EXCELERATE Deliverable 2.1

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	Creation of a database warehouse infrastructure for storing and organizing data for online performance assessment experiments	
WP No.	2	
Lead Beneficiary:	7 - CNIO	
WP Title	Benchmarking	
Contractual delivery date:	28 February 2017	
Actual delivery date:	21 February 2017	
WP leader:	Alfonso Valencia Søren Brunak	7 - CNIO 39 - DTU
Partner(s) contributing to this deliverable:	7 - CNIO 12 - BSC 25 - SIB	

Authors and Contributors:

Salvador Capella-Gutierrez (CNIO), Diana de la Iglesia (CNIO), Jürgen Haas(SIB), Josep Lluís Gelpi (BSC), José María Fernández (CNIO),

Table of contents

1. Executive Summary

Appropriate benchmark design is a requirement for assessing and improving bioinformatics methods and tools, and needs of best practices to pass over from raw data to valuable knowledge for decision-making. In order to build a continuous automated benchmarking infrastructure which hosts different - emerging or existing - benchmark efforts, bioinformatics tools and data-types, it is crucial to discuss storage, analysis, comparison and sharing of large heterogeneous data sets. Finally, the most appropriate format to communicate results by either exposing them to third parties resources e.g. tools registries, and/or directly via a web-portal.

The increasing complexity of the constantly growing body of biological data e.g. unstructured description of resources, non-standardized input and output formats, lack of appropriate metadata and/or deprecated software source codes, represents a tremendous challenge. Thus, we must develop new technical solutions or adapt existing ones to leverage existing data and to better prepare for future challenges around the scientific, technical and functional evaluation of bioinformatics methods and tools.

Deliverable 2.1. defines the data warehouse infrastructure needed to host different benchmark initiatives from a broad range of bioinformatics fields, provides a reference implementation, and oversee the overall ELIXIR Tools and Data Services Registry integration and operation.

On this report we will examine the following aspects of a data warehouse:

- **Background** about the need of an automated benchmark infrastructure.
- An **architecture overview** with an emphasis in the data warehouse as central component.
- An explanation about the **database warehouse design and implementation**.
- A final review about **future works** in the context of the benchmark infrastructure.

Horizon 2020 grant n. 676559

2. Impact

Not applicable

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 subareas. (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 X No

5. Adjustments made

No adjustments have been made.

6. Appendix 1: Database warehouse infrastructure for storing and organizing benchmark data

Horizon 2020 grant n. 676559

Database warehouse infrastructure for organizing and storing benchmark data.

Summary.

Appropriate benchmark design is a requirement for assessing and improving bioinformatics methods and tools, and needs of best practices to pass over from raw data to valuable knowledge for decision-making. In order to build an automated benchmarking infrastructure which hosts different - emerging or existing - benchmark efforts, bioinformatics tools and data-types, it is crucial to discuss storage, analysis, comparison and sharing of large heterogeneous data sets. Finally, the most appropriate format to communicate results by either exposing them to third parties resources e.g. tools registries, and/or directly via a web-portal.

The increasing complexity of the constantly growing body of biological data e.g. unstructured description of resources, non-standardized input and output formats, lack of appropriate metadata and/or deprecated software source codes, represents a tremendous challenge. Thus, we must develop new technical solutions or adapt existing ones to leverage existing data and to better prepare for future challenges around the scientific, technical and functional evaluation of bioinformatics methods and tools.

Deliverable 2.1. defines the data warehouse infrastructure needed to host different benchmark initiatives from a broad range of bioinformatics fields, provides a reference implementation, and oversee the overall ELIXIR Tools and Data Services Registry integration and operation.

Background.

Critical benchmarking of bioinformatics tools adds value to different research communities by providing objective metrics in terms of scientific quality, technical reliability, and functionality [Jackson et al. 2011, Friedberg et al. 2015]. At the same time, target criteria agreed within a community are an effective way to stimulate new developments by highlighting areas which require improvements [Costello and Stolovitzky 2013]. This is especially relevant when progress can be measured close to real-time by continuous automated benchmark services.

