





## OPINION ARTICLE

# Creating cloud platforms for supporting FAIR data management in biomedical research projects. [version 1; peer review: 1 approved, 1 approved with reservations]

Marcel Jentsch, Valentin Schneider-Lunitz, Ulrike Taron, Martin Braun, Naveed Ishaque , Harald Wagener, Christian Conrad, Sven Twardziok 

Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Center of Digital Health, Berlin, 10117, Germany

**V1** First published: 03 Jan 2024, 13:8  
<https://doi.org/10.12688/f1000research.140624.1>  
 Second version: 22 Mar 2024, 13:8  
<https://doi.org/10.12688/f1000research.140624.2>  
 Latest published: 29 Apr 2024, 13:8  
<https://doi.org/10.12688/f1000research.140624.3>

## Abstract

Biomedical research projects are becoming increasingly complex and require technological solutions that support all phases of the data lifecycle and application of the FAIR principles. At the Berlin Institute of Health (BIH), we have developed and established a flexible and cost-effective approach to building customized cloud platforms for supporting research projects. The approach is based on a microservice architecture and on the management of a portfolio of supported services. On this basis, we created and maintained cloud platforms for several international research projects. In this article, we present our approach and argue that building customized cloud platforms can offer multiple advantages over using multi-project platforms. Our approach is transferable to other research environments and can be easily adapted by other projects and other service providers.

## Keywords

Cloud, data management, data science, bioinformatics, platforms, FAIR



This article is included in the [Research on Research, Policy & Culture gateway](#).

## Open Peer Review

Approval Status ? ✓

	1	2
<b>version 3</b> (revision) 29 Apr 2024		
<b>version 2</b> (revision) 22 Mar 2024	? view ↑	✓ view ↑
<b>version 1</b> 03 Jan 2024	? view	✓ view

1. **Anna Bernasconi** , Politecnico di Milano, Milan, Italy
2. **Joseph Bonello** , University of Malta, Msida, Malta

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the **Bioinformatics** gateway.

**Corresponding author:** Sven Twardziok ([sven.twardziok@bih-charite.de](mailto:sven.twardziok@bih-charite.de))

**Author roles:** **Jentsch M:** Writing – Review & Editing; **Schneider-Lunitz V:** Writing – Review & Editing; **Taron U:** Visualization; **Braun M:** Methodology, Resources; **Ishaque N:** Project Administration, Writing – Review & Editing; **Wagener H:** Resources, Supervision, Writing – Review & Editing; **Conrad C:** Project Administration, Supervision, Writing – Review & Editing; **Twardziok S:** Conceptualization, Methodology, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The authors thankfully acknowledge the computer resources and the technical support provided by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B). This study was supported by the European Commission with the projects EOSC-life (no. 824087, Horizon 2020), EASI Genomics (no. 824110, Horizon 2020), ESPACE (no. 874710, Horizon 2020) and environMENTAL (no. 101057429, HORIZON-HLTH-2021-STAYHLTH-01).

**Copyright:** © 2024 Jentsch M *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Jentsch M, Schneider-Lunitz V, Taron U *et al.* **Creating cloud platforms for supporting FAIR data management in biomedical research projects. [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2024, 13:8 <https://doi.org/10.12688/f1000research.140624.1>

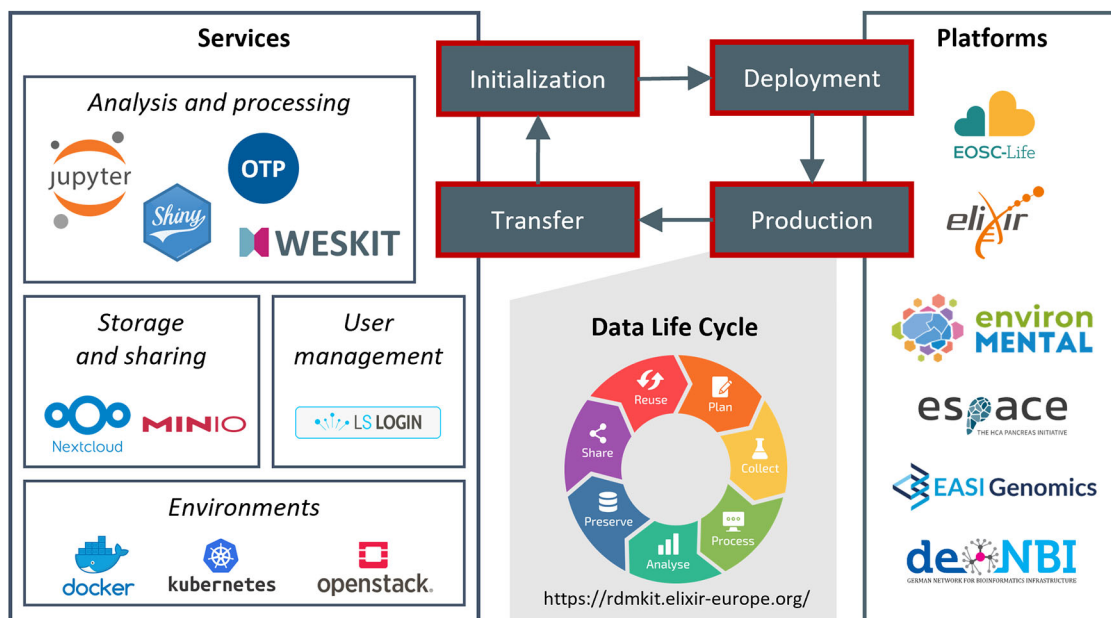
**First published:** 03 Jan 2024, 13:8 <https://doi.org/10.12688/f1000research.140624.1>

