CONFERENZA GARR 2025

FRONTIERE DIGITALI

Infrastrutture condivise e indipendenti per il futuro della ricerca

Bari, 13-15 maggio 2025

SELECTED PAPERS

Consortium GARR

Conferenza GARR 2025 - Frontiere Digitali
Infrastrutture condivise e indipendenti per il futuro della ricerca
Bari, 13-15 maggio 2025

ISBN 978-88-946629-4-8
DOI 10.26314/GARR-Conf25-proceedings

Editore: Associazione Consortium GARR
Via dei Tizii, 6, 00185 Roma, Italia
www.garr.it

Curatori editoriali: Marta Mieli, Carlo Volpe
Progetto grafico: Carlo Volpe
Impaginazione: Marta Mieli

Prima stampa: Settembre 2025
Numero di copie: 400
Stampa: Tipografare
Via della Magliana, 1098, 00148 Roma (RM)

Tutti i materiali relativi alla Conferenza GARR 2025 sono disponibili all'indirizzo:
conf25.garr.it/it/

Indice

Comitato di programma

Claudio Allocchio - GARR
Nicola Barbuti - Università di Bari
Claudia Battista - GARR
Roberto Bellotti - Università di Bari
Elis Bertazzon - GARR
Tommaso Boccali - INFN
Fabio Calefato - Università di Bari
Mauro Campanella - GARR
Massimo Carboni - GARR
Alessandro Cardini - INFN
Antonio Cisternino - Università di Pisa
Rocco De Nicola - Laboratorio Nazionale di Cybersecurity del Cini E IIT-CNR
Sara Di Giorgio - GARR
Francesca Lisi - Università di Bari
Marta Mieli - GARR
Giuseppe Navanteri - IFO Roma
Gabriella Paolini - GARR
Graziano Pesole - CNR-Ibiom e Università di Bari
Michele Ruta - Politecnico di Bari
Giada Sciarretta - FBK
Sabrina Tomassini - GARR
Davide Vaghetti - GARR
Simona Venuti - GARR
Carlo Volpe - GARR
Gloria Vuagnin - GARR

Tutte le presentazioni e maggiori informazioni
sono disponibili sul sito dell'evento:
conf25.garr.it/it/

DOI: 10.26314/GARR-Conf25-proceedings-01

La nuova disciplina europea delle reti di telecomunicazioni: il Digital Network Act

Innocenzo Genna

European Digital Policy and Regulation, Roma-Bruxelles

Abstract. La Commissione europea intende modificare l'attuale regime europeo di regolamentazione delle telecom, basato sulla focalizzazione sugli operatori dominanti, con un nuovo regime, sostanzialmente deregolato, dove sarà ancora possibile un mero accesso simmetrico (cioè applicabile a qualsiasi operatore, indipendentemente dalle dimensioni) alle sole infrastrutture civili. Si tratterebbe di un cambiamento epocale per le telecomunicazioni europee, pensato per ridurre la concorrenza nel settore e far aumentare i prezzi. La futura riforma, denominata "Digital Network Act" (DNA), appare però basata su premesse problematiche, in particolare circa lo stato della connettività in Europa e l'applicazione del Gigabit Infrastructure Act, una normativa adottata recentemente per scopi diversi da quelli ora dichiarati con il DNA. Vi sono inoltre travisamenti circa lo stato finanziario delle telco europee ed il confronto tra i settori telecom USA ed europeo. La proposta di DNA sarò oggetto della procedura legislativa di codecisione a partire dal 2026.

Keywords. Digital Network Act; Decade Digitale; Telecomunicazioni; Fibra Ottica; 5G

1. Introduzione

La Commissione europea si appresta a presentare, entro la fine del 2025, una proposta di riforma legislativa del framework europeo delle telecomunicazioni, denominata "Digital Network Act". Le linee generali di tale riforma sono desumibili da precedenti posizioni espresse dalla stessa Commissione, in particolare nel White Paper sulla connettività (Commissione europea, 2024), e più recentemente in un documento di consultazione, pubblicato il 6 giugno 2025, denominato "Call for Evidence" (CFE")(Commissione europea, 2025). La CFE include, tra gli altri elementi, proposte altamente dirompenti in materia di regolamentazione dell'accesso alle reti degli operatori dominanti. In sostanza si propone di declassare l'attuale regime pro-concorrenziale, basato sulla designazione e regolamentazione ex ante di operatori aventi "Significativo Potere di Mercato" ("SMP"), con la deregolamentazione di principio accompagnata da un regime "light" di regolamentazione simmetrica (cioè indipendente dalle dimensioni degli operatori) applicabile alle sole infrastrutture civili (tubi, condotte, pali ecc).

Queste proposte si basano su premesse errate e problematiche, tra cui:

(i) l'idea che attuale framework sarebbe datato e non più adatto ai tempi;

(ii) una rappresentazione fuorviante secondo cui la connettività digitale europea non sarebbe all'altezza degli obiettivi da raggiungere, in particolare per quanto riguarda la fibra

Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

ottica e la diffusione del 5G;
(iii) una forzata e parziale reinterpretazione del ruolo e dello scopo del Gigabit Infrastructure Act ("GIA"), un recente strumento normativo volto a facilitare la diffusione delle infrastrutture, non idoneo però a sostituire la regolamentazione dell'accesso alle reti.

2. Un framework regolamentare flessibile ed ancora adeguato

Il quadro regolamentare europeo è stato modificato ben 3 volte (2003, 2009 e 2018) per adattarsi ai cambiamenti di mercato e tecnologici. La revisione del 2018 è stata focalizzata proprio sull'esigenza di accompagnare le vecchie tecnologie basate sul rame verso le reti ad altissima velocità (fibra e 5G). Alcuni paesi europei sono ora leader mondiali nella connettività ad altissima velocità e vantano performance superiori ai paesi asiatici.

3. Lo stato della connettività nella UE

La posizione della DG Connect (la direzione della Commissione europea che si occupa di digitale e connettività) riflette una narrativa proveniente da operatori incumbent ed ambienti finanziari ad essi interessati, secondo cui il settore telecom europeo sarebbe talmente in crisi da non essere più capace di investire. In verità, si tratta di una problematica finanziaria prima che industriale, portata avanti dagli operatori incumbent e mobili quotati in borsa, che non riescono a soddisfare le aspettative dei propri investitori finanziari, anche per via di costi organizzativi e di legacy ancora legati ai monopoli telefonici ed alla bolla Internet del 2000. Invece le aziende non quotate ed in particolare i new entrants (ad esempio le società che hanno investito in nuove reti con il modello wholesale-only) sembrano avere una visione del mercato più fattuale e realistica, meno legata allo story-telling della crisi. Una attenta analisi dei bilanci delle principali telco europee (Telecompaper, 2024; PTS 2025)[1] è sufficiente poi a smontare la narrativa della crisi.
Ad ogni modo, per quanto gli investimenti in infrastrutture di ultima generazione siano importanti, non vi è evidenza che le telco europee siano finanziariamente impossibilitate ad investire in reti, anche tenuto conto del fatto che fondi pubblici sono disponibili in caso di necessità. Gli ultimi dati sulla copertura europea in FTTH pubblicati con il Rapporto 2025 sulla Decade Digitale mostrano che vi sono stati significativi miglioramenti e che la maggior parte dei paesi UE è in grado di raggiungere i propri obiettivi di copertura. In particolare:

- Copertura con reti ad altissima capacità: 82,5% (media UE)
- Copertura in fibra FTTH : 69,2% degli edifici (in aumento rispetto al 63,9% a fine 2023)
- Copertura 5G complessiva: 94%
- 5G nella banda 3,4-3,8 GHz: 67,7% (in aumento rispetto al 51% a fine 2023)

Questi dati riflettono un progresso solido e duraturo nell'implementazione delle infrastrutture di connettività, non una crisi che giustifichi lo smantellamento dell'attuale modello europeo di regolamentazione.

1 The Financial Health of the European Telecoms Operators, Telecompaper, 2024; Telecom operators and return on investment, PTS, 2025.

4. Il nuovo regime dell'accesso alle reti

La Commissione propone di relegare la regolamentazione basata su SMP a "ultima istanza", applicabile solo dopo l'applicazione di misure simmetriche (GIA o altre forme di accesso simmetrico già applicabili). Ciò rappresenta un radicale allontanamento dai principi normativi consolidati, che andrebbe a sostituire interventi mirati e basati sull'evidenza per affrontare posizioni dominanti di mercato comprovate, con un approccio generalizzato applicabile a tutti gli operatori, indipendentemente dal loro potere di mercato. Inoltre, questa proposta di dare priorità alla regolamentazione simmetrica contraddice i principi fondamentali del diritto della concorrenza dell'UE, che riconoscono agli operatori con significativo potere di mercato una responsabilità specifica nel non distorcere la concorrenza. L'abbandono di tale principio introduce incertezza giuridica e rischia di violare i principi di proporzionalità e certezza normativa sanciti dal diritto dell'UE.

Il GIA è stato concepito come uno strumento di riduzione dei costi, volto a supportare l'implementazione delle reti, non come uno strumento per risolvere i problemi strutturali di concorrenza. Il suo ambito di applicazione limitato e le sue esenzioni rischiano di consentire agli operatori dominanti di consolidare le proprie posizioni di mercato, anziché garantire un accesso equo e aperto a tutti.

Sarebbe quindi un gravissimo errore passare alla regolamentazione simmetrica, come regola di default, dopo che per tanti anni la UE ha assicurato ad operatori ed investitori che senza SMP non sarebbero stati regolati. Per di più, le reti degli operatori new entrants sarebbero ora accessibili agli operatori incumbent, che prima erano regolamentati. Si tratterebbe di uno dei più clamorosi cambiamenti di policy regolamentare mai visti nel mondo.

5. Il confronto con il mercato USA

La narrativa della crisi del mercato telecom europeo, che ha ispirato molte recenti posizioni della DG Connect, è basata anche sull'idea, poco scrupolosa, che il mercato telecom europeo sia indietro rispetto a quello USA.

Le telecom europee sono effettivamente meno redditizie di quelle americane per gli investitori finanziari. I prezzi retail americani sono effettivamente più alti di quelle europei, a causa da una differenza storica di partenza. Per lungo tempo gli operatori europei sono stati remunerati, oltre che con le tariffe retail, con il roaming internazionale e la terminazione mobile. La cancellazione in Europa di queste voci regolamentate (cancellazione dovuta al fatto che si trattava di costi dannosi per il mercato o iniqui per gli utenti) non è stato successivamente compensata da un aumento delle tariffe da parte degli operatori. Di conseguenza, le tariffe europee sono rimaste più basse di quelle USA ed il gap di redditività tra i rispettivi mercati è aumentato considerevolmente.

Tuttavia, a questa minore remunerativitá del mercato UE si contrappone un maggiore welfare per gli utenti europei, che possono godere, rispetto agli USA, di maggiore pluralismo, concorrenza, prezzi bassi ed infrastrutture di alto livello, in taluni casi superiori a quelle USA. Il direttore generale di DG COMP, Oliver Guersent, ha pubblicamente smentito il primato del mercato USA nel corso della presentazione dello studio "Protecting competition in a changing world" il 27 giugno 2024 (il video è disponibile su Youtube)

Peraltro, non è corretto affermare che il mercato USA sia concentrato con solo 3 operatori. Accanto ai grandi carrier continentali negli Stati Uniti operano numerosi operatori locali e di medie/piccole dimensioni, come ad esempio GCI Wireless in Alaska e Cellular One nel Mid-West. Si tratta di centinaia di operatori, non solo 3, come invece predica la narrativa sulla crisi delle telecoms (Broadbandnow, 2025).
Creare degli operatori continentali sarebbe peraltro possibile anche in Europa, perchè non esistono norme antitrust in Europa che impediscano le fusioni cross-borders. Persino una fusione tra Orange e Deutsche Telekom, i due più grandi operatori europei, appare ammissibile.

6. Conclusioni

Dalle liberalizzazioni ad oggi la regolamentazione europea delle telecomunicazioni ha prodotto risultati encomiabili, prezzi più bassi e una migliore qualità delle reti e dei servizi, investimenti sostenuti e innovazione continua. L'Europa eccelle rispetto ai suoi omologhi globali quando si tratta di combinare l'implementazione delle reti gigabit, la loro adozione da parte di consumatori e utenti professionali, nonché l'accessibilità economica e l'inclusione per i consumatori e le imprese europee. Il suo successo deriva da una combinazione equilibrata di forze di mercato e di una regolamentazione proporzionata e basata sulle evidenze.
Abbandonare questo modello a favore di alternative vaghe e non testate, come quelle proposte con la CFE su ispirazione di ambienti finanziari insoddisfatti, comprometterebbe le ambizioni digitali e la competitività globale dell'Europa. L'UE dovrebbe mantenere ciò che funziona e riformare parti del framework solo laddove sia dimostrabile la necessità.

BIBLIOGRAFIA

Broadbannow, 2025, The Complete List of Internet Companies in the US
Commissione europea, febbraio 2024, White Paper - How to master Europe's digital infrastructure needs?
Commissione europea, 6 giugno 2025, Call for evidence for an impact assessment - Ares(2025)4545535
PTS, 2025, Telecom operators and return on investment
Telecompaper, 2024, The Financial Health of the European Telecoms Operators

Autore

Innocenzo Genna è un avvocato specializzato in strategia, public affairs e regolamentazione europee nel settore delle telecomunicazioni e di Internet. Scrive per quotidiani e riviste nonché sul blog professionale RadioBruxellesLibera. Dal 2023 è esperto giuridico del sottosegretario alla Trasformazione Digitale. E' Vice Chair dell'IBA Communications Law Committee ed è stato consigliere di Euroispa, MVNO Europe, ECTA ed EIF, General Counsel per Tiscali e partner presso Ughi e Nunziante a Roma. E' cofondatore di Digit@lians, la community degli specialisti italiani del digitale operante a Bruxelles.

DOI: 10.26314/GARR-Conf25-proceedings-02

Connectivity Evolution at Concordia: Enabling Science in the Heart of Antarctica

Alessandro Mancini[1], Erik Geletti[2]

[1]Consiglio Nazionale delle Ricerche – Istituto di Informatica e Telematica,
[2]Istituto Nazionale di Oceanografia e di Geofisica Sperimentale

Abstract. In the heart of Antarctica, Concordia Station is a year-round research hub whose extreme remoteness makes connectivity essential for both science and daily operations. Until 2023, its only link to the outside world was a limited, high-latency geostationary VSAT connection. In December 2023, the station installed its first Starlink terminal, dramatically improving bandwidth and latency thanks to LEO satellite technology. This work outlines the ways in which Starlink has improved scientific data transmission, operational workflows, and the well-being of overwintering staff. It also examines routing behavior toward European research networks, the role of IPv6, and the challenges related to cybersecurity, network modernization, and resilience in extreme conditions. Connectivity at Concordia is now a true enabler of real-time research and collaboration—transforming one of the world's most isolated outposts into an active node of the global scientific community.

Keywords. Antarctic connectivity, VSAT, Starlink, satellite connectivity, network impact

Fig. 1
Concordia Station with halo effect

Introduction

Concordia Station, jointly operated by Italy (PNRA) and France (IPEV), is one of the most remote and scientifically valuable research bases in Antarctica. Situated at over 3,200 meters on the Antarctic plateau (Dome C), more than 1,000 km from the coast, its extreme isolation makes connectivity essential not only for scientific activities but also for logistics, operations, and the well-being of personnel.

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

Over the past few years, low Earth orbit (LEO) satellite constellations have begun to revolutionize research conducted in remote environments. These new technologies offer higher bandwidth and lower latency compared to traditional geostationary systems, enabling real-time data exchange and enhanced collaboration, while also introducing technical and organizational challenges.

Concordia hosts year-round operations, including both automated and manned observatories. It contributes to key global studies in atmospheric science, seismology, astronomy, glaciology, and human physiology under extreme conditions. Data flows vary in frequency and urgency: while high-priority streams like seismic or meteorological data are sent in near real-time, other datasets are processed locally and transmitted later. Stable and reliable connectivity is thus a fundamental enabler for Concordia's scientific mission.

2. VSAT Connectivity: Constraints and Limitations of an Essential Link

For many years, Concordia Station's only internet connection was a geostationary broadband satellite link based on VSAT technology. The system operated via Intelsat-19, one of the few satellites visible from the station's extreme latitude (75°S). The link used C-band, with a 3.8-meter Prodelin dish housed in a heated shelter and equipped with redundant components to ensure service continuity.

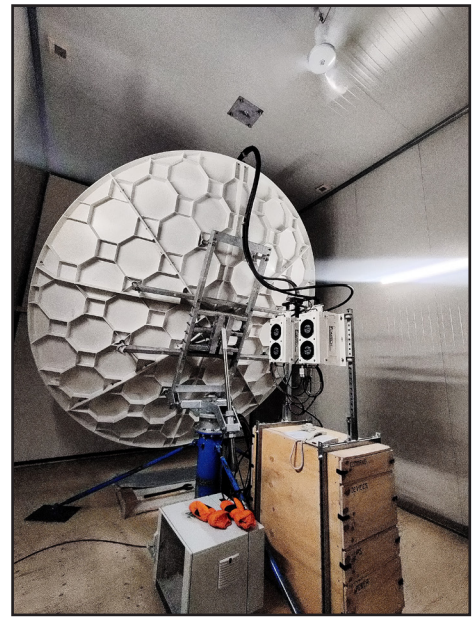


Fig. 2
VSAT
system

- Band: C-band (with L-band modem)
- Redundancy: 1:1
- Modems: 2× Comtech CDM-625
- Redundancy Switch: Comtech CRS-170A
- Amplifiers: 2× LPOD PS2-C 125W
- LNBs: 2× Norsat 3020XN
- Antenna: Prodelin 3.8 m, 1385 Series

The very low elevation angle—just 2.2°—was a major constraint: the satellite hovered near the horizon, making the signal vulnerable to obstructions and atmospheric interference. At even more extreme sites like Amundsen-Scott Station, geostationary satellites are entirely out of view.

Performance-wise, the link delivered symmetric bandwidth between 1 and 4 Mbps, but with high latency (≥ 800 ms RTT unloaded), due to the long signal path: from Concordia to a ground station in Napa, California, then via GRE tunnel to the Italian ISP's POP in Milan. This severely hindered interactive services, video calls, and large data transfers—

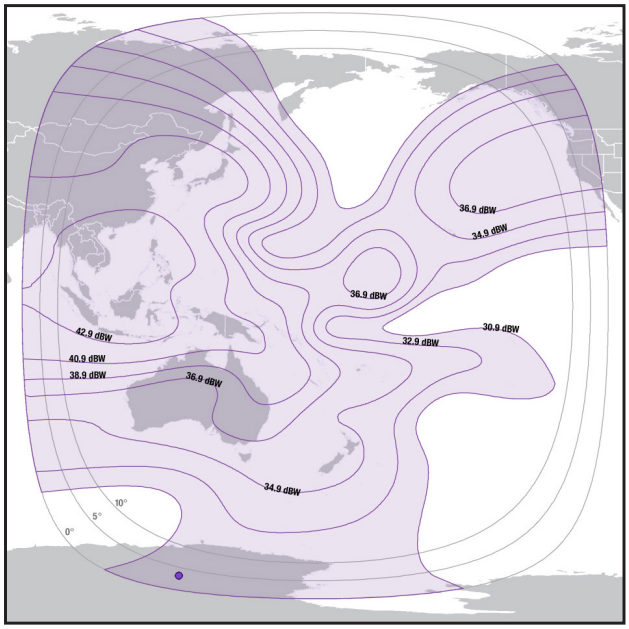


Fig. 3
IS-19 C-band West Hemi Beam

impacting both daily operations and scientific output.
Despite these limitations, the VSAT link remained Concordia's only digital lifeline for years, underscoring the vital role of telecom infrastructure in even the most remote places on Earth.

3. Introducing Starlink at Concordia

In December 2023, during the Antarctic summer campaign, the first Starlink terminal was installed at Concordia—marking a major step forward in site connectivity. Initially deployed as an experimental system alongside the VSAT link, the goal was to assess performance, reliability, and adaptability to Antarctica's extreme conditions.
Unlike geostationary satellites, Starlink uses a low Earth orbit (LEO) constellation, offering lower latency and better fault tolerance. While satellite density remains limited south of 50°S, preliminary tests at Concordia delivered promising outcomes. Traffic exits the constellation via the Sydney ground station, ~4800 km away, after 4–6 inter-satellite hops. Handover events can cause brief interruptions or jitter, but overall performance is consistently superior with far greater bandwidth and lower latency.
December was chosen for installation due to relatively mild temperatures (–25°C to –30°C), suitable for outdoor work. Cable installation must occur in such periods, as extreme cold makes materials prone to cracking.

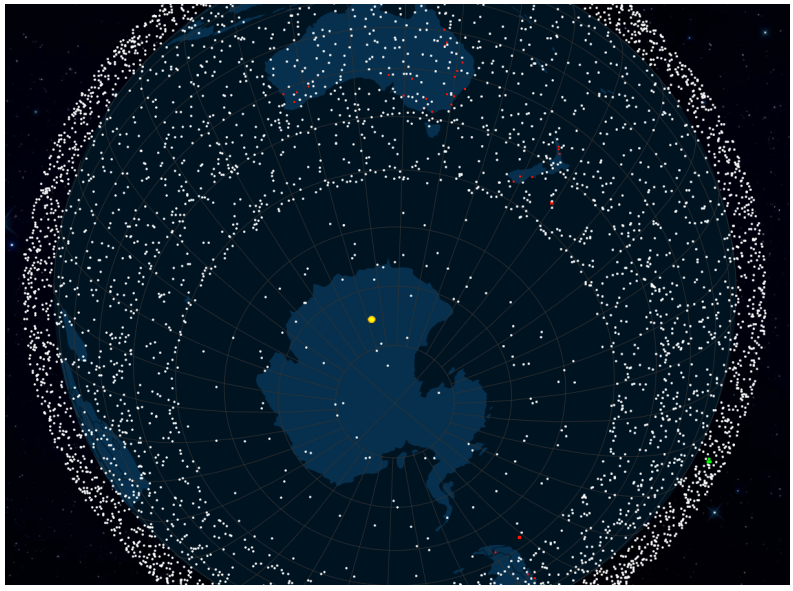


Fig. 4
Starlink satellite map

Fig. 5 Starlink antenna inside the insulation enclosure

A major challenge was ensuring continuous winter operation, when temperatures can drop to –80°C. Consumer-grade electronics are not designed for such extremes, so protection was required to shield the antenna without significantly degrading signal quality.

Tests with different cover materials were conducted to evaluate signal attenuation. A practical solution was adopted: a protective box made of wood and Styrodur (a high-performance insulating foam), equipped with an air-heating system to keep internal temperatures above critical levels.

Fig. 6 Insulation enclosure at -70°C

Installed in an open, unobstructed area with optimal satellite visibility, the heated enclosure operated continuously throughout the winter. The antenna remained fully functional, suffering no damage or prolonged outages—even in the coldest months—demonstrating the effectiveness of the design.

Thanks to this setup, the base maintained fast, stable connectivity during the isolation period. This improved scientific data transfers and greatly enhanced the quality of life for overwintering staff. Starlink has not fully replaced the traditional link, but it has proven to be a vital complementary technology with strong potential for future Antarctic campaigns.

4. Routing to Europe: Starlink and the Research Network

With the activation of Starlink connectivity at Concordia Station, a systematic monitoring campaign was launched to analyze network behavior and evaluate the impact on access to European scientific infrastructure. The goal was to understand how data routed through the Starlink constellation reaches key research entities, with particular attention to the GARR and GÉANT backbones used daily by institutes such as CNR, ENEA, and OGS.

Performance tests were conducted over IPv6, Starlink's native protocol. Although dual-stack support is available, IPv4 traffic is managed through CG-NAT, which limits inbound connectivity and traceability. The use of IPv6 enables better routing and consistency.

Traceroutes to selected nodes—including GARR sites in Italy and universities connected to GÉANT—revealed that Starlink traffic from Concordia first reaches the ground station located at the NEXTDC S1 datacenter in Sydney. From there, packets are routed primarily via commercial operators.

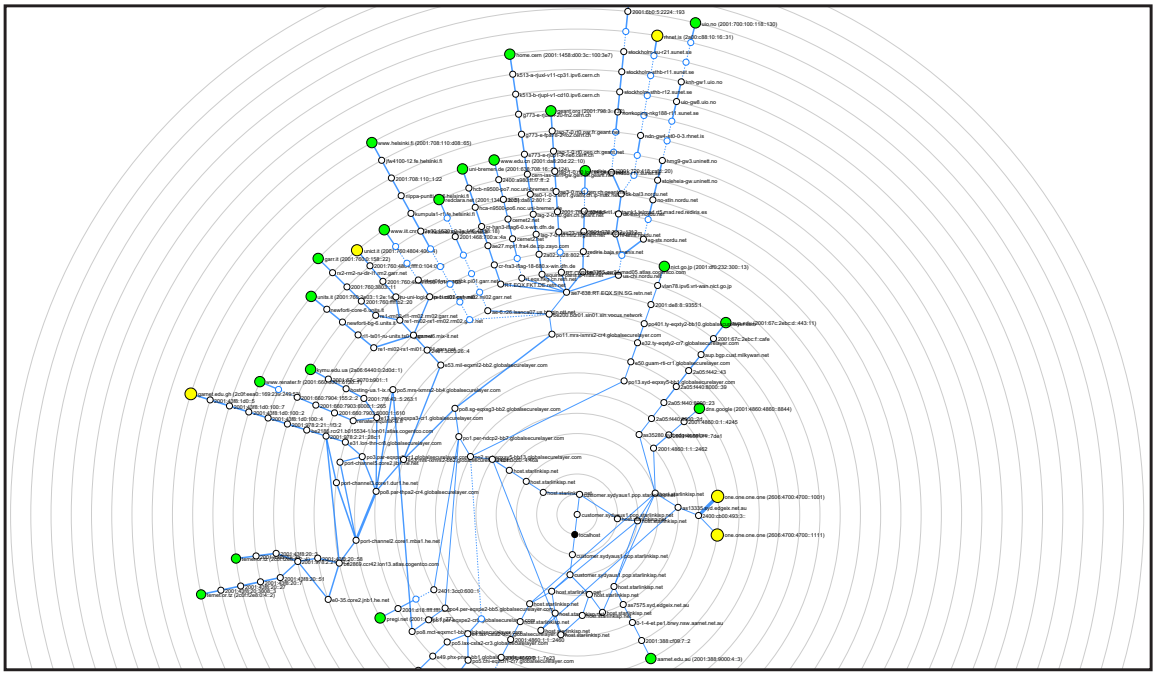


Fig. 7
Zenmap nmap traceroute from Concordia

The most immediate benefit has been the reduction in latency compared to the previous VSAT connection. While geographic distance remains a constraint, round-trip times to European nodes have significantly decreased, enabling near real-time interaction, especially with systems connected to the GARR network.
A latency comparison between Concordia and a GARR node at CNR-IIT in Pisa showed expectedly higher RTT from Antarctica. However, an interesting observation emerged:

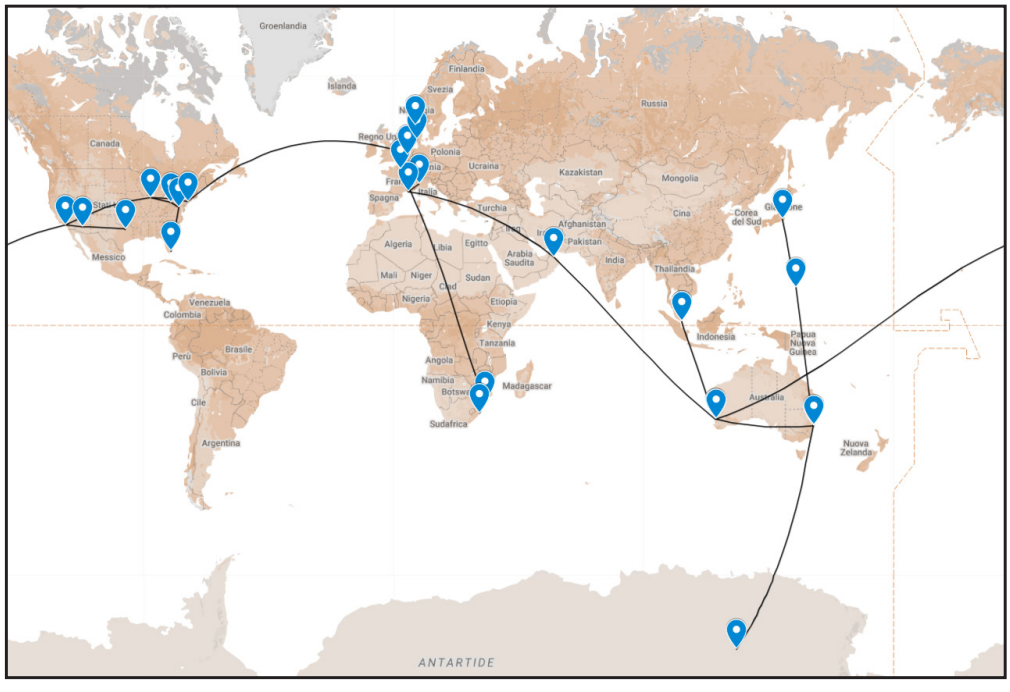


Fig. 8
Primary network routes from Concordia

as destination nodes shift eastward—toward Asia—the latency between Concordia and those regions tends to align more closely with values observed from Italy. This behavior highlights how geography continues to shape network performance.
Furthermore, the presence of AARNet (Australia's research and education network), which peers directly with Starlink at Sydney via the EdgeIX exchange, provides extremely low-latency routes within the region. This opens opportunities for future collaboration and traffic optimization based on research-oriented peering strategies.

5. Technological Renewal and New Connectivity Challenges

The introduction of high-speed broadband via Starlink has triggered a major shift in IT management at Concordia Station. A faster connection doesn't just mean quicker downloads—it demands a complete rethinking of the technological setup, affecting cybersecurity, local network structure, and service management.

