e.g. uploaded to a server, and/or automatically e.g. response via APIs implemented in systems like BeCalm. Unless previously agreed, participant data sets are often kept private to participants and/or communities. It would be recommendable 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.

- ***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 that participants from adjusting their systems to have the best performance for very specific data sets, which is often referred as overfitting. Overfitting may render systems useless and not-fit-to-purpose and, therefore, it is highly discouraged. Depending on the nature of the *Metrics Reference data sets*, those can be either "*Gold data sets"* or "*Silver data sets*". It is not an uncommon to have both types of data sets as part of a *Benchmarking event*. When available, *Golden data* is desirable because represent the ultimate data that any system should aim to produce. For instance, in the case of Protein Structure Predictions the experimental data deposited in the Protein Data Bank (PDB) is considered to be the *"Gold data"* for the benchmarking activities carried out by communities such as CAMEO, CASP, and CAPRI. In the absence of a gold standard, benchmarking efforts have to resort to "*Silver data*". For instance, synthetic and/or simulated datasets generated in silico following previous experiences[8] or with data generated using unsupervised learning approaches, based on the consensus

[8]Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics, 14(1), 184. https://doi.org/10.1186/1471-2105-14-184

among different —i.e. algorithmically independent — methods[9]. For the latter, naive methods e.g. Bayesian networks, can provide a baseline allowing assessors to measure relative performance between methods with, on average, moderate to good accuracy. Such consensus data is referred to as *"Silver data"*. However, data from silver standards should be used with caution as it needs to be revised regularly to adequately evaluate new developments in the field. 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 e.g. Q50, 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 participants 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. Even when participant data sets are not available, assessment data sets can be very useful to measure the performance evolution of different systems versions for the same challenge and/or the complexity of different reference metrics data sets for the same system. Ideally, assessment data sets would allow to track the evolution of both reference metrics data sets and systems versions. However, it would be nearly impossible to deconvolute the impact of each variable into the final results.

- ***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 an unified framework to compare participants performance among themselves for a specific scientific challenge and/or the evolution of individual participants along time. *Challenge data sets* allow data bundling and 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 2. In the case of continuous benchmarking systems, the red arrow at

[9] Elsik, C. G., Mackey, A. J., Reese, J. T., Milshina, N. V., Roos, D. S., & Weinstock, G. M. (2007). Genome Biology, 8(1), R13. https://doi.org/10.1186/gb-2007-8-1-r13

figure 2 indicates the start of the subsequent cycles which often tend to keep the same metrics and change the *Reference Metrics data sets* e.g. CAMEO.

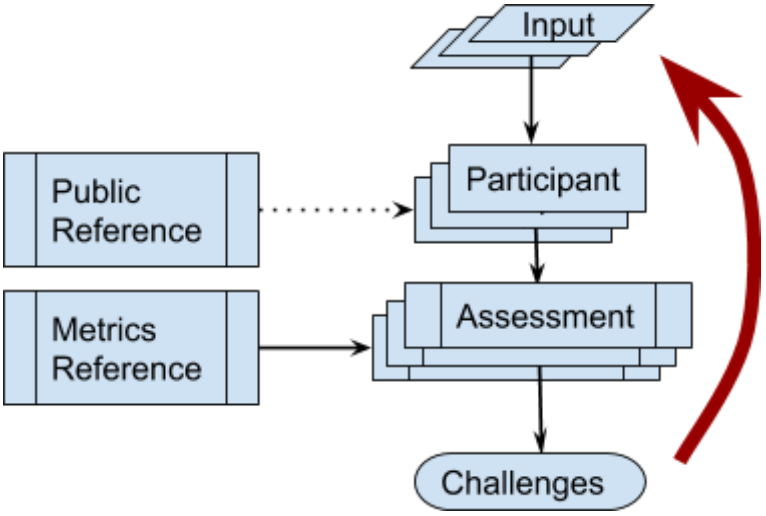

Figure 2. OpenEBench definition of datasets and how they relate to each other.

When considering a system like CAMEO [], the above mentioned six data sets map to this effort as identified in table 1.