## Introduction

Utilizing cloud platforms to link researchers, combine data, and manage data throughout the entire data life cycle is advantageous for biomedical research initiatives. It has become common practice to employ big data analytics to stratify and identify biomarkers in life science research ranging from biodiversity, development, and diseases such as cancer, mental health, or rare diseases. The information pertaining to biological samples from an individual donor has become incredibly complex and may comprise various scientific domains, types, and levels of biological organization.<sup>1</sup> Examples include data from neuroimaging, genomics, proteomics, or even wearables and electronic health records. The different modalities of data can represent different aspects and levels of a complex biological system. Harmonization and integration of these diverse types of data within large research projects is a major challenge.<sup>1</sup> Managing the flow of large datasets within a project and between different processing and analysis steps creates many technical and regulatory challenges when data is distributed across institutes, states, and countries.

The data management requirements of research projects can be represented by the different phases of the data life cycle. The research community has access to numerous definitions of the data life cycle. Here, we adhere to the ELIXIR RDMkit's<sup>2</sup> description, which classifies the data life cycle in research projects into the following seven phases (Figure 1). Each phase has different requirements on technical solutions:

1. **Plan:** The planning phase involves the formalization of a data management plan. This very important and often mandatory document enables projects to organize data flows and processes within the project. Especially for larger projects, this phase requires the planning of data infrastructures that support data management during all the following phases of the data life cycle.
2. **Collect:** The collection phase involves performing multiple experiments and generation of data by e.g., application of instruments and running measurements. This phase requires technical solutions that connect the many different involved sites and move data to further processing facilities.
3. **Process:** Processing data sets typically requires the linking of several tools into a workflow that can transform input data into processed files.<sup>3</sup> An example is the alignment of Next Generation Sequencing (NGS) reads to a reference sequence followed by the detection of genomic variants. Popular workflow systems such as Nextflow<sup>4</sup>



**Figure 1. Platform life cycle.** The schema describes the four phases of our platform development and project support. Starting with project initialization and selection of relevant services, we install a customized project platform in the deployment phase. In the production phase of the platform, projects use the services for data management activities covering the whole data life cycle. Finally, in the transfer phase, knowledge and experiences feed back into our service portfolio.

and Snakemake<sup>5</sup> support and structure the creation of workflows. Depending on their complexity, research projects may then require environments that manage and monitor many workflow executions.

4. **Analyze:** The analysis phase involves the use of applications, scripts, and notebooks to analyze data sets. Examples include applying machine learning methods, calculating statistics, summarizing data, or creating graphs. This task is usually performed interactively by an interdisciplinary, cross-institutional data science team. This requires an environment where different people can collaboratively access data as well as develop, share, and apply analysis scripts and notebooks.
5. **Preserve:** Data preservation is a critical process aimed at guaranteeing the long-term safety, integrity, and accessibility of data, spanning several decades if required. This covers a range of strategies and procedures to mitigate data loss risk, corruption, or obsolescence over time.
6. **Share:** Biomedical research projects often involve many partners from different institutes and countries. While accessing institutional computer centers is regularly restricted for external scientists, projects require solutions for sharing data between scientists. This also requires federated access functionalities.
7. **Re-use:** Providing computational access to data facilitates integration of the project data with external data sets from the external communities and reanalysis in new approaches.

Cloud technologies offer solutions for research data management by providing scalable resources for data processing, enabling controlled access to data through federated user management, and facilitating collaborative and interactive data analyses through shared workspaces.<sup>6</sup> There are different approaches to implement cloud platforms for biomedical research projects and a basic distinction is made between cloud types and service models.<sup>7</sup> Centralized platforms based on the System as a Service (SaaS) model provide data processing or data management functionalities for many users and multiple research projects in parallel. Hereby the platforms offer many established tools for the end-users, which support data management and can be used to answer project-specific questions. Important examples of multi-user platforms are Galaxy-Europe<sup>8</sup> and Anvil.<sup>9</sup> The creation of customized platforms that are subsequently run on a cloud Infrastructure as a Service (IaaS), or Platform as a Service (PaaS) architecture is a counter design to the centralized platforms.