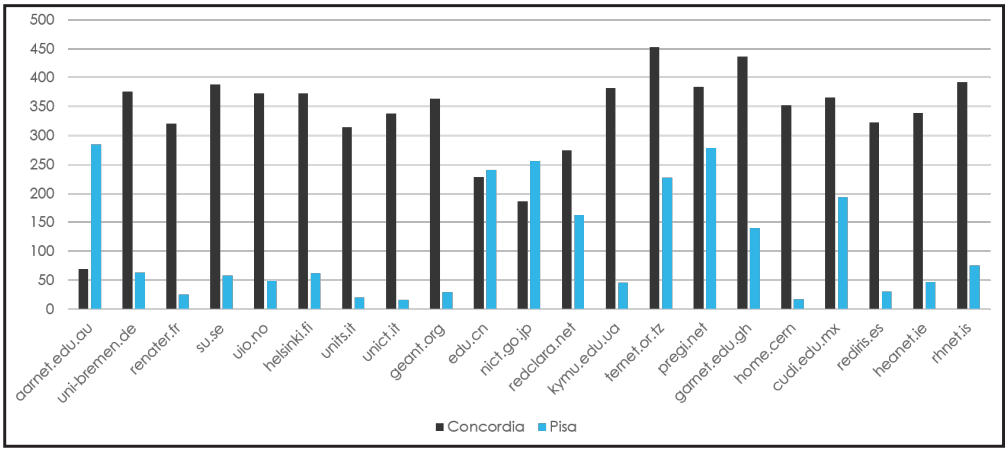


Fig. 9
RTTs from Concordia and Pisa

To harness this potential, a modernization plan is underway, including new fiber optic links between labs to boost internal data transfer, support high-throughput instruments, and enable remote access to computing and storage services.
Meanwhile, there's a clear need a more reliable wireless network is needed for indoor and, where possible, outdoor coverage. In Antarctica, hardware must withstand extreme cold, and antenna placement must account for ice and limited winter access.
Network expansion has also revealed new cybersecurity issues. Previously, only a

Fig. 10
Wi-Fi and mmWave antennas outside

few devices were granted internet access; now, with improved connectivity, many more systems—including personal devices used by the staff—are online. This shift has highlighted instances of "shadow IT," such as ad hoc cloud service usage or unauthorized devices. While these trends raise concerns, they also offer insights into the evolving daily needs of station personnel, pointing toward more adaptive, participatory, and sustainable IT service models.

As connectivity becomes more critical, ensuring continuous internet access is essential. To prevent outages and maintain resilience, a backup connection, ideally with a different provider, should be in place.

Following other Antarctic programs, a key challenge is adopting IPv6, which provides more address space, better routing, and modern network support. The United States Antarctic Program (USAP), for example, is already moving toward IPv6 deployment across its infrastructure[1], and doing the same at Concordia would boost compatibility, future-proofing, and performance for science and operations in this extreme environment.

6. Conclusions

In an extreme environment like Concordia, having a faster internet connection doesn't just mean increased download capacity. It means accelerating science, enabling collaboration, and making researchers' daily work more efficient and secure.

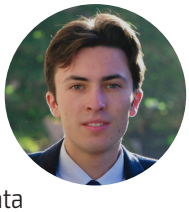
In one of the most remote and hostile places on Earth—where, until just a few years ago, the station was practically silent and isolated from the global network—it is now possible to conduct real-time scientific activities, communicate with colleagues thousands of kilometers away, exchange knowledge, and actively engage with the global scientific community. In this context, connectivity is not just a technical support tool—it is a true enabler of research.

Authors

Alessandro Mancini alessandro.mancini@iit.cnr.it

Senior technologist at the Institute of Informatics and Telematics of CNR in Pisa, where he works on the design, security, and management of network and telephony infrastructures. He teaches in the First-Level Master's Program in Cybersecurity at the University of Pisa. He has participated in multiple Antarctic expeditions at Concordia Station, including one as winter-over, managing all IT and telecommunications services in one of the most extreme environments on Earth.

Erik Geletti egeletti@ogs.it



IT Technician at the National Institute of Oceanography and Applied Geophysics (OGS) and one of the 13 winter-over personnel at Concordia Station for the 2025 Antarctic mission. He specializes in system administration, networking, endpoint management, network device monitoring, IT process automation, and DevOps CI/CD web development. He has also contributed to managing oceanographic data for the Italian National Oceanographic Data Center (NODC).

1 https://www.usap.gov/technology/4756/

DOI: 10.26314/GARR-Conf25-proceedings-03

BioRepository@ELIXIR-IT: a computational environment for storing and sharing human genetic data

Claudio Lo Giudice[1*], Giorgia Miniello[1*], Guido Cauli[1*], Francesco Rubino[8], Gianluca Cecinato[1], Marco Moscatelli[7], Sharon N. Cox[2], Nadina Foggetti[3], Francesca De Leo[3], Angelo S. Varvara[2], Bruno Fosso[2], Ermes Filomena[2], Pietro D'Addabbo[2], Marco A. Tangaro[3], Roberto Cilli[4], Giacinto Donvito[6], Federico Zambelli[3,5], Ernesto Picardi[2,3], Flavio Licciulli[1], Graziano Pesole[2,3]

[1]Institute of Biomedical Technologies, National Research Council, 70126 Bari, Italy, [2]Department of Biosciences, Biotechnology and Environment, University of Bari A. Moro, 70126 Bari, Italy [3]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, 70126 Bari, Italy, [4]Department of Physics, University of Bari A. Moro, 70126 Bari, Italy, [5]Department of Biosciences, University of Milan, 20133 Milan, Italy, [6]National Institute for Nuclear Physics (INFN), 70126 Bari, Italy, [7]Research Area Milan 4, National Research Council, 20054 Segrate, Italy, [8]Ruder Boskovic Institute, Department of Medicine, Bijenička cesta 54, 10000 Zagreb

(*) Contributed equally to the work and are recognized as co-first authors

Abstract. Nucleic acid sequencing is becoming more accessible, opening doors for new healthcare applications like precision medicine and pharmacogenomics. These could greatly improve treatments for conditions such as cancer and genetic diseases. However, to make the most of this, we need to address complex technical, legal, and ethical issues, regarding data management. This paper introduces BioRepository, a new integrated service by ELIXIR-IT. BioRepository is designed to manage human genetic data from its collection to its storage, supporting the use of genetic information in research and healthcare

Keywords. Genomic Data, Data Management, FAIR Data principles, Elixir, Secure Data Access

1. Introduction

Biorepositories can be defined as structured services designed to collect, store, manage, and distribute biological specimens associated data and metadata for research and clinical applications. These repositories play a pivotal role in biomedical research, enabling large-scale studies in genomics and precision medicine. By collecting high-quality omics sampling data — derived from genomics and proteomics analysis and sequencing — and linking them to relevant metadata, biorepositories can promote reproducibility, interope-

rability, and data sharing.
With the widespread adoption of cost-effective sequencing technologies, volume and complexity of genetic data have increased dramatically. This trend necessitates computing infrastructures capable of ensuring secure storage, controlled access, traceability, and compliance with regulatory frameworks and laws. The sensitive nature of genetic data, in particular, requires robust mechanisms for authentication, encryption, and enforcement of access policies.

2. BioRepository@ELIXIR-IT Service Overview

In order to meet these demands, ELIXIR-IT (1) has developed an integrated service for the management of human genetic data, encompassing the entire data lifecycle from sequencing to deposition. These services are built on a computational environment based on virtual machines (VM) infrastructure, resulting in a secure, scalable and user-friendly environment that can be tailored to the needs of researchers and clinicians. The system leverages the ReCaS (2) data center in Bari (Italy), part of the Italian Computing and Data Infrastructure (ICDI) (3) and the European Open Science Cloud (EOSC) (4). The platform adheres to the FAIR principles (Findable, Accessible, Interoperable, and Reusable) and is compliant with the European General Data Protection Regulation (GDPR) (5).

3. Core Requirements

To fully support genomic data management, the BioRepository service addresses several technical needs that are critical for secure and reliable operation. These include robust data security, scalable infrastructure, and the ability to support high-quality data processing pipelines within a reproducible and transparent framework.

Data Security

Sequenced data are encrypted and digitally signed using the CRYPT4GH (6) encryption suite, while all data transfers are secured via asymmetric-encrypted SSHv2 tunnels (using public and private keys in ED25519 format), ensuring that only authorized recipients can access the data. Digital signatures verify both data integrity and source authenticity. The platform ensures that private keys remain within a secure internal environment and that all access events are logged and auditable.

Scalability

The infrastructure is designed to support projects ranging from small cohort studies to large-scale national initiatives. It uses a high-capacity redundant Parallel Storage System (DELL Isilon – PowerScale), OpenStack (7), and Proxmox Virtual Environment (8) to manage virtualized environments. An underlying Ceph infrastructure provides distributed, fault-tolerant storage across multiple nodes, supporting high availability and elastic scalability.

Processing Quality

The infrastructure ensures that all components, from data uploading to final storage support processing reliability, traceability, and compliance with quality standards.

4. System Architecture

The infrastructure consists of two main components detailed in Fig.1. Each component is designed to address the challenges of handling sensitive data while ensuring global accessibility and full compliance with data protection regulations such as the GDPR. The key parts are represented by a secure BioRepository and a virtualized analysis environment. The BioRepository, hosted at CNR-ReCaS in Bari, offers 5 PB raw space for encrypted storage with geo-redundant backups at CNR-ITB (Milan, Italy) and CNR-ICAR (Naples, Italy). Both external (such as uploads of raw omics data from sequencer workstations and download of processed data to the recipient) and internal data transfers are protected via SSH tunnels with asymmetric keys encryption. The repository also hosts curated, versioned reference datasets which can be essential for bioinformatics pipelines.

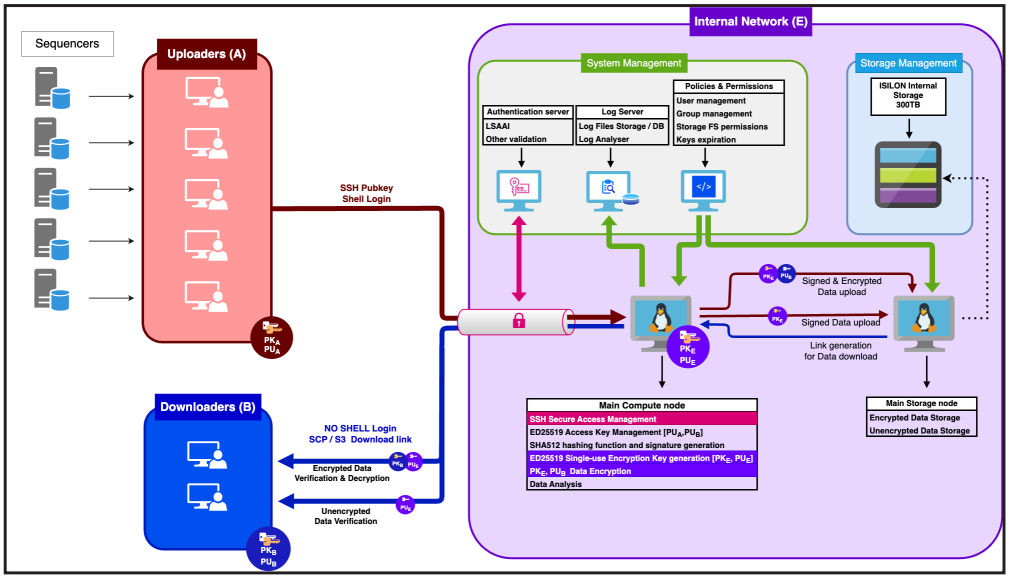


Fig. 1

BioRepository@ELIXIR-IT system architecture: visual representation of the secure data workflow across operational roles. Uploaders (A) send raw sequencing data through encrypted channels; the Internal Network (E) manages key generation, encryption, signing, logging, and storage; Downloaders (B) retrieve processed or raw data with integrity and authenticity validation. The infrastructure integrates policy enforcement, user management, and scalable encrypted storage

5. Secure Data Workflow

The system defines three operational roles: Uploaders (A), Downloaders (B), and the Internal Network (E).

- Uploaders submit raw data obtained from sequencers.
- Downloaders access processed data upon request.
- The Internal Network handles encryption, processing, and secure storage.

For each operational project, data workflow phases include:

1. Preparation: all users involved in an operational project must upload their ED25519

public keys to the system. Meanwhile, a CRYPT4GH public-private key pair PU(E) and PK(E) is created into the internal network for that operational project by the system administrator.

2. Upload: an encrypted transfer tunnel is established from the uploading workstation to the system network gateway. Data is transferred via SSH and then verified using SHA512 hash fingerprints.
3. Optionally, uploaded data can be processed using internal computing facilities, according to the specific agreement upon Institutes.
4. Hosting: sensitive data are encrypted using the download recipient public key PU(B), and digitally signed with a private key PK(E) generated into the system and unique for any different operational project, in order to grant data authenticity.
5. Download: authorized Downloaders acquire the public key PK(E) of the operational project, then they retrieve the processed data using a SSH secure tunnel. In the case of sensitive data, Downloaders can decrypt them using their own private key PK(B), while data integrity and authenticity is verified using the internal network Public Key PU(E). For non-sensitive data, the decryption part is skipped while integrity and authenticity can be verified using PU(E) as said.

Each processed dataset receives a unique identifier to support version control and traceability. All user actions are logged, and the logs are securely retained for auditing purposes. Once the download operation is complete or the intended hosting period of the operational project ceases, data and associated internal key pairs are securely deleted.

6. Infrastructure and Resources

While originally based on proprietary VMware ESXi [9], the Biorepository infrastructure currently relies on Proxmox VE, an open-source virtualization platform based on KVM/QEMU. Ceph [10] integration allows for efficient, resilient storage using OSDs, MONs, CRUSH maps, and Logical Volume Management (LVM). The system supports live migration of VMs, resource-aware scheduling, and automated failover.

7. Virtualized Analysis Environment

This environment also supports the deployment of customized, scalable bioinformatics pipelines executed on tailored VMs. VMs utilize encrypted filesystems and provide shared access to the BioRepository. High-availability configurations ensure minimal downtime, while snapshots and backup mechanisms can grant recovery and auditability of critical services.

Each VM is provisioned with:

- Up to 20 vCPUs (Intel Xeon E5/E7 with AVX-512)
- 200 GB DDR4 ECC RAM
- 500 GB local scratch storage
- 30 TB shared storage (NFS/iSCSI)

VMs can be equipped with one or two NVIDIA A100 GPUs to enable accelerated execution

of compute-intensive workflows such as Parabricks, DeepVariant, and Guppy.
Analytical environments can be replicated to support collaboration and benchmarking.

8. Software and Workflow Management

The platform supports containerized execution and environment management using:

- LXC and Docker for container isolation
- Conda and Mamba for dependency resolution and environment replication

To ensure processing consistency and traceability, bioinformatics workflows are containerized, version-controlled, and subject to rigorous quality assurance protocols. All analytical pipelines adhere to standardized sequencing data formats (e.g., FASTQ, BAM, VCF), and utilize metadata schemas aligned with the FAIR principles.
Audit trails and validation checks are embedded throughout the execution environment, ensuring reproducibility and transparency across different analyses. Containers are maintained in secure internal registries with automatic version tracking and security validation.

9. Conclusions

The BioRepository@ELIXIR-IT platform offers a secure, scalable, and standards-compliant solution for the full lifecycle management of human genetic data. By integrating open-source technologies such as Proxmox VE, Ceph, and CRYPT4GH, and by enforcing strict data protection policies, the platform ensures high levels of performance, interoperability, and trust.
Its modular and containerized architecture also supports a wide range of bioinformatics workflows, while GPU acceleration, live migration, and automated backups enhance computational efficiency and resilience. This platform serves as a forward-looking model for infrastructures supporting precision medicine, collaborative genomic research, and ethical data sharing in biomedical science.

Acknowledgments

The BioRepository service has been fully established thanks to funding from CNR.BiOmics — "National Research Center in Bioinformatics for Omics Sciences" (PON R&I 2014-2020, PIR01_00017) and PNRR ELIXIRxNextGenIT — “ELIXIR x NextGenerationIT: Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics” (IR0000010).
Angelo Sante Varvara is a PhD student within the European School of Molecular Medicine (SEMM).

References

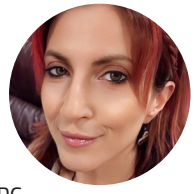
[1] https://elixir-italy.org/
[2] https://www.recas-bari.it/
[3] https://www.icdi.it/it/
[4] https://eosc.eu/
[5] https://gdpr-info.eu/

[6] https://crypt4gh.readthedocs.io/en/latest/
[7] https://www.openstack.org/
[8] https://www.proxmox.com/en/products/proxmox-virtual-environment/overview
[9] https://www.vmware.com/
[10] https://ceph.io/

Authors

Claudio Lo Giudice claudio.logiudice@cnr.it

Dr. Claudio Lo Giudice is a bioinformatician with a PhD in Cell Biology and Biotechnology. He conducted proteomic research in Finland and taught Bioinformatics at the University of Bari. Currently a technologist at CNR-ITB in Bari, he works on Linux infrastructure, scientific data management, and sensitive data handling. He is the author of REDIdb and UTRdb 2.0, databases for transcriptome and UTR region studies. His interests include bioinformatics, big data, cloud, RNASeq, and alternative splicing.



Giorgia Miniello giorgia.miniello@cnr.it

Dr. Giorgia Miniello holds a Ph.D. in Particle Physics from the University of Bari, with research conducted at the LHC-CMS on Higgs boson production and dark matter searches. She also earned a Master's in HPC Data Center Management, focusing on Big Data and monitoring systems. Currently a technologist at CNR-ITB, she develops cloud-based solutions for scientific infrastructures. Her main interests include particle physics, Big Data, HPC, and machine learning.

Guido Cauli guido.cauli@cnr.it

Student in Computer Science for Digital Businesses and graduated in Cinema, Photography and Audiovisual media. System administrator for GNU/Linux and Unix-like operating systems, mainly focusing on server and workstation management through Proxmox VE clustering and OpenStack systems, LXC and Docker container operations, enterprise networking and IT security aimed at the production, processing and secure storage of bioinformatics data.

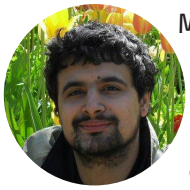
Francesco Rubino frubino@irb.hr

Francesco Rubino studied Biological Science in Bari, specialising in Bioinformatics. After experiences in industrial contexts, he defended his PhD thesis at Aberystwyth University working on ruminants' microbiota. He then worked at University of Queensland on marine microbiota. Returning to work on rumen microbiota at Queen's University Belfast, he contributed to Covid studies in wastewater for early diagnosis. Now he's at Ruder Boskovic Institute in Croazia as Research Associate.

Gianluca Cecinato gianluca.cecinato@cnr.it

Currently working at the Bari branch of the Institute for Biomedical Technologies (ITB) with the role of 'Technical Collaborator for Research Bodies (CTER)'. Responsibilities include the management of Unix-like open-source operating systems, networking, firewall administration, virtualization systems, and hardware maintenance of computing infrastructures

within the framework of the enhancement project 'ELIXIRxNextGenerationIT'.



Marco Moscatelli marco.moscatelli@cnr.it

Marco Moscatelli earned a Master's degree in Bioinformatics, during which he developed a bioinformatics platform to study the relationship between atmospheric particulate matter and human health. He attended courses on Red Hat System Administration and Ansible Essentials, which introduced him to system operations automation. He has experience in using cloud computing with Openstack and highlights the importance of this technology in providing elastic, scalable, and virtualized resources.

Sharon N. Cox sharonnatasha.cox@uniba.it

Sharon Natasha Cox, PhD in Biotechnology for Organ and Tissue Transplants, is a researcher who applies omics sciences to human health and disease. Her expertise includes, complex statistical and computational analysis of WES gene expression profiling, bioinformatics, variant identification, and the study of mtDNA–nDNA interactions in neurodegeneration. She is currently developing pipelines for long-read whole-genome and differential methylation analysis, with increasing focus on epigenomics.

Nadina Foggetti nadina.foggetti@cnr.it

A lawyer with a Ph.D. in EU and International Law, she is a Contract Professor in Cybersecurity, IT, and Biotech Law at Uniba. Technologist at CNR IBIOM and ELSI Officer for ELIXIR-IT, she has contributed to national and international projects on cybercrime, privacy, biotech, and digital law. Within ELIXIRxNextGenIT, she focuses on the Access Program, applying Open Science, FAIR data, and Open Access principles.

Francesca De Leo francesca.deleo@cnr.it

Francesca De Leo is Technology Director at CNR and Deputy Head of the ELIXIR-IT Node, based at the Institute of Biomembrane, Bioenergetics and Molecular Biotechnologies. She coordinates the ELIXIRxNextGenIT project funded by MUR and leads the Industry & Impact and Communication Offices within ELIXIR-IT. She holds a degree in Biological Sciences and a PhD in Biochemistry and Molecular Biology, with expertise in innovation, technology transfer, and research infrastructure management.

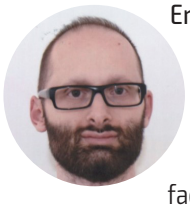
Angelo S. Varvara angelo.varvara@unimi.it

Angelo Sante Varvara is a PhD Student in Computational Biology at SEMM (European School of Molecular Medicine). Based in Bari, his expertise lies in computational analysis of human WGS and mitochondrial data, variant identification and prioritization, biological database creation and management, virtual machine administration, and biological workflow development. He is currently involved in human projects using long-read whole-genome and differential methylation analysis, with an increasing focus on rare human diseases.

Bruno Fosso bruno.fosso@uniba.it

Bruno Fosso is Associate Professor at the University of Bari “Aldo Moro”. The metagenomic investigation of host-associated microbiomes and environmental prokaryotic com-

munities is the main topic of his research activities. During the last 10 years, he developed tools and databases for both metabarcoding and shotgun metagenomics investigation of microbial communities. He participated in several Italian (MICROMAP, OMICS4FOOD) and European projects (BIOVEL, EMBRIC, LIFEWATCH, EXCELERATE and ELIXIR) and he is the coordinator of the ELIXIR-IT tools platform.

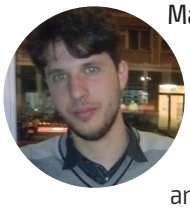


Ermes Filomena ermes.filomena@uniba.it

Bioinformatician passionate about free and open-source software, especially GNU/Linux systems. Since the beginning of his career, he has worked on NGS experiments, focusing on gene expression analysis from bulk and single-cell RNA-seq data. In the past two years, he has managed storage and primary analysis of data produced by the sequencing facility led by Prof. Pesole.

Pietro D'Addabbo pietro.daddabbo@uniba.it

Pietro D'Addabbo holds a degree in Medical Biotechnology and a PhD in Molecular Cytodifferentiation from the University "Alma Mater Studiorum" of Bologna. He has a 20-years experience in technical-scientific support and bioinformatic analysis, mainly in the field of genomic research. He is currently a fixed-term research assistant (RTDa, founded by PNRR) in Molecular Biology and lecturer in the Laboratory of Molecular Biology and Bioinformatics course at the University "Aldo Moro" of Bari.

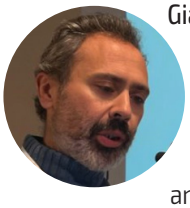


Marco A. Tangaro marcoantonio.tangaro@cnr.it

Currently a Researcher at CNR-IBIOM. Since 2015 he has been involved in the ELIXIR-IT community, developing Cloud services for bioinformatics and integrating new tools within the Galaxy workflow manager. In particular, he leads the development of the Laniakea platform, which allows the creation of on-demand Galaxy instances on the Cloud, and the UseGalaxy.it national Galaxy server.

Roberto Cilli roberto.cilli@uniba.it

Physicist and Data Analyst with experience in the information technology and services sector. Skilled in Geographic Information Systems, Machine Learning, and Spatial Analysis. Current research focuses on remote sensing—optical and SAR image processing, segmentation, registration, and quantitative metrics for decision support systems—and spatio-temporal data analysis for modeling and socio-economic applications.



Giacinto Donvito giacinto.donvito@infn.it

Giacinto Donvito, Senior Technologist, is an expert in distributed computing and cloud infrastructures for scientific research. He coordinates national and international projects in the fields of bioinformatics and omics data integration, leading the adoption of innovative technologies for the analysis, management, and enhancement of big data in life sciences and public-private partnerships.

Federico Zambelli

Federico Zambelli is an Associate Professor of Molecular Biology at the University of

Milan. His research focuses on the development of bioinformatics tools and algorithms for the analysis of sequencing data and the characterisation of gene expression regulation. As Technical Coordinator of ELIXIR-IT, he has contributed to several national projects aimed at building the technological infrastructure for biological data in Italy.

Ernesto Picardi ernesto.picardi@uniba.it

Ernesto Picardi is Full Professor of Molecular Biology at the University of Bari (Italy) and Research Associate at the Institute of Biomembranes and Bioenergetics (IBBE) of the National Research Council (CNR). His research activity focuses on bioinformatics and computational approaches to investigate co- and post-transcriptional molecular phenomena like alternative splicing and RNA editing by high-throughput sequencing technologies (including Illumina, PacBio, Oxford Nanopore). Further details are available at ORCID: http://orcid.org/0000-0002-6549-0114.

Flavio Licciulli flavio.licciulli@cnr.it

Master degree in Computer Science. Bioinformatician from 2001. Expertise in Research Data Management; expert in design and development of biological database and data integration tools; expert in application of FAIR principles for data and metadata standardization. Expert in Data Center management for the storage and processing of Life Science-oriented data. Competences in development of pipelines for the analysis of omics data.

Graziano Pesole graziano.pesole@uniba.it

Graziano Pesole is full professor of Molecular Biology in the University of Bari A. Moro and Associate Researcher of CNR-IBIOM, Director of "Consorzio Interuniversitario Biotecnologie (Trieste), Head of the Italian Node of ELIXIR, the European Research Infrastructure for Life Science (>400, h-index=84, total cites ≥30,000). His research activity is mostly focused on bioinformatics applications for the management and analysis of next generation sequencing data, also at single-cell resolution.

DOI: 10.26314/GARR-Conf25-proceedings-04

Verso una infrastruttura italiana di ricerca in fisica medica per lo sviluppo di Virtual Imaging Trials in diagnostica e terapia

[1]Barbara Caccia, [2]Giovanni Mettivier, [2]Paolo Russo, [3]Lidia Strigari

[1]Istituto Superiore di Sanità, Centro Nazionale Protezione dalle Radiazioni e Fisica Computazionale, Roma, Italy, [2]Università di Napoli "Federico II", Dipartimento di Fisica "E. Pancini", Napoli, Italy and INFN Napoli, [3]IRCCS Azienda Ospedaliero-Universitaria di Bologna, UOC Fisica Sanitaria-Dipartimento Malattie oncologiche ed ematologiche, Bologna, Italy

Abstract. I Virtual Imaging Trials (VITs) si sono affermati come strumenti innovativi per simulare e ottimizzare tecnologie diagnostiche e terapeutiche basate su radiazioni. In Italia, l'assenza di un'infrastruttura nazionale coordinata ha rappresentato un limite. Per colmare questa lacuna, ISS, INFN e AIFM stanno collaborando per produrre un prototipo di un framework integrato che combini risorse computazionali, dati clinici e metodologie in-silico. Questo framework punta a integrare capacità di calcolo, competenze in fisica medica e clinica, con un coordinamento orientato alle reali esigenze del Servizio Sanitario Nazionale. L'adozione di piattaforme sicure e interoperabili consente di avviare fin da subito la sperimentazione di nuove metodologie operative, gettando le basi per un prototipo efficace e funzionale. L'iniziativa rappresenta il primo passo verso la costituzione di una rete nazionale di VITs, a sostegno della medicina computazionale e personalizzata.

Keywords. Virtual Imaging Trial, Digital Twin, Simulazione, Intelligenza Artificiale

Introduzione

I Virtual Imaging Trials (VITs) si basano su modelli computazionali e simulazioni per riprodurre processi biologici e fisiologici, offrendo un metodo all'avanguardia per valutare dispositivi medici, farmaci e strategie terapeutiche attraverso l'uso di digital twin. I digital twin permettono di prevedere la risposta dei pazienti ai trattamenti e di ottimizzare l'efficacia delle procedure mediche. L'evoluzione e la diffusione dei VITs ha rivoluzionato, in particolare, il settore della fisica medica, offrendo strumenti di simulazione avanzata per la valutazione di tecnologie sanitarie che fanno uso di radiazioni sia per la diagnostica che per la terapia. Questo approccio, basato su modellazione computazionale, permette una valutazione accurata di immagini, dosi e parametri clinici in assenza di sperimentazioni dirette sul paziente, con evidenti benefici in termini di radioprotezione e rendendo più vicino l'obiettivo di una medicina personalizzata.

A livello internazionale, si stanno affermando diverse iniziative, in particolare negli Stati Uniti per l'implementazione dei VITs. Tra gli esempi più significativi vi è l'attività del Center for Virtual Imaging Trials (CVIT) della Duke University (Abadi et al. 2020), che

Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

rappresenta un hub in grado di riprodurre diverse tipologie di pazienti, scanner e lettori in ambito virtuale, offrendo piattaforme condivise, simulatori avanzati e dataset di qualità per promuovere la medicina computazionale e personalizzata.
Queste attività stanno dimostrando come i VITs possano anche guidare scelte regolatorie e tecnologiche. Sulla base di queste esperienze si è resa evidente la necessità di strutturare un framework nazionale coordinato e in grado di capitalizzare le competenze e le risorse presenti sul territorio.

1. Il contesto italiano: sinergia tra enti e comunità scientifica

Per accelerare la definizione del framework nazionale, dal 2024 sono state avviate delle attività di ricerca e sviluppo che hanno coinvolto l'Associazione Italiana di Fisica Medica (AIFM), l'INFN, e l'Istituto Superiore di Sanità. All'interno dell'Associazione Italiana di Fisica Medica (AIFM) è stato istituito un gruppo di lavoro dedicato allo sviluppo e alla promozione dei Virtual Clinical Trials (VCT) (Strigari et al. 2025). In parallelo, l'INFN ha avviato e finanziato un progetto triennale finalizzato alla realizzazione di un prototipo di framework nazionale per i Virtual Imaging Trials (Mettivier et al. 2025), in collaborazione con AIFM e con ISS. Un elemento chiave di questa iniziativa è il coinvolgimento dell'Istituto Superiore di Sanità (ISS), che partecipa con il proprio personale a entrambe le iniziative, in qualità di organo tecnico-scientifico del Servizio Sanitario Nazionale e di ente di ricerca pubblico. Il ruolo dell'ISS, anche in sinergia con il Ministero della Salute, sarà fondamentale per garantire coerenza con le priorità di sanità pubblica e l'armonizzazione con le normative e gli standard nazionali e internazionali.

