DOI: 10.26314/GARR-Conf23-proceedings-06

FOSSR-Fostering Open Science in Social Science Research: Building Italy's Innovative Open Cloud Infrastructure

Serena Fabrizio[1], Mario Sicuranza[2], Mario Ciampi[2], Emanuela Reale[1]

[1]Research Institute for Sustainable Economic Growth of the National Research Council of Italy, [2]Institute for High Performance Computing and Networking of the National Research Council of Italy

Abstract. FOSSR project is dedicated to creating an Italian Open Science Cloud, inspired by the European Open Science Cloud (EOSC). A central strategy of FOSSR revolves around leveraging Research Infrastructures (RIs) in the social sciences to ensure the availability of data adhering to the FAIR principles. FOSSR aligns with the broader mission of advancing Open Science in Italy, aiming to provide a framework of tools and services to support the social science academic community. This includes the integration of the Italian nodes of three existing RIs: CESSDA-ERIC, SHARE-ERIC, and RISIS, coordinated by the National Research Council of Italy (CNR). The FOSSR project strategically integrates cloud computing infrastructure and a Network of Data Centers, primarily situated in southern Italy which strengthens data processing capabilities. This network will enhance local computing capabilities and support the advanced data analytics needs of social science researchers. The federated platform offers various interfaces, and a Virtual Research Environment (VRE), promoting open, collaborative, and streamlined data access.

Keywords. Italian Open Science Cloud, Research Infrastructures, Cloud Computing, Social Science Research, Innovation

Introduzione

FOSSR - Fostering Open Science in Social Science Research aims to create an Italian Open Science Cloud, inspired by EOSC (European Open Science Cloud), in which to integrate cutting-edge tools and services dedicated to exploring issues pertinent to the evolution of contemporary societies' economic and social landscape. A key strategy to achieve this objective revolves around the exploitation of research infrastructures (RIs) in the social sciences, which facilitate the availability of data that adhere to FAIR (Findable, Accessible, Interoperable, Reusable) principles.

FOSSR embraces the overarching theme of advancing Open Science in Italy with the aim of establishing a framework of tools and services for the social science academic community. This involves the Italian nodes of three RIs in social sciences coordinated by CNR, including CESSDA-ERIC (Consortium of European Social Science Data Archives), SHARE-ERIC (Survey of Health, Ageing and Retirement in Europe), and RISIS (Research Infrastructure for Science and Innovation policy Studies), as well as ISTAT (Italian Institute for Statistics).

FOSSR primary objective is to promote widespread knowledge and awareness of data and methods used in empirical social science among multiple audiences. This is accomplished by providing (i) systematic and organized knowledge about available social science data resources in Italian data archives, particularly CESSDA Archive, which was the subject of a major infrastructural proposal; (ii) resources to support methodological advancements in data collection and analysis, which are particularly important for RISIS to understand the design, implementation, and outcomes of research and innovation policies, thereby improving the robustness of empirical evidence produced for policymakers and addressing new research questions; and (iii) tools and services to make advanced probability panels for longitudinal analyses publicly available to support important surveys such as SHARE, complemented by a network of online laboratories. The integration of these resources will directly contribute to the realization of Open Science for social science scholars, accompanied by a significant scientific training program for the production and analysis of social science based on FAIR empirical data.

1. FOSSR strategic nodes

Building a thematic network of existing RIs is a unique opportunity to improve their quality with high innovative services and resources, not existing in Italy, which can have a special value to support research and innovation in line with the main objectives of the PNRR – Action M4C.2.

The project will integrate hardware and software tools together with methodologies aligned with research approaches rooted in e-science, behavioral economics, and computational social sciences. From a practical point of view, it foresees significant progress in features such as data collection, integration, curation and sharing, the establishment of a survey facilitator, the development of a social listening framework, and the launching of an artificial population simulation facility.

In the operational context, this framework aims to build a unified knowledge-sharing platform, serving as a centralized gateway for all the tools and services provided by the Italian nodes of social science infrastructures. To achieve this objective, a nationwide platform will be conceptualized and constructed, leveraging a distributed cloud computing infrastructure. This platform will facilitate the establishment of a unified and cohesive system, encompassing national infrastructure nodes (such as CESSDA, RISIS, and SHARE), in addition to internally generated data and statistical information from ISTAT.