Customized platforms that adhere to a modular microservice design can strengthen the FAIR principles and improve the interoperability of the research data by providing Application Programming Interfaces (APIs) for access to data slices and data summaries.<sup>10</sup> The FAIR principles are a set of guidelines for making data Findable, Accessible, Interoperable, and Reusable.<sup>11</sup> These principles aim to improve the ability of researchers and other stakeholders to locate, understand, and (re-) use data, which is of importance within large projects as well as for sharing data with the scientific community. Although it has many advantages, creating customized cloud platforms is a complex task and requires expert knowledge, experienced technical staff, and guidelines.

Here, we present our approach for creating customized cloud platforms for biomedical research projects and argue that such platforms offer many advantages for managing and processing data over the whole data life cycle and according to the FAIR principles. With similar use cases in mind, the framework is easily adaptable to various research initiatives and infrastructures. Our methodology offers a flexible framework for the management of knowledge and experiences while enabling the effective creation of cloud platforms.

## Approach

Our approach is based on managing a portfolio of supported services, which we have established within our cloud infrastructure. Our cloud is operated as a part of the German de. NBI network,<sup>12</sup> but this approach applies as well to other commercial or public providers. For our portfolio services, we document best practices, deployment scripts and default settings internally using our instance of GitLab. Using git enables collaborative file editing, allows versioning, and supports integration of deployment pipelines. We also provide general guidelines and best practices for the research community in the de. NBI Cloud Wiki (<https://cloud.denbi.de/wiki/>). For developing our platforms, we follow a microservices architecture and manage our services as container images, which supports deploying these services in container environments. Our cloud environment mainly supports the container environments Docker and Kubernetes, which we apply depending on the requirements of the respective services.

During project initialization (Figure 1), we consult projects, analyze which services are relevant for the respective project and design a platform framework. The consultation is generally performed in close coordination with the project management and scientists. The project-specific services are then installed on a cloud infrastructure in the deployment

phase. We then maintain the project platforms throughout the whole production phase to support the data life cycle in the projects. Finally, knowledge and experiences of a project are transferred back into the portfolio and are then available for any other new project initialization. Hereby, we update deployment scripts and best practices as well as add new services, versions, and functionalities to our documentation. Applying this approach results in an evolving set of parallel running platforms over time, where any new platform profits from experiences and knowledge from all previous projects.

## Services

Our service portfolio consists of a set of core services (Figure 1), which are regularly required by our collaboration partners. The selection of these services is based on our experiences and reflects the requirements of our research bubble. Anyhow, the general concepts can be easily transferred to other tools and services. Here, we describe how these services contribute to supporting the data life cycle and to the application of the FAIR principles. We generally recommend our projects to validate the application of FAIR criteria through self-assessment, as e.g., applied by de. NBI network.<sup>13</sup>

### *OpenStack Manila*

In most projects, we provide a shared network files system in our OpenStack cloud platform using the Manila service such as used in other OpenStack implementations.<sup>14</sup> All project-relevant data can be stored, managed, and shared in such a central file system. The other services can thus access the same database. This allows files to be shared with the scientists, who can access the data via the respective other services. The central data structure also ensures that the data can be clearly identified via paths within the project, which facilitates communication, increases the reproducibility of the results, and ensures the findability of the data.

### *LS Login*

Life Science Login (LS Login) is a service which is developed by EOSC-life including the ELIXIR infrastructure to provide a unified user management and authorization infrastructure for European researchers.<sup>15</sup> It enables researchers to use their home organization credentials, community or other identities (e.g. Google, LinkedIn) to sign in and access data and services they are authorized to access. Thereby, it supports the OIDC standard which can easily be integrated by many other tools such as a user login system. By using LS Login users do not have to maintain multiple different user accounts for logging into the cloud platform services, which support the accessibility of data. By implementing LS Login, service providers don't need to maintain their own user management system.

### *JupyterHub*

JupyterHub is a multi-user platform, which allows for running Jupyter Notebook in a cloud environment.<sup>16</sup> Data scientists can use JupyterHub for interactive data analysis, and for sharing analysis scripts and results. Within the portal, the users have access to private home storage and optionally to shared cloud data, including the results of experiments, the shared user storage space, and a common code base. Optionally, JupyterHub can be configured to provide different compute resource profiles, such as high-memory machines for extensive data analysis and processing steps. Jupyter supports specifically the explorative analysis of research data.

### *R-Shiny server*

R-Shiny is a framework that allows for rapid and intuitive development of interactive data visualization applications. We support serving of R-Shiny apps using the open-source Shiny Server.<sup>17</sup> R-Shiny apps can be used for interactive data analysis by members of a project as well as by external users. To ensure quality of services for multiple users, parallel processing over multiple servers as containers using Docker or Kubernetes is possible. A specific example of an R-Shiny app is the iSEE application,<sup>18</sup> which provides interactive analysis of single-cell data. The iSEE app can e.g., support annotation of cell types in single-cell data for members of a consortium by displaying gene expression in selected cell clusters. Providing interactive access to data supports data analysis and increases the accessibility of data for people without programming knowledge.