1.1 Il gruppo di lavoro AIFM4VCT

Nel 2024 è stato istituito il gruppo di lavoro AIFM4VCT (AIFM4VirtualClinicalTrials), che coinvolge fisici medici e ricercatori da ospedali, università e centri INFN. Scopo del gruppo è armonizzare le attività nel settore dei Virtual Clinical Trial, promuovere l'uso di digital twins in ambito clinico e favorire lo sviluppo di metodologie computazionali condivise. Il primo dato prodotto dal gruppo di lavoro è stata una revisione sistematica della letteratura scientifica (84 articoli con affiliazione italiana). È stata condotta una revisione sistematica della letteratura degli ultimi cinque anni utilizzando le principali banche dati scientifiche (PubMed, Scopus, Embase). Le parole chiave utilizzate per la ricerca includevano l'affiliazione italiana e poi le keywords virtual imaging, computational twins e in-silico medicine. Gli

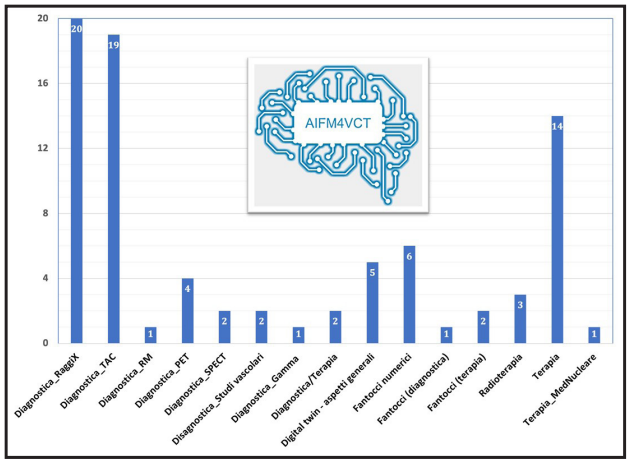


Fig. 1 Distribuzione dei lavori scientifici esaminati in base all'argomento trattato

ambiti di ricerca identificati riguardavano sia tecniche diagnostiche che di screening, sviluppo e utilizzo di fantocci digitali 3D, applicazione di strumenti di intelligenza artificiale, terapie personalizzate e medicina di precisione. L'analisi ha evidenziato una crescente produzione di studi su fantocci digitali, AI, imaging e radioterapia.
In figura 1 è riportata la distribuzione dei lavori esaminati in base all'argomento trattato. Lo studio è stato poi approfondito categorizzando gli argomenti in base al distretto anatomico e alle applicazioni specifiche.

1.2 Il progetto triennale VITA5 INFN

In sinergia con il gruppo di lavoro AIFM4VCT, l'INFN ha finanziato per il 2025-2027, un progetto di ricerca (VITA – Virtual Imaging TriAls in Medicine) nell'ambito della CNS5 – Commissione Nazionale per le ricerche tecnologiche, interdisciplinari e di fisica degli acceleratori. Il progetto intende costruire il prototipo di un'infrastruttura integrata in grado di combinare risorse computazionali, dati clinici, competenze fisiche e mediche, e strumenti di simulazione avanzata.
Il progetto prevede lo sviluppo di un'infrastruttura condivisa per il calcolo e lo storage, la realizzazione di una libreria di fantocci realistici anche tramite AI, e l'ottimizzazione di simulatori per esami diagnostici (es. Breast CT, cone-beam CT, brachiterapia oculare). Il progetto coinvolge INFN, AIFM e ISS, con l'obiettivo di creare il prototipo di un archivio nazionale di risorse e strumenti accessibile a clinici e ricercatori. L'approccio si basa su metodologie all'avanguardia, incluso l'uso di reti neurali generative e tecniche di segmentazione per la generazione di immagini sintetiche, promuovendo così un modello di medicina personalizzata supportato da simulazioni su larga scala.

2. Conclusioni

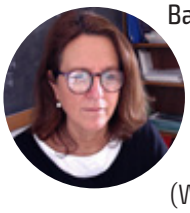
Queste iniziative rappresentano un passo strategico verso la realizzazione di una infrastruttura nazionale per i Virtual Imaging Trials, capace di integrare competenze, tecnologie e risorse distribuite. In questo contesto la collaborazione con GARR, che già offre servizi e supporto per la connettività a nodi strategici del Servizio Sanitario Nazionale come gli IRCCS, rappresenta una opportunità. La collaborazione tra la comunità dei fisici medici, gli enti di ricerca, come INFN e ISS, in una connessione coordinata con il Servizio Sanitario Nazionale può offrire una opportunità di crescita e collaborazione per la medicina computazionale di precisione, con impatti diretti su diagnosi, terapia, formazione e ricerca traslazionale.

Riferimenti bibliografici

Abadi E. et al. (2020), Virtual clinical trials in medical imaging: a review, J.Med.Imag. (7:4), pp. 1-40.
CVIT sito web http://cvit.duke.it
Mettivier G. et al. (2025), VITA: The Italian AIFM, INFN and ISS scientific collaboration towards a national center for virtual imaging trials, VITM2025, Manchester (UK)
Strigari L. et al. (2025), Horizon scanning of virtual clinical trial platforms and digital

twins in Italy, VITM2025, Manchester (UK)

Autori



Barbara Caccia barbara.caccia@iss.it

Fisico, dal 1985 svolgo attività di ricerca all'Istituto Superiore di Sanità. Sono Ricercatrice senior nel Centro Nazionale per la Protezione dalle Radiazioni e Fisica Computazionale, dove coordino il laboratorio di simulazioni Monte Carlo per applicazioni mediche e ambientali. Collaboro con l'INFN, di cui sono associata, e sono membro del network EURADOS (WG6). Dal 2018 contribuisco al WHO Collaborating Centre on Radiation and Health dell'ISS, dove coordino le attività sull'uso medico delle radiazioni.

Giovanni Mettivier giovanni.mettivier@na.infn.it

Sin dal 1997, collaboro con il gruppo di Fisica Medica del Dipartimento di Fi-sica dell'Università "Federico II" di Napoli all'interno di numerose sigle fi-nanziate dall'INFN per lo sviluppo e studio di strumentazione medicale. Dal 2006, ho iniziato ad interessarmi allo studio di scanner prototipali per la to-mografia della mammella non compressa. Il passaggio verso l'utilizzo di strumenti di AI e di simulazione Monte Carlo è avvenuto nel 2020. Attualmente sono il Responsabile Nazionale dell'esperimento VITA dell'INFN.

Paolo Russo paolo.russo@na.infn.it

Professore di Fisica Medica all'Università di Napoli Federico II, è esperto nello sviluppo di tecnologie per l'imaging medico, in particolare sistemi digitali basati su rivelatori a semiconduttore e simulazioni Monte Carlo per raggi X. Ha ricoperto incarichi internazionali in EFOMP e IOMP contribuendo alla definizione di standard professionali e formativi per i fisici medici. È stato Editor-in-Chief di Physica Medica (2013–2018). Ha coordinato progetti innovativi con INFN e collaborato a livello europeo, contribuendo allo sviluppo di tecnologie diagnostiche avanzate.

Lidia Strigari lidia.strigari@aosp.bo.it

La Dott.ssa Lidia Strigari dirige la UOC Fisica Medica presso l'IRCCS Azienda Ospedaliero-Universitaria di Bologna, in Italia. La sua attività di ricerca si concentra principalmente sull'applicazione delle radiazioni ionizzanti nelle procedure diagnostiche e terapeutiche, con un'attenzione particolare all'utilizzo di tecnologie avanzate per migliorare la personalizzazione dei trattamenti e l'accuratezza diagnostica. È attivamente coinvolta in iniziative di ricerca clinica sia a livello nazionale che internazionale, inclusi progetti che integrano l'intelligenza artificiale e le tecnologie del gemello digitale (Digital Twin) e modelli predittivi degli effetti delle radiazioni.

DOI: 10.26314/GARR-Conf25-proceedings-05

Ripensare l'intelligenza artificiale: dall'autonomia alla simbiosi

Donato Malerba

Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro

Abstract. In un momento storico in cui l'Intelligenza Artificiale (IA) sta ridefinendo in profondità il nostro rapporto con la tecnologia, si afferma con crescente rilevanza il paradigma dell'IA simbiotica: un modello fondato su forme di cooperazione autentiche, affidabili e adattive tra esseri umani e sistemi artificiali. Questo approccio supera la concezione dell'IA come semplice strumento autonomo e si avvicina a un'idea di agentività condivisa. Una tale visione, capace di rispondere in modo costruttivo a molte delle preoccupazioni odierne sull'IA, trova una concreta attuazione nello Spoke 6 del progetto FAIR (Future Artificial Intelligence Research), interamente dedicato all'intelligenza artificiale simbiotica.

Keywords. IA autonoma, IA simbiotica

Introduzione

Per decenni, lo sviluppo dell'Intelligenza Artificiale (IA) è stato guidato dall'ambizione di creare agenti autonomi, capaci di risolvere problemi in modo indipendente, senza necessità di interazione o supervisione umana. Non a caso, il paradigma emergente dell'IA agentica si riferisce a sistemi autonomi progettati per perseguire obiettivi complessi con un intervento umano minimo (Acharya et al., 2025). L'IA autonoma (Fig. 1) ha prodotto risultati straordinari, e i veicoli a guida autonoma rappresentano una delle applicazioni più visibili con sistemi in grado di percepire l'ambiente circostante e di navigare senza input umano (Van Brummelen et al., 2018). Questo orientamento verso l'autonomia è rafforzato anche dai modelli prevalenti dell'IA: l'apprendimento supervisionato mira a generalizzare da esempi, senza necessità di ulteriori istruzioni, il reinforcement learning si fonda sull'autoapprendimento attraverso l'interazione con l'ambiente, e le architetture agent-based sono concepite per operare in modo indipendente in contesti complessi. La visione cartesiana della mente come calcolo logico, separata dal corpo e dal contesto, ha fornito una giustificazione epistemologica al paradigma dell'IA autonoma. A ciò si aggiungono motivazioni industriali altrettanto forti, grazie alla promessa di risparmi operativi e di soluzioni che sono scalabili e replicabili su larga scala.

1. IA Autonoma: i limiti

Tuttavia, oggi è sempre più evidente che l'IA autonoma presenta limiti strutturali, teorici ed etici. Anzitutto, l'autonomia totale si rivela spesso un'illusione: anche i sistemi più avanzati dipendono da dati etichettati da esseri umani, e da scelte che riflettono valori,

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

priorità e vincoli umani. L'intervento umano è dunque implicito, sebbene spesso invisibile, in ogni fase dello sviluppo e dell'impiego di questi sistemi. Inoltre, la rincorsa all'autonomia ha spesso trascurato il potenziale trasformativo della collaborazione tra esseri umani e sistemi intelligenti. Questa viene ostacolata dall'opacità dei modelli appresi dai dati, che rende difficile comprenderne il funzionamento interno e verificarne le decisioni o attribuire responsabilità in caso di errore. L'esclusione degli esseri umani dal processo decisionale impedisce un loro intervento in situazioni di rischio con eventuale correzione delle decisioni prese dalla macchina. L'IA autonoma è spesso fragile quando si confronta con situazioni nuove o non previste in fase di addestramento. Peraltro, anche quando l'IA funziona correttamente dal punto di vista tecnico, può produrre risultati in contrasto con norme sociali, principi etici o bisogni concreti delle persone. Infine, all'IA autonoma manca una competenza essenziale nelle interazioni tra esseri umani: la capacità di negoziare significati, di comprendere ambiguità o di interpretare situazioni aperte e sfumate.

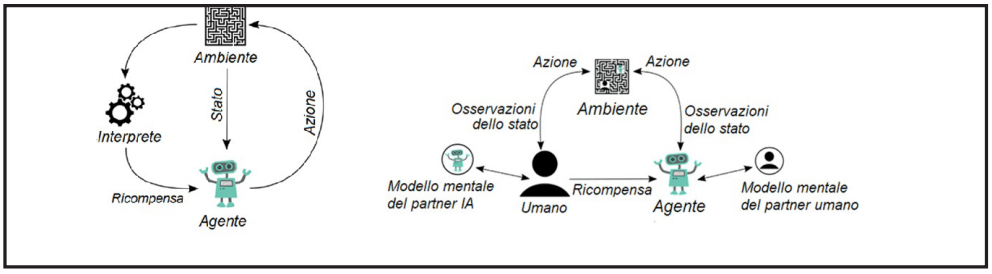


Fig. 1

A sinistra: l'IA autonoma, in cui un agente artificiale opera da solo per raggiungere il proprio obiettivo in modo efficace e accurato. L'agente percepisce lo stato dell'ambiente, agisce in autonomia e riceve una ricompensa in base alle sue prestazioni. A destra: l'IA simbiotica, in cui l'agente artificiale collabora con esseri umani, condividendo con loro un obiettivo. Grazie alla capacità di ragionare sulle azioni e sugli stati mentali delle persone, l'agente supporta decisioni condivise. I partner umani, a loro volta, comprendono il processo decisionale dell'IA e ne interpretano scopi e vincoli.

2. Un'alternativa: l'intelligenza artificiale simbiotica

Il paradigma dell'IA simbiotica (Fig. 1) propone un cambio di prospettiva radicale: anziché sostituire l'essere umano, la macchina ne potenzia le capacità cognitive, lo assiste in modo adattivo. Si tratta di costruire squadre miste di intelligenze umane e artificiali, in cui entrambe apprendono e si adattano reciprocamente. L'obiettivo non è solo raggiungere un risultato, ma farlo in modo tale che tutte le parti coinvolte comprendano le ragioni e i vincoli delle scelte compiute. Ciò significa che le sfide non si limitano alla risoluzione del compito, ma si estendono alla qualità dell'interazione e alla comprensione reciproca tra i membri della squadra eterogenea. Affinché tale collaborazione risulti efficace, è essenziale che l'agente artificiale sia in grado di comprendere e ragionare sulle azioni umane, tenendo conto degli stati mentali, delle intenzioni e delle conoscenze pregresse delle persone, adottando, cioè, una forma di theory of mind computazionale. Al contempo, è necessario che l'essere umano possa accedere e interpretare in modo trasparente il processo decisionale dell'IA, comprendendone motivazioni, vincoli e finalità. Solo attraverso questa bidirezio-

nalità cognitiva – la capacità dell'IA di adattarsi al modello mentale dell'utente e viceversa – è possibile costruire forme di cooperazione autentiche, affidabili e adattive, superando la visione dell'IA come semplice strumento e avvicinandosi a un modello di agentività condivisa.

Un esempio illuminante proviene dal mondo degli scacchi. Nella variante chiamata freestyle chess, il giocatore può consultare l'IA durante la partita. Sorprendentemente, si è osservato che le squadre vincenti non sono né gli esseri umani né le macchine da sole, ma la combinazione delle due (De Cremer, Kasparov 2021).

Il crescente impatto dell'IA nella vita quotidiana sta spostando il focus della ricerca dalle capacità tecniche delle macchine alla qualità della relazione uomo-IA. In questo scenario, il paradigma dell'IA simbiotica non rappresenta solo una prospettiva futura, ma una risposta necessaria alle dinamiche psicologiche, sociali ed etiche emerse nell'evoluzione dell'interazione con le tecnologie intelligenti. Come evidenziato in recenti studi (Prescott, 2024), le persone tendono a proiettare caratteristiche umane sugli agenti artificiali – un fenomeno noto come antropomorfismo – dando vita a interazioni che assomigliano sempre più a relazioni sociali, con effetti tangibili sul benessere individuale. Tutto ciò ha senso solo se l'IA non è concepita come un sostituto dell'essere umano, ma come un partner in grado di co-evolvere con lui, adattandosi ai suoi bisogni, emozioni e valori. In questo senso si inquadra anche una visione costruttivista del rapporto "simbiotico" tra esseri umani e sistemi artificiali (Carnevale et al., 2024).

3. Il progetto FAIR e lo Spoke 6: Symbiotic AI

Nell'ambito del progetto PNRR FAIR – Future Artificial Intelligence Research (https://fondazione-fair.it/), lo Spoke 6 affronta il tema della IA simbiotica, individuando cinque sfide fondamentali da affrontare per la realizzazione di sistemi davvero cooperativi tra esseri umani e macchine:

1. Design umano-centrico. I sistemi devono essere progettati per interagire con le persone, non solo per risolvere compiti. Questo implica la considerazione di emozioni, intenzioni, bisogni e modelli mentali degli utenti. È necessaria una progettazione orientata all'esperienza, che integri principi dell'interazione persona-macchina, dell'IA emotiva e dei modelli cognitivi del comportamento.
2. Comprensione dell'essere umano. Un sistema simbiotico deve saper interpretare chi ha di fronte, comprendendo testi, gesti, emozioni e intenzioni. L'IA deve acquisire competenze percettive e interpretative su base multimodale – testo, audio, video – per adattarsi in modo flessibile a contesti reali. Tecnologie chiave includono il natural language processing, il deep learning e l'analisi di dati eterogenei.
3. Input umano come risorsa. L'interazione non è un ostacolo, ma un'opportunità. L'IA deve essere in grado di integrare input provenienti dagli esseri umani, anche se incompleti o incerti. Servono approcci ibridi in cui i modelli statistici si combinano con rappresentazioni simboliche, grafi della conoscenza, inferenza e semantica.
4. Fiducia e trasparenza. La cooperazione richiede fiducia. Le decisioni dell'IA devono essere

comprensibili, spiegabili, accettabili e sostenibili. Oltre allo sviluppo di tecniche di XAI (eXplainable AI), è fondamentale integrare valori etici, principi giuridici e considerazioni ambientali e sociali. L'IA simbiotica deve essere progettata per l'empowerment, non per la sostituzione dell'essere umano.

5. Infrastrutture e scalabilità. Infine, la sfida tecnologica. Come sviluppare sistemi di IA simbiotica che siano efficienti, scalabili e distribuiti? Occorrono soluzioni capaci di operare su dati grandi ed eterogenei, anche in ambienti decentralizzati (edge computing), senza compromettere le capacità collaborative e adattive.

Con il coinvolgimento di sei università, un ente di ricerca nazionale e tredici imprese, lo Spoke 6 di FAIR ha avviato numerose ricerche dedicate all'IA simbiotica. Sono stati finanziati diciannove progetti su temi strettamente correlati, accomunati da una visione condivisa: non un'IA che lavori al posto dell'essere umano, ma insieme a lui. Una co-evoluzione possibile, secondo il paradigma della simbiosi.

4. Conclusioni

La sfida dell'IA, oggi, non è più imitare l'intelligenza umana, ma imparare a convivere con essa. Occorre costruire sistemi che crescano insieme agli esseri umani, anziché svilupparsi in completa autonomia. L'IA del futuro deve essere simbiotica, altrimenti rischia di risultare disallineata rispetto ai valori, ai bisogni e alle aspettative delle persone. Questa prospettiva rappresenta una risposta concreta a molte delle attuali preoccupazioni legate all'IA: può attenuare il timore della sostituzione occupazionale, ridurre bias e discriminazioni, rafforzare trasparenza e controllo, migliorare la tutela della privacy e della sicurezza, e infine promuovere un'innovazione realmente responsabile. Per realizzare tutto questo, è necessario affrontare numerose sfide di ricerca, come quelle già intraprese dallo Spoke 6, interamente dedicato all'IA simbiotica all'interno del progetto FAIR.

Riferimenti bibliografici

Deepak Bhaskar Acharya, Karthigeyan Kuppan, B. Divya (2025), Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey, IEEE Access (vol. 13), pp. 18912-18936.

Antonio Carnevale, Antonio Lombardi, Francesca A. Lisi (2024), A human-centred approach to symbiotic AI: Questioning the ethical and conceptual foundation, Intelligenza Artificiale (vol. 18, n.1), pp. 9-20

David De Cremer, Garry Kasparov (2021), AI should augment human intelligence, not replace it, Harvard Business Review (18 marzo).

Tony Prescott (2024), The psychology of artificial intelligence, Routledge, London.

Jessica Van Brummelen, Marie O'Brien, Dominique Gruyer, Homayoun Najjaran (2018), Autonomous vehicle perception: The technology of today and tomorrow, Transportation Research Part C: Emerging Technologies (vol. 89, aprile 2018), pp. 384-406.

Autore

Donato Malerba donato.malerba@uniba.it

Professore ordinario di Sistemi di elaborazione delle Informazioni presso l'Università di Bari. Direttore del Dipartimento di Informatica (2015-2022) e del CINI National Lab on Big Data (2014-2021). Componente del Consiglio Direttivo della Big Data Value Association e della PPP Big Data dell'UE (2015-2016). Responsabile scientifico dello Spoke 6 – Symbiotic AI – del progetto FAIR (Future Artificial Intelligence Research). I suoi interessi di ricerca riguardano la data science e l'intelligenza artificiale.

DOI: 10.26314/GARR-Conf25-proceedings-06

Intelligenza Artificiale e Innovazione Didattica: Prospettive Future nel Campo della Formazione Medica

Federico Siracusa[1], Floriana Vindigni[1], Federico Abate Daga[2], Elisabetta Galoppini[1], Vito Moscato[1], David Lembo[2]

[1]Direzione Sistemi informativi, Portale, E-Learning, [2]Dipartimento di Scienze Cliniche e biologiche, Università di Torino

Abstract. MedSim Academy è un'iniziativa formativa innovativa promossa dal Dip. di Scienze Cliniche e Biologiche e il Centro di Simulazione Medica Avanzata di Orbassano. Il progetto prevede la creazione di un percorso formativo bilingue (IT/EN) fruibile via web app, per la formazione propedeutica alle attività del centro. Con contenuti brevi, interattivi e accessibili, progettati secondo i principi del microlearning, del just-in-time learning e dell'Universal Design for Learning (UDL), il progetto risponde ai bisogni formativi emergenti nel settore sanitario, con attenzione alla diversità e all'inclusione. L'IA, integrata nel modello GPT "Design4Practice", supporta la progettazione del percorso modulare, ottimizzando tempi e risorse. Microcredenziali verranno rilasciate al termine di ogni modulo e un Open Badge, valido 1CFU, al termine del percorso. L'iniziativa promuove formazione continua, transizione digitale e sviluppo di competenze

Keywords. Simulazione Medica, Web App, Microlearning, Formazione Continua, IA

Introduzione

L'innovazione tecnologica sta trasformando la formazione medica, rendendo sempre più efficace e accessibile l'acquisizione di competenze pratiche. L'Intelligenza Artificiale (IA) sta giocando un ruolo chiave in questo processo, facilitando la creazione di contenuti didattici interattivi e personalizzati. Il progetto MedSim Academy nasce per rispondere a un'esigenza crescente di innovazione nella formazione medica, con l'obiettivo di preparare in modo più efficace studenti, specializzandi e personale sanitario all'uso consapevole delle tecnologie avanzate nei centri di simulazione. Il percorso formativo, fruibile attraverso una web app, è concepito per fornire competenze teoriche e pratiche propedeutiche alle attività del Centro di Simulazione Medica Avanzata, nel rispetto del principio "Don't Stop Training". Al centro dell'innovazione didattica proposta c'è Design4Practice, un GPT (Generative Pretrained Transformer) sviluppato per affiancare docenti e formatori nella progettazione di microcorsi digitali dinamici, brevi, accessibili e fortemente orientati alla pratica. Il GPT non è un semplice assistente tecnico, ma un vero e proprio facilitatore progettuale: aiuta a definire obiettivi concreti, sviluppare contenuti multimediali interattivi e personalizzare l'esperienza formativa per diverse tipologie di utenti.

1. MedSim Academy: Struttura del Percorso Formativo

Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

Il progetto MedSim Academy sarà strutturato in moduli didattici autoconsistenti, in italiano e inglese, che copriranno aspetti teorici e pratici della simulazione medica.
Ogni modulo includerà (Denny P. et al., 2023):

- Un test iniziale per valutare le preconoscenze
- Contenuti didattici interattivi, che integreranno video (Leiker D. et al., 2023), quiz e simulazioni basate su scenari clinici realistici
- Un test finale, il cui superamento garantirà l'accesso al Centro di Simulazione Medica
- Il rilascio di micro-credenziali e di un Open Badge al termine del percorso.

Per garantire un alto livello di qualità e pertinenza, l'IA verrà addestrata utilizzando i manuali tecnici degli strumenti presenti nel Centro di Simulazione, che forniranno informazioni dettagliate sul funzionamento e l'uso delle apparecchiature mediche. Inoltre, verranno integrati i contenuti teorici forniti dai docenti del corso di Medicine and Surgery, assicurando che il materiale generato sia allineato con le più recenti conoscenze accademiche e pratiche cliniche (Ma Y. et al., 2024).
L'IA sarà in grado di produrre testi formativi aggiornati e adattati alle esigenze degli studenti, garantendo chiarezza espositiva e coerenza concettuale. Questo processo permetterà di creare materiali didattici efficaci, che faciliteranno la comprensione degli argomenti trattati e supporteranno l'apprendimento degli studenti in modo dinamico e personalizzato.
Inoltre, l'IA permette:

- La creazione automatizzata di quiz e test adattivi per valutare le conoscenze pregresse e personalizzare il percorso di apprendimento.
- La sintesi e trascrizione automatica di contenuti video, migliorando l'accessibilità per tutti gli studenti.
- L'elaborazione di materiali multimediali interattivi, come infografiche e video esplicativi generati automaticamente, per arricchire l'esperienza formativa.
- La traduzione automatica e l'adattamento linguistico dei contenuti, garantendo un accesso inclusivo alla formazione per un pubblico internazionale.

2. L'IA come alleato nella progettazione formativa

Al centro dell'innovazione proposta da MedSim Academy c'è Design4Practice. Si tratta di un GPT progettato per accompagnare docenti e formatori nella creazione di percorsi didattici digitali. La sua funzione va ben oltre quella di un semplice assistente tecnico: è un vero e proprio partner progettuale, capace di facilitare l'intero processo di ideazione dei corsi.
Grazie a questo strumento, è possibile costruire microcorsi brevi, dinamici, interattivi e soprattutto orientati alla pratica. Il GPT guida gli utenti nella definizione di obiettivi formativi chiari e misurabili, li aiuta a generare contenuti multimediali personalizzati – come video, quiz, checklist – e ad adattare l'esperienza formativa alle caratteristiche e ai bisogni di diversi profili di apprendimento.
Alla base di questo approccio innovativo c'è una riflessione profonda su come rendere

l'apprendimento realmente efficace nei contesti di simulazione medica, che ha portato alla definizione di tre domande:

- Come fare in modo che chi accede al centro di simulazione sia subito in grado di agire in modo sicuro e autonomo?
- Come offrire contenuti davvero utili al momento giusto, evitando il sovraccarico informativo?
- Come garantire che ogni risorsa sia sempre accessibile, personalizzabile e aggiornata?

A partire da queste domande, Design4Practice costruisce percorsi formativi agili, inclusivi e flessibili, sfruttando modelli pedagogici collaudati.
L'architettura di questo strumento si regge su cinque pilastri fondamentali, che ne definiscono l'efficacia e l'originalità:

1. Missione – Offrire supporto concreto a chi progetta corsi digitali brevi, modulari e pratici, pensati per un apprendimento immediatamente applicabile.
2. Approccio integrato – Il modello combina in modo sinergico diversi principi pedagogici: micro-learning per favorire la focalizzazione, Just In Time per l'immediatezza, Universal Design for Learning per l'inclusività e il framework ADDIE per una progettazione sistematica.
3. Funzionalità operative – Il GPT è in grado di generare diversi tipi di contenuto (video, quiz, FAQ, checklist, valutazioni) e supportare nella definizione di micro-credenziali.
4. Metodo di lavoro –lo strumento chiede informazioni su ambito, target e obiettivi, per aiutare, attraverso un'interazione guidata, nella progettazione.
5. Filosofia progettuale – Tutto ruota attorno alla concretezza, alla personalizzazione e al miglioramento continuo attraverso il feedback: i corsi non sono mai statici, ma si evolvono in base all'esperienza d'uso.

MedSim Academy rappresenta un modello evoluto di formazione in simulazione medica, in cui l'IA si configura come un catalizzatore per l'innovazione didattica, capace di coniugare personalizzazione, accessibilità e aggiornamento continuo in un ecosistema formativo scientificamente fondato.

3. Un esempio pratico

Design4Practice è stato preliminarmente testato nella progettazione di un modulo formativo a partire dal manuale di uno degli strumenti presenti nel Centro di Simulazione, il simulatore per pazienti cardiologici MW10 Cardiology Patient Simulator "K" ver.2.
Il GPT ha permesso di sviluppare un'intera micro-unità di apprendimento.
Grazie al supporto guidato, sono stati generati:

- uno storyboard completo per due microvideo (setup tecnico e uso clinico), quiz di preconoscenza e valutazione finale con feedback formativi,
- una checklist operativa accessibile,
- una sezione di FAQ rapide,
- un learning design brief strutturato.

Il processo ha incluso l'allineamento agli obiettivi formativi secondo la Tassonomia di Bloom e ai Descrittori di Dublino, assicurando coerenza con standard universitari e professionali (ESCO). Lo scambio con l'IA ha facilitato la rapidità nella creazione di contenuti, l'attenzione all'accessibilità e la trasformazione di contenuti tecnici in esperienze formative brevi e immediatamente applicabili.

4. Prospettive Future

MedSim Academy si propone di diventare un punto di riferimento nell'utilizzo dell'IA per la creazione di contenuti didattici nella formazione medica. Il percorso formativo, attualmente in fase di progettazione e sviluppo, mira a migliorare l'efficacia dell'apprendimento e a ottimizzare l'uso delle risorse disponibili. L'integrazione di tecnologie basate su IA consentirà di sviluppare materiali più interattivi, aggiornati e personalizzati, offrendo agli studenti un'esperienza didattica innovativa e su misura.

Il progetto punterà al perfezionamento degli strumenti di generazione automatica dei contenuti, con l'obiettivo di affinare la qualità delle informazioni prodotte e garantire un aggiornamento costante dei materiali formativi. L'introduzione di strumenti di IA come Design4Practice nella formazione medica rappresenta un salto qualitativo verso modelli di apprendimento più personalizzati, scalabili e centrati sulle reali esigenze degli utenti. La collaborazione tra centri clinici, esperti di didattica digitale e sviluppatori di IA apre la strada a un ecosistema formativo che combina rigore scientifico e flessibilità operativa, rendendo la formazione continua una prassi concreta e sostenibile.

Riferimenti bibliografici

Denny P., Khosravi H., Hellas A., Leinonen J., & Sarsa S. (2023). Can we trust AI-generated educational content? comparative analysis of human and AI-generated learning resources. arXiv preprint arXiv:2306.10509.

Leiker D., Gyllen A. R., Eldesouky, I. & Cukurova, M. (2023). Generative AI for learning: Investigating the potential of synthetic learning videos. arXiv preprint arXiv:2304.03784.

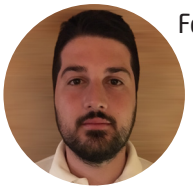
Ma Y., Song Y., Balch, J. A. Ren, Y. Vellanki, D. Hu, Z., ... & Shickel, B. (2024). Promoting AI competencies for medical students: a scoping review on frameworks, programs, and tools. arXiv preprint arXiv:2407.18939.

Sitografia

https://mondodigitale.aicanet.it/lutilizzo-dellintelligenza-artificiale-nellinsegnamento-a-medicina/

https://www.agendadigitale.eu/scuola-digitale/apprendimento-personalizzato-con-lai-ecco-gli-strumenti-piu-utili/

Autori



Federico Siracusa federico.siracusa@unito.it

Federico Siracusa è IT Specialist presso la Direzione Sistemi Informativi dell'Università di Torino dal 2018. Opera nella sezione ICT del Polo di Medicina S. Luigi, dove gestisce l'infra-

struttura informatica curandone in particolare gli aspetti legati a sicurezza, continuità operativa e automazione. Si occupa di monitoraggio, gestione centralizzata dei sistemi e digitalizzazione dei servizi in ambito didattico e clinico.