The mentioned RIs are also important for non-academic users, such as stakeholders and policymakers. Despite the number of users coming from Italy is relevant, there is a need to improve the services and data provided to address new emerging research questions related to the changes of the economy and society, as well as to the transformative turn of science and innovation.

The integration of this pool of resources will concretely contribute to the realization of Open Science for social science scholars, accompanied by a major scientific training program on social science research methods and tools based on FAIR empirical data.

The main innovation for all the RIs involved in the thematic network are:

- Promoting the use of data for the social sciences among scholars and non-academic users and assisting data users to exploit data and resources.
- Providing innovative tools to collect, explore, analyze, and harmonize data.
- Building a distributed set of data centers to store and distribute data and resources derived from the combined activities of the thematic network.
- Training new generations of researchers/data users and disseminate the results deriving from projects developed by using FOSSR resources, developing Master courses and funding PhD positions for research on economic and societal change (in particular in the South of Italy), which allow the creation of a new pool of scholars in Data Science.

FOSSR aims to produce in this way a strong impact on the Italian communities located in the South, also through an important investment in large data centers that put the Open Cloud in this part of the country. This way, it might become a number of infrastructural resources at disposal of other public research organizations, Universities and possibly private operators and firms.

2. The heart of FOSSR: Cloud Computing Infrastructure and Network of Data Centers

The Open Cloud is developed on the top of a network of distributed data centers located in 4 strategic areas including 2 medium-large nodes (Naples and Palermo) and 2 small nodes (Turin, and Catania).

2.1 Cloud Computing Infrastructure

Through the cloud computing infrastructure, researchers will be empowered with a user-friendly, consolidated entry point that grants them access to computational services. It will also provide the means to archive and reach virtual laboratories, enabling the utilisation, processing, and public dissemination of datasets and outcomes. This unified platform will enable users to effortlessly access all the available resources and tools via a solitary entry point, employing a Single Sign-On (SSO) identification mechanism, by offering a single access point and a series of innovative services for the collection, mana-

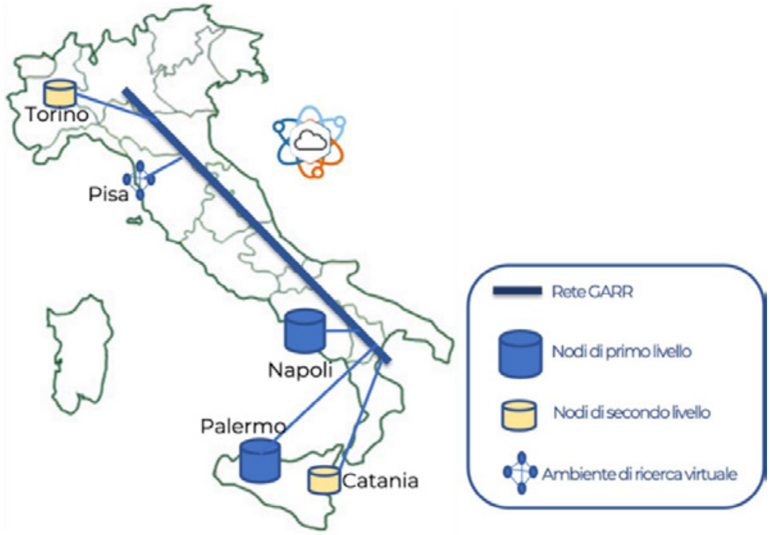


Fig. 1
Network of
Data Centers

gement, and analysis of economic and social data in compliance with the FAIR principles (Wilkinson, 2016).

As shown in Fig. 1, the platform will provide an avenue to access data in an open, collaborative, and streamlined manner facilitated by shared interfaces.