### *MinIO*

Findability requires registration of metadata in public data registries and assigning unique identifiers to datasets. Using an S3 (Simple Storage Service) object storage, all files are structured into buckets and further grouped into folders. Single files or groups of files can then be identified via unique URIs. We usually set up an S3 endpoint via the open-source software MinIO.<sup>19</sup> The identifiers are globally unique since the domain name of the website is included. The unique

identifier can then also be used to reference data in publications and in public data registries. An S3 server also provides an Application Programming Interface (API) access to the data, which allows access to all data that is to be published or shared with external users. The access to the S3 server can be public or protected by authentication.

#### *WESkit*

WESkit<sup>20</sup> is a workflow execution service implementing the GA4GH WES API.<sup>21</sup> It was developed to manage the execution of bioinformatics workflows at the German Cancer Research Center (DKFZ) and Charité Universitätsmedizin Berlin, but it can also be used in a cloud framework as well. The WESkit software provides features such as workflow monitoring, logging, and provenance tracking. It directly supports the processing of data. Hereby WESkit provides an interoperable environment for the execution of Snakemake<sup>5</sup> and Nextflow<sup>4</sup> workflows. The documentation of workflow executions supports the reusability of resulting data.

#### *OTP*

The platform “One Touch Pipeline” (OTP) was initially developed by the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ)<sup>22</sup> as a part of the International Cancer Genome Consortium (ICGC) for management and processing of sensitive cancer genomics data. Since 2020, OTP is used at the DKFZ and Berlin Institute of Health for automatic execution of different workflows to process NGS data coming from whole genome sequencing (WGS), exome sequencing, RNA sequencing (RNA-seq), whole genome bisulfite sequencing (WGBS), and single-cell RNA sequencing (10x scRNA-seq) data. OTP has also been used for management of single-cell genomics data from COVID-19 patients.<sup>23</sup>

#### *Nextcloud*

Nextcloud is a suite of client-server software for creating and using file hosting services. It is enterprise-ready with comprehensive support options. Being free and open-source software, anyone is allowed to install and operate it on their own private server devices.<sup>21</sup> It combines the comfort and user-friendliness of cloud solutions such as Dropbox or Google Drive with the requirements for security, data protection and control. The service runs on our infrastructure so that data does not need to leave the house. This allows users to share, for example, intermediate results and analysis scripts as part of data analysis. Access is granted to all members of a project and each user receives an initial 50 GB of storage.

### Implementation in projects

Multiple projects implemented the presented approach for creating customized cloud platforms to support the daily data life cycle.

#### *ESPACE*

The ESPACE project merged three prior Human Cell Atlas (HCA) early pilot studies to build a first version of the Human Cell Atlas of the Pancreas. The HCA project is an international effort to map all the cells in the human body and understand their functions.<sup>24</sup> The ESPACE cloud platform consists of shared storage, a JupyterHub portal, an R-Shiny Server for providing interactive applications and MinIO as an S3 backend for providing computational access to the data. The JupyterHub portal (<https://espace-cloud.bihealth.org>) is actively used by more than ten data scientists throughout Europe for data processing and interactive data analysis. It is planned that the ESPACE data will soon be available to the research community via the ESPACE cloud platform.

#### *environMENTAL*

The environMENTAL project investigates how some of the greatest global environmental challenges, climate change, urbanization, and psychosocial stress affect mental health across the lifespan. The project aims to identify underlying molecular mechanisms and develop preventions and early interventions. Cohort data of over 1.5 million EU citizens and patients, enriched with deep phenotyping data from large-scale behavioral neuroimaging cohorts, are used to identify brain mechanisms related to environmental adversity underlying symptoms of depression, anxiety, stress, and substance abuse. The Berlin Institute of Health (BIH) (<https://www.bihealth.org/>) provides a cloud platform to support data management in the environMENTAL project and all phases of the data life circle. The platform comprises a Nextcloud instance as a central storage for respective environmental data sets, a JupyterHub Portal as an interactive workspace and WESkit for running data processing on the integrated data sets.

### *EASI-Genomics*

The EASI-Genomics consortium aims to provide translational access to cutting-edge sequencing technologies and data analysis methodologies to researchers, adhering to ethical and legal requirements, as well as FAIR and secure data management. A JupyterHub-based cloud platform was utilized to provide users with translational access to example datasets and Jupyter notebooks to provide guidelines for data analysis for single-cell, multi-omics and spatial transcriptomics data in both R and Python programming languages (<https://easi-genomics-cloud.bihealth.org>).

### *SpaceHack*

The JupyterHub-based cloud platform facilitated the SpaceHack project at the ELIXIR Germany BioHackathon 2022. The project's objective was to generate benchmarking data sets by leveraging the expertise of researchers with both biological and technical backgrounds to evaluate the performance of segmentation and cell assignment tools in the context of tissue-specific challenges.<sup>25</sup> Over a week, the platform was utilized by more than 60 users, enabling virtual and onsite participants to collaborate effectively.