Floriana Vindigni floriana.vindigni@unito.it

Instructional Designer presso l'Università di Torino, ha una consolidata esperienza nella progettazione didattica digitale, nell'innovazione metodologica e nella formazione su tecnologie educative. Dopo un dottorato in Chimica e anni di ricerca, oggi si occupa di e-learning, innovazione educativa, IA applicata alla didattica, accessibilità e progettazione di contenuti multimediali. Ha partecipato a progetti nazionali ed europei ed è relatrice in conferenze internazionali su didattica e tecnologie.

Federico Abate Daga federico.abatedaga@unito.it

Tecnico della ricerca presso il Dipartimento di Scienze Cliniche e Biologiche dell'Università di Torino. Si occupa di simulazione medica applicata alla formazione sanitaria e di metodologia dell'attività fisica adattata per persone con patologie croniche.

Elisabetta Galoppini elisabetta.galoppini@unito.it

Laureata in Lettere antiche e con una Laurea magistrale in Scienze Linguistiche, ha collaborato per anni con l'Università di Torino nel campo dell'e-learning. Dal 2023 è entrata a far parte della Direzione Sistemi informativi Portale E-learning e svolge la propria attività presso gli uffici Web ed E-learning del Polo di Medicina Orbassano e Candiolo. In sinergia con il corso di laurea in Medicine and Surgery, ha partecipato a diversi progetti riguardanti la simulazione medica avanzata.

Vito Moscato vito.moscato@unito.it

Vito Moscato è capo area presso l'Università di Torino, nell'Area Servizi ICT, Web e-learning dei poli di Medicina di Orbassano e Candiolo. Dal 2006 coordina sessioni sperimentali con piattaforme informatiche per studi sul comportamento economico. Referente per l'adozione di soluzioni digitali a supporto della didattica e della ricerca, garantisce servizi ICT integrati alla comunità universitaria.

David Lembo david.lembo@unito.it

David Lembo è Professore Ordinario di Microbiologia presso l'Università di Torino, dove dirige il Laboratorio di Virologia Molecolare. Laureato con lode in Scienze Biologiche, Dottore di ricerca in Microbiologia Medica (Pisa) e post-doc alla Hoffmann-La Roche, ha ideato saggi antivirali brevettati e fondato startup biotech. Autore di oltre 120 pubblicazioni e due manuali, è membro EIC e docente in master internazionali.

DOI: 10.26314/GARR-Conf25-proceedings-07

ARGOS: A Retrieval-augmented GeneratiOn approach for Scientific communication

Daniele Di Bella[1,2], Pietro Roversi[1,2]

[1]Consiglio Nazionale delle Ricerche–Istituto di Biologia e Biotecnologie Agrarie (CNR–IBBA), [2]Fondazione Telethon

Abstract. Effectively communicating biological research to non-specialist audiences remains a critical challenge. Within the Broad-Spectrum Rescue-of-Secretion project, we want to explore the potential of Retrieval-Augmented Generation (RAG) in life sciences communication. We hence developed ARGOS, a Python-based pipeline leveraging OpenAI's GPT-4.1 combined with bibliographic retrieval from Zotero libraries, to generate Wikipedia-style summaries tailored for diverse audiences. Expert evaluations of ARGOS-generated texts in English and Italian showed high scores for correctness and readability, though completeness was somewhat limited by dataset scope and prompt design. Overall, ARGOS proved to be a good instrument to conduct further studies

Keywords. large language models, retrieval-augmented generation, scientific communication, public outreach, rare diseases

Introduction

Effectively communicating biomedical research to non-specialist audiences—particularly patients affected by rare diseases—remains a critical yet insufficiently addressed challenge in contemporary science communication. One such effort was undertaken within the framework of the two-year project Broad-Spectrum Rescue-of-Secretion of Tdark Glycoprotein Mutants, funded by the Telethon Foundation1. The project investigates whether modulating the endoplasmic reticulum (ER) quality control enzyme UGGT can promote the secretion of misfolded yet functional ("responsive") glycoprotein mutants. These mutants are implicated in rare congenital diseases, and the approach aims to evaluate UGGT inhibition or deletion as a broad-spectrum therapeutic strategy, with particular focus on poorly characterized Tdark glycoproteins. The project also considers potential cellular risks associated with targeting such a central checkpoint in the ER quality control system. To support public engagement with the research, our team initiated a science communication effort focused on improving access to information about the ten Tdark glycoproteins studied in the project. Specifically, we are undertaking the creation of Wikipedia entries for each protein to provide accurate, accessible, and broadly disseminated summaries of existing knowledge for affected individuals and the wider public.

This initiative also presented an opportunity to investigate the potential of recent advances in natural language processing to support the dissemination of life sciences research. In fact, Large language models (LLMs), which are capable of generating fluent, gramma-

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

tically correct text in multiple languages, show promise in this domain, even though they are known to generate "hallucinations"—plausible but incorrect statements. Retrieval-Augmented Generation (RAG) addresses this limitation by combining LLMs with information retrieval systems that guide generation using external, user-specified sources, thus improving factual reliability (Lewis et al., 2021).

Wanting to assess whether RAG-based systems can meaningfully enhance life sciences communication, we developed ARGOS (A Retrieval-augmented GeneratiOn approach for Scientific communication), which we present in this contribution. ARGOS generates Wikipedia-style summaries using bibliographic sources retrieved from a user's Zotero library, and it is not an end in itself. It serves as an experimental tool to explore the broader question of how RAG architectures might improve the experience of public facing biomedical content.

In the following sections, we present the reasons that led us to think RAG tools can be beneficial for science communication, their inherent contradictions, ARGOS' workflow, and an initial assessment of its performance.

1. Intentions and contradictions

1.1 Intentions

Despite the rapid pace of discovery in fields such as biology, researchers often lack both the time and incentives to engage in public outreach (Nerlich 2017). Moreover, those who attempt to do so frequently encounter peer pressure and professional disincentives, discouraging sustained communication efforts (Rose et al. 2020). These systemic barriers contribute to a widening gap between the scientific community and the broader public.

Language further compounds this divide. English remains the dominant language of science communication, restricting access for non-English-speaking populations and reinforcing a singular cultural perspective in the interpretation and dissemination of knowledge (Márquez and Porras 2020). Consequently, scientific knowledge often remains both linguistically and culturally inaccessible to large segments of the global population and, together with the aforementioned issues and the intrinsic difficulty of scientific matters, exacerbates that separation between the scientific community and the general public that often leads them to perceive each other not as entities in continuity, but as different entities.

Providing tools that enable scientists to share their work more easily, accurately, and inclusively could foster the ongoing process of shaping the scientific community's social vocation, that sees it in dialogue with the rest of the global community, of which it is an integral part. RAG applications offer a promising solution for that. Such systems can assist researchers in creating accurate, accessible summaries of their work across multiple languages and cultural contexts.

1.2 Contradictions

Although the use of RAG has been proposed as a promising strategy for democratizing access to scientific knowledge, this approach presents inherent tensions, particularly con-

cerning linguistic equity and cultural representation. For instance, ARGOS relies on OpenAI's GPT-4.1 model, yet, the training data for GPT models are unevenly distributed in favour of English-language sources, limiting their performance in generating content in other languages and potentially introducing cultural bias into the output.
At present, ARGOS uses two standardized English-language prompts to generate both English and non-English outputs. The first prompt instructs the model on the desired output language and audience, and requests it to tailor the output text to the relevant cultural context, while the second orders the model to proofread what it has produced. This method highlights a paradox: a tool with linguistic and cultural biases is being used to address the very inequities it may perpetuate.
Unfortunately, due to the computational and financial costs associated with training LLMs from scratch, most developers—including our team—must rely on pre-trained models. Consequently, we are constrained by the design choices and implicit biases embedded by those who created such models.
Moreover, while AI technologies may appear low-cost, their affordability is relative, often excluding users in low- and middle-income countries (LMICs). This raises a concern: can reducing disparities be attempted through the use of technologies that are themselves products of those disparities? Encouragingly, a growing number of initiatives in LMICs are developing language models tailored to local non-European languages, social contexts, and computational constraints2. These efforts open new possibilities for the use of RAG systems in scientific communication.

2. ARGOS workflow and first validations

Currently ARGOS is built around a Python pipeline that can be called through a command line interface, and that follows the workflow depicted in Fig. 1.

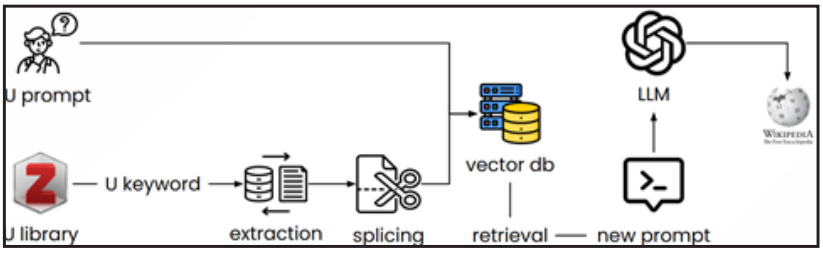


Fig. 1
ARGOS
workflow

Firstly, the user provides a prompt (U prompt) and some keywords (U keyword). The keywords are used to browse the user's Zotero library in search of items that are extracted and spliced in chunks. Each chunk is vectorized through OpenAI's text-embedding-3-large model and stored in a vector database. U prompt is vectorized as well and used to launch a similarity search that selects the chunks more likely to contain information related to the user's request. These chunks are then used to create a new prompt which is submitted to the LLM in charge to generate the output text (in our case, GPT-4.1). Before providing this output to the user, the same LLM is asked to correct any error according to the grammar and syntax of the user's desired language, and to adapt the tone and the style of the parts

of the text that aren't meeting the communication needs of the user's selected audience.
To validate the application, we decided to generate 10 texts, 5 in English and 5 in Italian, about 5 of the 10 glycoproteins of our interest, and submit them to three colleagues, experts in these three proteins. For each text, we asked the colleagues to rank its scientific accuracy (correctness), the presence of relevant information about the protein described (completeness) and their reading experience (readability) on a scale ranging from 1 to 10. Results are presented in Fig. 2.

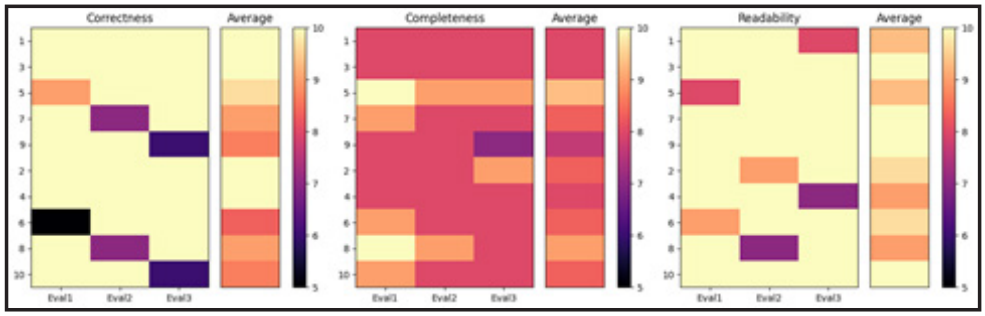


Fig. 2
First validation of ARGOS

The numbers labelling the rows of the matrices are the texts' identification codes: even numbers indicate Italian texts, while odd numbers are for English texts. In all three cases, 1 and 10 indicate very bad and very good ratings, respectively.
As can be seen from Fig.2, the correctness and readability of the texts are generally highly ranked. Both parameters present very small differences that identify Italian texts as slightly less correct and readable, which, for the readability, may be explained by the worse performances of GPT models in non-English languages. However, given the small size of the sample of experts, such differences are likely non-meaningful fluctuations.
On the other hand, while completeness as well is highly scored, it is visibly worse, and this could be due to our inexperience on the proteins ARGOS wrote about. In fact, during the creation of the Zotero datasets, we may have omitted some papers that the expert colleagues considered of primary importance. Moreover, our inexperience may have led us to formulate questions differently than the experts would have done. In fact, as underlined in the interesting contribution of Wong and colleagues (Wong et al. 2025), a key feature of RAG systems is the dependence on the user's prompt. This element leads such systems to select particular information from nonparametric memory (i.e., from the sources to which they are given access) and it is dependent on the user's starting beliefs, which is why Wong and colleagues warn against this feature of RAG systems. We probably approached ARGOS believing that the 5 selected proteins could be described by some features, while experts in those molecules would choose others, and this led the system to respond in a way that was assessed as not complete.
Overall, ARGOS produced texts that passed the experts' evaluation positively and even allowed us to find answers about a protein of our interest that had been sought for some time. We consider it sufficiently good to be used in the next steps of our project, which may include experiments with other RAG systems and broad groups of annotators.

References

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021), Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, arXiv, (arXiv:2005.11401). https://doi.org/10.48550/arXiv.2005.11401

Márquez, M. C., & Porras, A. M. (2020), Science Communication in Multiple Languages Is Critical to Its Effectiveness. Frontiers in Communication, (5). https://doi.org/10.3389/fcomm.2020.00031

Nerlich, B. (2017), Time and science communication. Making Science Public. https://blogs.nottingham.ac.uk/makingsciencepublic/2017/04/07/time-science-communication/

Rose, K. M., Markowitz, E. M., & Brossard, D. (2020), Scientists' incentives and attitudes toward public communication, Proceedings of the National Academy of Sciences, 117(3), pp 1274–1276. https://doi.org/10.1073/pnas.1916740117

Wong, L., Ali, A., Xiong, R., Shen, S. Z., Kim, Y., & Agrawal, M. (2025), Retrieval-augmented systems can be dangerous medical communicators, arXiv. https://doi.org/10.48550/ARXIV.2502.14898

Links

1https://www.fondazionetelethon.it/en/what-we-do/research/projects-funded/broad-spectrum-rescue-of-secretion-of-tdark-glycoprotein-mutants/
2https://www.nature.com/articles/d41586-025-01546-6

Autori

Daniele Di Bella daniele.dibella@ibba.cnr.it
Daniele Di Bella is a computational biologist at the IBBA-CNR Institute in Milan. From March 2023 to March 2024 he worked on his thesis at the Alfred Wegener Institute of Bremerhaven, Germany. After his graduation at the University of Milan, in 2024, he started working as a research fellow at IBBA-CNR, where he focuses on AI and bioinformatics.

Pietro Roversi pietro.roversi@cnr.it
Pietro Roversi is a structural biologist at the IBBA-CNR Institute in Milan. From 1996 to 2021, he worked in the UK at Cambridge, Oxford and Leicester. Since 2012 he leads a research project focusing on the potential of secretion-rescue strategies for the therapy of congenital rare disease due to a responsive missense mutation in a secreted glycoprotein gene.

DOI: 10.26314/GARR-Conf25-proceedings-08

CYMEDSEC: Cybersecurity for Medical Infrastructures to remove the barriers in emerging digital technologies

Francesco Ricciardi, Michela Falcone, Francesco Giuliani

Innovation and Research Unit, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo

Abstract. The CYMEDSEC project addresses the growing cybersecurity challenges of digital healthcare in Europe. As aging populations increase service demand and healthcare workforce shrinks, digital solutions like connected medical devices (IoMT) become essential. However, these bring cybersecurity risks. CYMEDSEC develops security-by-design frameworks, benefit-risk tools, and ethical guidelines to support safe, resilient, and compliant IoMT systems. With pilots in two hospitals, the project advances Europe's strategic autonomy and fosters trust in digital health through cross-sector collaboration, innovation, and regulatory alignment.

Keywords. Cybersecurity, IoMT, Regulation, Benefit-risk analysis

Introduction

Projections from Eurostat indicate that European Union will face a demographic decline, with a reduction of the population of 6.1% by 2100 (EUROPOP2023, 2023). This will result in an absolute reduction of the number of healthcare professionals along with an increase of the population over 65. Digitalization could be one of the ways to face the shortage of healthcare operators. Digitalization could help mitigate this shortfall by enabling alternatives to hospital admission and facilitating earlier discharge (Gilbert et al., 2024). As healthcare infrastructures become increasingly digitized, the security of digital and connected medical devices and systems is a pressing concern with broad social, economic, and geopolitical implications (Longras et al., 2023). This security gap could hinder the adoption of digital diagnostic and therapeutic innovations in a world with a growing demand of digital health services (Coventry and Branley, 2018).

The European research project CYMEDSEC (Cybersecurity for Medical Devices and Connected Healthcare Systems, GA 101094218, https://cymedsec.eu/) is addressing this challenge by developing robust security frameworks for Internet of Medical Things (IoMT) technologies. By integrating cybersecurity-by-design principles, risk assessment tools, and ethical considerations, CYMEDSEC seeks to protect patient data, ensure system resilience, and support regulatory compliance across Europe. Under this light, the CYMEDSEC project contributes to remove cybersecurity barriers from emerging digital technologies.

1. Approaching Cybersecurity in Medical Infrastructures

Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

Modern healthcare systems heavily rely on digital infrastructures, including electronic health records (EHRs), telemedicine platforms, and AI-driven diagnostics. IoMT devices, such as connected pacemakers, insulin pumps, and remote monitoring sensors, enhance patient care but also introduce cybersecurity vulnerabilities. These technologies are susceptible to cyber threats, including ransomware attacks, data breaches, and system disruptions that can have life-threatening consequences. Safety and security are both concerns in a connected healthcare world (Skierka, 2018).

The complexity of securing IoMT in Europe is exacerbated by regulatory fragmentation, varying levels of cybersecurity awareness, and the rapid evolution of cyber threats. Healthcare providers, technology manufacturers, and policymakers must work together to develop and implement effective security solutions that balance innovation with patient safety.

The CYMEDSEC research project adopts a multi-faceted approach to strengthening cybersecurity in medical infrastructures. The project focuses on three key pillars: Cybersecurity by Design, Benefit-Risk Toolbox for IoMT Security, and Societal and Ethical Considerations. CYMEDSEC promotes the integration of security mechanisms from the earliest stages of medical device development. This includes implementing encryption standards, secure authentication protocols, and real-time anomaly detection to protect against cyberattacks. By embedding security at the design level, the project aims to reduce vulnerabilities before devices are deployed in healthcare environments. At the same time, the project will provide a comprehensive framework for assessing the risks and benefits of cybersecurity interventions in IoMT. This toolbox supports stakeholders in making informed decisions by evaluating the trade-offs between security measures, usability, and system performance. It also facilitates compliance with EU regulations, including the Medical Device Regulation (MDR) and the Network and Information Security Directive (NIS2). Beyond technical solutions, CYMEDSEC addresses ethical and legal challenges associated with cybersecurity in healthcare. Patient data privacy, informed consent, and the ethical implications of automated decision-making are central to the project's objectives. By incorporating stakeholder perspectives, including those of patients, clinicians, and regulators, CYMEDSEC ensures that security policies align with societal values and public trust. The outputs of the project will be validated in two hospitals in a one-year long pilot. Casa Sollievo della Sofferenza is in charge of the coordination of the project case studies.

2. Results and Impacts

The CYMEDSEC project started in November 2023 and will last 4 years. Up to now a systematic review of both the best practices in cybersecurity of connected medical devices as well as the risks and gaps in current Benefit-Risk Analysis approaches has been completed. An analysis of the legal and ethical challenges that are of relevance for the project, a review of the fleet management challenge in the world of connected IoMT and the preparation of the demonstrators that will be showcased in the case studies have been already delivered. The preparation of a minimised and hardened system architecture for a highly secure gateway is under development while, a completed generalized attack model will guide the

upcoming case studies activities.

Cybersecurity in healthcare extends beyond technical risks. Side by side with ethical and privacy considerations there are significant geopolitical and economic ramifications. Cyberattacks on healthcare institutions can disrupt essential services, compromise sensitive patient information, and erode public confidence in digital health solutions. In a globally connected world, weak cybersecurity in one region can have cascading effects, making international collaboration essential. CYMEDSEC contributes to Europe's strategic autonomy in cybersecurity by fostering innovation in security technologies, enhancing regulatory coherence, and promoting best practices across borders.

3. Future Challenges and Research Directions

While CYMEDSEC represents a significant step forward, several challenges remain. The rise of AI-driven cyber threats, the complexity of securing legacy medical systems, and the ethical implications of continuous patient monitoring require ongoing research and adaptation. Future efforts must also focus on harmonizing cybersecurity standards across global healthcare networks and developing dynamic security models capable of responding to emerging threats in real-time. Interoperability in healthcare will be also a topic to address in the near future in light of cybersecurity considerations.

4. Conclusions

Digitalization and connected medical devices could represent the future in a scenario where the healthcare will be characterized by a reduction of healthcare workforce and in an increasing demand of services. The intersection of medical infrastructures and digital innovations presents both opportunities and challenges. CYMEDSEC project plays a crucial role in addressing these challenges by advancing cybersecurity frameworks that protect patients, healthcare providers, and broader society. As digital healthcare continues to evolve, collaborative efforts between researchers, policymakers, and industry leaders will be essential in ensuring that technological advancements do not come at the cost of safety, security, or trust. Through projects like CYMEDSEC, Europe is leading the way in shaping a resilient, secure, and ethically grounded digital healthcare ecosystem.

Bibliography

Coventry L., Branley D. (2018), "Cybersecurity in healthcare: A narrative review of trends, threats and ways forward," Maturitas, vol. 113, pp. 48-52

EUROPOP 2023, Population projections in the EU, https://ec.europa.eu/eurostat/statistics-explained/index.php?oldid=497115

Gilbert S., Ricciardi F., Mehrali T., Patsakis K. (2024), "Can we learn from an imagined ransomware attack on a hospital at home platform?", npj Digital Medicine 7, 65

Longras A., Pereira T., Amaral A. (2023), "Cybersecurity Challenges in Healthcare Medical Devices". In: Pereira, T., Impagliazzo, J., Santos, H. (eds), Internet of Everything, IoECon 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 458

Skierka I.M. (2018) "The governance of safety and security risks in connected healthcare," Living in the Internet of Things: Cybersecurity of the IoT - 2018, pp. 1-12, London

Fundings

This work was supported by the European Commission under the Horizon Europe Program, as part of the project CYMEDSEC (101094218). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

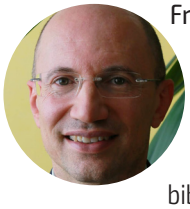
Authors

Francesco Ricciardi f.ricciardi@operapadrepio.it

Francesco Ricciardi is a senior executive engineer at the Innovation and Research Unit of IRCCS Casa Sollievo della Sofferenza. He is involved in the ideation and conduction of research projects in the fields of cybersecurity, healthy ageing and assistive robotics. He works also on digitalization initiatives and collaborates with the institution's management on organizational matters related to the outpatient services.

Michela Falcone m.falcone@operapadrepio.it

Michela Falcone is an automation engineer working at the Innovation and Research Unit at IRCCS Casa Sollievo della Sofferenza. She is involved in projects focused on assistive robotics, virtual reality, exoskeletons, and artificial intelligence applied to healthcare.



Francesco Giuliani f.giuliani@operapadrepio.it

Francesco Giuliani is the Head of the Innovation and Research Unit at Casa Sollievo della Sofferenza Research Hospital (IRCCS), where he leads a team of professionals working in the field of digital transformation of healthcare. The main areas of interest include assistive robotics, artificial intelligence, data analysis, eHealth, active and healthy aging, bibliometrics, science communication, virtual and augmented reality.

DOI: 10.26314/GARR-Conf25-proceedings-09

TrasparenzAI: piattaforma open-source per il monitoraggio della trasparenza amministrativa

Ivan Duca, Dario Elia, Claudia Greco, Massimo Ianigro, Cristian Lucchesi, Marco Spasiano

Consiglio Nazionale delle Ricerche

Abstract. TrasparenzAI è una piattaforma open source sviluppata da Consiglio Nazionale delle Ricerche (CNR) e l'Autorità Nazionale Anticorruzione (ANAC) per l'analisi e la consultazione della trasparenza amministrativa delle Pubbliche Amministrazioni italiane. Basata su architettura a microservizi, consente il monitoraggio automatizzato e sistematico delle sezioni "Amministrazione Trasparente" dei siti delle Pubbliche Amministrazioni, facilitando l'accesso, la verifica della conformità normativa e il supporto all'adeguamento degli standard di trasparenza

Keywords. web scraping, web crawling, open source, trasparenza amministrativa

Introduzione

La trasparenza amministrativa costituisce un pilastro essenziale per la fiducia nelle istituzioni pubbliche, favorendo la prevenzione della corruzione e una maggiore accountability. In Italia, il Decreto Legislativo 33/2013 impone alle Pubbliche Amministrazioni (PA) l'obbligo di pubblicare informazioni, dati e documenti nella sezione "Amministrazione Trasparente" dei loro siti web, con l'obiettivo di garantire un accesso semplice e immediato ai cittadini, agli operatori economici e agli organismi di controllo. Tuttavia, l'eterogeneità delle modalità di pubblicazione e la mancanza di standard uniformi hanno reso difficile sia la gestione delle informazioni da parte delle PA, sia la consultazione e il monitoraggio da parte degli utenti e degli enti di controllo.

In risposta a queste criticità, nasce il progetto TrasparenzAI, una piattaforma open source realizzata dal CNR in collaborazione con ANAC. TrasparenzAI si propone di semplificare e rendere più efficiente il monitoraggio e l'accesso alle informazioni pubblicate dalle PA, offrendo strumenti avanzati per l'analisi automatizzata e la verifica della conformità normativa.

La piattaforma è accessibile all'indirizzo: https://www.trasparenzai.it

1. Uniformità nella consultazione delle informazioni pubblicate dalle PA

TrasparenzAI garantisce l'uniformità nella consultazione delle informazioni pubblicate dalle Pubbliche Amministrazioni (PA) attraverso un sistema automatizzato e centralizzato che si pone l'obiettivo di superare la frammentazione delle modalità di pubblicazione adottate dai singoli enti. La piattaforma effettua una scansione sistematica delle sezioni

Conferenza GARR 2025 - Frontiere Digitali Selected Papers
Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

"Amministrazione Trasparente" dei siti delle PA, verificando la presenza, la struttura e la conformità delle informazioni secondo regole definite e configurabili.
Questo processo si basa su:

- Crawling e analisi automatizzata: TrasparenzAI utilizza crawler e strumenti di rendering per acquisire e analizzare i contenuti delle sezioni "Amministrazione Trasparente", indipendentemente dalla struttura tecnica adottata dal singolo ente.
- Regole di verifica uniformi: le regole di conformità sono centralizzate e definite tramite il Rule Service, che applica criteri omogenei a tutti i siti analizzati, garantendo così una valutazione standardizzata della trasparenza.
- Aggregazione e presentazione dei risultati: i risultati delle verifiche vengono aggregati e resi disponibili tramite dashboard, mappe e report, offrendo una visione comparabile e accessibile della situazione nazionale e facilitando la consultazione da parte di cittadini, operatori e organismi di controllo.
- Supporto all'adeguamento: la piattaforma non si limita a rilevare le difformità, ma fornisce anche indicazioni alle amministrazioni per adeguarsi agli standard richiesti, promuovendo così un progressivo allineamento delle pratiche di pubblicazione.
- In sintesi, TrasparenzAI riduce la disomogeneità delle informazioni pubblicate, offrendo un'interfaccia e strumenti di analisi che rendono la consultazione uniforme, efficiente e conforme agli obblighi normativi su scala nazionale.

2. Descrizione della piattaforma TrasparenzAI

2.1 Obiettivi e contesto normativo

La piattaforma TrasparenzAI nasce per rispondere alle esigenze di uniformità e controllo nell'applicazione degli obblighi di trasparenza. Il contesto normativo italiano, pur prescrivendo la pubblicazione di dati e documenti, non ha previsto un meccanismo centralizzato né standard tecnici vincolanti, generando una frammentazione delle soluzioni adottate dalle diverse amministrazioni. Ne consegue una notevole difficoltà sia per i cittadini, che devono orientarsi tra strutture e denominazioni diverse, sia per ANAC, che deve monitorare su scala nazionale il rispetto degli obblighi di pubblicazione.

2.2 Architettura tecnica

TrasparenzAI è basata su un'architettura modulare a microservizi, sviluppata principalmente in Java (Spring Boot) e Python (FastAPI), con una interfaccia frontend realizzata in TypeScript (Angular). Tutti i componenti sono rilasciati come software open source e comunicano tramite API REST. L'infrastruttura utilizza esclusivamente tecnologie open source, tra cui Keycloak per l'autenticazione OAuth2, PostgreSQL per la gestione dei dati, MinIO per lo storage S3-like, e Selenium Grid per il web-crawling parallelizzato delle pagine web.

2.3 Componenti principali

La piattaforma si compone di diversi microservizi, ciascuno con specifiche funzionalità:

- Public Sites Service: gestisce le informazioni sugli enti pubblici italiani, integrando

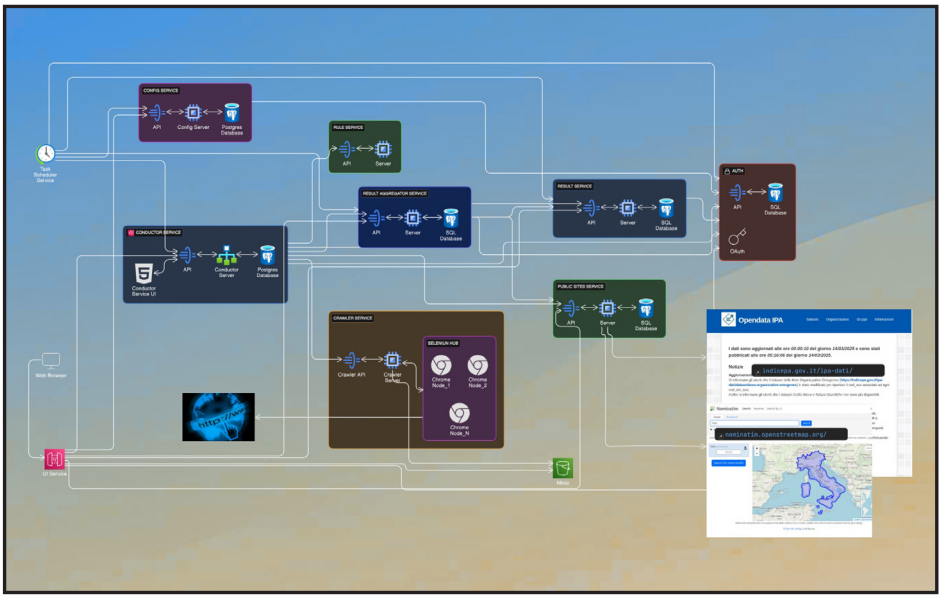


Fig. 1 Architettura della piattaforma TrasparenzAI

dati da IndicePA e ISTAT, e fornisce servizi di geolocalizzazione e consultazione tramite API REST.