Data Set Types	**Applied to CAMEO-3D**
Public Reference	Deposited Fixed Benchmarking Sets (usually at the Protein Data Bank)
Metrics Reference	Subset of weekly PDB released structures
Input	Sequences of target structures selected from the PDB weekly release
Participant	Different predicted structures obtained from participant's servers
Assessment	Results from applying a list of accepted metrics to predict
Challenge	Integrated assessment for each Benchmarking Event. Cumulative monthly, yearly reports

Table 1. Mapping of CAMEO data sets to the already described benchmarking data sets.

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 incentive open discussions and decisions within community around which scientific challenges are relevant.

Horizon 2020 grant n. 676559

Moreover, those efforts can be re-used by other communities and maximizing the data added value. For some communities in the health sector, it is accepted that (some) reference data sets are private, and therefore they cannot be made publicly available for ethical 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 general rule, data should follow the FAIR data principles[10] [Wilkinson et al. 2016], which states how to make data Findable, Accessible, Interoperable and Re-usable. This is part of a general movement in favor of implementing the principles around Open Science, Open Data and Open Source.

When defining reference data sets the data ownership is an important aspect. In order to avoid systems overfitting communities might decide to conduct specific experiments to generate *Input* and/or *Metrics Reference data sets*, which are used for specific benchmarking events. In those circumstances and until data is publicly released e.g. via a scientific publication, data is private to the organizers and benchmarking participants should honor that. Thus, a legal mechanism to regulate data ownership and use is highly relevant. Specifically, participants should accept a legal binding agreement which prevent them to use accessed data for purposes different to participating in the benchmarking activities at hand. CAMI (Critical Assessment of Metagenome Interpretation) already implements such policy to guarantee that participants honor such agreement. However, their system cannot change the status easily, given that there is a manual validation of scanned documents step before participants gain access to data.

A1.5. Data accessibility for Community-led Scientific benchmarking activities.

Another important aspect for supporting benchmarking activities carried out for scientific communities is how data is accessed and shared through OpenEBench and associated APIs. As stated in the previous section, data should be made publicly through the data life cycle unless ethical and/or legal aspects prevent that. However, the system should be flexible enough to offer scientific community members, organizers and participants control over how data is accessed and distributed at any point in time. Thus, we propose four different data accessibility models in OpenEBench:

- **Private**. This is the most restrictive accessibility model in OpenEBench. In this mode, only the data owner have access to this data as well as the data derived from it e.g. *Assessment data* obtained when processing participants data. This accessibility model will facilitate participants to compare themselves with already existing data in a specific Benchmarking event, and might be useful at the initial stages of benchmarking challenges when it is needed to make sure that submitted data is behaving as expected.

- **Restricted access**. This accessibility model allows to share data sets using URLs. This is a very convenient mechanism to foster collaborations among developers of distributed systems as well as to communicate results with restricted audiences e.g. among peers when a scientific manuscript is submitted.

[10] 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

- **Community based**. This is the default accessibility model when a *Benchmarking event* is on-going. This model allows participants to share and/or compare their system performance, e.g. *Assessment* and/or *Challenge data sets*, on real time among community members. This will facilitate open and transparent discussions among community members and it also can facilitate the detection of potential flaws in the setting up of the on-going event.

- **Publicly available**. This is the default accessibility model for already closed Benchmarking events. This visibility mode allows different stakeholder to have access to data e.g. *Assessment* and/or *Challenge data sets*, and data transformations associated to them e.g. transitions between experts and non-experts views applying different classification algorithms. Making publicly available data is not constraint to finalized Benchmarking event because participants and/or events organizers can make data under their responsibility public. Importantly, once a data set is made public, it should be maintain as such to avoid potential confusion across stakeholders.

Independently of the visibility mode, data should follow the FAIR principles e.g. use of persistent and unique identifiers, because it should be possible to change the visibility mode among available ones e.g. private data could be made available to a whole community; restricted access data can be made publicly available, etc. Moreover, data should be interoperable at any time in and outside OpenEBench to facilitate their access, secondary analysis and/or further re-use by communities running scientific benchmarking activities. OpenEBench will work closely with the ELIXIR Data platform to identify the most suitable long-term data repositories for data generated at the platform.

A1.6. Community-led Scientific benchmarking activities. A use-case perspective.