One of the most productive and sustainable ways to create a community-driven benchmark initiative is to organize it as a challenge-based competition with clear participation rules, a scientific sound set of questions, and previously agreed common data sets. Several challenges have been organized over the last two decades with fruitful competitions and successful results, e.g. CASP (Critical Assessment of Techniques for Protein Structure Prediction) [Moult *et al*. 1995], BioCreAtIvE (Critical Assessment of Information Extraction in Biology) [Hirschman *et*

Horizon 2020 grant n. 676559

al. 2005], CAFA (Critical Assessment of Functional Annotation) [Radivojac *et al*. 2013] and QfO (Quest for Orthologs) [Altenhoff *et al*. 2016], among others. Our focus at ELIXIR-EXCELERATE WP2 is on developing and making sustainable over time a benchmark infrastructure which can be used for existing communities and newly created ones. In particular we foster benchmarking efforts, which can be automated, and could potentially run continuously. We envision that many scientific communities would benefit from a stable, generic and efficient infrastructure devoted to host unattended, periodic and continuous benchmark services. Such infrastructure will be in charge of gathering participants data, measure performance, and produce metrics on-demand (Figure 1). Periodic updates on benchmark data sets will contribute to advance in the development of bioinformatics methods and tools by reflecting new challenges in each field which need to be tackled by new developments.

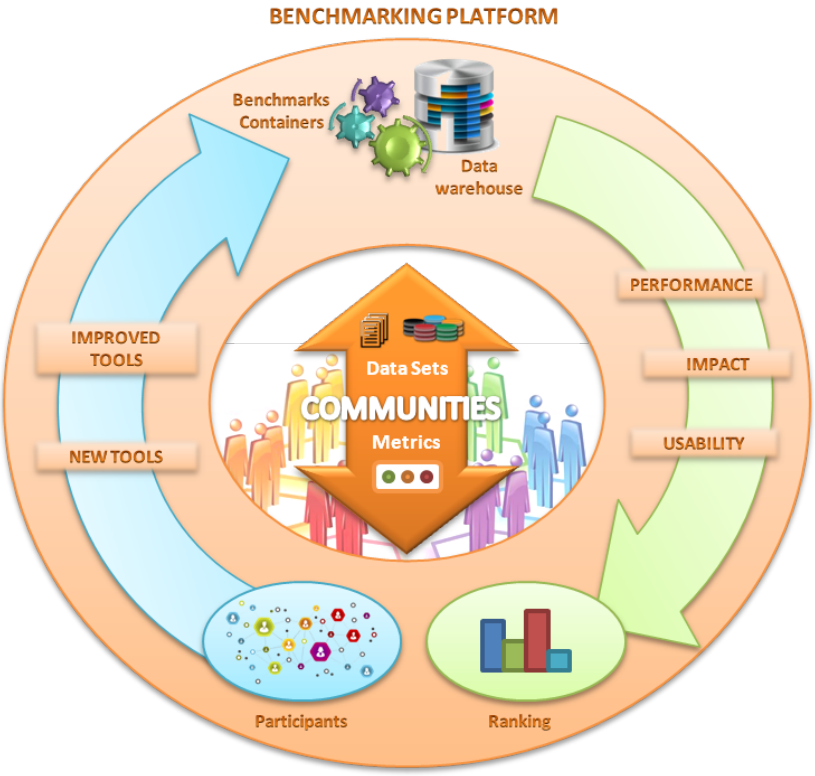

Figure 1. Continuous benchmarking process and components which allow developers to implement new functionalities and tackle relevant questions as the scientific domain evolves.

Architecture overview.

The key challenge tackled at WP2 is the integration of highly heterogeneous data sets and data models from a multitude of bioinformatics fields into a single infrastructure. In order to stock this infrastructure with data we need to be able to integrate biomedical data-sets from existing benchmarking challenges which often

use different formats and semantics. Moreover, the infrastructure should facilitate the direct integration of newly created benchmark efforts. Thus, the main goal is to define a framework in which most (if not all) benchmarking services created by different communities can fit.

The system architecture has been defined based on existing efforts and after an extensive revision of successful and failed benchmark experiences (Milestones 2.1 and 2.2, unpublished). The system, depicted at figure 2, is designed around automated continuous community-driven benchmarking efforts. The key component of such unified infrastructure is the data warehouse. The main aim of this warehouse is to store all the metadata related to each edition of a given benchmark e.g. participants, pointers to input and output data sets, metrics and results, etc. in order to make them accessible to users, and tools developers (Task 2.4). As previously shown, the use of a data warehouse improves data quality and interoperability, and significantly shortens data collection and handling times compared to a manual process [Watson *et al*. 2002].