- Config Service: permette la configurazione centralizzata dei parametri di sistema.
- Result Service e Result Aggregator Service: raccolgono e aggregano i risultati delle scansioni e delle verifiche effettuate sui siti delle PA.
- Task Scheduler Service: pianifica e avvia periodicamente le scansioni dei siti, coordinando i flussi di lavoro eseguiti tramite il servizio Conductor.
- Rule Service: gestisce le regole di verifica della conformità normativa, configurabili tramite file JSON.
- Conductor Service: coordina i workflow di analisi e verifica, garantendo flessibilità e adattabilità a evoluzioni future.
- UI Service: offre un'interfaccia web intuitiva per la consultazione dei dati, la visualizzazione delle mappe e dei grafici, e la gestione delle attività amministrative.
- Web Scraping Service: riceve dalla piattaforma gli URL per i quali effettuare lo scraping, acquisisce la pagina simulando una sessione browser e restituisce sia i contenuti codificati BASE64 che lo screenshot grafico della stessa. Il trasferimento delle informazioni agli altri microservizi avviene attraverso l'inserimento nell'object store MinIO.

2.4 Funzionalità e flusso operativo

Il funzionamento della piattaforma si basa su un ciclo automatizzato di scansione e verifica:

1. Il Task Scheduler Service avvia, con cadenza configurabile, la scansione dei siti delle PA.
2. Il Public Sites Service fornisce l'elenco aggiornato degli enti e dei relativi siti web.
3. Il Conductor Service coordina i workflow di crawling, analisi e salvataggio dei risultati.

Le pagine HTML vengono acquisite da un insieme di crawler 'firefox-based' che lavorano in parallelo all'interno di un framework basato su Selenium Grid per garantire scalabilità e velocità nell'esecuzione dello scraping dai vari siti.
Le regole di verifica definite nel Rule Service controllano la presenza, la struttura e la conformità delle sezioni "Amministrazione Trasparente" rispetto alla normativa vigente, dando la possibilità, attraverso l'interfaccia web, di definire dinamicamente il comportamento.
I risultati delle verifiche vengono aggregati e resi disponibili tramite dashboard, mappe geografiche e report esportabili.
Al momento, la piattaforma è in grado di operare un'analisi di 23.663 siti web, verificando in automatico la presenza di tutte le sezioni previste dal D.Lgs. 33/2013 in circa 20 ore.

2.5 Sicurezza e accesso

L'accesso alle funzionalità della piattaforma è regolato tramite autenticazione e autorizzazione basate su OAuth2, con ruoli predefiniti gestiti tramite token JWT. Ogni microservizio espone endpoint protetti, garantendo la sicurezza dei dati e delle operazioni.

2.6 Interfaccia utente e strumenti di consultazione

TrasparenzAI offre una interfaccia web ricca e intuitiva, che consente:

- Ricerca e consultazione delle amministrazioni monitorate.
- Visualizzazione delle mappe geografiche delle PA italiane.
- Analisi grafica dei risultati per regola e per ente.
- Accesso allo storico dei controlli e alle evidenze delle scansioni.
- Esplorazione delle sezioni e delle regole applicate.
- Gestione delle autorizzazioni e delle configurazioni di sistema.

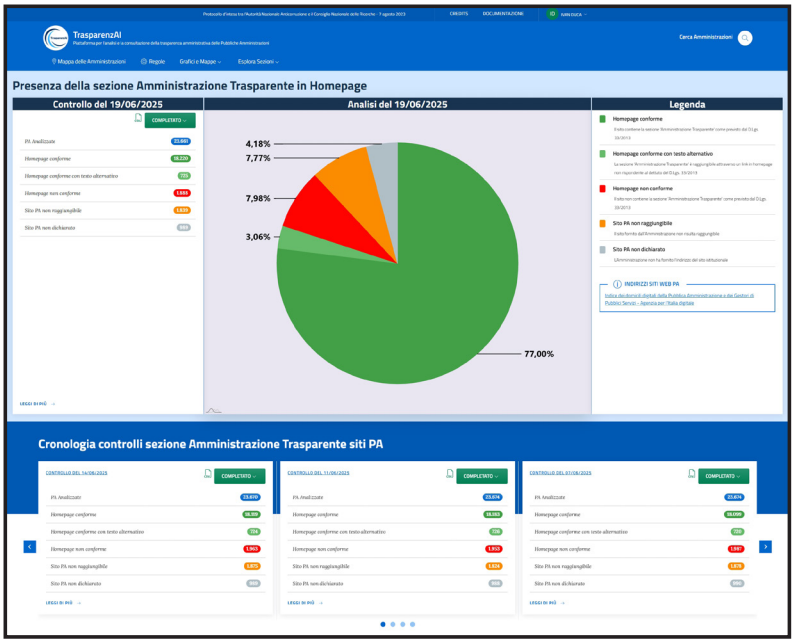


Fig. 2 Homepage con profilo Utente Avanzato

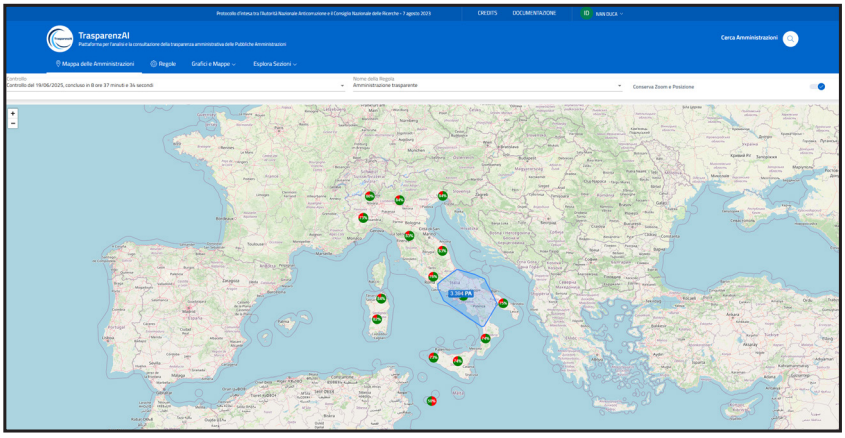


Fig. 3 Visualizzazione geografica per macro-aree dei risultati di scansione

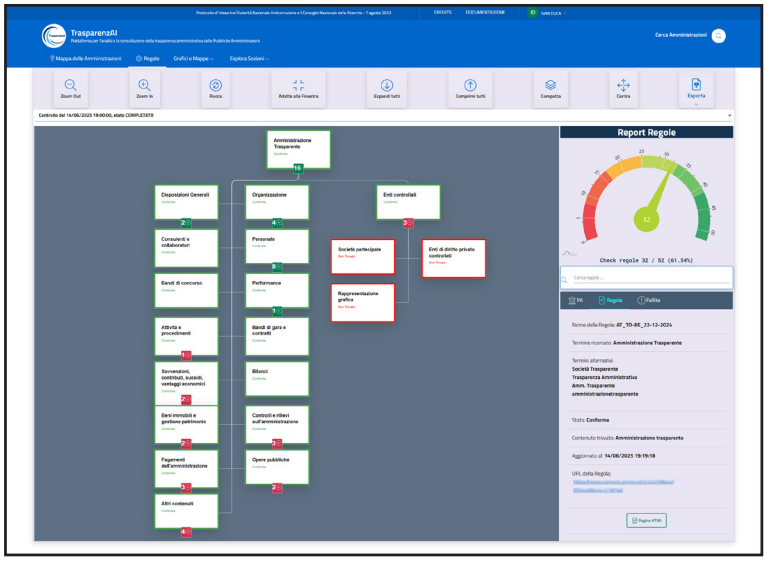


Fig. 4 Visualizzazione dei risultati di scansione per singola Amministrazione

2.7 Riusabilità e apertura

Uno degli aspetti distintivi di TrasparenzAI è la sua piena apertura e riusabilità: il codice sorgente è pubblicato su GitHub (https://github.com/trasparenzai) e i componenti sono progettati per essere adattabili a progetti analoghi di crawling, analisi e monitoraggio di siti web pubblici. La documentazione tecnica e amministrativa è disponibile pubblicamente, favorendo la diffusione e l'adozione della piattaforma da parte di altre amministrazioni e della comunità open source.

3. Sviluppi futuri

L'evoluzione futura della piattaforma TrasparenzAI prevede l'integrazione progressiva di intelligenza artificiale (AI) a supporto del controllo qualitativo dei contenuti pubblicati dalle pubbliche amministrazioni. L'obiettivo è superare l'attuale livello di verifica formale della struttura dei siti web istituzionali, ampliando le capacità della piattaforma verso un'analisi più concreta, sostanziale e contestuale delle informazioni rese disponibili online.

4. Conclusioni

TrasparenzAI rappresenta un significativo passo avanti nell'innovazione digitale della pubblica amministrazione italiana, offrendo una soluzione tecnologica avanzata per il monitoraggio e la promozione della trasparenza amministrativa. La piattaforma affronta in modo sistematico le criticità derivanti dall'eterogeneità delle modalità di pubblicazione, semplificando l'accesso alle informazioni e supportando le amministrazioni nell'adeguamento agli standard normativi.

L'approccio modulare, l'utilizzo esclusivo di tecnologie open source e la piena apertura del codice garantiscono scalabilità, adattabilità e sostenibilità nel tempo. La collaborazione tra CNR e ANAC testimonia l'importanza di sinergie istituzionali per la costruzione di strumenti efficaci a servizio della collettività.

In prospettiva, TrasparenzAI pone le basi per una nuova stagione della trasparenza, in cui il controllo e la vigilanza sulle attività delle PA possano essere esercitati in modo efficiente, flessibile e orientato all'evoluzione digitale, rafforzando la fiducia dei cittadini nelle istituzioni e promuovendo una cultura della legalità e dell'accountability.

Autori

Ivan Duca ivan.duca@cnr.it

Informatico del Consiglio Nazionale delle Ricerche con oltre trent'anni di esperienza nello sviluppo di infrastrutture digitali e progetti di ricerca ICT. Ha collaborato con Commissioni parlamentari e autorità indipendenti per la valorizzazione documentale e il contrasto alla criminalità, contribuendo a iniziative nazionali su trasparenza, strumenti digitali per il contrasto alle infiltrazioni criminali e la corruzione, digitalizzazione della PA e innovazione tecnologica.

Dario Elia dario.elia@cnr.it

Consulente parlamentare e dipendente Consiglio Nazionale delle Ricerche, ha coordinato progetti su temi di strumenti innovativi per il contrasto alle infiltrazioni criminali e la corruzione per la Commissione parlamentare antimafia. Esperto in sinergie istituzionali, ha guidato tavoli tecnici con INPS sui fenomeni socio-economici e partecipato a progetti per la valorizzazione dei patrimoni documentali storici delle Commissioni parlamentari e la trasparenza nelle PA.

Claudia Greco claudia.greco@cnr.it

Avvocato, segue un percorso accademico post-laurea orientato al management pubblico e il diritto dell'informatica. Nel Consiglio Nazionale delle Ricerche ha ricoperto incarichi nel campo della protezione dei dati personali, ha partecipato a progetti di ricerca ed attività di formazione, focalizzandosi sulla sicurezza dei dati personali e contribuendo a sviluppare tecnologie innovative e formare personale su normative specifiche.

Massimo Ianigro massimo.ianigro@cnr.it

Primo tecnologo del Consiglio Nazionale delle Ricerche, è laureato in Informatica e si occupa di ICT dal 1992. Responsabile dei servizi telematici del CNR di Bari, si occupa di

cybersecurity, forensics e privacy. Ha collaborato con GARR e forze dell'ordine, ha brevetti e pubblicazioni in vari ambiti (robotica, image processing, cybersecurity) e ha partecipato a progetti europei e nazionali oltre a occuparsi di docenze e valutazione progetti. Ha contribuito allo sviluppo di piattaforme per l'Antimafia e l'ANAC.

Cristian Lucchesi cristian.lucchesi@cnr.it

Laureato in informatica all'Università di Pisa. Dal 2002 lavora per l'Istituto di Informatica e Telematica del Consiglio Nazionale delle Ricerche dove ricopre il ruolo di Responsabile dell'Unità Servizi e Sviluppo applicazioni IIT. I suoi principali interessi scientifici sono nelle aree dei sistemi informativi, architetture distribuite, continuous integration and deployment, sviluppo software, opensource, ingegneria del software. La sua passione lo ha sempre portato ad utilizzare strumenti open source ed essere parte attiva della community, anche con software scritti e rilasciati open source da lui.

Marco Spasiano marco.spasiano@cnr.it

Primo tecnologo del Consiglio Nazionale delle Ricerche, dove ricopre il ruolo di responsabile della Sezione Sviluppo Software dell'Ufficio Agenda Digitale dell'Amministrazione Centrale. Ha una lunga esperienza nel campo del software e ha contribuito, guidato e sviluppato progetti open source per la Gestione della Contabilità e nell'ambito dell'E-Government e della Dematerializzazione.

DOI: 10.26314/GARR-Conf25-proceedings-10

Autenticazione Sicura e Moderna: MFA e Passkey nella Migrazione a Shibboleth IdP 5

Andrea Garzena[1], Federico Cucinella[2]

[1]Politecnico di Torino, [2]Politecnico di Torino, Present S.p.A.

Abstract. Il Politecnico di Torino ha intrapreso un progetto di migrazione alla versione 5 di Shibboleth Identity Provider per rispondere alle nuove esigenze di sicurezza, interoperabilità e scalabilità nella gestione delle identità digitali. Il nuovo sistema introduce il supporto nativo all'autenticazione multifattoriale (MFA) tramite eduMFA e metodi passwordless, in linea con gli standard SAML2, FIDO2 ed eIDAS. L'infrastruttura, containerizzata con Docker e orchestrata in ambiente ad alta affidabilità, prevede l'impiego di più nodi interconnessi per migliorare la scalabilità. Un set di applicativi sviluppati in-house consente la gestione delle sessioni, dei fattori MFA e delle credenziali utente, offrendo un'esperienza coerente e sicura. L'intervento documenta le scelte tecnologiche e architetturali adottate, proponendo un caso concreto utile ad altri enti operanti in contesti federati

Keywords. FedOps, IAM, Shibboleth, Passkey, MFA

Introduzione

Nel contesto accademico attuale, una gestione sicura e interoperabile delle identità digitali è fon-damentale per garantire l'accesso protetto ai servizi universitari. Il Politecnico di Torino, in linea con le migliori pratiche e le esigenze di cybersicurezza, ha avviato un aggiornamento completo del proprio Identity Provider (IdP), basato su Shibboleth.

La precedente infrastruttura, sebbene affidabile, non soddisfaceva più i requisiti moderni come l'autenticazione multifattoriale, l'integrazione con identità federate e sistemi nazionali, o l'uso di metodi passwordless (es. Passkey). L'aumento dei servizi digitali e della varietà di utenti richiede-va una soluzione aggiornata, conforme agli standard attuali (SAML2, FIDO2, eIDAS), capace di ga-rantire sicurezza e continuità.

Questo lavoro descrive la migrazione a Shibboleth IdP v5, le tecnologie impiegate e le strategie per assicurare un passaggio fluido per gli utenti. L'obiettivo è offrire un esempio concreto utile an-che ad altre istituzioni operanti in contesti federati.

1. Tecnologie Utilizzate

La nuova infrastruttura dell'Identity Provider (IdP) del Politecnico di Torino è stata progettata per garantire modularità, scalabilità, affidabilità e facilità di manutenzione. La scelta delle tecnologie è stata guidata dalla necessità di adottare strumenti consolidati nel panorama open source, facil-mente integrabili tra loro e compatibili con le specifiche del protocollo SAML2.

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

1.1 Shibboleth Identity Provider

Il nucleo del sistema è rappresentato da Shibboleth Identity Provider v5, la cui esecuzione è affi-data a un container Docker che integra Apache Tomcat come servlet container. Questa configura-zione consente una gestione più efficiente del ciclo di vita dell'applicazione (build e deploy grazie ad un approccio infrastructure-as-code), semplifica la scalabilità orizzontale e migliora la portabi-lità tra ambienti di sviluppo, staging e produzione. L'intera configurazione è orchestrata in modo da rispettare i requisiti di alta affidabilità, con più istanze del container esposte dietro un bilan-ciatore di carico.

L'uso di Docker ha facilitato anche l'automazione della configurazione e l'orchestrazione futura in contesti a maggiore complessità, oltre a semplificare l'aggiornamento e la gestione delle dipen-denze.

1.2 eduMFA e fudiscr: gestione MFA

Come backend dell'autenticazione multifattoriale è stato adottato eduMFA, un progetto open source sviluppato specificamente per il contesto universitario, derivato da PrivacyIDEA. eduMFA supporta metodi di autenticazione moderni, tra cui TOTP, WebAuthn e Passkey, ed è stato integra-to direttamente nel flusso SAML di Shibboleth IdP.

La comunicazione tra Shibboleth IdP ed eduMFA è stata resa possibile grazie all'impiego del plu-gin fudiscr, sviluppato per Shibboleth e integrato nativamente con eduMFA. Questo componente consente a Shibboleth di demandare l'intero flusso di autenticazione MFA a eduMFA, ricevendo in risposta l'esito dell'autenticazione avanzata. fudiscr opera come interfaccia tra IdP e sistema MFA, garantendo flessibilità e compatibilità con l'infrastruttura esistente.

1.3 Shibboleth Service Provider

Parallelamente all'IdP, è stato implementato uno Shibboleth SP (v3) dedicato ad autenticare l'u-tente verso gli applicativi di gestione. Questo SP consente l'accesso autenticato a un'applicazione web interna per:

- gestione delle sessioni attive;
- verifica e configurazione dei metodi MFA;
- cambio password;

L'adozione di Shibboleth SP per questo scopo ha permesso di mantenere un sistema coerente con il modello federato, pur garantendo funzionalità locali avanzate.

1.4 MariaDB per il Persistence Layer

Per la persistenza delle informazioni legate alle sessioni utente e alle configurazioni di eduMFA è stato utilizzato MariaDB, un sistema di database relazionale open source che offre buone presta-zioni e facilità di integrazione. I database gestiti includono lo storage delle sessioni di Shibboleth IdP (storageservice), la persistenza di eduMFA ed un punto di appoggio per l'autenticazione esterna (utilizzata per il login con SPID/CIE, eduGAIN, ecc.)

1.5 Apache HTTP Server come Reverse Proxy

A monte di tutti i servizi è stato posizionato come reverse-proxy Apache HTTPD, con il compito di:

- terminare le connessioni HTTPS,
- distribuire il traffico verso i container Tomcat o le interfacce web protette da SP,
- fungere da punto unico di ingresso e controllo, facilitando l'applicazione di regole di sicurezza a livello di URL, intestazioni HTTP o politiche di accesso.

Questa configurazione garantisce flessibilità nel bilanciamento del carico, nella gestione dei certi-ficati e nell'applicazione di policy di sicurezza, oltre a mantenere pulita la separazione tra logica applicativa e gestione dell'accesso.

1.6 Applicativi di Gestione In-House

A complemento dell'infrastruttura centrale di autenticazione, sono stati sviluppati internamente una serie di applicativi web dedicati alla gestione operativa dell'identità utente. Questi strumenti sono stati realizzati in PHP.
Le applicazioni sono eseguite dietro Apache tramite PHP-FPM e integrate con il Service Provider Shibboleth, garantendo accesso protetto e coerente con il sistema di autenticazione federato.
Le principali funzionalità offerte da questi applicativi includono:

- Gestione del secondo fattore (MFA): l'interfaccia consente agli utenti di configurare, aggiun-gere, modificare o rimuovere i metodi di autenticazione secondaria. Il sistema si appoggia a eduMFA come backend, sfruttandone l'API REST per interrogazioni e aggiornamenti dello sta-to MFA. Si è preferito utilizzare questo approccio anziché esporre il frontend di eduMFA in modo da favorire la user experience (considerata la complessità del frontend nativo di eduM-FA) ed aumentare la sicurezza.
- Gestione delle sessioni attive: è possibile per l'utente visualizzare l'elenco delle sessioni at-tualmente aperte, con possibilità di revoca manuale in caso di sospetta compromissione o fi-ne utilizzo. I dati sono ricavati direttamente dallo StorageService di Shibboleth, con l'ausilio di una vista materializzata che consente interrogazioni efficienti per-utente.
- Cambio e reset password: è disponibile un flusso sicuro per il cambio della password. Per il reset in caso di perdita delle credenziali, è prevista un'autenticazione alternativa (via codice OTP su cellulare o SPID/CIE nei prossimi sviluppi), sempre nel rispetto delle policy dell'ateneo.
- Consultazione degli attributi utente: gli utenti possono accedere a un'interfaccia informativa che elenca gli attributi di identità rilasciati dal sistema (es. nome, matricola, ruoli, affiliazioni, ecc.). Questo servizio è utile per diagnosticare problemi di autorizzazione e per verificare la correttezza dei dati personali utilizzati nei flussi federati.

Lo sviluppo di questi strumenti in-house ha permesso una personalizzazione fine delle funzionali-tà, garantendo coerenza con le politiche interne di gestione dell'identità e maggiore reattività nel rispondere a esigenze emergenti o segnalazioni da parte dell'utenza.

2 Infrastruttura

La nuova infrastruttura dell'Identity Provider (IdP) del Politecnico di Torino è stata progettata per garantire alta disponibilità, resilienza e modularità operativa. La configurazione attuale si basa su un cluster virtualizzato ridondato, distribuito su due zone distinte. L'infrastruttura dell'IdP è in esecuzione su tale cluster ed è composta, al momento, da tre nodi virtuali coordinati e intercon-nessi.

2.1 Architettura ad Alta Affidabilità (HA)

Il cluster di virtualizzazione – su cui sono in esecuzione in nodi dell'IdP – è strutturato per assicu-rare la tolleranza ai guasti e la continuità operativa anche in caso di malfunzionamenti o manu-tenzioni pianificate. Le due zone fisiche su cui il cluster è distribuito rappresentano una separa-zione fisica pensata per mitigare i rischi legati a fault localizzati, blackout o indisponibilità tempo-ranea di una delle zone.

2.2 Ruoli dei nodi

L'infrastruttura è attualmente composta da tre nodi, ciascuno con un ruolo specifico all'interno del sistema IdP:

• Nodo 1 (nodo principale): è il nodo centrale della configurazione attuale e ospita tutti i com-ponenti applicativi, ad eccezione di Shibboleth IdP.
Su questo nodo risiedono:

- Il reverse proxy Apache, punto d'ingresso dell'intera architettura;
- Il Service Provider Shibboleth locale, utilizzato per la gestione delle sessioni, dei pro-fili e delle funzioni di autenticazione secondarie;
- Il sistema di gestione MFA (eduMFA);
- Il database MariaDB, utilizzato per la persistenza delle sessioni e delle configurazioni MFA.
- Gli applicativi di gestione sviluppati in-house in PHP, in esecuzione su PHP-FPM

• Nodo 2 e 3 (nodi worker): questi due nodi gemelli sono dedicati esclusivamente all'esecuzione di Apache Tomcat con Shibboleth IdP.

- Entrambi i Shibboleth IdP afferiscono allo stesso DB (in esecuzione sul primo nodo) e quindi condividono le sessioni SSO: possono in questo modo a rispondere simulta-neamente alle richieste di autenticazione.
- Il bilanciamento del traffico tra i due nodi IdP è gestito dal reverse proxy in esecuzio-ne sul nodo principale.

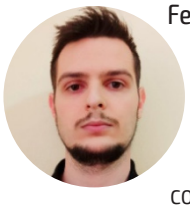
Questa suddivisione consente di mantenere separati i carichi applicativi dalle componenti core del servizio di autenticazione, migliorando l'affidabilità e semplificando gli interventi di aggior-namento o scaling.

Autori

Andrea Garzena andrea.garzena@polito.it

Esperto di sistemi informativi al Politecnico di Torino, ha una solida esperienza nell'IT ap-

plicata al contesto accademico. Lavora nel Presidio E-learning e in passato ha guidato l'Ufficio Piattaforme Applicative, occupandosi di integrazione e gestione di sistemi complessi. Laureato in Ingegneria Informatica, è certificato Red Hat e ha competenze avanzate in Linux, clustering, Moodle e Shibbo-leth. Ha contribuito allo sviluppo delle infrastrutture per la didattica a distanza e per l'autenticazione federata dell'Ateneo.



Federico Cucinella federico.cucinella@polito.it

Laureato in Ingegneria Informatica (Computer Networks & Cloud Computing) presso il Politecnico di Torino, ha sviluppato la sua tesi nell'ambito del progetto CrownLabs. Dopo la laurea, ha iniziato a collaborare con il Politecnico per l'aggiornamento e la gestione di diversi sistemi applicativi, prima di entrare in Present S.p.A. da cui continua a operare come consulente, con un focus sui si-stemi cloud ed informatici dell'università, in contesti applicativi orientati alla didattica ed infra-strutturali.

DOI: 10.26314/GARR-Conf25-proceedings-11

EHDS e Data Governance: Verso un'Europa della Condivisione dei Dati Sanitari

N. Foggetti, G. Pesole, F. De Leo, B. Fosso, M.A. Tangaro, A. Cestaro, F. Licciulli, C. Lo Giudice, M. Chiara, G. Cauli, M. D'ambrosio

Abstract. L'Unione europea ha recentemente adottato una serie di atti legislativi per garantire l'applicazione dei principi della scienza aperta e per creare un quadro giuridico solido per la condivisione dei dati. Questi interventi legislativi segnano un passaggio cruciale verso un'infrastruttura digitale paneuropea, in cui la gestione, la condivisione e il riutilizzo dei dati si inseriscono in un modello regolamentato, sicuro ed efficiente. La strategia europea sui dati mira a costruire un mercato unico dei dati, rafforzando la competitività globale dell'Europa e garantendo la sovranità digitale. In questo contesto, si stanno sviluppando spazi comuni europei per i dati, con l'obiettivo di rendere disponibili più dati per la società e l'economia, preservando al contempo i diritti degli individui e delle aziende. Tra i pilastri di questa strategia emergono due regolamenti fondamentali: Il Regolamento 2025/327 sullo Spazio Europeo dei Dati Sanitari (EHDS) e Il Regolamento 2022/868, noto come Data Governance Act (DGA) Questi atti normativi sono centrali per disciplinare lo scambio di dati in ambito sanitario, promuovendo al contempo innovazione, ricerca e tutela della privacy. L'applicazione dei principi enunciati all'interno del nuovo quadro giuridico di riferimento, la necessità di garantire una compliance, avranno un impatto significativo sui progetti di ricerca quali GoE, GDI e B1+Million Genome che mirano a creare un'infrastruttura europea per la condivisione dei dati genetici e sanitari. Questo lavoro mira ad analizzare le problematiche connesse all'applicazione della nuova disciplina all'interno dei progetti di ricerca aventi ad oggetto la creazione di infrastrutture europee per la condivisione dei dati

Keywords. EHDS- Data Altruism - Opting-out - data protection - sovranità digitale

1. Introduzione

L'Unione Europea ha recentemente adottato una serie di atti legislativi per garantire l'applicazione dei principi della scienza aperta e per creare un quadro giuridico solido per la condivisione dei dati. Questi interventi legislativi segnano un passaggio cruciale verso un'infrastruttura digitale paneuropea, in cui la gestione, la condivisione e il riutilizzo dei dati si inseriscono in un modello regolamentato, sicuro ed efficiente. La strategia europea sui dati[1] mira a costruire un mercato unico dei dati, rafforzando la competitività globale dell'Europa e garantendo la sovranità digitale. In questo contesto, si stanno sviluppando spazi comuni europei per i dati, con l'obiettivo di rendere disponibili più dati per la società e l'economia, preservando al contempo i diritti degli individui e delle aziende. Tra i pilastri di questa strategia emergono due regolamenti fondamentali: Il Regolamento 2025/327 sullo Spazio Europeo dei Dati Sanitari (EHDS)[2] e Il Regolamento 2022/868, Data Gover-

1 Cfr. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_it#:~:text=La%20strategia%20europea%20per%20i,ricercatori%20e%20delle%20pubbliche%20amministrazioni

2 Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the Euro

Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

nance Act (DGA)[3].

2. Un Mercato Unico dei Dati per la Sovranità Digitale Europea

La Strategia Europea per i Dati mira a creare spazi comuni europei dei dati, ambienti sicuri e regolamentati in cui i dati possano essere resi disponibili per l'innovazione e il progresso scientifico, senza compromettere i diritti e le libertà dei cittadini. Il DGA stabilisce il quadro giuridico per consentire a individui e imprese di condividere volontariamente i propri dati a beneficio della società, attraverso organizzazioni di fiducia. Queste entità, denominate "organizzazioni per l'altruismo dei dati", possono registrarsi ufficialmente come intermediari riconosciuti nell'Unione, facilitando l'accesso a dati rilevanti per la ricerca e l'innovazione. Uno degli obiettivi principali del DGA è quello di supportare la ricerca scientifica, creando pool di dati su larga scala, accessibili per analisi avanzate e apprendimento automatico. Questo approccio garantisce che il valore informativo dei dati sanitari e genetici possa essere sfruttato per il progresso medico e tecnologico. L'EHDS è la prima iniziativa normativa dell'UE volta a regolamentare in modo specifico la gestione e il riutilizzo dei dati sanitari. L'EHDS introduce un doppio livello di utilizzo dei dati sanitari, il primo è l'uso primario dei dati che ha l'obiettivo di garantire che ogni cittadino possa avere il controllo dei propri dati sanitari, favorendo un sistema interconnesso per l'erogazione delle cure sanitarie in tutta l'UE[4]. L'uso secondario dei dati si pone l'obiettivo di promuovere un modello regolamentato per il riutilizzo dei dati sanitari in ambiti come la ricerca scientifica, l'innovazione tecnologica, l'elaborazione delle politiche sanitarie e la regolamentazione. Il DGA non ha creato per gli enti pubblici alcun obbligo di consentire il riutilizzo dei dati, tantomeno li ha esentati dagli obblighi di riservatezza ove previsti ai sensi del GDPR. Il concetto di "riutilizzo" deve essere inteso come limitato a "determinate categorie di dati" disciplinate nel DGA e possibile solo in presenza di alcuni "requisiti specifici", in particolare, la trasparenza, il non riutilizzo degli stessi per scopi differenti e l'equità delle tariffe ove applicate[5]. Un'altra importante novità introdotta dal DGA è il concetto di "altruismo dei dati" che si sostanzia nella possibilità concreta di operare una "condivisione volontaria dei dati" sulla base del consenso accordato dagli interessati, ovve-

pean Health Data Space, OJ L, 2025/327, 5.3.2025, ELI: http://data.europa.eu/eli/reg/2025/327/oj
Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724, OJ L 152, 3.6.2022, p. 1–44, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868

3 Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724, OJ L 152, 3.6.2022, p. 1–44, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868

4 https://health.ec.europa.eu/ehealth-digital-health-and-care/digital-health-and-care/electronic-cross-border-health-services_it

5 M. Amendola, 2024, Il Principio solidaristico e il Data Governance Act, in Iura & Legal Systems, p. 53-58, (http://elea.unisa.it/bitstream/handle/10556/7122/Amendola.%20Il%20principio%20solidaristico%20e%20il%20data%20governance%20act.pdf?sequence=5) ; C Iurilli, 2024, la tutekla del dato personale alla prova del Data Governance Act. Data sharing, reclamo e tutela giurisdizionale effettiva, Judicium, Il processo civile in Italia e in Europa, (https://art.torvergata.it/bitstream/2108/355423/1/VERSIONE%20PER%20ASN%20SAGGIO%20DGA%20IURILLI.pdf) ; F. Bravo, 2021, Intermediazione di dati personali e servizi di data sharing dal GDPR al Data Governance Act, in Contratto e Impresa Europa, p. 208 ss.

ro in virtù di autorizzazione di altri titolari dei dati, senza la richiesta o la ricezione di un compenso per obiettivi di interesse generale previsti per legge nazionale[6].

