1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.400

Published: 07 Sept. 2023

ICT Infrastructure supporting the Italian Research Infrastructure on Microbial Resources MIRRI-IT

Marco Beccuti1[https://orcid.org/0000-0001-6125-9460], Antonio d'Acierno2[https://orcid.org/0000-0003-0516-0794], Simone Donetti1[https://orcid.org/0000-0001-7204-4078], Sandro Gepiro Contaldo1[https://orcid.org/0009-0007-3889-8644], Paolo Romano3[https://orcid.org/0000-0003-4694-3883], and Giovanna Cristina Varese1[https://orcid.org/0000-0002-1455-6208]

1 University of Turin, IT

2 ISA - CNR, Italy

3 IRCCS Ospedale Policlinico San Martino, Genoa, Italy

Keywords: Research Infrastructure, Microbial resources.

Extended Abstract

In 2022, the SUS-MIRRI.IT project (Strengthening the MIRRI Italian Research Infrastructure for Sustainable Bioscience and Bioeconomy, www.sus-mirri.it) was funded with ~17M€ by the Italian government on the NextGeneration EU-funded Recovery and Resilience National Plan (PNRR) – Research Infrastructure - to strengthen the Italian Research Infrastructure for Microbial Resources (MIRRI-IT) and ensure its long-term sustainability.

The main objectives of this project are:

(a) to implement MIRRI-IT's organisation and set up its operative procedures and quality standards;
(b) to improve quality of Italian microbial Biological Resource Centers databases and conceive MIRRI-IT's services based on the partners' expertise and genetic resources;
(c) to set up a single entry point platform to promote MIRRI-IT's resources in terms of data, services, cutting-edge technologies and expertise.

In this contest, the new needs for data integration and system interoperability, stressed by the FAIR approach to data sharing, have become evident and urgent thus leading to the identification of four ICT macro-activities which will be carried out in the project.

(1) The activity related to the Italian Collaborative Working Environment (ItCWE) platform foresees its implementation based on WordPress [1], a very popular content management system. It will provide four main functionalities, namely access to (i) data and services, (ii) expert consulting, (iii) TransNational Access (TNA) program, and (iv) training. The WordPress multisite instance will be configured taking as a reference the platform recently implemented for the Microbial Resource Research Infrastructure MIRRI.
(2) The activity related to the Italian Culture Collections Catalogue (ItCCC) aims to develop it as the main access point to information on microbial resources available at the national level. As such, it must be as FAIR (Findable, Accessible, Interoperable, Reusable) as possible, while taking into account the existing limitations of microbial information semantics. The ItCCC will consider the data model agreed upon by the ICT

Task Force of MIRRI at the European level and defined in the context of the IS_MIRRI21 EU project (https://ismirri21.mirri.org/) as the reference dataset for the MIRRI Information System (MIRRI-IS), to facilitate the upload of data from Italian collections into the ItCCC and, possibly, the MIRRI-IS. Data will be also made available to researchers through dynamic web pages, while interoperability will be ensured using standard technologies (such as REST APIs, JSON). ItCCC will be served by the Apache HTTP server [2] and based on PostgreSQL [3] and PHP scripts [4].

(3) The activity related to the Microbial Biological Resource Center database (mBRCdb) will lead to the development of a standalone open-source application supporting the local management of their catalogue by single culture collections (CCs). Indeed, many CCs do not currently have an effective data management system for their collection and do not have IT staff able to support its implementation. This application will also guide the curators transforming data into a format compliant with the MIRRI-IS data model by implementing basic data quality controls. To maximize its portability on different operating systems and architectures, Java language was identified as an appropriate candidate for its development.

(4) The Dataverse proof-of-concept (PoC) activity is related to the implementation of a Harvard Dataverse [5] instance to investigate how the semantic annotation of microbial information could be suitably stored by exploiting this open-source tool that is reputed as one of the most effective tools for improving data FAIRness. Its configurable system of roles and nested repositories will be investigated to see if it could allow the enforcement of data curation, preservation, and publishing workflows. Moreover, its ability to integrate with established user identification platforms, such as ORCID, and digital objects, such as DOI and FDO, will be investigated. Its capacity to share and discover data through semantic metadata will also be investigated..

The corresponding computational infrastructure that we implemented for hosting these ICT activities is reported in Fig.1 where the hardware level is highlighted in gray, the virtualization level in dark orange, and the application level corresponding to the macro-services in blue and the micro-services needed for each macro-services ii a light orange.