An agile approach is being taken to define and implement the whole infrastructure, and in particular the data warehouse. A first prototype is presented in this deliverable, tagged as https://github.com/inab/benchmarking-data-model/releases/tag/20170220. A initial version of the data access portal, just for demonstration purposes is available at http://elixir.bsc.es/benchmarking. The data warehouse and its data model will receive modifications as the common data model evolves to incorporate new benchmarking efforts with their technical peculiarities and data integration-related issues. This process reuses existing systems and services whenever possible, supports extensibility on the long-term, allows for integration of new efforts and will enable interoperability with the ELIXIR Tools and Data Services Registry (Tasks 1.2 and 2.2).

Regarding the rest of components of the infrastructure, we have adopted a modular design which makes an extensive use of REST APIs (Application Program Interface) to abstract the parallel development of different components. At the moment, the most developed component is the data model, a common component of the data warehouse and many of the APIs. The rest of components only have prototypes which will be subject of numerous modifications before achieving a stable status. For instance, data from existing initiatives are manually fetched to the data warehouse in order to understand i) where to obtain the data, and ii) make sure that all existing data is properly captured by the proposed data model (figure 3). The infrastructure is built making use of software containers, specifically Docker [Boettiger 2015], which facilitate reproducibility, easy deployment and flexible building of collections of tools and engines dedicated to handle a system, which will evolve over time to accommodate a myriad of different benchmark efforts and communities.

The platform will also provide uniform access, based on standard interfaces, to relevant external (public) sources of data and tools that need to be integrated in our benchmarking platform, such as the *bio.tools* registry [Ison et al. 2016]. The system

will provide benchmark results primarily to the ELIXIR Tools and Data Services Registry for enriching the user experience in a single end-point. However, the modular design of the infrastructure allows to provide data to other tools registries, and even provide directly results to end-users, and tools and methods developers via a web portal. Results available via the web portal will have a mechanism to allow tools developers to decide whether they want to share them publicly or not. This is an important aspect to ensure community engagement by allowing a fully customizable experience [Friedberg *et al*, 2015].

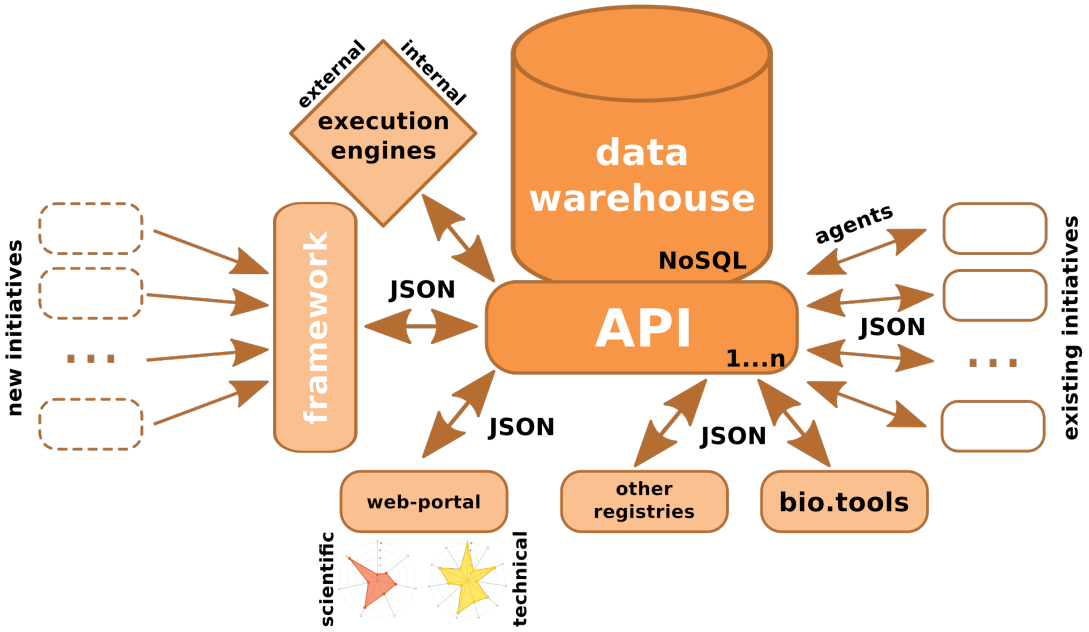

Figure 2. Schematic representation of the benchmarking platform for storing and organizing data from existing and newly created efforts. The data warehouse is the key component of the infrastructure and make an extensive use of APIs to interact with different internal and external components abstracting the development of the whole system.