3. Opt-out: tra diritto alla privacy e deroghe per interesse pubblico

Uno degli aspetti più controversi del Regolamento EHDS è l'introduzione della regola dell'opting-out, che stabilisce il diritto per i cittadini di opporsi alla condivisione dei propri dati per uso secondario. L'accordo introduce un approccio armonizzato all'opt-out, che impone agli Stati membri di adottare un meccanismo di opt-out accessibile e di facile comprensione. Ciò significa che, una volta che un individuo si oppone alla condivisione dei propri dati sanitari per utilizzi secondari, tali dati non potranno essere resi disponibili né trattati. Tuttavia, il diritto di opt-out può essere superato se un organismo di sanità pubblica, un'istituzione o un ufficio dell'Unione richiede i dati per finalità di ricerca scientifica di rilevante interesse pubblico. Questa possibilità di deroga è molto ampia e comporta il rischio di un'implementazione disomogenea del regolamento nei diversi Stati membri. Il legislatore europeo fa, quindi, rientrare le finalità di ricerca scientifica tra quelle di interesse pubblico per la quali, ai sensi dell'articolo 50 del GDPR permane il diritto di obiezione, ma decade l'obbligo del consenso[7].

Inoltre, gli Stati membri mantengono il diritto di introdurre ulteriori garanzie a livello nazionale per specifiche categorie di dati, come i dati genomici, i dati provenienti da applicazioni per il benessere e le biobanche/banche dati. In applicazione del GDPR, il diritto di opposizione deve essere esplicitamente portato all'attenzione dell'interessato e presentato in modo chiaro e separato da qualsiasi altra informazione al momento della prima comunicazione con l'interessato. L'interessato può esercitare questo diritto attraverso mezzi automatizzati e specifiche tecniche (articolo 21, paragrafo 4). L'istituto dell'opt-out è disciplinato in modo diverso nei sistemi giuridici degli ordinamenti nazionali, infatti il Comitato europeo per la protezione dei dati (EDPB) ha precisato che nei casi in cui il consenso non sia la base giuridica per il trattamento dei dati personali, esso può comunque essere utilizzato come garanzia per il trattamento. Inoltre, il diritto di opporsi al marketing diretto ai sensi dell'articolo 21, paragrafo 2, del GDPR è un diritto incondizionato che consente di vietare l'uso dei dati personali per finalità di marketing diretto ed è generalmente definito opt-out. Sulla base di questa interpretazione, l'opt-out potrebbe essere considerato una misura di salvaguardia per il trattamento delle categorie particolari di dati personali all'interno dell'EHDS, nei casi in cui il consenso non sia richiesto.

Al fine di garantire un trasferimento sicuro dei dati personali, le Faq che sono state pub-

6 E. Cremona, 2023, Quando i dati diventano beni comuni: modeli di data sharing e prospettive di riuso, in Rivista italiana di inforamtica e diritto, fascicolo 2, (https://www.rivistaitalianadiinformaticaediritto.it/index.php/RIID/article/view/178/151)

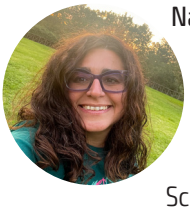
7 D. Corti, 2025, Consenso (a fasi progressive), interesse pubblico per l'IA, opt-out per l'uso secondario: le nuove regole per la ricerca scientifica sui dati sanitari, BioLaw Journal, fascicolo 1, p. 275-290; G. Zanfir-Fortuna, 2020, Article 21. Right to object,in C.Kuner, L.A. Bygrave, C. Docksey (eds.), op. cit., 2020, 519; M. Fraioli, 2019, Il diritto di opposizione e la revoca del consenso, R. Panetta (a cura di), Circolazione e protezione dei dati personali tra libertà e regole del mercato, 2019, Milano, 243-244.

blicate all'indomani dell'entrata in vigore del Regolamento[8], precisano che i dati, prima di confluire negli Health Data access Bodies (HDAB), organismi per la condivisione di dati, debbano essere sottoposti a processo di pseudonimizzazione. Non esiste ad oggi un modello unico di pseudonimizzazione e questo rischia di generare un'applicazione a geometria variabile dell'applicazione degli obblighi previsti dal Regolamento. Un ulteriore aspetto di problematicità riguarda la situazione in cui il titolare non è più in grado di collegare un set di dati pseudonimizzati ad un uno specifico individuo per mancanza di elementi di identificazione, in questa prospettiva il rispetto del diritto dell'opt-out non può essere garantito. Alla luce di quanto evidenziato, la cooperazione internazionale in grado di promuovere la redazione di buone prassi o linee guida, potrebbe garantire un primo passo verso l'armonizzazione normativa, pur attingendo ai paradigmi della soft law[9].

4. L'Impatto sulla Ricerca: Il Caso di GoE, GDI e B1+Million Genome

L'applicazione delle nuove regole avrà un forte impatto su progetti di ricerca chiave come GoE (https://genomeofeurope.eu/), GDI (https://gdi.onemilliongenomes.eu/) e B1+Million Genome (https://b1mg-project.eu/), che mirano alla creazione di un'infrastruttura europea per la condivisione dei dati genetici e sanitari. In particolare, il progetto Genome Data Infrastructure (GDI) prevede la creazione di un organismo centrale di accesso ai dati sanitari per la ricerca scientifica. L'integrazione delle norme EHDS in questo contesto ridefinirà le dinamiche di governance dei dati, influenzando la sostenibilità dei modelli giuridici esistenti. L'EHDS e il DGA segnano un punto di svolta nella regolamentazione della condivisione dei dati sanitari in Europa. Se da un lato si aprono nuove opportunità per l'innovazione e la ricerca, dall'altro emergono importanti sfide giuridiche legate alla protezione dei dati, all'equità nell'accesso e alla sicurezza delle informazioni sensibili. La capacità di bilanciare diritti individuali e interessi collettivi sarà cruciale per il successo di questi regolamenti e per il futuro dello spazio europeo dei dati.

Authors



Nadina Foggetti nadina.foggetti@cnr.it

A lawyer with a Ph.D. in EU and International Law, she is a Contract Professor in ICT Law, IT, and Biotech Law at Uniba. Technologist at CNR IBIOM and ELSI Officer for ELIXIR-IT, she has contributed to national and international projects on cybercrime, privacy, biotech, and digital law. Within ELIXIRxNextGenIT, she focuses on the Access Program, applying Open Science, FAIR data, and Open Access principles.

Graziano Pesole graziano.pesole@uniba.it

Graziano Pesole is full professor of Molecular Biology in the University of Bari A. Moro

8 Commissione europea, Frequently Asked Questions on the European Health Data Space, 5 March 2025, 32 (https://health.ec.europa.eu/latest-updates/frequently-asked-questions-european-health-data-space-2025-03-05_en)

9 Di recente, l'EDPB ha adottato un documento sulla pseudonimizzazione (si v. in particolare il paragrafo 22) EDPB, Guidelines 01/2025 on Pseudonymisation, 16 January 2025 (https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-012025-pseudonymisation_en)

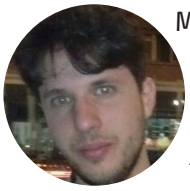
and Associate Researcher of CNR-IBIOM, Director of "Consorzio Interuniversitario Biotecnologie (Trieste), Head of the Italian Node of ELIXIR, the European Research Infrastructure for Life Science (>400, h-index=84, total cites ≥30,000). His research activity is mostly focused on bioinformatics applications for the management and analysis of next generation sequencing data, also at single-cell resolution.

Francesca De Leo francesca.deleo@cnr.it

Francesca De Leo is Technology Director at CNR and Deputy Head of the ELIXIR-IT Node, based at the Institute of Biomembrane, Bioenergetics and Molecular Biotechnologies. She coordinates the ELIXIRxNextGenIT project funded by MUR and leads the Industry & Impact and Communication Offices within ELIXIR-IT. She holds a degree in Biological Sciences and a PhD in Biochemistry and Molecular Biology, with expertise in innovation, technology transfer, and research infrastructure management.

Bruno Fosso bruno.fosso@uniba.it

Bruno Fosso is Associate Professor at the University of Bari "Aldo Moro". The metagenomic investigation of host-associated microbiomes and environmental prokaryotic communities is the main topic of his research activities. During the last 10 years, he developed tools and databases for both metabarcoding and shotgun metagenomics investigation of microbial communities. He participated in several Italian (MICROMAP, OMICS4FOOD) and European projects (BIOVEL, EMBRIC, LIFEWATCH, EXCELERATE and ELIXIR) and he is the coordinator of the ELIXIR-IT tools platform.



Marco A. Tangaro marcoantonio.tangaro@cnr.it

Currently a Researcher at CNR-IBIOM. Since 2015 he is involved in the ELIXIR-IT community, developing Cloud services for bioinformatics and integrating new tools within the Galaxy workflow manager. In particular, he leads the development of the Laniakea platform, which allows the creation of on-demand Galaxy instances on the Cloud, and the UseGalaxy.it national Galaxy server.

Alessandro Cestaro a.cestaro@cnr.it

Alessandro Cestaro received his master degree in Biology at University of Padova in 2000 working in the Biophysics field; during this work he began to explore the applications of Computer Science in Biology. Going further in that direction he gained his PhD in 2005, still at University of Padova, with a thesis about genome sequencing and annotation of deep sea bacterium P. profundum.

Since 2005 he was at the Edmund Mach Foundation as bioinformatician specialized in genome data analysis and genome data management; he had worked, mainly, at the gene prediction and functional annotation of several plant genomes: grapevine, apple, pear, strawberry, lemon among others. Since 2023 he is infrastructure manager and Node coordinator of ELIXIR-IT at the National Research Council (CNR) of Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM) of Bari.

Claudio Lo Giudice claudio.logiudice@cnr.it

Dr. Claudio Lo Giudice is a bioinformatician with a PhD in Cell Biology and Biotechnology. He conducted proteomic research in Finland and taught Bioinformatics at the University of Bari. Currently a technologist at CNR-ITB in Bari, he works on Linux infrastructure, scientific data management, and sensitive data handling. He is the author of REDIdb and UTRdb 2.0, databases for transcriptome and UTR region studies. His interests include bioinformatics, big data, cloud, RNASeq, and alternative splicing.

Matteo Chiara matteo.chiara@unimi.it

Born in Manerbio (Brescia) on August 9, 1984, graduated in Functional Genomics and Bioinformatics from the University of Milan in 2009. PhD in Biological and Molecular Sciences from the same university in 2012. His research activities have spanned several years between the Department of Biosciences at the University of Milan and the IBIOM Institute of the National Research Council. Since 2023, Associate Professor of Molecular Biology in the Department of Biosciences at the University of Milan. His research group focuses on the development and implementation of algorithms for analyzing genomic data produced by modern high-throughput NGS sequencing technologies.

Guido Cauli guido.cauli@cnr.it

Student in Computer Science for Digital Businesses and graduated in Cinema, Photography and Audiovisual media. System administrator for GNU/Linux and Unix-like operating systems, mainly focusing on server and workstation management through Proxmox VE clustering and OpenStack systems, LXC and Docker container operations, enterprise networking and IT security aimed at the production, processing and secure storage of bioinformatics data.

Flavio Licciulli flavio.licciulli@cnr.it

Master degree in Computer Science. Bioinformatician from 2001. Expertise in Research Data Management; expert in design and development of biological database and data integration tools; expert in application of FAIR principles for data and metadata standardization. Expert in Data Center management for the storage and processing of Life Science-oriented data. Competences in development of pipelines for the analysis of omics data.

Marilena D'Ambrosio m.dambrosio@cnr.it

Marilena D'Ambrosio is a biologist with a PhD in Public Health and a residency in Microbiology and Virology. She has worked in the medical device industry, including as Business Development Manager at Medtronic Italia, and holds a Master's in Nutrition. At the National Research Council, she developed expertise in research infrastructure impact assessment, combining skills in biology, microbiology, and project management across clinical, environmental, and scientific research domains.

DOI: 10.26314/GARR-Conf25-proceedings-12

Dall'addestramento necessario alla "malnutrizione" dei modelli: degenerazione dell'IA e questioni di diritto d'autore

Massimo Farina

Professore Associato di Informatica giuridica, Università degli Studi di Cagliari

Abstract. Il paper analizza la "malnutrizione" dei modelli di intelligenza artificiale, un fenomeno tecnico-giuridico derivante dall'uso indiscriminato di dati sintetici o raccolti tramite scraping massivo, che innesca degrado cognitivo (Model Autophagy Disorder) e violazioni dei diritti esclusivi dell'autore. Esplorando il nesso tra robustezza scientifica e legittimità normativa, il contributo propone soluzioni basate sulla tutela negoziale, che integrano metadati e licenze machine-readable per garantire una "dieta" sostenibile dei modelli, bilanciando innovazione e tutela autoriale

Keywords. malnutrizione dei modelli, Model Autophagy Disorder, diritto d'autore, text and data mining, licensing-by-design

1. Introduzione: dalla fame di dati alla "malnutrizione" cognitiva

L'insaziabile necessità di dati per l'addestramento dei modelli di intelligenza artificiale (IA), in particolare quelli linguistici di grandi dimensioni, sembra alimentare un progresso tecnologico senza sosta. Tuttavia, l'accumulo indiscriminato di contenuti – spesso generati sinteticamente o raccolti tramite pratiche di scraping massivo prive di adeguata validazione – conduce a una forma di malnutrizione cognitiva che compromette la robustezza dei modelli (E. M. Bender et al., 2021). L'assimilazione di dati di scarsa qualità, paragonabili a "calorie vuote", riduce la ricchezza semantica e innesca fenomeni di autoreferenzialità, richiamando riflessioni filosofiche e logiche sulla ricorsione (B. Russell, 1908; K. Gödel, 1931; D. Hofstadter, 1979). Tale processo degenerativo, noto come Model Autophagy Disorder (MAD), si manifesta attraverso un degrado qualitativo, amplificazione di bias e produzione di "allucinazioni" (I. Shumailov et al., 2023; L. Palazzani, 2020; G. Sartor, 2022; G. Ziccardi, 2024). Sul piano giuridico, la voracità di dati collide con i diritti esclusivi dell'autore, tutelati dalla normativa sul diritto d'autore, solo parzialmente mitigata dalle eccezioni per il text and data mining previste dalla Direttiva (UE) 2019/790 (R. Caso, 2020; S. Dell'Arte, 2023; P. B. Hugenholtz, 2019). Ne emerge un circolo vizioso: la carenza di fonti lecite aggrava la "malnutrizione" tecnica, mentre l'uso sregolato di dati incide sulla sostenibilità e sull'affidabilità dell'IA (M. A. Lemley, B. Casey, 2021; M. Farina, 2024). Analizzare questo nesso significa collocare la "dieta" dei modelli al confine tra informatica giuridica e filosofia del diritto, con l'obiettivo di individuare soluzioni che preservino con-

Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

giuntamente la solidità scientifica e la legittimità normativa dell'intelligenza artificiale.

2. "Malnutrizione" tecnica e patologia dei dati riciclati

La carestia cognitiva delineata in apertura si concretizza, sul versante tecnico, in un processo degenerativo che emerge quando la quota di testo sintetico presente nel corpus oltrepassa la seguente soglia di contaminazione. Il modello, riutilizzando il proprio output come input, avvia una spirale di autofagia che erode in modo misurabile la diversità lessicale e la capacità di generalizzare. Simulazioni iterative hanno mostrato che, già dopo poche generazioni, le distribuzioni linguistiche collassano verso forme stereotipate, con perdita delle "code rare" e progressiva amplificazione dei bias pre-esistenti (I. Shumailov, et al., 2023). L'effetto, registrato sia in domini testuali sia in quelli multimodali, viene oramai identificato come Model Autophagy Disorder (S. Alemohammad, et al., 2023) e risulta tanto più marcato quanto più il ciclo di addestramento ignora procedure di data refreshing o di filtraggio qualitativo.

Il problema non si limita alle prestazioni del modello, ma la contaminazione raggiunge anche i set di benchmark, alterando artificialmente le metriche di valutazione e mascherando la reale entità del degrado (C. Deng, 2023). Approcci "data-centrici" suggeriscono di spostare l'attenzione dall'ottimizzazione architetturale alla cura del dato, prevedendo pipeline di validazione, deduplicazione e rebalancing capaci di ripristinare varietà semantica e copertura dei fenomeni rari (S. E. Whang, et al., 2023). In tale direzione, esperienze recenti in ambito nazionale hanno indicato strategie di augmentation semi-supervisionata mirate a reintegrare porzioni di lingua minoritaria e registri specialistici spesso trascurati nei corpora generalisti (P. Bruno, et al., 2025).

In siffatto contesto, residua un vincolo strutturale, relativamente all'accesso a fonti human-made di qualità, il cui approvvigionamento risulta sempre più costoso sul piano giuridico e commerciale. Il rischio è che la scarsità di materiale lecito funga da incentivo per il ricorso a dati riciclati, intensificando proprio quella "malnutrizione" tecnica che si vorrebbe scongiurare. Da qui il necessario raccordo con la dimensione giuridica e la carenza di "nutrienti informativi" genuini, che non potrà essere sanata se l'ecosistema regolatorio non offrirà soluzioni equilibrate al problema della riproduzione delle opere protette, di cui si dirà nel successivo paragrafo, dedicato proprio alla "malnutrizione" giuridica.

3. "Malnutrizione" giuridica e diritti esclusivi

Il circolo vizioso evidenziato nella dimensione tecnica trova un corrispettivo speculare sul piano normativo, ogniqualvolta vi sia carenza di contenuti "nutritivi" genuini. In tali casi, di sovente, l'addestramento dei modelli avviene mediante scraping massivo, con il conseguente rischio di usurpare i diritti esclusivi dell'autore (G. Sartor, 2022; O. Pollicino, P. Dunn, 2024) che insistono su ogni copia di un'opera protetta, anche se destinata a processi computazionali e non alla fruizione umana. In tal senso, è bene tener presente che la deroga introdotta dagli articoli 3 e 4 della Direttiva (UE) 2019/790, che consente il text and data mining per finalità scientifiche, è subordinata al rispetto di condizioni rigorose e, soprattutto, al potere di opt-out riconosciuto ai titolari delle opere. Ne risulta un "defi-

cit nutritivo" giuridicamente indotto, laddove gli autori e gli editori esercitano l'opzione di esclusione. In presenza di tale condizione, infatti, diminuisce la quantità di dati leciti disponibili, costringendo i modelli a ricorrere a materiale di provenienza dubbia e aggravando, di riflesso, la "malnutrizione" tecnica di cui già si è detto.
La compressione dei diritti esclusivi, e in particolare del diritto di riproduzione, emerge con particolare evidenza nei contenziosi avviati da organi di stampa (F. Di Tano, 2011), fotografi e illustratori contro i fornitori di LLM, in occasione dei quali si invoca, a giustificazione dell'uso, la natura «trasformativa» del training per sottrarre la copia alla sfera esclusiva. In termini più espliciti, si tenta la riconduzione delle copie di training alla categoria delle riproduzioni «meramente tecniche» necessarie all'analisi, facendo leva sulle eccezioni previste dagli articoli 5.1 e 5.5 della Direttiva 2001/29/CE (M. A. Lemley, B. Casey, 2021). A tal proposito, è utile evidenziare che l'istituto del fair use, tipico dell'ordinamento statunitense (ma non soltanto) non trova un equivalente funzionale nell'ordinamento europeo, incentrato su un sistema di eccezioni tipizzate e di diritti patrimoniali non derogabili per via pretoria (D. J. Gervais, 2020; M. L. Montagnani, A. Trapova, 2020; G. Sartor, 2022; N. W. Netanel, 2011; D. L. Burk, J. E. Cohen, 2001; P. B. Hugenholtz, M. Senftleben, 2011; A. Stazi, 2015).
Rinviando i tanti necessari approfondimenti ad altra sede, è doveroso, qui, evidenziare che questa "malnutrizione" giuridica non è soltanto il riflesso di un conflitto distributivo tra creatori e sviluppatori di IA, bensì è anche il sintomo di un disallineamento sistemico fra infrastrutture digitali globali e regimi nazionali di proprietà intellettuale. Il paragrafo successivo, conclusivo, si concentrerà proprio su tale disallineamento, esplorando le prospettive tecniche e regolatorie che – fra watermark, tracciabilità dei dataset e modelli di licenza collettiva – potrebbero ristabilire un equilibrio sostenibile tra libertà di ricerca e tutela degli autori.

4. Prospettive per una "dieta" sana.

Da quanto brevemente illustrato nei paragrafi precedenti, si evince che all'origine della "malnutrizione" dei modelli di IA non sussiste soltanto la scarsa affidabilità delle fonti di approvvigionamento, ma anche la violazione dei diritti esclusivi dell'autore dovuta talvolta a scarsa conoscenza della disciplina di settore, talaltra ad una deliberata scelta, in mala fede, di non rispettarla. I rimedi contro questa spirale patologica non possono certamente essere così drastici da fermare il progresso tecnologico o addirittura farlo arretrare (considerato l'attuale stato dell'arte), ma devono comunque essere posti in atto prima che i diritti degli autori possano ritenersi compromessi "senza ritorno". In tal senso, merita certamente un'attenta riflessione il fatto che la tutela autorale non si esaurisce nella cornice legale nazionale o sovranazionale, che di per sé, sola, risulta insufficiente, ma si completa con la - troppo spesso trascurata - tutela negoziale fatta di licenze d'uso o, più in generale, di accordi negoziali (S. Ricketson, J. C. Ginsburg, 2022; J. C. Ginsburg, 2016). Nel dominio digitale i contratti rappresentano la prima linea di difesa degli autori, se si considera la rapidità con cui i dataset vengono copiati e ricombinati, che rende inattuabile una verifica ex post del rispetto di norme positive, peraltro da azionarsi su impulso di parte. L'autonomia

negoziale e il diritto di disporre dei diritti esclusivi, per lo meno quelli di natura patrimoniale, potrebbe certamente rappresentare l'elemento chiave di effettivo irrobustimento della tutela. Nello specifico, si potrebbe erigere una comune solida infrastruttura "licensing by design" (M. Rodriguez, 2019), che permetterebbe di pre-calibrare diritti e obblighi delle parti coinvolte prima dell'ingresso l'opera nel ciclo di addestramento. Si tratta di una soluzione già percorribile con gli strumenti resi disponibili, seppur con peculiarità proprie, dagli ordinamenti nazionali e sovranazionali ma che necessita di un ruolo attivo degli autori e ancor prima di consapevolezza del proprio status. Così, sulla base del principio del consenso dell'autore, la buona prassi di informare ogni potenziale utilizzatore, sulle regole di circolazione dell'opera, così come decise dall'autore, potrebbe risolvere, alla fonte, quella porzione di "malnutrizione" dei modelli dovuta alla raccolta libera e incontrollata.
Da questa consapevolezza prendono le mosse alcune iniziative[1], ciascuna con le proprie caratteristiche, rivolte all'obbiettivo comune del rafforzamento della protezione delle opere dell'ingegno destinate all'addestramento dei modelli di IA.
Tra le iniziative di questo genere, il progetto HOLMES[2], dell'Università di Cagliari, prevede l'incorporazione nell'opera digitale di un passaporto di metadati, leggibile dalla macchina, contenente la volontà negoziale dell'autore; l'automazione del controllo di liceità, mediante incorporazione nei modelli di un sistema di lettura delle licenze tale da assicurare che l'opera venga utilizzata solo nel rispetto delle condizioni stabilite dall'autore. Il sistema mira a creare un meccanismo di protezione dinamico, ossia una tutela negoziale continua, che restituisce agli autori un controllo effettivo sulle modalità di sfruttamento delle loro opere all'interno delle pipeline di addestramento, senza paralizzare la ricerca.
Soluzioni di questo tipo risultano in linea con l'orientamento licensing-by-design (M. Rodriguez, 2019) – di integrazione di protocolli di tracciabilità, licenze collettive estese e watermark crittografici –, che rappresenta la via più promettente per spezzare il circolo vizioso della "malnutrizione", garantendo al contempo robustezza cognitiva e liceità.

Bibliografia

Alemohammad S., Casco-Rodriguez J., Luzi L., Humayun A. I., Babaei H., LeJeune D., Siahkoohi A., Baraniuk R. G. (2023), Self-Consuming Generative Models Go MAD, ArXiv:2307.01850.

Bender E. M., Gebru T., McMillan-Major A., Shmitchell S. (2021), On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, pp. 610-623.

Bruno P., Calimeri F., Filice F., Marte C., Perri S. (2025), IDADA: A Blended Inductive-Deductive Approach for Data Augmentation, in Artale A., Cortellessa G., Montali M. (a cura

1 Data Provenance Initiative (DPI) - https://www.dataprovenance.org/ ; Content Authenticity Initiative (CAI) - https://contentauthenticity.org/; Fairly Trained e Certification for Ethical Data Use - https://www.fairlytrained.org/; altri progetti di di ricerca (elencati in Kirchenbauer et al., 2023), che sviluppano tecniche di watermarking crittografico per tracciare i contenuti generati o utilizzati dai modelli di IA.
2 Per i dettagli del progetto HOLMES "Harmonizing Ownership and Legal Measures for Ethical AI Systems", si veda il portale dedicato https://sites.unica.it/holmes/

di), AIxIA 2024 – Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 15450, Springer, pp. 79-91.
Burk D. L., Cohen J. E. (2001), Fair use infrastructure for rights management systems, in Harv. J.L. & Tech, pp. 41-83.
Caso R. (2020), Il conflitto tra diritto d'autore e ricerca scientifica nella disciplina del text and data mining della direttiva sul mercato unico digitale, in Il Diritto Industriale, vol. 2, pp. 118-126.
Dell'Arte S., (2023), Fondamenti di diritto d'autore nell'era digitale, Key Editore, Milano.
Deng C., Zhao Y., Tang X., Gerstein M., Cohan A. (2023), Investigating Data Contamination in Modern Benchmarks for Large Language Models, ArXiv:2311.09783.
Di Tano F. (2011), Diritto d'autore e aggregatori di notizie online: spunti dal caso Federazione Italiana Editori Giornali vs. Google News Italia, in Ciberspazio e diritto, vol. 2, pp. 193-220.
Farina M. (2024), Degenerazione e rischio creativo dell'Intelligenza Artificiale "forte": forme di prevenzione e tutela complementare, in L'Ircocervo, 23 (1), pp. 122-138.
Gaudenzi Sirotti A. (2024). Il nuovo diritto d'autore: la tutela della proprietà intellettuale nell'era dell'intelligenza artificiale, Maggioli, Santarcangelo di Romagna.
Gervais D. J. (2020), The Machine as Author, in Iowa Law Review, vol. 105, pp. 2053-2106.
Ginsburg J. C. (2016), Overview of Copyright Law, in R. Dreyfuss, J. Pila (eds.), Forthcoming, Oxford Handbook of Intellectual Property, Columbia Public Law Research Paper No. 14-518, Available at SSRN.
Gödel K. (1931), Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, in Monatshefte für Mathematik und Physik, vol. 38, pp. 173-198.
Hofstadter D. (1979), An Eternal Golden Braid, Basic Books, New York.
Hugenholtz P. B. (2019), The New Copyright Directive: Text and Data Mining (Articles 3 & 4), in Kluwer Copyright Blog, 24 luglio 2019.
Hugenholtz P. B., Senftleben M. (2011), Fair use in Europe: in search of flexibilities, in SSRN:1959554.
Lemley M. A., Casey B. (2021), Fair Learning, in Texas Law Review, 99 (4), pp. 743-785.
Montagnani M. L., Trapova A. (2020), US and EU: diverging or intertwined paths?, in O. Pollicino G. M. Riccio (eds), Copyright and Fundamental Rights in the Digital Age, Edward Elgar, pp. 188-215.
Netanel N. W. (2011), Making sense of fair use, in Lewis & Clark Law Review, vol. 15, pp. 715-771.
Palazzani L. (2020), Tecnologie dell'informazione e intelligenza artificiale: Sfide etiche al diritto, Edizioni Studium, Roma.
Pollicino O, Dunn P. (2024). Intelligenza artificiale e democrazia: opportunità e rischi di disinformazione e discriminazione, EGEA, Milano.
Ricketson S., Ginsburg J. C. (2022), International Copyright and Neighbouring Rights: The Berne Convention and Beyond (3 ed.), Oxford University Press, Oxford.
Rodriguez M. (2019), Licensing by design: A systematic approach, in The Serials Librarian, vol. 76 (1-4), pp. 178-184.

Russell B. (1908), Mathematical Logic as Based on the Theory of Types, in American Journal of Mathematics, 30 (3), pp. 222-262.
Sartor G. (2022), L'intelligenza artificiale e il diritto, Giappichelli, Torino.
Shumailov I., Shumaylov Z., Zhao Y., Gal Y., Papernot N., Anderson R. (2023), The Curse of Recursion: Training on Generated Data Makes Models Forget, ArXiv, abs/2305.17493.
Stazi A. (2015), La tutela del diritto d'autore in rete: bilanciamento degli interessi, opzioni regolatorie europee e "modello italiano", in Il diritto dell'informazione e dell'informatica, 30(1), pp. 89-110.
Whang S. E., Roh Y., Song H., Lee J. (2023), Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective, in VLDB Journal, 32, pp. 791-813.
Ziccardi G. (2024), Dati avvelenati. Truffe, virus informatici e falso online, Raffaello Cortina, Milano.

Autori

Massimo Farina m.farina@unica.it
Professore Associato di Informatica Giuridica (con abilitazione scientifica nazionale alle funzioni di Professore di Prima Fascia) presso l'Università degli Studi di Cagliari. Avvocato Cassazionista del Foro di Cagliari e Responsabile della Protezione Dati dell'Ateneo Cagliaritano. Coordinatore del Laboratorio Universitario ICT4Law&Forensics e Componente del Direttivo ANDIG (Associazione Nazionale Docenti di Informatica Giuridica e diritto dell'informatica). Autore di monografie e articoli scientifici su varie tematiche di Informatica Giuridica e Diritto dell'Informatica.