Within WP2 we have successfully contacted 12 community efforts and initiated collaborations with some of them. While CAMEO[11] and QfO[12] have always been very closely following the OpenEBench development, data from TCGA[13] and workflows from CocoBench are currently being imported into OpenEBench

Tables 2, 3 and 4 contain a comprehensive overview of real world data and its correspondence with the above identified data types for benchmarking in general.

[11] Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., … Schwede, T. (2017). Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins: Structure, Function, and Bioinformatics, 86, 387–398. https://doi.org/10.1002/prot.25431

[12] Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., … Dessimoz, C. (2016). Standardized benchmarking in the quest for orthologs. Nature Methods, 13(5), 425–430. https://doi.org/10.1038/nmeth.3830

[13] Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., … Mariamidze, A. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell, 174(4), 1034–1035. https://doi.org/10.1016/j.cell.2018.07.034

Communities	Data sets type	
	Public Reference	**Metrics Reference**
CAMEO	PDB protein structures in mmCIF format	latest release of PDB protein structures in mmCIF format [1], structures not available
CASP	PDB protein structures	protein structures withheld from publication by PDB, structure can be known by assessors
QfO	Uniprot Reference Proteomes	1) Manually curated and/or community agreed evolutionary relationships among protein pairs. 2) Functional annotation of specific protein families.
BioCreative	Public text corpora, like PubMed and/or results from previous BioCreative challenges	1) Training datasets, provided before the challenge starts. 2) Manually curated text corpora used as input for the participants
CAPRI	PDB protein structures in mmCIF format Published Benchmark Datasets	Delayed PDB structures in agreement with authors
CAFA	UniProt protein database (sequences + annotations), GOA	Manually and automatically curated GO and HPO annotations obtained later from the input
MD	Trajectory DBs (MoDEL, BNS)	Consensus Good Quality Trajectories
TCGA	MC_3 Working Group produced a MAF file from 9,079 exome samples from a variety of cancer types.	Lists of cancer driver genes per each analysed cancer type (33). Lists of driver mutations associated to each cancer type for a total set of 3,437 unique mutations.
GMI	No training data provided.	Ingroup definitions from analysis of real outbreaks using classical methods and tree topologies (as stored in the trees[14]), and known

[14] https://github.com/WGS-standards-and-analysis/datasets

Horizon 2020 grant n. 676559

		ground truth from lab experiment design[15] [Ahrenfeldt et al. 2017]
CoCoBench RNAseq	No training data provided.	Generated ground truth upon which sampling was based e.g. gene lists of up- and downsampled between samples.

Table 2. Reference data used to train participants systems as well as to evaluate their performance.

Communities	Data sets type	
	Input Reference	**Participant**
CAMEO	Protein Sequences in FASTA format, Protein Structure Models for quality estimation and refinement	protein structures produced by modeling servers [2], quality estimations produced by servers and standalone packages,contact predictions produced by servers
CASP	Protein Sequences in FASTA format, Protein Structure Models for quality estimation and refinement	protein structures produced by modeling servers [2], quality estimations produced by servers and standalone packages,contact predictions produced by servers
QfO	Specific - in terms of release - set of UniProt Reference Proteomes.	1) Predicted orthologous pairs; 2) Annotated proteins.
BioCreative	Text corpora	1) Detected biologically significant entities (names) such as gene and protein names and their association to existing database entries. 2) Detected entity-fact associations (e.g. protein - functional term associations).
CAPRI	Several protein sequences in FASTA format	protein - protein complexes produced by protein - protein docking methods
CAFA	Anonymized UniProt sequences and specific GO release to be used for the annotations	Scored GO terms annotating each anonymized sequence

[15]Ahrenfeldt, J., Skaarup, C., Hasman, H., Pedersen, A. G., Aarestrup, F. M., & Lund, O. (2017). Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. BMC Genomics, 18(1). https://doi.org/10.1186/s12864-016-3407-6

Horizon 2020 grant n. 676559

MD	3D structures (from PDB) and defined simulation conditions	Set of MD Trajectories
TCGA	Same as Public Reference data set to avoid confounding effects from using a different variant calling protocol	Cancer driver genes and/or mutations specific per cancer type.
GMI	Reference datasets from real outbreaks[16] and wetlab experiments [Ahrenfeldt et al. 2017]	List of strains that belong/do not belong to outbreak and reconstructed tree topology.
CoCoBench RNAseq	Raw read data from publically available RNA-seq experiments subsampled so as to generate a known ground truth within a real dataset.	List of up- and downregulated genes.