Database warehouse design and implementation.

Within the benchmarking work package, the initial design, implementation and validation of the data warehouse has been closely related to existing benchmarking initiatives from different bioinformatics domains e.g. CAMEO [Haas *et al*, 2013], CAFA [Radivojac *et al*, 2013], and Quest for Orthologs [Altenhoff *et al*, 2016]. Feedback coming from other partners in ELIXIR will contribute to refine and/or extend the initial model by bringing in other domains e.g. text-mining (WP3) or the marine metagenomics use case (WP6)(Task 2.5). Moreover, ongoing collaborations with other European H2020 projects such as openMINTED (http://openminted.eu/)

will allow us to show the clear role of ELIXIR as a stable pan-European infrastructure.

As mentioned before, the data warehouse stores data and metadata from past and ongoing initiatives submitting data to the platform. To ensure reproducibility and enable interoperability across different editions of the same benchmark, and initiatives potentially sharing data sets, three important steps should be conducted for each data source prior to its inclusion within the data warehouse: i) data extraction, ii) data transformation, iii) data-model mapping between the data source and the data warehouse. To achieve this we need:

- An standards-based semantic core dataset to provide a common vocabulary for data models.
- A mapping format to integrate data models from different benchmark experiments into the common data model by establishing the correspondence between the concepts, attributes and relations of both models.
- A semantic interoperability framework to retrieve data from uniform queries from different sources of benchmarking data.

Special care will be taken to ensure that the entire process of data extraction and integration is in line with all applicable data protection requirements, and follow the FAIR principles for data management and stewardship put forward by other ELIXIR partners [Wilkinson *et al*. 2016].

Database system.

Despite of the wide use of relational databases, these systems were not designed to cope with the large-scale and heterogeneity requirements posed by biomedical data nowadays, in terms of performance and scalability. Thus, to design and implement the data warehouse we have used a non-relational approach based on the implementation of a NoSQL database (http://nosql-database.org/). The features provided by non-relational databases appear to be promising for analyzing and managing the large volumes of data needed to successfully carry continuous automated benchmarks of bioinformatics methods and tools. Specifically, NoSQL databases are open-source, distributed, horizontally scalable and easy to replicate. In addition, they are schema-free, which means that the original data model could be modified dynamically in a simple manner to include new peculiarities of the data to be represented and stored. Such characteristics make NoSQL databases an ideal candidate to structure and store heterogeneous data about benchmarking experiments which will evolve over time. From the numerous available NoSQL systems, we have chosen MongoDB (https://www.mongodb.com) to implement the data warehouse. The MongoDB document-centric architecture makes it the most suitable choice for the system. Moreover, MongoDB stores data in document-like structures which encode information using standard formats e.g. JSON (JavaScript Object Notation). This is essential to ensure data interoperability and the use of non-proprietary formats. In addition, MongoDB lets users perform structured and *ad hoc*

queries, allowing the creation and combination of many types of questions, which is a feature that fulfills the search needs envisioned for the benchmarking platform.

Data interoperability.

The main issue to be considered when developing a database for storing heterogeneous data is interoperability. Linking and integrating data sets from different sources requires identifying common concepts, attributes and relationships in the data sets that refer to the same real-world entity but use a different representation and/or format. Thus, in order to reduce structural and contextual differences among data provided by different benchmark communities, we have created a mapping between the core dataset and the information models corresponding to each benchmark, both at the schema and instance level. Despite the highly automated nature of the infrastructure, this step is fully manual to ensure the correct mapping of data source concepts to the ones in the platform. An important aspect for levering data in the platform is to make it FAIR [Wilkinson et al. 2016]. FAIR principles establish a guide about how to make data, and especially meta-data, **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable always taking into account the access policies established by the original communities and/or our platform. In order to facilitate interoperability, we will make an extensive use of the EDAM ontology (http://edamontology.org) which is strongly linked to the ELIXIR Tools and Services registry (WP1). However, the system is flexible enough to make use of other ontologies developed by ELIXIR partners e.g. biosharing (https://biosharing.org), and non-ELIXIR partners e.g. meta-share (http://www.meta-share.org).