These interfaces will accommodate various tiers of access, specifically mar-ketplace, programmatic interfaces, query language, as well as a Virtual Re-search Environment (VRE):

- Marketplace. Developed within a multilingual web portal, the FOSSR mar-ketplace will afford social science researchers the capacity to access and distribute applications, services, informative resources, and datasets: it al-lows functionalities for searching, storing, retrieving, and processing the content through a well-defined catalogue. Users will be able to publish, an-alyse, and explore content via a comprehensive catalogue.
- Programmatic Interfaces: The information encompassing data, metadata, and service catalogue can be interacted with by external applications through machine-to-machine communication protocols rooted in OpenAPI/Swagger REST APIs.
- Query Language: The cloud infrastructure will facilitate the querying of both data and metadata through the utilization of standardized languages and protocols. As a foundational language and protocol for this purpose, SPARQL will be employed, enabling the querying of data and metadata over HTTP(s).
- Virtual Research Environment (VRE): VRE will serve as a catalyst for collaborative research and the promotion of open science principles, employing an array of datasets and analytical tools, supplemented by corresponding services. The FOSSR VRE will be constructed upon the foundation of the established electronic infrastructure known as D4Science, a venture overseen by CNR-ISTI.

The upcoming Cloud Platform will serve as a valuable tool for unifying servers and hardware resources to construct a cohesive system. This federated infrastructure will be established upon the existing hardware infrastructure, specifically the data center network. This network enhancement will be carried out through initiatives aimed at upgrading the four data centers.

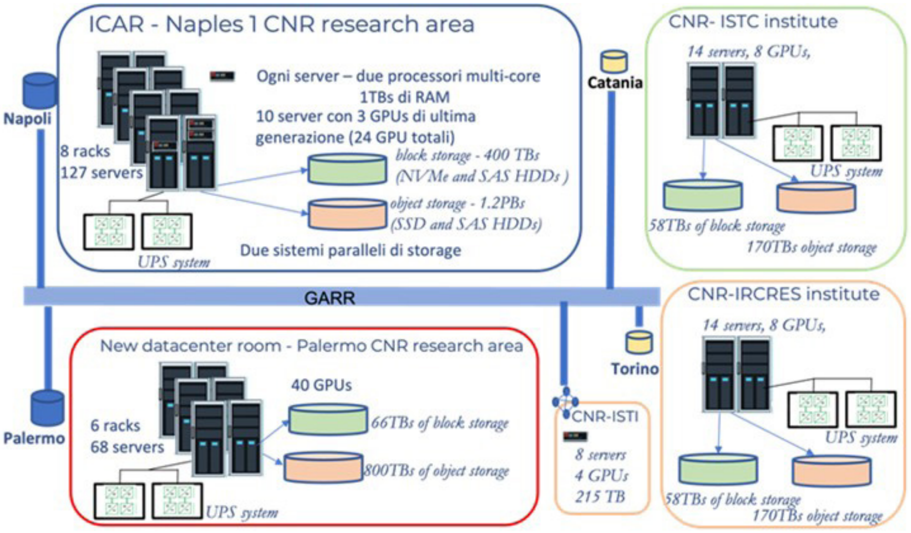
2.2 Network of Data Centers

The network of data centers includes 2 medium-large nodes (CNR-ICAR, Naples and CNR-ICAR, Palermo) and 2 small nodes (CNR-IRCrES, Turin and CNR-ISTC, Catania). The former is referred to as First Level Nodes (1LNs), while the latter as Second Level Nodes (2LNs). Additional hardware will be purchased in Pisa to deliver a Virtual Research Environment (VRE) on a dedicated IT infrastructure. The two 1LNs will be designed to provide access to hardware and software resources for high-performance computing and big data storage by means of cloud computing technologies.

The distinction between first and second level mainly regards the size and number of machines, where the application services made available to the infrastructure are similar. The main difference is that the first-level nodes will deal more with services for external exposure (web portal, marketplace, access point and security), but all nodes will have the same levels of security and Service Level Agreements (SLAs).