DOI: 10.26314/GARR-Conf25-proceedings-13

SIGMA, un nuovo approccio al trattamento statistico dei dati

Doriana Frattarola, Simona Spirito

ISTAT – Direzione Centrale per la Raccolta Dati

Abstract. Negli ultimi anni, l'ISTAT ha incrementato l'integrazione dei dati per migliorarne la qualità e ridurre l'onere per i rispondenti. La necessità di adeguarsi alle disposizioni europee in materia di protezione dei dati personali ha spinto l'ISTAT a implementare il sistema SIGMA. Esso prevede che ogni trattamento statistico di dati personali sia svolto all'interno di una determinata area (Dominio Specifico di Integrazione), dove sono presenti soltanto i dati necessari per una specifica finalità statistica, per il tempo strettamente necessario, e a cui hanno accesso soltanto le persone autorizzate

Keywords. Protezione dei dati, Pseudonimizzazione, Domini specifici di integrazione

Introduzione

Per migliorare la qualità dell'informazione statistica e ridurre al contempo l'onere statistico per i rispondenti, l'ISTAT integra fonti di natura statistica e non statistica in modo sempre più consistente. In tale contesto la simultanea necessità di adeguarsi alle disposizioni europee e nazionali in materia di protezione dei dati personali (GDPR – Regolamento UE n.679/2016 e Decreto legislativo n.101/2018) ha spinto l'ISTAT a adottare un nuovo approccio per il trattamento statistico dei dati.

Tradizionalmente, il trattamento dei dati in ISTAT avviene – laddove possibile – con l'utilizzo degli pseudonimi in sostituzione delle variabili con alto livello di identificazione. Nel 2020, ai fini di una maggiore tutela degli interessati (persone fisiche alle quali i dati si riferiscono), l'Autorità garante per la protezione dei dati personali ha dato all'ISTAT delle avvertenze (Provvedimento n.10 del 23.01.2020), esprimendo la necessità di adottare idonee tecniche di pseudonimizzazione per garantire due principi fondamentali del GDPR: minimizzazione e limitazione della conservazione dei dati personali trattati nell'ambito dei lavori statistici.

In particolare, l'Autorità ha rilevato che l'utilizzo di pseudonimi invarianti nel tempo e nello spazio non può garantire il rispetto dei due principi, poiché non sono previste né la rigenerazione né la differenziazione degli pseudonimi in relazione alle diverse finalità statistiche perseguite: una persona fisica viene identificata con lo stesso pseudonimo in tutte le fonti di dati utilizzate in Istituto, permettendo agli utilizzatori interni di non avere limiti nelle integrazioni di dati a prescindere dalla finalità statistica da perseguire.

Per rispondere alle disposizioni dell'Autorità, è stato implementato il nuovo sistema SIGMA (SIstema di Gestione dei Microdati Amministrativi e statistici). Esso prevede che ogni

Conferenza GARR 2025 - Frontiere Digitali Selected Papers
Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

trattamento statistico sia svolto all'interno di una determinata area di lavoro, denominata Dominio Specifico di Integrazione (DSI), in cui sono presenti soltanto i dati necessari a conseguire una specifica finalità statistica e a cui hanno accesso esclusivamente le persone autorizzate a svolgere il trattamento dei dati.
In un DSI, ciascuna persona fisica è identificata da uno pseudonimo specifico del dominio: quest'ultimo impedisce l'integrazione dei dati presenti in un dominio con quelli presenti in altri domini. Inoltre, ciascun DSI ha una scadenza stabilita sulla base della finalità statistica. Tale scadenza non pone limiti ad eventuali analisi longitudinali sui dati: prima della data di scadenza si definisce un nuovo DSI (nuova area di lavoro) in cui è possibile elaborare i dati con nuovi pseudonimi specifici (rigenerati).

1. I processi in cui interviene SIGMA

I processi interni in cui è previsto l'utilizzo di SIGMA sono i seguenti:

1. Raccolta dei dati;
2. Trattamento primario dei dati raccolti;
3. Rilascio dei dati ai DSI;
4. Trattamento statistico dei dati nei DSI.

Il funzionamento di SIGMA si basa sul forte collegamento tra dati e metadati, ossia le informazioni relative ai metadati guidano l'effettivo trattamento dei dati nei vari processi.

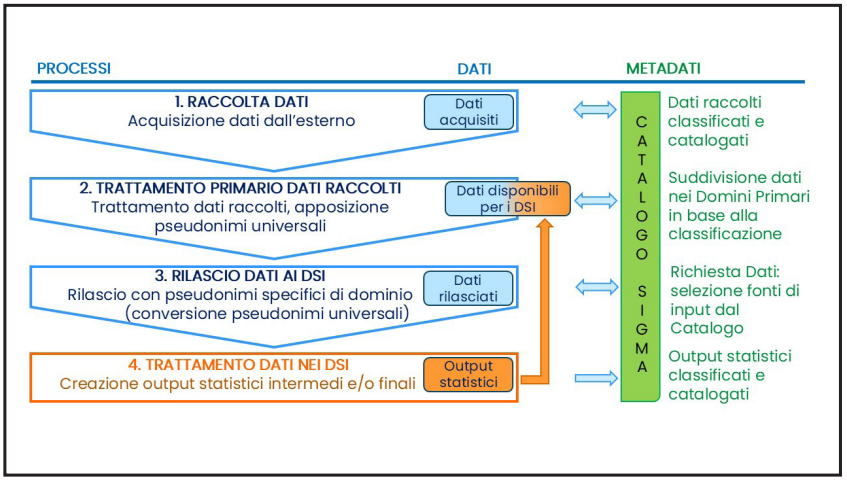


Fig. 1
SIGMA: processi, dati, metadati

1.1 Raccolta Dati

Per il perseguimento delle diverse finalità statistiche, l'ISTAT acquisisce sia dati amministrativi e statistici forniti da Enti esterni sia dati di indagine provenienti dalla raccolta diretta. Ciascuna fonte di dati acquisita viene catalogata, ossia le informazioni che caratterizzano la fonte sono inserite nel Catalogo dei metadati di SIGMA.
Nel Catalogo le informazioni relative a ciascuna fonte sono strutturate su più livelli: fonti, tracciati, variabili, forniture. Ogni fonte è composta da uno o più oggetti (dataset) con un proprio tracciato. Per ogni tracciato sono definite le caratteristiche di classificazione

a livello di variabile e sono indicate una o più forniture (la fornitura fa riferimento a un differente periodo di riferimento dei dati o a differenti stati di aggiornamento del dataset). La classificazione consiste nel fornire, per ciascuna variabile, le seguenti informazioni:

- tipologia di variabile: l'informazione è strutturata su più livelli gerarchici. Il primo livello distingue tra: variabile identificativa, tematica relativa a particolari categorie di dati (GDPR, art.9), tematica relativa a condanne penali e reati (GDPR, art.10), tematica «comune» o di processo, pseudonimo. I successivi livelli di classificazione offrono un maggiore dettaglio della tipologia di variabile;
- popolazione di riferimento;
- modalità di rilascio (distinguendo tra: in chiaro, in forma pseudonimizzata, non rilasciabile).

1.2 Trattamento primario dei dati raccolti

I dati acquisiti sono suddivisi, sulla base delle informazioni di classificazione contenute nel Catalogo, nei seguenti domini primari (DP), ciascuno corrispondente ad uno schema Oracle:

- DP-ID, contenente dati identificativi;
- DP-S, contenente dati tematici rientranti in particolari categorie;
- DP-TEM, contenente dati tematici comuni o di processo.

Nei primi due DP i dati sono conservati in modo criptato.

Non è previsto il DP dei dati relativi a condanne penali e reati perché tali dati al momento non possono essere trattati dall'ISTAT, secondo quanto disposto dall'Autorità Garante.

In uno schema Oracle dedicato, avviene il processo di pseudonimizzazione primaria, ossia ad ogni unità statistica viene assegnato un codice univoco, denominato pseudonimo universale di SIGMA. Tale pseudonimo garantisce la connessione tra i tre domini primari.

Una volta trattati nei domini primari, i dati sono disponibili per il rilascio nei DSI, in cui avviene l'elaborazione dei dati da parte degli utenti interni, necessaria al conseguimento delle specifiche finalità statistiche.

1.3 Rilascio dei dati ai DSI

Il rilascio dei dati ai DSI avviene a fronte di un'apposita richiesta, la cui compilazione viene effettuata dagli utenti interni tramite l'interfaccia di SIGMA. Durante la compilazione si selezionano le fonti necessarie al conseguimento della finalità statistica direttamente dal Catalogo di SIGMA (ulteriori dettagli al cap. 2).

A fronte di ciascuna richiesta, i dati sono rilasciati nel DSI con gli pseudonimi specifici di dominio: gli pseudonimi universali, generati nel processo precedente di trattamento primario, sono convertiti dal sistema in pseudonimi specifici, in modo tale che sia possibile l'integrazione delle fonti soltanto all'interno del DSI. Infatti, una stessa unità statistica è identificata da uno pseudonimo specifico diverso a seconda del DSI di riferimento.

1.4 Trattamento statistico dei dati nei DSI

Il trattamento dei dati viene svolto all'interno di un dominio specifico di integrazione

(DSI), utilizzando gli pseudonimi specifici di dominio.
I dati trattati in un determinato DSI producono output statistici intermedi e/o finali che, se necessario, possono essere messi a disposizione di altri DSI. Per renderli disponibili, occorre che i relativi metadati siano inseriti nel Catalogo, in modo tale che altri utenti interni possano utilizzarli.

2. L'utilizzo dell'interfaccia per il rilascio dei dati ai DSI

Per avviare il trattamento statistico nel DSI, occorre effettuare le seguenti operazioni mediante l'interfaccia di SIGMA.

- Definizione del DSI, mediante l'inserimento delle principali caratteristiche (denominazione, periodo di validità, finalità statistica, referente, ecc.).
- Compilazione delle richieste di dati, che avviene in modo guidato e si basa su un meccanismo di selezione delle fonti di dati (necessarie per la propria finalità statistica) dal Catalogo, scendendo al livello di singola variabile. Per un DSI possono essere effettuate una o più richieste.

3. Conclusioni

Considerando tutti gli aspetti relativi al funzionamento e all'utilizzo del sistema (dalla costruzione del sistema dei metadati, alla definizione dei DSI fino al rilascio dei dati), SIGMA rappresenta lo strumento che consente di mettere in atto misure tecniche e organizzative per la protezione dei dati personali fin dalla fase di progettazione e per tutta la durata del trattamento statistico dei dati (privacy by design). In particolare, tramite la predisposizione delle richieste di dati, SIGMA permette di individuare i dati personali strettamente necessari al trattamento prima di effettuare il rilascio al DSI (privacy by default).
Il sistema è in continua evoluzione per soddisfare le esigenze degli utenti e migliorare l'efficienza dei processi. Alcune nuove funzionalità sono attualmente in fase di test. Ad esempio, verrà implementata una funzionalità per consentire una più efficiente gestione del Catalogo.
Un'ulteriore funzionalità, denominata "filtro delle unità statistiche", è in fase di test per rafforzare la minimizzazione dei dati. Attualmente le fonti richieste sono rilasciate effettuando una selezione dei dati a livello di variabile (riduzione "verticale" dei dati): questa funzione consentirà di selezionare, per ogni fonte, il set di unità statistiche di interesse (riduzione "orizzontale" dei dati). In questo modo, sarà possibile rilasciare sottoinsiemi di una o più fonti, contenenti solo le unità di interesse.

Autori

Doriana Frattarola frattarola@istat.it

Laureata in Economia, specializzata in Metodi statistici per l'analisi dei sistemi economici. È Ricercatrice presso ISTAT. Ha lavorato nell'ambito dell'indagine su reddito e condizioni di vita (EU-SILC) maturando esperienza nell'integrazione tra le indagini EU-SILC e Spese delle famiglie (HBS). Dal 2022 lavora nella Direzione Centrale per la Raccolta Dati. È responsabile del coordinamento delle funzioni trasversali alla Raccolta dei dati per il transito delle indagini nel nuovo sistema SIGMA.

Simona Spirito spirito@istat.it

Laureata in Scienze Statistiche, specializzata in Ricerca Operativa e Strategie Decisionali. È Primo Tecnologo presso ISTAT. Ha lavorato nella produzione delle statistiche su previdenza e assistenza sociale e al 15-mo Censimento generale della popolazione e delle abitazioni. Dal 2016 lavora nella Direzione Centrale per la Raccolta Dati. È responsabile della progettazione di nuove piattaforme per l'acquisizione e il trattamento dei dati e dell'interoperabilità tra sistemi di gestione dei dati.

DOI: 10.26314/GARR-Conf25-proceedings-14

Introducing the Elettra Scientific Data Lake: Concepts, Architecture and Select Applications

Roberto Pugliese, Matteo Billè, Marco De Simone, Iztok Gregori, Daniele Favretto, Francesco Guzzi, Aljosa Hafner, Fulvio Bille', and George Kourousias
IT Group, Elettra Sincrotrone Trieste, Italy

Abstract. The Elettra scientific Data Lake (EDL) represents a tailored adaptation of modern data lakehouse architecture for synchrotron facilities. By combining the flexibility of data lakes with the governance of data warehouses, EDL addresses the unique challenges of scientific data management including format heterogeneity, FAIR compliance, and real-time processing requirements. Built on heterogeneous on-site infrastructure spanning edge computing to HPC clusters, EDL supports custom web-based applications that transform raw experimental data into scientific insights while maintaining ISO27001 security standards

Keywords. data lake, scientific data, synchrotrons, data analysis, data management

1. Elettra Sincrotrone Trieste and Data Lakes

Elettra Sincrotrone Trieste is a multidisciplinary research infrastructure center operating two advanced light sources: the Elettra synchrotron, a third-generation electron storage ring (2/2.4 GeV) operational since 1993, and the FERMI free-electron laser. The facility serves 32 beamlines covering spectroscopy, diffraction, scattering, and imaging techniques, supporting researchers from over 50 countries. The upcoming Elettra 2.0 upgrade increases coherence by approximately 50 times, increasing X-ray brilliance and by more than two orders of magnitude [1]

Modern data lakes provide scalable repositories for storing vast amounts of raw data in native formats. Data lakehouses extend this concept by combining data lake flexibility with data warehouse performance and governance features. This hybrid architecture offers an optimal foundation for managing the complex, heterogeneous data streams generated by synchrotron experiments.

2. Scientific Data and the Elettra Data Lake (EDL)

Scientific data differs fundamentally from business data in its complexity and heterogeneity. Elettra experiments generate diverse formats including TIFF images, CIF crystallographic files, raw detector streams, CSV logs, and proprietary formats from specialized equipment. This diversity reflects the varied scientific techniques employed across beamlines.

The facility embraces open science through FAIR (Findable, Accessible, Interoperable, Reusable) principles. HDF5 serves as a primary container format for complex scientific

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

data while maintaining crucial metadata. Digital Object Identifiers (DOIs) ensure persistent identification and citation of datasets, transforming experimental output into citable research products.

EDL adapts commercial data lakehouse concepts for scientific workflows by preserving native formats while building sophisticated metadata layers enabling cross-dataset discovery and analysis. The architecture supports streaming ingestion for real-time monitoring, automated quality assessment, and comprehensive provenance tracking linking raw data to processed results.

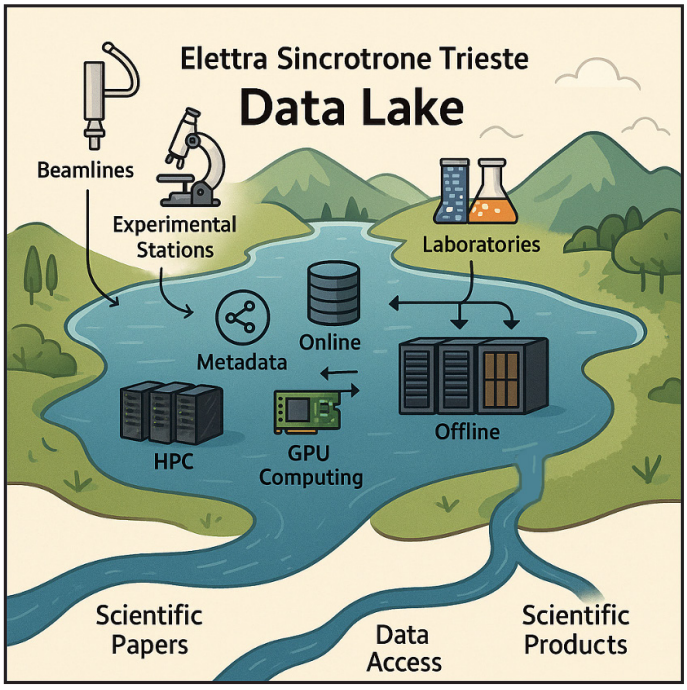


Fig. 1
Schematic representation of Elettra Data Lake

3. EDL Hardware Infrastructure

The Elettra Data Lake operates entirely on local infrastructure, ensuring data sovereignty and microsecond-level latencies critical for experimental workflows. This heterogeneous ecosystem spans from edge computing devices at beamlines handling multi-gigabyte-per-second data streams to centralized HPC resources.

The on-site data center houses diverse computational resources including high-memory nodes for large-scale processing, GPU-accelerated systems for machine learning and reconstruction, and specialized hardware for domain-specific calculations. Storage employs a tiered architecture with NVMe for hot data, disk arrays for active datasets, and tape libraries for long-term preservation.

For offline data archiving, Elettra employs an IBM Spectrum Archive 4500 tape library equipped with 8 LTO-8 drives, providing a substantial 14 petabytes of uncompressed storage across 1200 LTO-8 tapes. This system leverages IBM LTFS alongside a custom, in-house developed software solution. Built on RESTful APIs with Python and utilizing Celery workers for job distribution, this software is fully Dockerized and features a scalable architecture designed for robust scientific data archiving. It ensures data integrity through double-copy storage and SHA512 checksums for verification. The custom archiving system is seamlessly integrated with the Virtual User Office (VUO) [2]. Raw scientific data is automatically saved in a dedicated tape pool in duplicate copies immediately after

production. Principal Investigators or beamline scientists can initiate the restoration of raw data copies from offline storage at any time. Furthermore, they have the autonomy to move entire investigations to offline storage, freeing up valuable space on their online storage. Both raw scientific data and full investigations are saved in double copies within dedicated tape pools.

The environment supports MPI for distributed processing across hundreds of cores and extensive GPU computing on both desktop workstations and server-grade accelerators. Sophisticated scheduling systems unify this heterogeneous ecosystem while ensuring experimental deadlines are met.

4. Custom EDL Applications for Scientific Data

The VUO web application serves as a comprehensive platform that collects and manages all information related to an experiment, from the initial request (proposal) through to the subsequent data collection. In addition to this core function, the application provides a unified login system for all internal company services and supports the implementation of the FAIR principles. Built on top of this ecosystem are dozens of specialized applications tailored to specific needs.

EDL's effectiveness manifests through custom web-based applications that transform raw data into insights. These tools provide intuitive interfaces while incorporating advanced user management, role-based access control, and audit trails aligned with Elettra's ISO27001 certification.

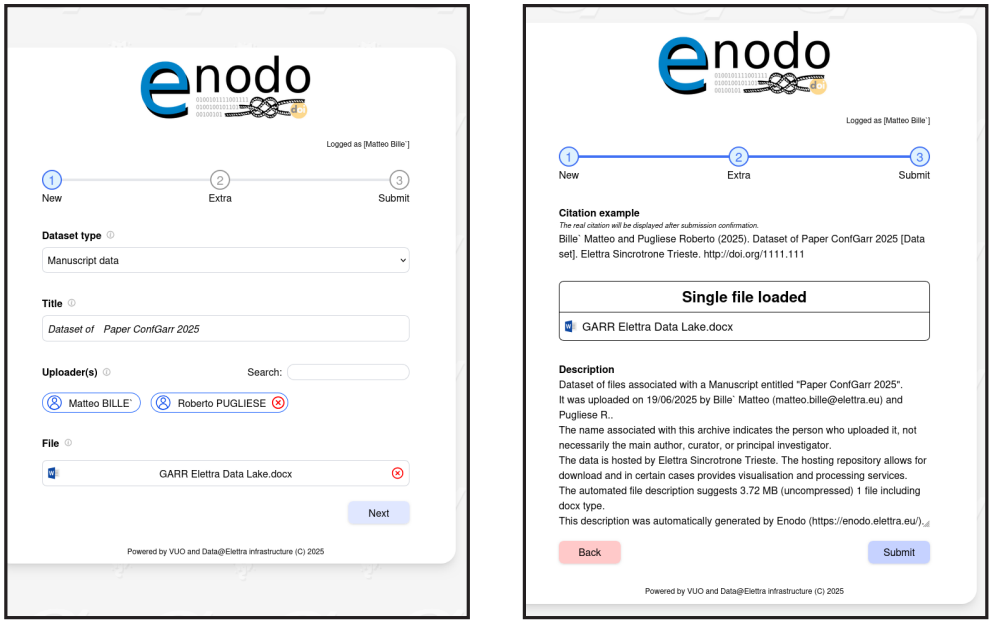


Fig. 2a - 2b

a) Enodo interface showing the complete set of input fields required to associate a dataset with a DOI.

b) pre-submission step, you where you can see the summary of your insertion

Some of these applications are:

- Enodo is a novel system inspired by Zenodo and WeTransfer. It allows users to upload manuscript datasets and associate them with a DOI (Digital Object Identifier), enabling citation in scientific publications. It ensures that datasets remain freely accessible and are persistently stored within the Elettra Data Lake, a FAIR and standardised repository. Publicly available on enodo.elettra.eu
- -XRFitVis provides an interactive environment for visualizing the results of XRF (X-ray Fluorescence) experiments. Built using web technologies, it allows researchers to access the tool both on-site and remotely. Publicly available on vuo.elettra.eu/go/xrfitvis
- -STP3 supports a dedicated micro-tomography beamline and operates on specialized hardware due to the computational demands of reconstruction. Its interface allows users to define optimal parameters and obtain full reconstructions of 100GB datasets in under 10 minutes. Used by the SYRMEP beamline.
- 4DHorizon is designed for visualizing ARPES (Angle-Resolved Photoemission Spectroscopy) data. It supports both 2D and 3D datasets and offers multiple adjustable parameters to modify the LUT and histogram, perform k-space transformations, extract the spectrum of the current slice, and enable smooth volume slicing for intuitive navigation through the volume. Used by the BaDElPh beamline.

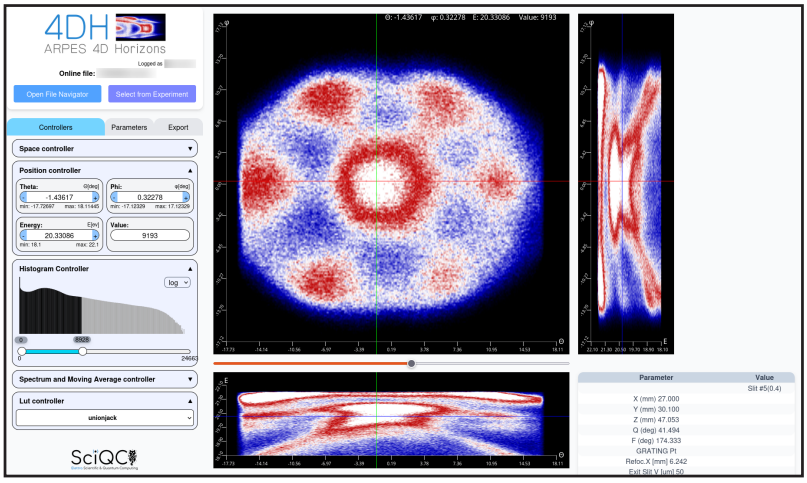


Fig. 3
Visualization of a 3D volume in the application. The left panel displays various visualization controls, while the right panel shows the rendered images along with key parameters and a navigation cursor for exploring the data

- Darkiver acts as a platform for file compression and decompression services. It offers a variety of conversion options, enabling users to upload files and quickly retrieve them in the desired output format. R&D in the context of PANOSC EOSC EU Node.
- eAI is a collective of experimental services and applications based on local LLMs. They meant to explore locally deployed services similar to ChatGPT but also Elettra specific applications for translation, summarization, scientific report generation and similar tasks. Available to Elettra personnel at BETA on eai.elettra.eu

Each application underwent co-development with scientific staff, ensuring interfaces match experimental workflows. Web-based architecture enables remote collaboration and real-time monitoring, proving invaluable for international research teams.

5. Conclusions and Future Perspectives

Synchrotron facilities generate data volumes that challenge traditional management approaches. EDL demonstrates that successful scientific data infrastructure requires deep integration with experimental workflows and flexibility to evolve with emerging methodologies. The heterogeneous hardware ecosystem provides the computational diversity necessary for the full spectrum of scientific analysis.

Custom applications showcase the importance of domain-specific tools in democratizing access to sophisticated analysis capabilities. As Elettra transitions to Elettra 2.0, the infrastructure must evolve correspondingly. Machine learning and Artificial Intelligence will play increasingly prominent roles in both analysis and experiment optimization. Through continued innovation, Elettra is establishing a model for transforming the data deluge into accelerated scientific discovery.

Acknowledgments

EDL requires competence and contributions from personnel beyond the list of authors. We acknowledge the contribution of the whole IT Group and of many beamline scientists of Elettra Sincrotrone Trieste.

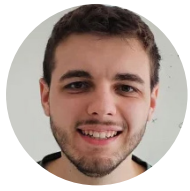
References

[1] https://www.elettra.eu/it/lightsources/elettra-2-0/elettra-2-0.html
[2] https://vuo.elettra.eu

Authors

Roberto Pugliese roberto.pugliese@elettra.eu

Roberto Pugliese, is the Deputy General Coordinator and IT Director at Elettra Sincrotrone Trieste. He holds a Ph.D. in Management, an MSc in Computer Science, an MBA, and PMP certification. Innovation Manager and Singularity University alumnus and ambassador, his research spans AI, robotics, telepresence, and data science, with publications in top journals like JSR and NIM. He supervises and directs the Elettra Data Lake.



Matteo Billè matteo.bille@elettra.eu

Matteo Billè is a scientific software engineer in the Scientific and Quantum Computing unit at Elettra Sincrotrone Trieste (quantum.elettra.eu). He is involved in data analysis projects with a focus on advanced visualization, in the development of the scientific data lake Data@Elettra, and in AI projects on local LLMs.

Scicomp Group sci.comp@elettra.eu

DOI: 10.26314/GARR-Conf25-proceedings-15

Real-World Federation of Autonomous Kubernetes in an Interconnected Continuum

Giuseppe Zangari[1], Fulvio Risso[2]

[1]ArubaKube, [2]Politecnico di Torino

Abstract. High-performance computing (HPC) and GPU clusters often suffer from inefficiencies of underutilized resources. Studies have shown that many HPC nodes and accelerators run well below full capacity, with CPUs and memory frequently only half-used and GPU memory largely untapped. Such underutilization translates into sunk costs and idle investments, even as other organizations struggle with insufficient compute capacity. Peaks in demand can overwhelm local clusters—researchers and engineers face queue backlogs and delays when their on-premises resources are saturated. This combination of underused hardware in one place and unmet needs in another highlights a critical inefficiency in the status quo. This paper explores how a federated Kubernetes-based approach can turn these inefficiencies into opportunities. By leveraging Kubernetes and Liqo, independent clusters can securely and transparently share compute resources while maintaining full autonomy over their infrastructure. The solution enables organizations to "burst" workloads to remote clusters on demand, resolving capacity shortfalls without costly hardware over-provisioning. At the same time, it allows those remote clusters to share or utilize their idle cycles, improving overall utilization. This federated model preserves cluster sovereignty: each participant retains control through policies and isolation, ensuring that sharing does not compromise security or autonomy. In essence, the AGER initiative demonstrates a real-world "computing continuum" that mitigates waste and scarcity by interconnecting cloud and HPC resources across institutional boundaries. This federated continuum unlocks innovation and operational value. Liqo and Kubernetes provide the cloud-native, secure foundation for this continuum, enabling seamless resource sharing "without borders" and establishing a new paradigm of collaborative computing at scale

Keywords. Computing-Continuum, Liqo, HPC, Efficiency, Offloading

Introduction

Despite widespread cloud and edge computing adoption, the global computing landscape remains fragmented. Organizations operate isolated clusters—on-premises HPC systems, private clouds, or edge nodes—that run independently. This isolation causes resource fragmentation: surplus capacity in one cluster cannot meet demand in another, leading to underutilization and unmet needs. Studies on NERSC's Perlmutter supercomputer show that most jobs used only a fraction of allocated resources, with ~50% of GPU jobs consuming just a quarter of GPU memory (Li et al. 2023). Meanwhile, organizations lacking HPC/GPU capacity face slowdowns and job queues, delaying critical R&D work. This imbalance highlights structural inefficiencies in modern research and enterprise computing.

Conferenza GARR 2025 - Frontiere Digitali Selected Papers
Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

A federated cloud-native infrastructure addresses this by connecting isolated clusters into a computing continuum (Iorio et al. 2023), conceptualize this as "liquid computing," where applications dynamically find execution venues across federated resources. This approach improves performance and flexibility while preserving decentralization and ownership: no single party controls all resources. Each participant—university, corporate cloud, or edge site—remains autonomous, sharing resources under its own policies.

Complementing this is Europe's focus on data sovereignty and federated data sharing. Marino et al. (2023) propose infrastructure-level data spaces, where clusters securely exchange and process data under agreed rules. Using Kubernetes-based federation (Liqo), flexible, on-demand data spaces span multiple domains, ensuring that providers retain sovereignty over infrastructure and data. Initiatives like Gaia-X further stress the importance of federation with autonomy and security.

Within this context, the AGER initiative demonstrates a real-world federated cloud-native infrastructure. AGER links independent Kubernetes clusters across multiple organizations into a resource continuum, operationalizing Iorio's vision with open-source tools. Using Liqo, each cluster can peer with others, securely advertising and consuming resources without altering internal configurations. Workloads flow to available capacity, embodying the "liquid computing" model.

AGER spans diverse environments—university HPC clusters, industrial research sites, and cloud providers—forming a nationwide Kubernetes continuum in Italy across Turin, Bologna, and Bergamo. Its mantra, "research without walls," reflects its ability to run workloads across sites seamlessly, bypassing traditional scheduling and cluster boundaries. AGER remains policy-first and cloud-agnostic, with each site defining sharing rules. This paper details AGER's method and value across academic, industrial, and enterprise contexts, showing federated Kubernetes infrastructure as a practical model for innovation and resource efficiency.

