Facilitating the use and re-use of OpenEBench FAIR data through EOSC services

José Mª Fernández [1,2], Meritxell Ferret [1,2], Laia Codó [1,2], Josep Ll. Gelpí [1,2,3] and Salvador Capella-Gutierrez [1,2]

[1] Spanish National Bioinformatics Institute (INB/ELIXIR-ES); [2] Barcelona Supercomputing Center (BSC); [3] Dept Biochemistry and Molecular Biomedicine, University of Barcelona (UB)

ELIXIR All Hands 2022, 7-10 June, Amsterdam, Netherlands

OpenEBench is the ELIXIR open data infrastructure for supporting community-led scientific benchmarking and technical monitoring of bioinformatics methods and services. It is led by the Barcelona Supercomputing Center (BSC) in collaboration with partners within ELIXIR and beyond.
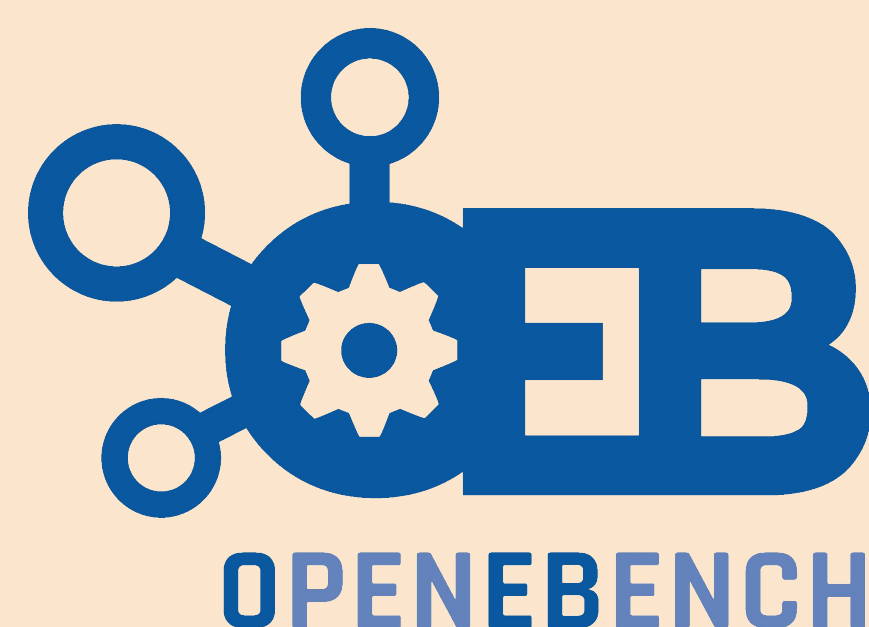
FAIR principles are high-level governance guidelines for ensuring scientific digital objects sustainability. They promote transparency in benchmarking assessment, results reproducibility, and pipeline and dataset reuse. OpenEBench use several strategies to further enforce the FAIR-by-design data strategies of the platform.

OpenEBench Data Model

OpenEBench Benchmarking Data Model defines the structure of a whole scientific benchmarking process. Using a set of JSON Schemas plus some extensions, the model represents the concepts used for scientific communities during a benchmarking life cycle.

- **Community:** The description of a benchmarking community, like CASP, CAFA, Quest for Orthologs, etc...
- **BenchmarkingEvent:** A benchmarking event is defined as a set of challenges coordinated by a community, either attended or unattended.
- **Challenge:** A challenge is the evaluation strategy defined by the community. It can be defined by a set of one or more metrics, reference datasets and test actions, related to the participants involved in the challenge.
- **Contact:** A reference contact of a community, tool, metrics or any other object.
- **Reference:** A bibliographic reference, used to document a community, a contact, a tool, a dataset, a benchmarking event or metrics.
- **Tool:** Software which can be used in the lifecycle of one or more benchmarking communities. Can be a participant in a particular benchmarking challenge, or software used to perform the benchmark itself.