Standards.

Data standards are crucial for collaborations and exchange of data in general, and in particular for scientific and technological disciplines. Using a common formalism to represent elements within a domain guarantees the creation of dynamic and interoperable information systems that share data models based on standard terminologies. The adoption of standards is, hence, mandatory for the long-term operation of the benchmarking infrastructure, but the current lack of standards with regards to statistical assessment of tools, data set preparation and data sharing is impeding this important aspect. Hence, a task within WP2 (Task 2.5) is dedicated to the identification, evaluation and selection of appropriate standards, based on the community recommendations and the experience acquired in each challenge. The standards will also facilitate the end-users interpretation of the results.

Major format standards for representing biological data are based on variation of plain text formats e.g. CSV, TSV, XML, JSON, RDF, FASTA, PHYLIP, STOCKHOLM, SAM, FASTQ, among others, due to the fact that text is the exchange data format of the World Wide Web. Over the years, binaries and/or compressed versions of existing plain text formats have become available e.g. BAM and

FASTQ.gz, in order to improve tools efficiency and save physical space. The protein structure community for example is currently investing into extending the mmCIF standard towards hybrid modeling reflecting the constant evolution and grooming of existing and emerging standard formats. Thus, we have considered major standard formats for the data warehouse and will make sure to implement conversion tools for those cases where non-standard formats are used by any benchmark community. The later will be done in close collaboration with bio.tools for proper format mapping using EDAM as part of task 2.4.

Data model.

In order to achieve data interoperability at the schema level, we have first designed a data model to guide its development (Figure 3). This data model is implemented as a virtual model layer where data models from each benchmark initiatives are mapped to it as part of the data warehouse implementation.

There are three important considerations about the data model.

1. It allows to represent the native data models from each benchmarking effort e.g. input and output data sets, results and metrics, providing communities members with unified access to the available data sets for each challenge and/or benchmarking event.
2. It serves as a framework to connect different communities and/or benchmark events, as well as to integrate the tools participating in the benchmarking experiments.
3. It makes the data shareable among different communities, promoting the adoption of FAIR principles and the implementation of open science policies.

The data model, based on JSON Hyper-Schema (http://json-schema.org), currently allows information exchange about engaged communities, participant tools, input and output data sets, metrics, benchmark events and results. The model also reflects the special characteristics of continuous benchmarking events e.g. the periodic nature of different benchmark editions and/or rounds. Finally, it connects the participant tools with their implementations at the ELIXIR Tools and Data Services Registry (*bio.tools*).

Figure 3 shows a conceptual representation of the data model. The main concepts contained in this model are as follow:

- **Community**: Represents of any engaged benchmark community e.g. CASP, CAFA, and Quest for Orthologs, which includes its name, a unique community acronym, a short description, related links (URIs) e.g. links to its main site, publications, data repositories, community status e.g. consolidated, emerging, abandoned, and contacts of responsible researchers in charge of its coordination.

- **Contact**: A reference contact of a community, tools or metrics, including name, email, comments about the contact, and links related to the person e.g. publications, LinkedIn account, ORCID, main site among others.