The data center at the Naples site of CNR-ICAR will be the largest among the first-tier data centers and will have to feature hardware components capable of managing the cloud system, guaranteeing its availability and functionality while at the same time providing a computing infrastructure adequate for the operation of services and advanced tools for data analysis using artificial intelligence algorithms, in particular based on Deep Learning. The Naples 1LN will be hosted in the ICAR data center located in the Naples 1 CNR research area. The data center will host 8 racks with 100 servers. Each server will be equipped with two multi-core processors and 1TBs of RAM allowing different kinds of computing workflows: data-intensive, CPU-intensive, parallel computing and virtualization. Some of these servers will host the latest generation Graphics Processing Units (GPUs) to run applications to handle massive AI and DL workloads. Each of these servers will host 4 units to allow their use in parallel for training deep neural networks and shorten the time required for data-intensive applications. The computing server cluster will be connected to two separate parallel storage systems: one dedicated to block storage with 400 TBs composed of nvme and SAS HDDs and a second one dedicated to object storage with 1.2PBs with SSD and SAS HDDs. The servers and the storage systems will be interconnected with a high throughput and low latency network technology specialized for HPC systems. The presence of two 1LNs will provide the cloud infrastructure with enough redundancy to allow high availability of its cloud functionalities and the possibility to replicate data across the sites for disaster recovery; it will also enable the development of scalable applications, capable of handling high loads and traffic spikes, with multi-region architecture for resilience to local outages. The complete configuration is schematized in Fig. 2.

Fig. 2
Data Centers Configuration



2.3 GARR Connection and Management Network

As mentioned, the different nodes will be directly connected by the GARR network, each node within it being configured with different networks.

As shown in Fig. 3, the In-Band Management Network (IBN) allows high-speed interconnection of the storage nodes (of the Ceph servers) of the Cloud partition as well as data traffic for communications useful for infrastructure management. This network is configured in a fully redundant spine-leaf topology

The OOBN interconnects the Baseboard Management Controllers (BMCs) of all nodes and the out-of-band (OOB) management interfaces of all other infrastructure components (storage systems, network equipment, PDUs, etc.). This network is configured in a spine-leaf topology and is redundant at the spine level.

The LLN interconnects all the computing nodes of the HPC partition of the server cluster. This network is realized with RockPort technology.

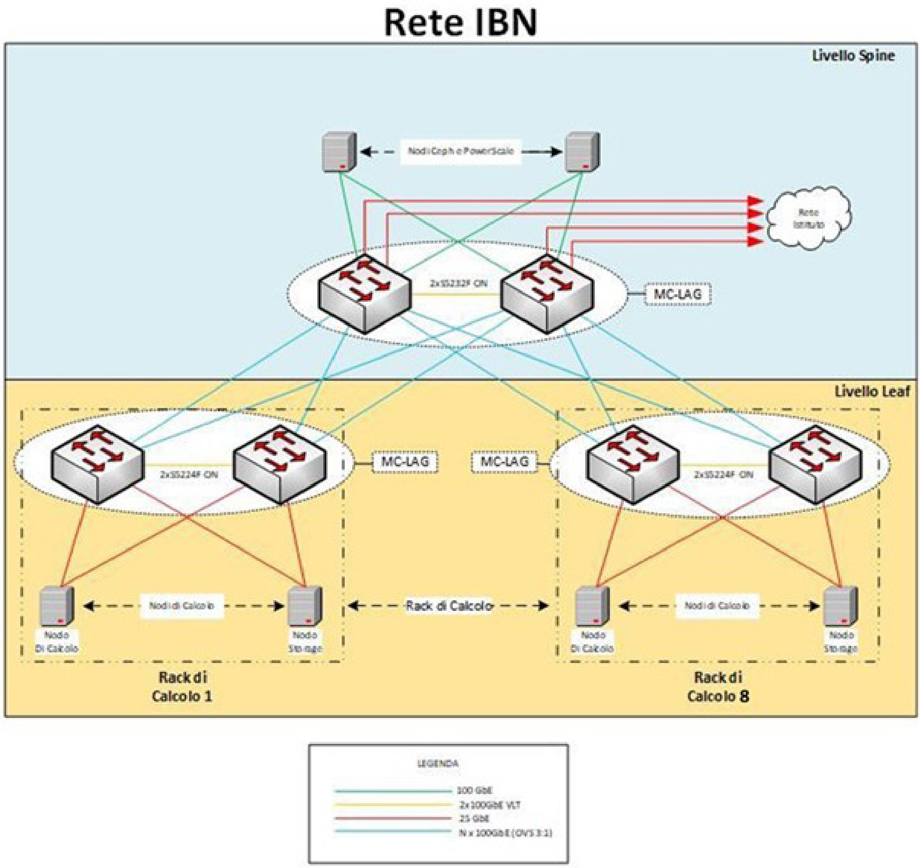


Fig. 3
In-Band Management Network

3. Conclusions

The FOSSR project represents a significant step forward in the field of Open Science in Italy and will help to foster the development of social sciences in a fair, transparent, and accessible way. It will also help to address some of the challenges that social science researchers face when it comes to data sharing and open-access publishing.

