OpenBio.eu: An Open and Social Research Object Repository and Workflow Management System

Computational Biology, Systems Biology and Bioinformatics are three crucial branches of the life sciences that have been particularly struck by the reproducibility crisis. Efforts to dissect the main causes of this crisis, resulted in bringing forward some central, systemic and deeply rooted issues in quantitative research. These issues include (1) unwillingness to share data, methods and source code, (2) lack or inadequacy of online environments and computational tools to help researchers share the complete methodology in a rapidly reproducible manner, and (3) lack - or hesitant adoption - of standards and metrics that could indicate the reproducibility merit of papers published in scientific journals and conferences. What is more intimidating is that these issues can be easily faked so that it looks like scientists have resolved them. For example, Data can be partially published in a repository that has an expiration date, source code can be cited in a repository that is practically impossible to compile/install/reuse, and papers can be considered “open” just by making the manuscript available free of charge.

Aiming to address the above matters, we developed and launched the openbio.eu initiative. OpenBio.eu is a novel online environment where researchers can create, edit, share, re-combine, export, execute, rate and collaborate on Tools, Data and Workflows (called hereafter Research Objects, ROs). OpenBio assumes minimum IT knowledge, does not impose any Domain Specific Language (DSL), and allows users to use any programming / scripting language. Also, every RO can be linked to publications and a user’s profile; this acts as a multi-directional connection between reproducible ROs, scientific publications and user profiles. Users get credit every time a RO that they “own” is reused, thus providing a much missing incentive for RO sharing in academia. It is stressed here that ROs are not simple descriptions or annotations of “real-life” Tools, Data or Workflows; they are software components, directly downloadable and executable in any modern computation environment.

Starting from Tools, users can create a tool by simply importing the commands that install it. These commands are the same with these that install the tool in any computer with a BASH shell terminal. The same tool can have many versions. Different users can import Tools that have the same name and version. Apart from the BASH commands, a Tool can be accompanied with a rich formatted text in markdown. Users can define Tool dependencies with a simple drag and drop. Tools can be in a “Draft” or “Finalized” stage. “Draft” Tools may be edited in the future. Once users are confident that a Tool does not need any refinements, they can select to “Finalize” it. A “Finalized” tool cannot be edited further; this acts as an immutable object that can take part in a reproducible, future-proof pipeline. Nevertheless, any user can “Fork” a “Draft” or a “Finalized” version of a Tool. Forking is a concept borrowed from software development, where users can have an identical copy of a source code and edit it as they wish.

Contrary to other environments, OpenBio.eu does not make any semantic distinction between Tools and Data. Similar to Tools, Data have versions, commands that install/download them, and even dependencies from other Data (or Tools). This simplifies the object model and dissolves an unnecessary and old-fashioned separation.

Workflows have been a central concept in Bioinformatics. Strictly speaking, a Workflow is a series of computational steps, connected through dependencies of the form: “a step gets to run only when these other steps are completed”. Common Workflow Management Systems (WMS) in Bioinformatics like Galaxy, Nexftlow and Snakemake, make this abstraction. Although this flow-centric construction of workflows has a long history in Bioinformatics, it has been inherited from industrial design systems where rigidity is of paramount importance, and is not perfectly suited to the versatility and flexibility of modern bioinformatics research. Moreover, flow-centric construction of workflows requires IT skills above average. In OpenBio.eu, users construct Workflows by simply importing the commands that execute a step. Again, these commands are the same as those that they would use in a terminal supporting BASH. The main difference is that instead of implicitly defining step order through a dependency resolution algorithm, they can directly call steps from other steps. This abstraction follows the “function calls function” paradigm that is familiar to users with basic programming skills.

OpenBio.eu offers a user friendly GUI for Workflow composition. Tools, Data and other Workflows are simply imported with drag and drop. Workflows can be nested indefinitely, while iteration and conditional execution is natively supported. The versioning logic that guides Tools and Data, also guides Workflows. Namely, the “Edit”, “Fork”, “Finalize” and “Draft” semantics, also work for Workflows.

Workflows can be converted from “structure-centric” to “flow-centric” constructs with an algorithm that creates Directed Acyclic Graphs (DAGs). These DAGs can be described in Common Workflow Language (CWL) and consequently imported to common Workflow Execution Environments like Galaxy, Nextflow, Snakemake and Airflow. We also provide a Docker container that is connected to the OpenBio.eu web server and acts as a personalized execution environment. Users can have as many Execution Environments as they wish. Every Workflow execution creates a separate object (Report) that contains results and logs. These can be shared with other users.

OpenBio.eu also offers an argumentative discourse and collaborative knowledge building component, which may capture researchers’ tacit knowledge on a particular RO and facilitate the corresponding information analysis and decision making. This component enables the linking of ROs with semantically rich knowledge graphs, which maintain chains of views and arguments accompanied by the supporting data, and reflect the users’ collective intelligence. Through interactive visualization features, the overall positive or negative sentiment towards the value of a RO can be also deducted. Finally, OpenBio.eu offers a REST API, a citation reference manager and a Question and Answer (Q&A) component.

Overall, OpenBio.eu is an extrovert and inclusive Open Science Environment. It aims to become a hub for researchers that want to maximize the visibility and reproducibility of their research.

Availability: https://www.openbio.eu
Documentation: https://github.com/kantale/OpenBioC/tree/master/Documentation
Source code: https://github.com/kantale/OpenBioC BSD-3 Clause Licence