### *OTP2EOSC*

The OTP2EOSC project developed a cloud-ready data management and processing platform that sends analysis workflows to the data, avoiding transferring data between sites. This increases the efficiency of data management and data security. The project deployed a demonstration cloud platform (<https://otp-demo.bihealth.org>) for processing sensitive cancer-genomics data based on OTP, WESKit, and relevant ICGC cancer genomics workflows. The platform is available for interested users and to test the functionalities of the OTP software.

## **Discussion**

Collection, processing, analysis, storage, and access of biomedical research data require platforms for managing research data.<sup>26</sup> For this end, the use of a cloud platform can greatly support biomedical projects. There are already many existing platforms available and important examples of central multi-project platforms include Galaxy-Europe,<sup>8</sup> the AnVIL project,<sup>9</sup> the Human Cell Atlas data coordination platform<sup>24</sup> and the Cancer Genomics Cloud.<sup>27</sup> Such platforms can be used for specific use cases like the execution of Galaxy workflows or to work on specific data types like human cells or cancer genomics data. There are also several commercial providers including DNA Nexus and Seven Bridges, which offer the development and deployment of centralized multi-project platforms. Central platforms are generally operated via the SaaS model and the costs for infrastructure and resources are usually returned to the projects.<sup>7</sup> Alternatively, biomedical projects can also create their own customized platforms and run them using an IaaS or a PaaS approach.

From a data management perspective, there are several advantages and disadvantages to using a central multi-project cloud platform. These platforms come with a lot of functionality already built in and provide unified environments to support all stages of the data lifecycle, making it easier to define processes and train scientists across projects. However, biomedical scientists are typically involved in several different projects at the same time, and it can be challenging for them to work in many different cloud environments.<sup>10</sup> By using external central platforms, projects can worry less about technical infrastructure and have lower technical requirements. On the other hand, it can be a challenge to obtain funding for the use of external platforms<sup>7</sup> and integration into the local data access regulations and processes needs to be solved. Existing platforms do not necessarily provide all the desired functionality for specific projects and customization may be required. For multi-project platforms, there is a risk that changes and updates to individual components become more difficult when the number of ongoing projects increases, making it less flexible to respond to the needs of individual projects. Monolithic platforms also tend to suffer from lock-in effects and data becoming less interoperable.<sup>10</sup>

The operation of a project-specific cloud platform based on an IaaS or PaaS model can offer several advantages over using a centralized multi-project framework. A customized cloud platform can save funding while operating on the existing resources of the institutes data centers, has greater flexibility and adaptability, and has a limited timeframe for long-term maintenance. Furthermore, long-term archiving may be easier to achieve. A local platform can be adapted to local data protection regulations such as those required by EU law with the General Data Protection Regulation (GDPR). On the technical side, copying large sets of data to a central platform can also be an issue. The features of a project-specific platform can be easily adapted to the specific needs of the users and the individual project goals. By developing self-controlled project-specific platforms, data interoperability and interconnection of platforms can be achieved more easily.

## **Conclusions**

A customized cloud platform can improve data sharing within a consortium, support data analysis and provide access to data over the whole data life cycle according to the FAIR criteria. This is particularly useful in an environment where

many different partners within a consortium are analyzing a shared database. A common data structure also supports the sharing of analysis scripts and tools, making results reproducible within a project. In addition, data can be shared with external parties via a cloud platform as part of a review process, or more generally for reusing data by other researchers in new analyses. Hereby, our approach offers a structured and efficient way for developing platforms and serving multiple research projects.

The approach presented here of managing a portfolio of supported services has proven its worth in various projects at the BIH. A major benefit is the ease and speed with which new projects can be launched and supported. Due to management of know-how, existing platforms can be easily transferred to new projects within a few working days. By using existing software solutions and a microservice architecture, the individual components of the platform can be set up in a short time by a skilled cloud engineer. The microservices architecture makes the platform flexible and adaptable to individual user needs. In addition, the portfolio can be constantly updated during the transferring of services to new projects. However, this approach only makes sense if the potential projects have recurring software and service requirements. It is also a challenge to manage, update, extend and clean up the service portfolio on an ongoing basis.

Our supported projects benefited greatly from the provision of the de. NBI cloud for research in Germany and its integration into the European ELIXIR network. The de. NBI cloud is a federated cloud framework operated by the de. NBI network in Germany and is available for academic projects in Germany.<sup>12</sup> The de. NBI cloud is involved in the European Open Science Cloud (EOSC) project via the ELIXIR network with other European partners. Through EOSC, academic partners across Europe can also access cloud resources. The federated structure of the de. NBI cloud supports making platforms available where the data is stored. This has technical advantages with data transfer as well as legal advantages, e.g., if data is not allowed to leave an organization or to cross regional boundaries. It also has the advantage that local contact people are available on site. In addition, central services such as the central de. NBI project management system, established processes, and best practices as well as the LS Login can be used and accessed. The local availability of cloud technologies is of significant value for our project partners.