Horizon 2020 grant n. 676559

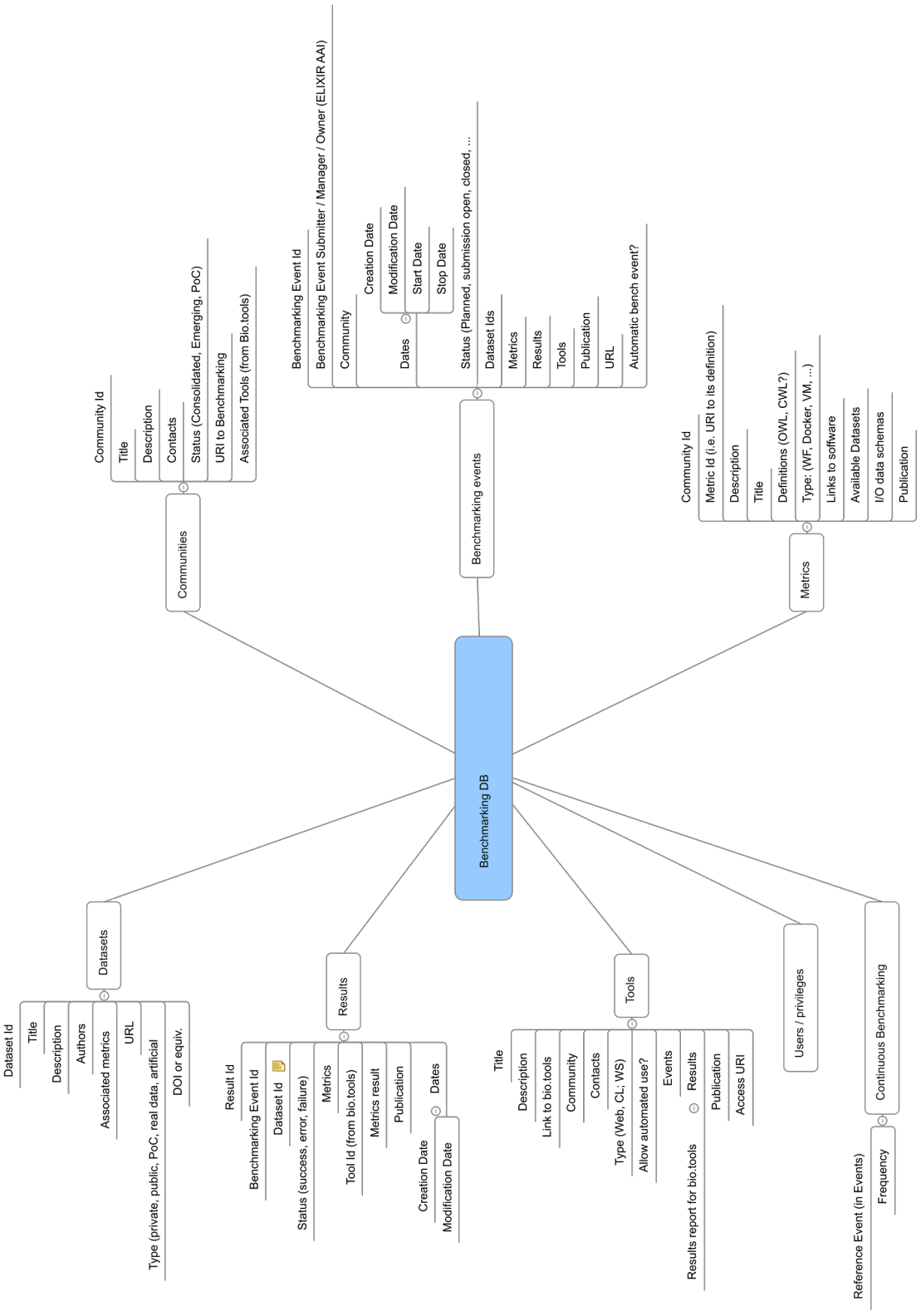

Figure 3. Conceptual data model of the data warehouse.

- **Reference**: A pre-print or publication reference, used to associate documents to a given community, contact, tool, dataset, benchmark event and/or metrics. It includes its title, DOI, PubMed identifier (when available), abstract, related links e.g. supplementary data, related resources, and relevant authors info.
- **Tool**: A tool which is benchmarked by one or more communities, in one of more test events from one or more benchmarking events. This entity includes tool name, a short description, a flag that indicates whether the tool can be automatically invoked, the community or communities where the tool belongs to, a reference contact, the status of the tool, references to the tool, the access type and a link to its access point, and a tools identifier linking it to the ELIXIR Tools and Data Services Registry.
- **Metrics**: Used metrics which can be applied over a dataset in one or more test events from one or more benchmarking events. It includes metrics name, description, creators and/or maintainers contacts, its formal definition, its execution type and data schema, and links and bibliographic references.
- **Data Set**: Either a reference one, often the benchmark input data set, or test event from a benchmarking event result which can be considered as output data set. It includes a data set short name, version or release date, a description, relevant dates for the data set e.g. creation and modification times, data set type, a link to the data set itself, contact details for creator/s, maintainer/s and/or curator/s data set, associated metrics resulting from the analysis of the data set, and references related to the data set in the data model, if any. Additional data sets types can be any kind of cumulative metrics (global statistics, prospective analysis, tools comparison, etc.) generated as part of the Benchmarking assessment or by the benchmarking platform.