- **Dataset:** Any one of the datasets involved in the benchmarking events lifecycle. So, they can be interrelated (for data provenance) and cross-referenced from the other concepts. There are 7 types of datasets defined in the model, which correspond to the specific data used in the different steps of a benchmarking event (e.g. metrics_reference, participant...) They are further explained in the data types section.
- **TestAction:** The involvement of a tool in a challenge, taking as input the datasets defined for the challenge, and generating the result datasets in the format agreed by the community. The generated datasets are later related to metrics datasets, which are the metrics agreed by the community for the challenge, used later to assess the quality of the result.
- **Metrics:** Defined metrics which can be computed from a dataset. Could be, for instance, the numerical values indicating some tools performance.
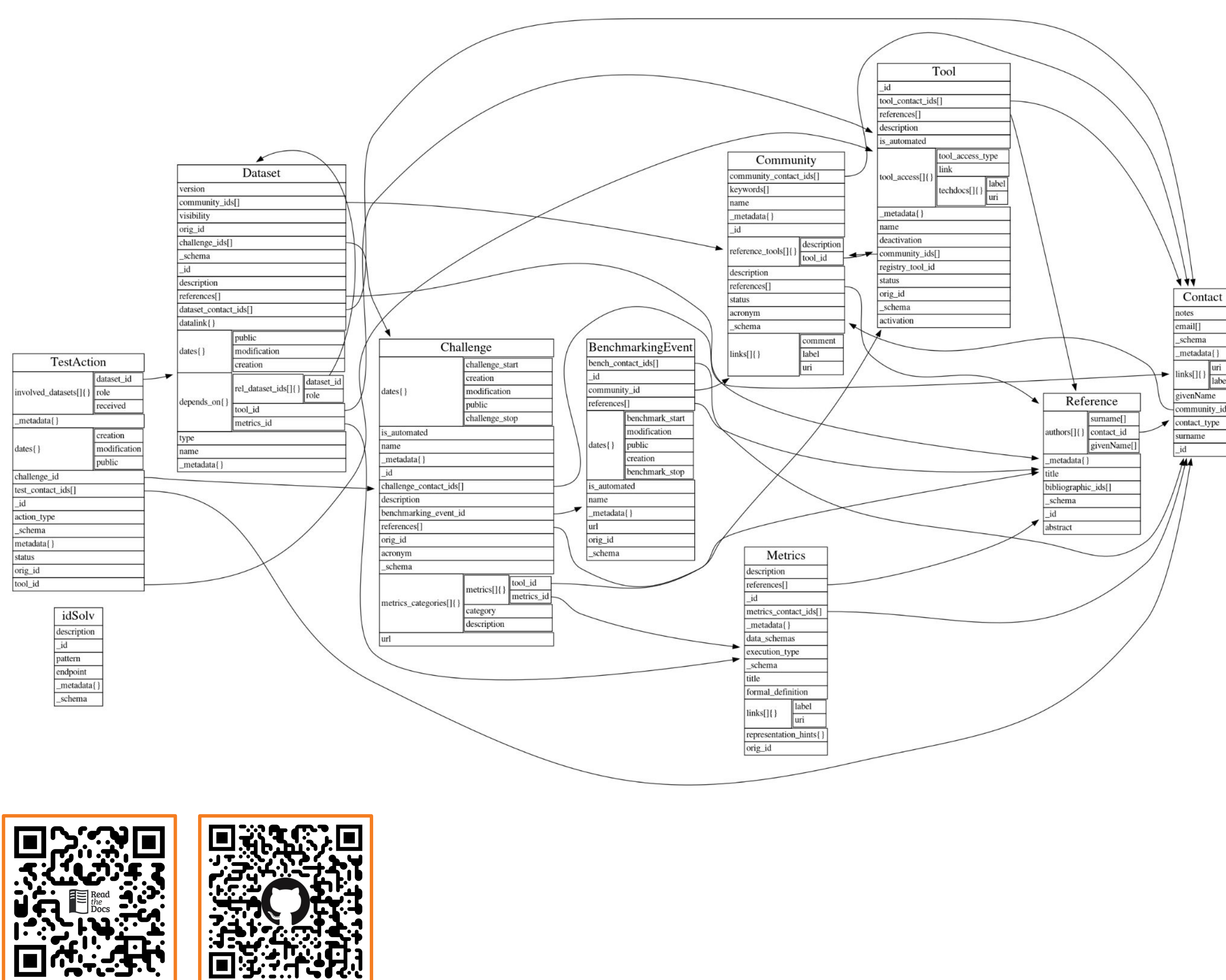
Benchmarking data publication to EUDAT

OpenEBench makes use of EUDAT (B2SHARE) for long-term availability and storage of scientific benchmarking datasets. To this end, OpenEBench is one of the research communities of EUDAT (4), benefiting from a particular extended metadata model.