Table 3. An overview of the data used for participants in each challenge and the data produced by them, which will be later evaluated.

Communities	**Data sets type**	
	Assessment	**Challenge**
CAMEO	Superposition-independent Scores e.g. lDDT[17]	Specific assessments, aggregated scores within a certain category, e.g. CAMEO QE, CAMEO 3D, CAMEO CP
CASP	Assessment based on manually assigned assessment units, overall ranking produced by combining several scores by individual assessors in each category	one round of assessment in each category
QfO	Precision VS recall assessment based in different units depending on challenge . e.g average Schlicker similarity between orthologs and functional annotations VS total number of orthologs relations predicted	Several assessment approaches such as discordance with species tree or functional similarity with GO annotations.

[16] https://github.com/WGS-standards-and-analysis/datasets

[17] Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics, 29(21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473

BioCreative	Manually annotated and curated text corpora. And probably precision/recall, comparing with the annotated results	On each edition, the number of tasks (different test events on the same benchmarking event) increases, so the original two categories (entity detection, entity-fact associations) have evolved into six in the last edition.
CAPRI	Manually compare the submissions to the experimental structure e.g. evaluate the models on criteria that depend on the geometry and biological relevance of the predicted interactions.	Each round of CAPRI.
CAFA	Semi-automated assessment, as there can be an initial phase where the metrics are fine-tuned. At the end, it is taken into account precision (how generic, precise or overfitted is the annotation), recall (were all the annotations provided?), and a minimum percentage of predicted annotations in each category (to avoid declaring winner some participant which only provide results for easy targets)	On each round, several categories, by organism, as well as the "moonlighting" dataset category
MD	Results of Trajectory quality analysis	Comparative assessment with reference data
TCGA	Precision and % of true positives per cancer type.	Performance summary across all cancer types for each participant.
GMI	Assessment of whether the used system was able to cluster the outbreak strains correctly e.g. no non-outbreak strains in the outbreak cluster and vice versa. This can be measured using a relative Robinson & Foulds distance for each case.	Assessment of comparability and robustness of different WGS-based outbreak detection and reconstruction methods on datasets containing different challenging characteristics
CoCoBench RNAseq	Characteristics of detection of up- and downregulated genes (e.g. ROC curves, cutoffs etc.).	Summary of different results over all used tools (e.g. F-score of TP rate at different cutoffs).

Table 4. Data produced once metrics are applied to participants submitted data considering reference metrics data, and how those data are considered in an ample context.

Horizon 2020 grant n. 676559

In summary, CAMEO, CASP and CAPRI are employing gold standards, while QfO, CAMI, GMI, and CoCoBench have to resort to synthetic benchmark sets and maintain a well-appreciated and rather large effort to keep these synthetic reference data sets up-to-date and distinct enough to qualify as reference in future assessments. Other efforts such as BioCreative and MD have some hybrid approximations as data sets can be considered gold standards once they have been validated by expert curators.

A1.7. Update on data model since deliverable D2.1

The OpenEBench data model[18] is currently in Version 0.4, reflecting the above described data types and their dependencies. At the highest level of hierarchy, the Benchmarking Event refers to a number of "Challenges" that initiate the data flow through a series of "Test Actions" (setupEvent, testEvent, metricsEvent, or statisticsEvent) to reflect the operations linking those data types (see Figure 2). The dataflow ends in a number of "Challenge data sets" referring to a collection of reports on the benchmarking process. Details of the data model, validation tools, and prototype data can be found at https://github.com/inab/benchmarking-data-model.

A1.8. Further uses of OpenEBench.

Attracting scientific-led benchmarking activities at the time of capturing their needs and specificities is the main driver behinds OpenEBench development. However, one of the mid-term objectives, together with the platform sustainability, is how to certify analytical workflows. There is an increased interest by bioinformatics groups working at healthcare settings on certifying their protocols in order to ensure that they consistently produce the expected results and incorporate the state-of-the-art technologies into their daily activities. Since OpenEBench will host most, if not all, of those analytical workflows it seems plausible to go a step further and provide the means to certify technical and/or scientific performance for a range of settings and input data sets.

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