The cloud platform for social sciences has the aim to provide innovative tools and services for the analysis of social and economic data from different systems. The utilisation of cloud computing infrastructure will facilitate the seamless integration and provisioning of services geared towards storing, analysing, querying, retrieving, and sharing extensive arrays of social science data, all aligned with a standardised data model. The mix of these tools and services with open science practices will help to increase the visibility and impact of social science research in Italy.

References

D4Science Infrastructure, Infrastructure for the management of scientific data, https://www.d4science.org/about-us

Garr Network, https://www.garr.it/it/infrastrutture/rete-nazionale/infrastruttura-di-rete-nazionale

Reale, Emanuela, Ciampi, Mario, Paolucci, Mario, Zinilli, Antonio, Nuzzolese, Andrea Giovanni, Cerulli, Giovanni, Spinello, Andrea Orazio, De Gregorio, Daniela, Saccone, Massimiliano, & Rosati, Maria Elisa. (2022, December 21). FOSSR Kick-off meeting poster and presentations. Zenodo. https://doi.org/10.5281/zenodo.7458327

Rockport Networks, https://rockportnetworks.com/

Stilo, Maria Alessandra, Fabrizio, Serena, & Fava, Alessia. (2023). FOSSR: Driving Fair and Open Social Science Research (Version 1). Zenodo. https://doi.org/10.5281/zenodo.8005487

Wilkinson M., Dumontier M., Aalbersberg I. et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 https://doi.org/10.1038/sdata.2016.18

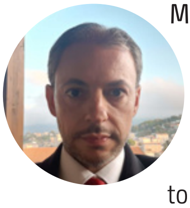
Authors

Serena Fabrizio serena.fabrizio@ircres.cnr.it

Serena Fabrizio PhD in Communication Science at Sapienza University of Rome is Researcher at CNR-IRCrES. Since 2015 with a research fellowship at CNR-IRCrES worked on several European and national research projects. In particular in the management of Training and Communication activities of RISIS project, and in the research team of a PRIN coordinated by CNR-IRCrES. Main research interests concern policies for higher education and governance systems, analysis of public funding on R&D, evaluation of impact of Social Science and Humanities research. She is involved in the FOSSR project, on Community building and training activities.

Mario Sicuranza mario.sicuranza@icar.cnr.it

Mario Sicuranza, PhD, MSc, is a Research Technologist of the Institute ICAR CNR. He received a Ph.D. degree in Information Engineering on Cybersecurity for Health Information System in 2016. His research interests include e-health, web services, and security architectures. He collaborates with the Italian Presidency of the Council of Ministers, the Ministry of Health, and the Ministry of Economy and Finance, on the National Interoperability Framework for the interoperability of the Electronic Health Record regional systems. He is co-author of more than 50 scientific papers. He leads the WP7 on developing the data center network in FOSSR.



Mario Ciampi mario.ciampi@icar.cnr.it

Mario Ciampi is a Senior Technologist at CNR-ICAR, where he works on technology research in the field of information systems, by leading the "System Interoperability and Management" technology group. He has held numerous leadership roles in research projects co-funded by European Commission, Italian Ministries, and Regions. He acts as an evaluator of research projects for the European Union, international advisory boards, and national authorities. He leads the WP6 on Cloud Computing in FOSSR.

Emanuela Reale emanuela.reale@ircres.cnr.it

Emanuela Reale is Director of CNR IRCrES. Her research area is the study of the public sector research institutions and policies, with particular reference to university policy, governance, funding mechanisms, methods and tools for university and research assessment, science and technology indicators. She has worked in numerous national and international projects as a principal investigator or coordinator. She is member of the Consortium Management Committee of RISIS, of the Board of Eu-SPRI, of the Board of CHER, and of the Board of ENID. She is the Scientific Coordinator of the FOSSR project.