Annotated according to OpenEBench data model (1), datasets are locally stored in the platform (2). With one click, the user can export them to B2SHARE (3), which will mint a versioned DOI (digital object identifier) for that dataset. A back cross-link with OpenEBench identifiers is kept so that published benchmarking data can be consumed from both B2SHARE and OpenBench platforms (5).
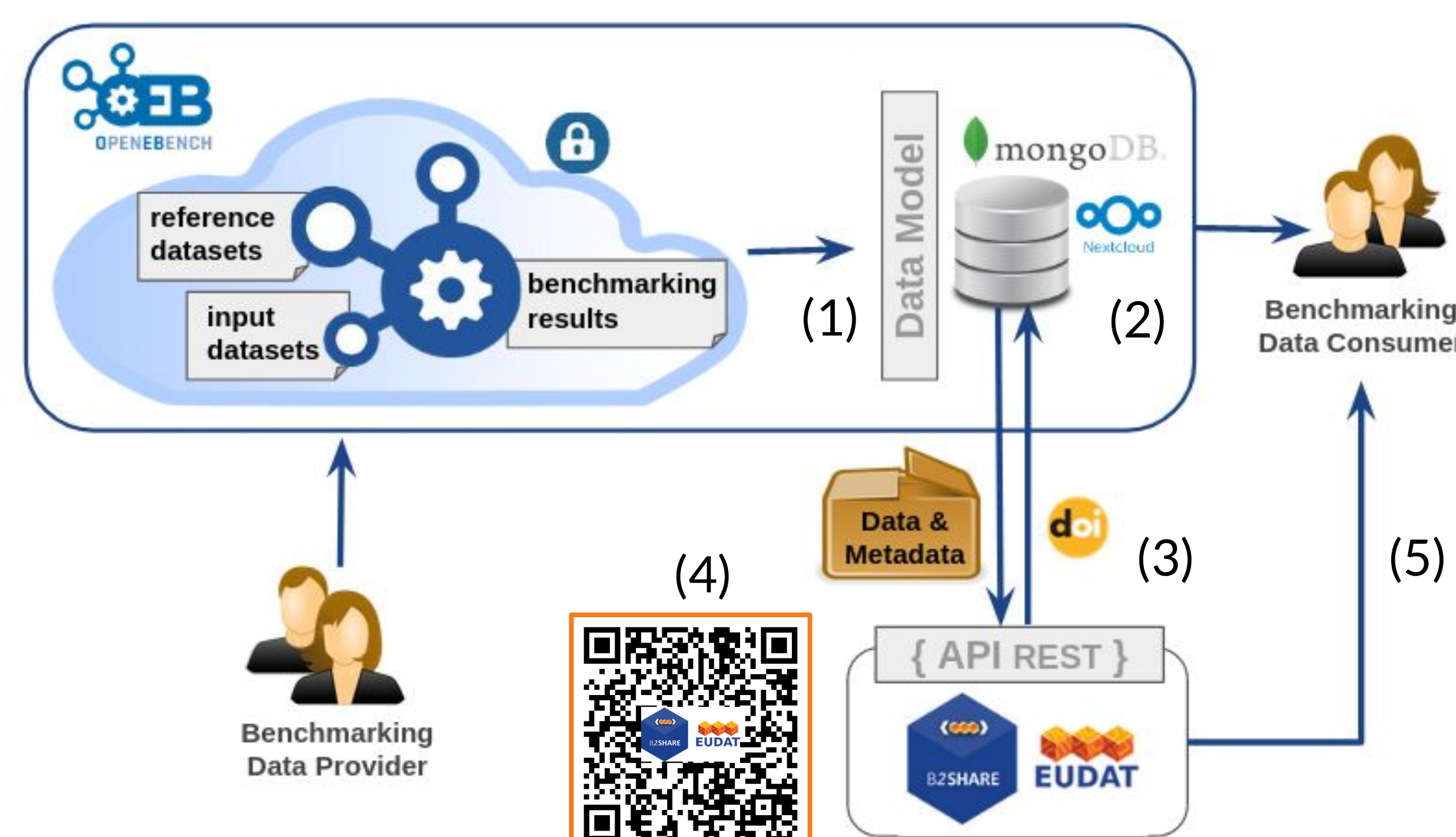
Contact

Meritxell Ferret
Junior Research Engineer, INB/ELIXIR-ES
openebench-support@bsc.es

@ELIXIREurope   /company/elixir-europe