### Data availability

No data are associated with this article.

### Acknowledgements

The authors would like to thank all members of our collaborating projects and consortia including EOSC-life, EASI Genomics, ESPACE and environMENTAL for useful discussions, feedback, and insight into their specific use cases and requirements. We acknowledge financial support from the Open Access Publication Fund of Charité – Universitätsmedizin Berlin and the German Research Foundation (DFG).

### References

- Cirillo D, Valencia A: **Big data analytics for personalized medicine.** *Curr. Opin. Biotechnol.* 2019; **58**: 161–167. [Publisher Full Text](#)
- ELIXIR: **Research Data Management Kit. A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075).** 2021. [Reference Source](#)
- Perkel JM: **Workflow systems turn raw data into scientific knowledge.** *Nature.* 2019; **573**(7772): 149–150. [PubMed Abstract](#) | [Publisher Full Text](#)
- Di Tommaso P, et al.: **Nextflow enables reproducible computational workflows.** *Nat. Biotechnol.* 2017; **35**(4): 316–319. [Publisher Full Text](#)
- Mölder F, et al.: **Sustainable data analysis with Snakemake.** *F1000Res.* 2021; **10**: 33. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Langmead B, Nellore A: **Cloud computing for genomic data analysis and collaboration.** *Nat. Rev. Genet.* 2018; **19**(5): 325. [PubMed Abstract](#) | [Publisher Full Text](#)
- Navale V, Bourne PE: **Cloud computing applications for biomedical science: A perspective.** *PLoS Comput. Biol.* 2018; **14**(6): e1006144. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Community, T.G: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update.** *Nucleic Acids Res.* 2022; **50**(W1): W345–W351. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Schatz MC, et al.: **Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space.** *Cell Genom.* 2022; **2**(1).
- Sheffield NC, et al.: **From biomedical cloud platforms to microservices: next steps in FAIR data and analysis.** *Sci Data.* 2022; **9**(1): 553. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wilkinson MD, et al.: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Belmann P, et al.: **NBI Cloud federation through ELIXIR AAI.** *F1000Res.* 2019; **8**: 842. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mayer G, et al.: **Implementing FAIR data management within the German Network for Bioinformatics Infrastructure (de. NBI) exemplified by selected use cases.** *Brief Bioinform.* 2021; **22**(5). [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Castro León J: **Advanced features of the CERN OpenStack Cloud.** *EPJ Web Conf.* 2019; **214**: 07026. [Publisher Full Text](#)
- Harrow J, et al.: **ELIXIR: providing a sustainable infrastructure for life science data at European scale.** *Bioinformatics.* 2021; **37**(16):



- 2506–2511.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. JupyterHub: **Zero to JupyterHub with Kubernetes**. 2023.  
[Reference Source](#)
  17. Shiny Server.  
[Reference Source](#)
  18. Rue-Albrecht K, *et al.*: **iSEE: Interactive SummarizedExperiment Explorer**. *F1000Res*. 2018; **7**: 741.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  19. MinIO: 2022.  
[Reference Source](#)
  20. Kensch PR, *et al.*: **Executing workflows in the cloud with WESkit**. *BioHackXiv*. 2023. February 20.
  21. Rehm HL, *et al.*: **GA4GH: International policies and standards for data sharing across genomic research and healthcare**. *Cell Genom*. 2021; **1**(2).
  22. Reisinger E, *et al.*: **OTP: An automatized system for managing and processing NGS data**. *J Biotechnol*. 2017; **261**: 53–62.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  23. Trump S, *et al.*: **Hypertension delays viral clearance and exacerbates airway hyperinflammation in patients with COVID-19**. *Nat Biotechnol*. 2021; **39**(6): 705–716.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  24. Regev A, *et al.*: **Science forum: the human cell atlas**. *elife*. 2017; **6**: e27041.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  25. Ishaque N, Long B, Kuemmerle L: **SpatialHackathon**. 2022.  
[Reference Source](#)
  26. Navale V, von Kaeppler D, McAuliffe M: **An overview of biomedical platforms for managing research data**. *J Data Inf Manag*. 2021; **3**(1): 21–27.  
[Publisher Full Text](#)
  27. Lau JW, *et al.*: **The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research**. *Cancer Res*. 2017; **77**(21): e3–e6.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

# Open Peer Review

Current Peer Review Status: ? ✓

---

## Version 1

Reviewer Report 06 March 2024

<https://doi.org/10.5256/f1000research.153998.r242776>

© 2024 Bonello J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Joseph Bonello** 

University of Malta, Msida, Malta

### Summary:

The project is about using a local microservice-based cloud architecture as a cost-effective solution over a per-project platform to support FAIR data management. The authors highlight the tools they have available in their environment and how projects can leverage this technology with the benefit of having project-specific expertise, built over several projects, available to them.