- **Benchmarking event**: It is defined as a specific challenge category inside a set of challenges e.g. CASP 8 which is about contact predictions or CAFA 3 which is about functional site residues identification, either attended or unattended. This concept comprises all related test events, one per tool involved in the challenge. Moreover, the benchmarking event entity includes events name, whether it is automated or not, relevant dates for the event e.g. creation, modification, starting date, due date, a public URL to the benchmark event site (if available), the community where the event belongs to, and related contacts and references, if any.

- **Test Event**: It defines each tool involved in a specific benchmark event. The tool takes for this specific test event an input data set, and generates as result an output data set. The generated data set can hold the values of several metrics related to the challenge e.g. in the case of CAFA, the output data set can be the raw results, and the associated metrics can be the official answer to the challenge. Alternatively, an assessment challenge can generate as output dataset a copy of the input one, augmented with the

quality metrics computed with the assessment tool. This entity includes the tool that is being tested, input and output data sets, the benchmarking event where the test event was generated, a report on the test results and relevant dates for the test event e.g. creation and modification times.

Future works.

Here we have introduced the approach followed for the creation of the database warehouse infrastructure for organizing and storing benchmarking data, by providing first an overview of the overall benchmarking architecture, followed by a detailed description of the data warehouse design and implementation. We have also illustrated our strategy for ensuring interoperability among data sets and tools from different benchmarks and how we plan to map existing data related to the different benchmark efforts to a reference data model common to all the initiatives. Such a model will be subject to further modifications in order to accommodate the requirements posed by each benchmark initiatives and/or community. The data model, together with examples, is publicly available at the GitHub repository where one can follow the iterative refinement of the model.

After establishing a common data model which is able to represent the main concepts, attributes and relationships of each benchmark initiative - continuous or not, we will deal with data heterogeneities at the instance level. This involves the records and/or entries conversion from their former data model to the common model. We will only import to the data warehouse valuable information for communities avoiding deprecated and/or obsolete data and metadata which may increase the model complexity with little to none impact. This is especially relevant in a constantly evolving field such bioinformatics where data sets, and methods and tools are evolving continuously.

After populating the database, we will work in two parallel processes. On one hand, we will keep incorporating existing and emerging benchmark initiatives which could provide new insights to the process of modelling continuous benchmarking experiments. On the other hand we will advance on the APIs implementation in order to interact with the data warehouse. Finally, we will explore extensions to the data model to incorporate technical monitoring data in contrast to the scientific ones.

References

Altenhoff, A. M. *et al.* Standardized benchmarking in the quest for orthologs. *Nat. Methods* **13,** 425–430 (2016).

Boettiger, C. An introduction to Docker for reproducible research. *ACM SIGOPS Oper. Syst. Rev.* **49,** 71–79 (2015).

Costello JC, Stolovitzky G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. Clin Pharmacol Ther. 2013 May;93(5):396-8.

Friedberg, I., Wass, M. N., Mooney, S. D. & Radivojac, P. Ten simple rules for a community computational challenge. *PLoS Comput. Biol.* **11,** e1004150 (2015).

Haas, J. *et al.* The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database* **2013,** bat031 (2013).

Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6 Suppl 1,** S1 (2005).

Ison, J. *et al.* Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res.* **44,** D38-47 (2016).

Jackson M, Crouch S, Baxter R. Software Evaluation: Criteria-based Assessment. Technical Report. Software Sustainability Institute. 2011.

Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Genet.* **23,** ii–iv (1995).

Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10,** 221–227 (2013).

Watson, H. J., Goodhue, D. L. & Wixom, B. H. The benefits of data warehousing: Why some organizations realize exceptional payoffs. *Inf. Manag.* **39,** 491–502 (2002).

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., *et al*. The FAIR Guiding Principles
for scientific data management and stewardship. *Sci Data*. 3: 160018 (2016).