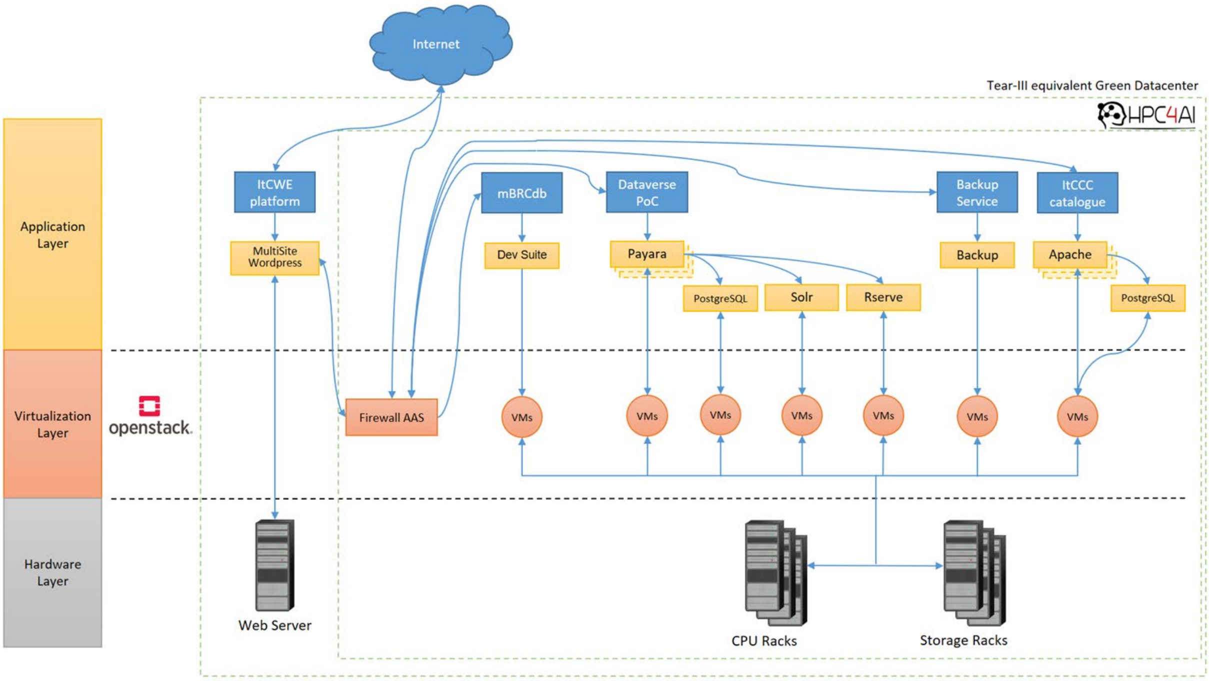

Fig.1: The schema of ICT infrastructure implemented for the SUS-MIRRI.IT project.

The Hardware level was set up on the HPC4AI data centre (https://hpc4ai.unito.it) based on HPC-cloud convergence architecture with 2200+ cores, 100+ GPUs, and four storage classes (i.e. 300TB SSD hyper-converged, 1PB hybrid multi-tenant HW-encrypted scale-out NAS, 1PB cold storage backup, 170TB deduplicated storage backup), 100GB/s SDN networking. It is organized in experimental and production islands hosted into 250KVA Tier-III adiabatic green datacenter with 16 racks.

The virtualization level is based on the open-source OpenStack cloud technology (co - developed with Canonical - https://canonical.com/, the company behind Ubuntu), which is currently the de facto standard for HPC-cloud convergence architecture. Indeed, OpenStack manages data centre resources through a set of modular services which interoperate in a service-oriented architecture. It provides abstractions for computational, storage, and networking resources and allows cloud administrators to provide them easily, dynamically, and securely to different tenants.

The application level implements the project macro-services in terms of their microservices exploiting the OpenStack modularity which allows us to easily add new services or rescale those already created according to new requirements that would arise during the project.

Initially all the VMs hosting the microservices were created with a medium profile of resource assignment (i.e. 4 vCPU, 8GB RAM), while the disk space allocation is managed dynamically according to the amount of data that will have to be stored and managed. All disk space used is provided by the cloud as virtual space, with redundant management of physical disks using different technologies. Depending on the type of disk chosen, the management policy changes and we pass from Ceph-based software systems with Replica 3 or Erasure Code[1] (k=2,m=1) to dedicated hardware appliance solutions with mixed-type disk storage or appliance with hardware deduplication. VMs use disks of all types depending on the needs and tasks to be performed.

Author contributions

All authors contributed equally

Competing interests

The authors declare that they have no competing interests.

Funding

SUS-MIRRI.it PNRR project.

References

1. WordPress Official Site. URL: https://wordpress.com/
2. Apache HTTP Server Project Official Site. URL: https://httpd.apache.org/
3. PostgreSQL Official Site. URL: https://www.postgresql.org/
4. PHP Official Site URL: https://www.php.net/

[1] Erasure coding uses storage capacity more efficiently than replication. The n-replication approach maintains n copies of an object (3x by default in Ceph), whereas erasure coding maintains only k + m chunks. For example, 2 data and 1 coding chunks use 1.5x the storage space of the original object.

5. Gary King, "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing", Sociological Methods & Research, vol. 36(2), pages 173-199, November 2007.