These are my detailed comments for the individual sections of the article.

### Abstract:

The abstract provides a summary of the project and highlights what the article is about sufficiently well.

### Introduction:

The introduction describes the use of cloud platforms in biomedical research. In particular, reference is made to the need for the analysis of large datasets that need to adhere to strict legal and ethical frameworks. The authors adhere to the ELIXIR RDMkit definition of the data lifecycle which they summarise for the benefit of the overall narrative. The authors summarise well the problem associated with researchers' use of data and provide a good overview of the FAIR principles.

### Approach:

The authors describe the approach as being based on Docker and Kubernetes and how they provision services based on a core set of services that facilitate the research process based on the existing needs of their users.

The authors then proceed to list and describe the list of services they make available on the platform they maintain. To complement this list, they provide several examples of projects where the technologies have been implemented.

*Comments:*

1. The authors mention the validation of adherence to FAIR principles through a self-assessment. It would be interesting if this self-assessment is defined better, to highlight the synergy between the provided services and adherence to the FAIR principles.
2. The list of tools and services provided on the platform is interesting. However, how these tools relate to the FAIR principles, and to which principles they relate, is not immediately clear. I suggest that the authors explicitly link the tool to the relevant FAIR principle for added clarity.

**Discussion:**

In the discussion section, the authors compare the proposed solution to multi-project platform, comparing the pros and cons of the two different solutions.

*Comments:*

1. In terms of costs, comparing the costs between the multi-project solution and using an in-house solution needs to consider not only the cost of purchasing and maintaining the hardware (including the salary costs, power, etc). It is important to acknowledge these costs, although I am aware that they may be difficult to quantify and compare.
2. I think that a short mention of the performance differences offered by the two solutions merits a mention. This is in view of the fact that in order to achieve high performance locally, you must acquire the hardware to match the needs (e.g. GPUs). This can be expensive and require expertise to maintain.

**Conclusion:**

In the concluding remarks, the authors highlight the benefits of having a local data structure that supports sharing resources within the project. They highlight the benefits of the proposed solution and provide examples where the solution has been used.

**General suggestions for improvement:**

- Capitalize machine learning (Machine Learning) and data science (Data Science).
- Provide a figure where each of the tools that support the data lifecycle is organized under the relevant heading of the RDMKit phases.
- Provide more details on the security aspects that a local solution offers, how it compares to multi-project solutions and how it helps researchers align with legal requirements.
- Show more clearly (with examples) how a local solution would encourage reproducibility of scientific study and how, in reality, can project groups share their data and research within a FAIR environment.

**Is the topic of the opinion article discussed accurately in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Are arguments sufficiently supported by evidence from the published literature?**

Yes

**Are the conclusions drawn balanced and justified on the basis of the presented arguments?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics; FAIR; Data Science

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 15 Mar 2024

**Sven Twardziok**

Dear reviewer,

Thank you very much for your constructive comments and suggestions. Our responses to your suggestions are in detail below and changes are included in the new version.

- The authors mention the validation of adherence to FAIR principles through a self-assessment. It would be interesting if this self-assessment is defined better, to highlight the synergy between the provided services and adherence to the FAIR principles.

We added an additional explanation of the term “self-assessment”. The reader will find more information in the cited publication.

- The list of tools and services provided on the platform is interesting. However, how these tools relate to the FAIR principles, and to which principles they relate, is not immediately clear. I suggest that the authors explicitly link the tool to the relevant FAIR principle for added clarity.

We modified the figure to make a connection between the phases of the data life cycle, the FAIR principles and the services.

- In terms of costs, comparing the costs between the multi-project solution and using an in-house solution needs to consider not only the cost of purchasing and maintaining the hardware (including the salary costs, power, etc). It is important to acknowledge these costs, although I am aware that they may be difficult to quantify and compare.

We added a sentence to make the reader aware of acquisition and operating costs; anyhow, our approach is not focused on running on institutes data centers, but an advantage is, that such existing hardware can be used.

- I think that a short mention of the performance differences offered by the two

solutions merits a mention. This is in view of the fact that in order to achieve high performance locally, you must acquire the hardware to match the needs (e.g. GPUs). This can be expensive and require expertise to maintain.

The presented approach is not limited to the use of own hardware and therefore the purchase of own hardware is not a requirement.

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 22 February 2024

<https://doi.org/10.5256/f1000research.153998.r242781>

© 2024 Bernasconi A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Anna Bernasconi**

Politecnico di Milano, Milan, Lombardy, Italy

The paper describes the choice of adopting customized cloud platforms over centralized multi-project platforms in the context of biomedical research. The authors explain that, so far, the main trend has preferred the first kind of infrastructure instead of the second, bringing several limitations.

I do believe that the author's experience is of great value and is worth being shared with a broader community. Therefore, I make some suggestions that should lead to improving the presentation of the work in the current manuscript.

Figure 1 should drive the reader along the whole paper; however, it is not well-used. E.g. seven phases in the data life cycle are not connected to the other parts of the figure. The four project phases are not described in the paper. Services have been partitioned into four areas in the figures but this organization is not reflected in the paper (Services section, pages 5-6).

Several parts of the Introduction and Discussion attempt to draw a line between centralized platforms based on SaaS (sometimes called central multi-project cloud platforms) and "customized cloud platforms" (based on IaaS/PaaS) such as the one proposed by the authors (see Page 4 and 7). This distinction is far from being well-explained and pragmatically justified. I urge the authors to better describe this, as it is fundamental for understanding their contribution.

The paper is focused on a specific experience, the one of the Berlin Institute of Health. The experience would be plausible in many other research institutes; however, the paper reads more like an experience paper/institute report rather than a research paper. This is even more evident when the author themselves speak about their research bubble ("The selection of these services is based on our experiences and reflects the requirements of our research bubble."). To overcome this, a more structured discussion on competitor / alternative platforms is needed. In this way, the paper would become understandable/useful also for whom is coming from different "bubbles".

The descriptions of Services sometimes provide details that are not properly justified (e.g., why explain that 'Access is granted to all members of a project and each user receives an initial 50 GB of storage.' if further scalability is not discussed? Is 50GB a limitation in practical scenarios?)

MINOR:

- Page 5: explain OIDC acronym
- don't --> do not
- The adherence to FAIR principles is evaluated by "self-assessment". Could this be elaborated further?
- Page 6, WESkit paragraph, remove 'as well'
- Bibliography, please provide URLs for entries 2, 16, 17, 19, 25

**Is the topic of the opinion article discussed accurately in the context of the current literature?**

Partly

**Are all factual statements correct and adequately supported by citations?**

Yes

**Are arguments sufficiently supported by evidence from the published literature?**

Yes

**Are the conclusions drawn balanced and justified on the basis of the presented arguments?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Data integration, Genomic data, Knowledge management, Bioinformatics, FAIR principles

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 15 Mar 2024

**Sven Twardziok**

Dear reviewer,

Thank you very much for your constructive comments and suggestions. Our responses to your suggestions are in detail below and changes are included in the new version.

- o Figure 1 should drive the reader along the whole paper; however, it is not well-used. E.g. seven phases in the data life cycle are not connected to the other parts of the figure. The four project phases are not described in the paper. Services have been partitioned into four areas in the figures but this organization is not reflected in the paper (Services section, pages 5-6).

Thank you very much for this comment. We adapted the description of the four project phases the text. We also We modified the figure to make a connection between the phases of the data life cycle, the FAIR principles and the services.

- Several parts of the Introduction and Discussion attempt to draw a line between centralized platforms based on SaaS (sometimes called central multi-project cloud platforms) and "customized cloud platforms" (based on IaaS/PaaS) such as the one proposed by the authors (see Page 4 and 7). This distinction is far from being well-explained and pragmatically justified. I urge the authors to better describe this, as it is fundamental for understanding their contribution.

We agree that the distinction between the terms "centralised platforms" and "customised platforms" was unclear and have removed these distinctions accordingly throughout the manuscript. The manuscript now focuses more on the distinction between IaaS/PaaS and IaaS at the project level, as these are common terms and projects ultimately have to choose between the models when building a cloud platform.

- The paper is focused on a specific experience, the one of the Berlin Institute of Health. The experience would be plausible in many other research institutes; however, the paper reads more like an experience paper/institute report rather than a research paper. This is even more evident when the author themselves speak about their research bubble ("The selection of these services is based on our experiences and reflects the requirements of our research bubble."). To overcome this, a more structured discussion on competitor / alternative platforms is needed. In this way, the paper would become understandable/useful also for whom is coming from different "bubbles".

We agree that this work reflects our own view. However, the framework of this work is an opinion paper and we believe that our experiences and the general concept in the present form will be useful for other readers as well.

- The descriptions of Services sometimes provide details that are not properly justified (e.g., why explain that 'Access is granted to all members of a project and each user receives an initial 50 GB of storage.' if further scalability is not discussed? Is 50GB a limitation in practical scenarios?)

Thanks for the hint. We have revised the description of the services and removed unnecessary information.

MINOR:

- Page 5: explain OIDC acronym

We added an explanation of the OIDC acronym.

- don't --> do not

done

- The adherence to FAIR principles is evaluated by "self-assessment". Could this be elaborated further?

We added an additional explanation of the term "self-assessment". The reader will find more information in the cited publication.

- Page 6, WESkit paragraph, remove 'as well'

done

- Bibliography, please provide URLs for entries 2, 16, 17, 19, 25

Urls are provided as links on the article website as well as in the pdf documents

**Competing Interests:** No competing interests were disclosed.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**