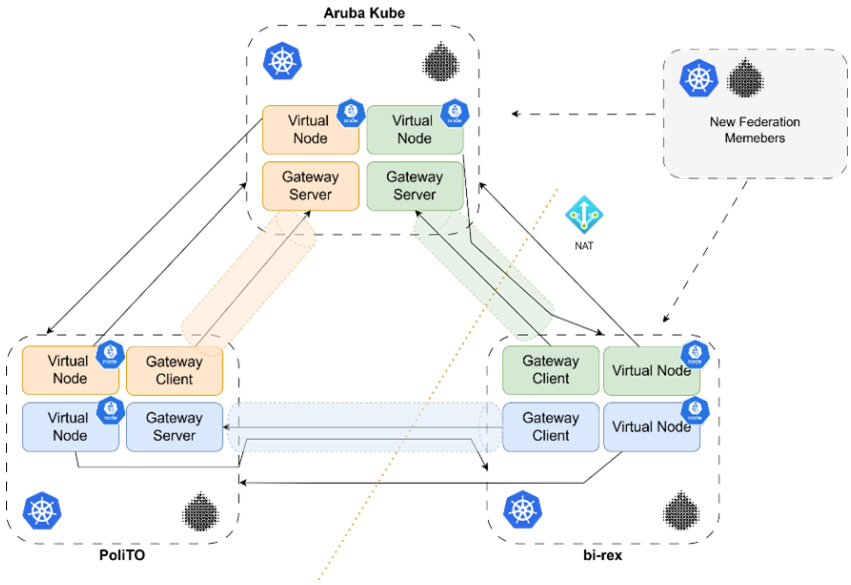
1. Federation Framework

AGER links independently managed Kubernetes clusters into a federated resource continuum using Liqo, an open-source extension designed for seamless multi-cluster Kubernetes federation. On each participating cluster, Liqo deploys a lightweight operator that handles federation tasks with minimal overhead. This operator:

- Advertises idle capacity (CPU, GPU, and memory) by creating a virtual node within the local Kubernetes API, representing the resources available from peered clusters. This abstraction allows local schedulers to see remote resources as if they were native nodes.
- Establishes encrypted network tunnels (typically over WireGuard or equivalent backends), ensuring that inter-cluster traffic remains private and secure. Importantly, Liqo retains the original service accounts, network policies, and namespace isolation, ensuring that identity and access control behave consistently even when pods are offloaded across organizational boundaries.

- Respects local quotas, priority classes, and preemption policies, so that no exported capacity jeopardizes critical home workloads. Clusters can dynamically adjust or revoke resource offers at runtime if local demand surges, offering real-time governance over shared capacity.

Fig. 1
AGER high level architecture



Scheduling remains native and fully transparent: Cluster A's scheduler operates as usual, and when it cannot place a pod locally (due to resource exhaustion or scheduling constraints), it targets the virtual node representing Cluster B. Liqo intercepts this scheduling decision and handles offloading the pod to Cluster B, ensuring that it runs in a sandboxed namespace mapped to the originating tenant. From an operator and developer perspective, the pod appears local—logs, metrics, monitoring hooks, and debugging tools (like kubectl logs and kubectl exec) function exactly as if the pod were on-premises.

Critically, federation is opt-in and namespace-scoped. This means that each organization retains strict control over what resources are shared, with whom, and under what conditions—key for addressing compliance, sovereignty, and governance mandates often imposed in both academia and enterprise. Policies can restrict federation by namespace, resource type, or workload class.

Joining the federation requires no disruptive changes (Marino et al. 2023): a single Helm chart installation of Liqo and a secure token exchange between clusters is sufficient. No "lift-and-shift," migration, or workload reconfiguration is necessary. Existing CI/CD pipelines, deployment scripts, and monitoring frameworks remain fully compatible, making AGER's approach a low-friction, production-ready solution for CTOs seeking scalable, policy-governed, and secure multi-cluster resource sharing across heterogeneous infrastructure environments.

2. AGER across sectors

AGER's federated Kubernetes continuum is not just a technical advancement—it represents a strategic enabler for innovation-driven organizations facing compute, budget, and time-to-market pressures. By breaking down infrastructure silos, AGER empowers institutions and enterprises to dynamically access, trade, and optimize distributed resources without compromising data governance or operational autonomy. This model fosters cross-institutional collaboration, accelerates research and product cycles, and transforms underutilized capacity into a business asset. The following use cases illustrate how federation drives measurable impact across academic research, industrial operations, and enterprise digital transformation.

2.1 Academic and Medical Research

Genome analytics, climate simulation, and large-language-model training surge unpredictably. With AGER, a university hospital can burst oncology pipelines to a spare resources of a national supercomputing centre during peaks, then re-claim resources when demand subsides. Turnaround time may drops from days to hours, grant-funded GPUs avoid idleness, and multi-institution collaborations proceed without data exfiltration using policies.

2.2 Industrial Optimization

Manufacturers, energy firms, and media studios face cyclical compute spikes. Instead of over-provisioning, they federate with AGER. A car maker, for exam-ple, runs crash-simulation sweeps on partner clusters overnight, returning re-sults before the morning stand-up. Capital expenditure falls, idle hardware may gains revenue as a traded asset, and production schedules are insulated from HPC bottlenecks.

2.3 Industrial Optimization

Global enterprises juggle dozens of Kubernetes deployments across clouds and edges. AGER federation converts these silos into a single elastic plane; latency-sensitive microservices drift to edge nodes while batch analytics migrate to available AGER resource. Governance domains remain intact because federa-tion respects jurisdictional boundaries encoded in policies. The net effect is lower total buffer capacity, predictable spending, and faster feature roll-outs

3. Conclusion and future work

AGER proves that federated Kubernetes can reconcile autonomy with collabora-tion. By exposing surplus capacity as a service, it elevates idle hardware from sunk cost to strategic asset, compresses time-to-insight in research, smooths in-dustrial production cycles, and sharpens enterprise competitiveness. Future re-search should explore fine-grained brokering—e.g., sub-GPU sharing—and in-tegrate market pricing to incentivise broader participation. Standardised trust frameworks (Gaia-X, IDSA) can further institutionalise policy exchange, ena-bling federations that span hundreds of clusters on a continental scale. The journey toward a durable computing continuum has begun; the next step is to

mainstream it, making compute-as-commons as ubiquitous as the internet itself.

Bibliographic References

Li J., Michelogiannakis G., Cook B., Cooray D., & Chen Y. (2023). Analyzing Re-source Utilization in an HPC System: A Case Study of NERSC's Perlmutter. Lecture Notes in Computer Science, 13948, 297-316.

Iorio M., Risso F., Palesandro A, Camiciotti L., Manzalini A. (2023) Computing without borders: The Way Thowards Liquid Computing, IEEE Transaction on Cloud Computing (vol. 11, no. 3), pp 2820-2838

Marino J., Camiciotti L., Cheinasso F., Olivero A., Risso F. (2023), Enabling Compute and Data Sovereignty with Infrastructure-Level Data Spaces, ESAAM '23: Proceedings of the 3rd Eclipse Security, AI, Architecture and Modelling Conference on Cloud to Edge Continuum (October 2023), pp 77-85

Authors

Giuseppe Zangari giuseppe.zangari@arubakube.cloud

Giuseppe Zangari (born in 1982) graduated from the Politecnico di Torino and holds an EMBA from the Graduate School of Management at Politecnico di Milano. He has held various leadership positions in global software organizations like Nokia and Pirelli, in Italian SMEs and in Politecnico di Torino, leading the development of business effective solutions with technologies ranging from IoT to cloud computing and AI. He is an expert of digital transformation, a startup mentor, and he also served as Innovation Lead. At ArubaKube, he is responsible for maximizing the software project's value, serving concurrently as Product and Business Development Lead.

Fulvio Risso fulvio.risso@polito.it

Fulvio Risso is full professor at Politecnico di Torino. Born in Saluzzo, Italy on November 15, 1971, he shares his birthday with the announcement of the Intel 4004 chip. Fulvio completed his BSc in Computer Engineering from Politecnico di Torino in July 1995 and got his PhD in Computer Engineering from the same institution in January 2000. His academic journey has been marked by significant contributions in the field of cloud computing, edge computing, network functions virtualization, and software-defined networking. He greatly contributed to open-source software, starting many successful project such as WinPcap, the de-facto packet capture library for Windows, and many others. He recently started the ArubaKube spin-off of Politecnico di Torino, where he serves as Chief Innovation Officer.

DOI: 10.26314/GARR-Conf25-proceedings-16

Collaborative and Reproducible science infrastructure: the Europlanet GMAP JupyterHub processing environment

Giacomo Nodjoumi[1,2,3], Carlos H. Brandt[4], Javier Suárez-Valencia[3], Erica Luzzi[5], Mario Valiante[6], Veronica Camplone[1,2], Edoardo Rognini[1,2], Angelo Pio Rossi[7], M. Giardino[1,8], A. Zinzi[1,8]

[1]Space Science Data Center (SSDC), Agenzia Spaziale Italiana (ASI), Via del Politecnico snc, 00133, Rome, Italy, [2]INAF/Osservatorio Astronomico di Roma (INAF-OAR), Via Frascati 33, 00078, Monte Porzio Catone (RM), [3]School of Science, Constructor University Bremen gGmbH, Bremen, DE, [4]EGI Foundation, Amsterdam, NL, [5]Mississippi Mineral Resources Institute, University of Mississippi, 114 Brevard Hall, University, USA, [6]Dipartimento di Ingegneria Civile, Università degli Studi di Salerno, Fisciano, Italy, [7]Earthgraph GmbH, Bremen, Germany, [8]Agenzia Spaziale Italiana (ASI) - Via del Politecnico snc, 00133, Rome, Italy

Abstract. Scientific research frequently uses advanced computational tools and skills to handle its large datasets. We developed a user-friendly, innovative solution based on standardized Dodgeville recipes and ipykernels to streamline interactive data processing environment deployment and management. Our solution is compatible with a series of widely used analytical software based on Python, and R, and specialized software for planetary sciences such as the USGS Integrated Software for Imager and Spectrometers (ISIS) and the NASA Ames Stereo Pipeline (ASP), providing a scalable and highly customizable solution for collaborative research, teaching, and specialized group.

Introduction

Research data are available in various formats, ranging from raw measurements to processed and calibrated products. Accessing such data is facilitated by dedicated archives and platforms such as NASA's Planetary Data System (PDS) (PDS Geosciences Nodes, 2020), ESA's Planetary Science Archive (PSA) (Planetary Science Archive, 2025) , Mars System of Information (MarSI) (Quantin-Nataf et al., 2018), SSDC-ASI's Multi-purpose Advanced Tool for Instruments of the Solar System Exploration (MATISSE) (Zinzi et al., 2016), ESA MAAP (Multi-Mission Algorithm and Analysis Platform) (Albinet et al., 2019) and general-purpose environments like Google Colab. However, turning these data into useful scientific knowledge usually needs special software and expertise in computing. Several initiatives aim to provide Analysis Ready Data (ARD) (Building a Lunar Spatial Data Infrastructure (SDI) - ADS, 2021; Knowledge Inventory of Foundational Data Products in

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

Planetary Science, 2021; Fergason et al., 2021) and interactive web-based analytical interfaces [10]. However, many advanced analyses still depend on sophisticated, open-source software packages such as the USGS Astrogeology Research Program's ISIS and (Sucharski et al., 2020) NASA Ames Stereo Pipeline (ASP) (Beyer et al., 2018). Despite their effectiveness, these tools require familiarity with UNIX systems and programming, presenting a barrier to broader adoption among researchers.

Proposed Solution

To address these challenges, we developed a versatile, scalable solution using Docker containers (Forde et al., 2018) and JupyterHub (Kluyver et al., 2016). Our solution, namely Europlanet GMAP JupyterHub (Nodjoumi et al., 2025), emphasizes standardized Dockerfile recipes enhanced with ipykernels, facilitating easy adaptation for specific tasks or research working groups. The core image is the ISIS-ASP-GISPY Docker image, which integrates ISIS, ASP, and a curated collection of Python packages (GISPY - Geospatial Python), allowing customization for diverse analytical requirements. Docker containers offer a lightweight, conflict-free environment, while JupyterHub provides a web-based platform for interactive computing through notebooks that integrate code, output, and explanatory text. A schematic of the architecture is presented in Figure 1.

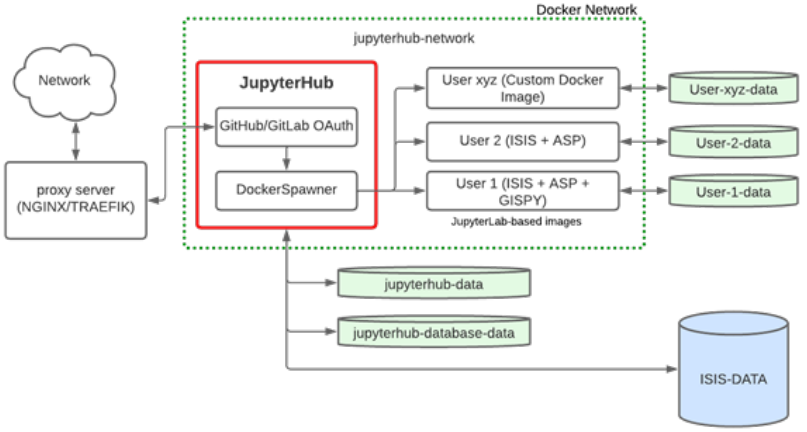


Fig. 1 Schematic of the proposed architecture. Boxes represent docker containers, while cylindres represent docker volumes.

Deployment and Management

The system is deployed through a semi-automated script handling Docker volume creation, networking, and image configuration. Additional parameters enable advanced settings for web service deployment, user access control, and resource allocation. Researchers interact with their computing environments through an accessible JupyterHub web interface, simplifying adoption and use. Shared folders for users can be configured as well as third-party services connections, such as GitHub/GitLab and cloud storages like Google Cloud.

Advantages and Impact

Our method presents substantial benefits over traditional computational workflows:

- Simplified Deployment and Maintenance: Dockerized environments reduce complexity and administrative burden.
- User-Friendly Interface: JupyterHub provides intuitive access, catering to various levels of technical expertise.
- Scalability: Easily adaptable for expanding user groups and/or workloads, suitable for collaborative research, educational settings, and workshops.
- Customization: The environment and software packages can be tailored to specific research needs, accommodating additional tools as required.

For instance, curated Docker images can be used on a laptop as standalone containers for development and then the developed code can be executed in an identical Docker container running on HPC nodes running JupyterHub without reproducibility issues.
Another example is a data centre composed of several teams, each with its own specialized docker image, customized with multiple ipykernels, each for specific task.
Initially developed within the Europlanet GMAP (Geological Mapping of Planetary Bodies) project at Constructor University, this solution was tested at international conferences such as COSPAR. Currently, it is being implemented at the Space System Data Center (SSDC) of the Italian Space Agency (ASI). After an initial assessment and integration with the other SSDC tools and services, we plan to gradually open and enable the JupyterHub service to the wider community of researchers and citizens. Final users will then have access to a user-friendly environment, remotely accessible, with embedded access to data, tools and pipelines running on high-end HPC, relieving the user to perform extensive computing tasks on personal computers and laptops.
A semi-production-ready implementation is publicly available on Zenodo and GitHub (Brandt et al., 2024). Continuous improvements are underway to enhance user experience, particularly focusing on session persistence and recovery functionalities critical for lengthy computational tasks. Further details about GMAP can be found at https://www.europlanet.org/.

Conclusion

This standardized and customizable computational environment significantly simplifies access to advanced analytical tools, promoting broader participation across diverse research communities. By lowering technical barriers, it supports collaborative, scalable, and effective data analysis, ultimately facilitating scientific analyses and innovation. Moreover, the HPC-based infrastructure, as well as the integration and accessibility of multiple datasets, services and tools already available at the SSDC, may also allow the application of novel processing technologies on even very large datasets and/or old datasets, further improving and widening the scientific outcome.

References

Albinet, C., Whitehurst, A. S., Jewell, L. A., Bugbee, K., Laur, H., Murphy, K. J., Frommknecht, B., Scipal, K., Costa, G., Jai, B., Ramachandran, R., Lavalle, M., & Duncanson, L. (2019). A Joint ESA-NASA Multi-mission Algorithm and Analysis Platform (MAAP) for Biomass, NISAR, and GEDI. Surveys in Geophysics, 40(4), 1017–1027. https://doi.org/10.1007/s10712-019-09541-z

Beyer, R. A., Alexandrov, O., & McMichael, S. (2018). The Ames Stereo Pipeline: NASA's Open Source Software for Deriving and Processing Terrain Data. Earth and Space Science, 5(9), 537–548. https://doi.org/10.1029/2018EA000409

Brandt, C. H., Nodjoumi, G., Rossi, A. P., Luzzi, E., & Suarez Valencia, J. (2024). europlanet-gmap/docker-jupyterhub: First Release (Versione 1.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.14555694

Building a Lunar Spatial Data Infrastructure (SDI)—ADS. (2021). https://ui.adsabs.harvard.edu/abs/2021LPICo2549.7054H/abstract

Fergason, R. L., Hunter, M. A., Laura, J. R., & Hare, T. M. (2021). Analysis Ready Data Available Through the SpatioTemporal Asset Catalog (STAC) Specification: Investigating the Application to Planetary Data. 2549, 7023. https://ui.adsabs.harvard.edu/abs/2021LPICo2549.7023F

Forde, J., Head, T., Holdgraf, C., Panda, Y., Nalvarete, G., Ragan-Kelley, B., & Sundell, E. (2018). Reproducible Research Environments with Repo2Docker. https://openreview.net/forum?id=B1lYOwuoxm

Kluyver, T., Ragan-Kelley, B., P&#233, Rez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., n, Abdalla, S., Willing, C., & Team, J. D. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In Positioning and Power in Academic Publishing: Players, Agents and Agendas (pp. 87–90). IOS Press. https://doi.org/10.3233/978-1-61499-649-1-87

Knowledge Inventory of Foundational Data Products in Planetary Science. (2021). https://essopenarchive.org/doi/full/10.1002/essoar.10501479.1

Nodjoumi, G., Brandt, C. H., Suárez-Valencia, J. E., Luzzi, E., Valiante, M., & Rossi, A. P. (2025). Collaborative and Reproducible Planetary Science Through the Europlanet GMAP JupyterHub Processing Environment. Earth and Space Science, 12(5), e2025EA004251. https://doi.org/10.1029/2025EA004251

PDS Geosciences Nodes. (2020). PDS Geosciences Node Orbital Data Explorer (ODE). https://ode.rsl.wustl.edu/

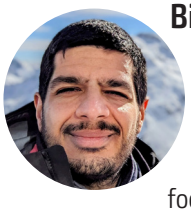
Planetary Science Archive. (2025). https://psa.esa.int/psa/#/pages/home

Quantin-Nataf, C., Lozac'h, L., Thollot, P., Loizeau, D., Bultel, B., Fernando, J., Allemand, P., Dubuffet, F., Poulet, F., Ody, A., Clenet, H., Leyrat, C., & Harrisson, S. (2018). MarsSI: Martian surface data processing information system. Planetary and Space Science, 150, 157–170. https://doi.org/10.1016/j.pss.2017.09.014

Sucharski, T., Mapel, J., Jcwbacker, Kristin, Lee, K., AgoinsUSGS, Shepherd, M., Combs, C. R., Stapleton, S., Dcookastro, Rodriguez, K., Becker, K. J., Sides, S., Cole, Jusflag, Wilson, T., Acpaquette, Williams, K., Jlaura, … Rsaleh57. (2020, luglio). USGS-Astrogeo-

logy/ISIS3: ISIS 4.2.0 Public Release (Versione 4.2.0). Zenodo. https://doi.org/10.5281/zenodo.3962369
Zinzi, A., Capria, M. T., Palomba, E., Giommi, P., & Antonelli, L. A. (2016). MATISSE: A novel tool to access, visualize and analyse data from planetary exploration missions. Astronomy and Computing, 15, 16–28. https://doi.org/10.1016/j.ascom.2016.02.006

Bibliography



Giacomo Nodjoumi

Giacomo Nodjoumi holds a Bachelor's degree in Geology and a Master's degree in Engineering Geology, Land Use Management and Georisks, both from Sapienza University of Rome. He earned a PhD in Geosciences from Constructor University Bremen, with a thesis focused on the automatic detection of pits, skylights, and cave candidates on the Moon and Mars using remote sensing and deep learning techniques and subsurface sounding radar data analyses. He is currently a research fellow at the Italian Space Agency's Space Science Data Center (ASI-SSDC), where he works on the development and deployment of scientific data services, such as the Europlanet GMAP JupyterHub, and other services on HPC systems. He contributes to several European projects, including the Horizon 2020 EXPLORE project, where he helped design tools like L-EXPLO and L-HEX for lunar data visualization and analysis, and Europlanet's GMAP initiative, supporting the creation of reproducible workflows for planetary surface mapping.

Carlos Brandt

Carlos Brandt is a Software Architect at the EGI Foundation, where he is part of the Technical Solutions team. With extensive expertise across multiple layers of the scientific software stack — including system administration, database modeling, numerical simulation, image processing, and distributed computing — Carlos brings a comprehensive approach to scientific computing and infrastructure design.
He holds a PhD in Astrophysics from the Sapienza University of Rome, building on a background that includes a degree in Physics from the Federal University of Rio Grande do Sul and a Master's in Numerical Simulations from the National Laboratory for Scientific Computing, both in Brazil. His academic and research trajectory led to roles in computational astrophysics at the Italian Space Agency's Data Center and as a lecturer in Data Engineering at Jacobs University Bremen.
Now based in Germany, Carlos has contributed to several European projects focused on space and geospatial data engineering. He is a strong advocate for open-source software and is passionate about initiatives that promote open access to scientific knowledge.

Erica Luzzi

Dr. Erica Luzzi is a planetary geologist specializing in Mars surface processes, ice detection, and terrestrial analogues. With international experience working at NASA, participating to ESA campaigns, and working in several universities across Europe and the US, she contributed to field campaigns in extreme environments and published widely on Martian geology, analogs for Enceladus, and resource mapping for future human exploration missions.

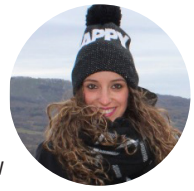
Javier Suaréz-Valencia

Geologist from the National University of Colombia and Ph.D. from Constructor University

Bremen in Germany, currently a postdoc in the University of Padova. He has worked for over eight years studying planetary surfaces using remote sensing techniques, mainly on Earth, Mars, and Pluto. Currently, for his postdoctoral project, he is analysing the geomorphology and composition lava tubes as potential analogues for lunar exploration missions.

Mario Valiante

Mario Valiante is a Research Fellow at the Department of Civil Engineering of the University of Salerno (Academic Discipline GEO/04 – Physical Geography and Geomorphology) since January 2022, where he also serves as lecturer in General Geology and Geomorphology. He obtained his Master's degree with honors in Engineering Geology, Land Use Management and Georisks from Sapienza University of Rome in 2015. In 2016, he qualified as a Professional Geologist (Regional Register of Campania, Section A).
In 2020, he earned a PhD in Earth Sciences from Sapienza University of Rome. In 2021, he held a postdoctoral research contract at the Department of Civil Engineering of the University of Salerno, working across the disciplines GEO/04 and GEO/05.
His main research interests focus on the analysis of landslide phenomena in geologically complex environments and on GIScience, particularly the development of data structures for geospatial information related to geological risk.



Veronica Camplone

Veronica Camplone is a geologist (PhD. in Earth and Environmental Sciences) specializing in remote sensing of Earth and planetary surfaces, with a focus on sedimentology and geomorphology. She also holds a postgraduate degree in Space Institutions and Policies, combining scientific and institutional expertise. Since 2021, she has been a research fellow at INAF in Rome, where she works on planetary data analysis and contributes to the development of the MATISSE tool for visualizing and comparing data from space missions.

Edoardo Rognini

Edoardo Rognini graduated in Physics with a specialization in Astronomy and Astrophysics and obtained a PhD working on radiative diffusion and levitation in low-mass stars. Since 2017, he has been living and working in Rome (initially at INAF-IAPS, then at ASI-SSDC), focusing on thermophysical modeling of airless bodies and thermal data analysis from space missions in the Solar System.

Angelo Pio Rossi

Angelo Pio Rossi is a planetary geoscientist, practitioner, and enterprenuer with 20+ experience in designing, implementing projects/programs on: Geoscience research, Remote Sensing, geospatial data handling, planetary mapping, training and education. More info on http://aprossi.eu and http://earthgraph.eu/

Marco Giardino

Marco Giardino holds a Master Degree in Computer Engineering. His main professional interests are software engineering and scientific data management, with a strong focus on FAIR principles and open science.

He has taken part in the scientific space missions Mars Express, Dawn, and ExoMars and has been involved in several others. He is currently working at the Italian Space Agency, where he leads the IT activities of the Space Science Data Center.

Angelo Zinzi

Angelo Zinzi has both Master Degree and PhD in Physics, with topics relevant to Planetary Sciences. He is now a staff technologies at the Italian Space Agency (ASI) and his main aims are comprised in the management of data of planetary exploration missions using FAIR principles and international standards, such as Virtual Observatory, and he developed the scientific webtool MATISSE. He has been or is involved in a series of space missions, such as ESA Rosetta, ASI LICIACube, ESA JUICE and participated to a series of international projects, such as NEOROCKS.

DOI: 10.26314/GARR-Conf25-proceedings-17

Open Science Near Real Time Data: an example of the application of FAIRness in an oceanographic context

Alexia Cociancich, Sebastian Plehan, Elena Partescano, Alessandra Giorgetti

National Institute of Oceanography and Applied Geophysics - OGS

Abstract. . Open Science is central to European research policy, promoting collaboration, quality, and accessibility. The FAIR Data Principles are essential for effective data management. The National Institute of Oceanography and Applied Geophysics applies these principles to oceanographic data from the Adriatic Sea, addressing challenges in data acquisition and transmission. Since raw data lacks metadata, the National Oceanographic Data Centre reconstructs metadata to ensure usability. ERDDAP facilitates standardized data access, supporting interoperability across European research infrastructures like EMSO ERIC. A geoportal enhances accessibility with intuitive visualizations. Achieving full FAIRness remains a challenge, but ongoing efforts align with European Open Science goals.

Keywords. FAIR, Open Science, Oceanography, Near RealTime

Introduction

Open Science is at the heart of European research policy. It aims to facilitate the exchange, collaboration and improvement of the quality of research and to place science at the centre of human and social development (https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science_en).

Open science policy and practice is changing the way research is conceived, managed, shared and evaluated.

Today, it is no longer enough to simply publish the results of publicly funded research without making the data as open and accessible as possible.

In this context, the FAIR Data Principles (Wilkinson et al. 2016) developed in 2016 recognize that good data management across the entire lifecycle is critical to the success of Open Science. Researchers who follow these principles produce data that are Findable, Accessible, Interoperable and Reusable (FAIR)

1. Open Science at OGS and the Challenge of Oceanographic Data

Open Science is one of the research and innovation missions that the National Institute of Oceanography and Applied Geophysics - OGS has identified in order to expand the user community of scientific data.

For more than 10 years, the OGS has been managing a network of oceanographic instruments in the Adriatic Sea that can transmit data in near real time (Partescano et al. 2017) . The difficulties that the environment poses to data acquisition and transmission significantly increase the challenges of subsequent data management. The heterogeneity of the

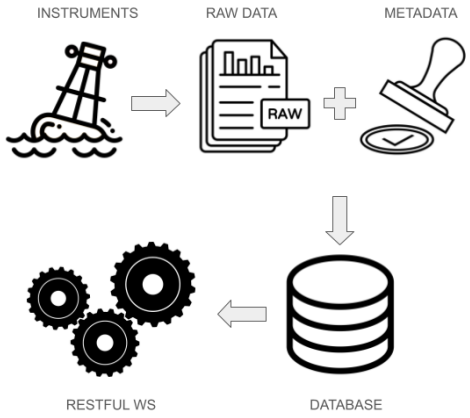
Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

parameters recorded and the instruments used requires a considerable effort in managing the data obtained.
Firstly, it should be noted that the transmitted raw data does not contain any metadata for reasons of efficiency. Although the latter must be associated after transmission to optimize throughput, without metadata to contextualise the data, it would be difficult to analyze, which is why the National Oceanographic Data Centre is committed to reconstructing the information content through the metadata process, because distributing data in near real time without respecting the basic principles of the FAIR approach would be of little use.

2. Ensuring FAIRness through Metadata and Data Services

To guarantee the FAIRness of the data, the correct assignment of metadata is crucial for its findability and the use of standard vocabularies is equally important to ensure a high level of interoperability. In this sense, the services provided by BODC (Thomas et al. 2015) are a fundamental support for the oceanographic research community.
To ensure data accessibility in a standardized and harmonized way, ERDDAP (https://nodc.ogs.it/erddap) was chosen as the preferred solution. As one of the most widely used open source software in the scientific field related to environmental data, it allows data managers to combine different sources and publish datasets in a powerful RESTful web service (Mendelssohn & Simons 2008) as well as a simple and intuitive web interface. This enables users to filter and export data in several formats, improving data accessibility and interoperability.

Fig. 1
Data flow schema



3. Harmonizing Data and Enhancing Accessibility for All Users

A further strength is the possibility to use ERDDAP in a federated way. For example, this has enabled the EMSO ERIC infrastructure (https://emso.eu) to make the datasets of different research centres distributed across the European territory via a single endpoint. The harmonization example carried out within the EMSO ERIC community (Martínez et

al. 2024), of which the OGS is a member, addressed the challenge of overcoming the complexity and heterogeneity of different data and metadata formats. The use of semantic artefacts such as vocabularies and standards together with the use of ERDDAP and common metadata standards are elements that have allowed to achieve a high level of FAIRness.
Finally, a geoportal (https://nodc.ogs.it/geoporta) has been developed to provide the most important information on the stations of the monitoring network in the Adriatic Sea in near real time to ensure an immediate and more intuitive graphical representation even for non-experts.

4. Conclusions

In conclusion, improving the FAIRness of data (Tanhua et al. 2019) is a crucial goal for all those involved in data management in order to meet the expectations of the European Commission. To obtain this, communities and data management infrastructures have faced numerous efforts and overcome challenges, but the path that leads to open science is not over yet, there are numerous initiatives to achieve a complete FAIRness of data, which is even more challenging in the context of the management of data acquired in near real-time.

5. Acknowledgements

The authors would like to thank the Technological development and support for acquisitions infrastructure (TEC) of the OGS and EMSO community as well as the financial contribution of ITINERIS project, the Italian Integrated Environmental Research Infrastructures System funded by the EU - Next Generation and PNRR Funds for the realisation of an integrated system of research and innovation infrastructures.

Bibliographic references

Martínez, E., Libes, M., Fratianni, C., Partescano, E., Cociancich, A., Pensieri, S., Paladini de Mendoza, F., Rodero, I., (2024) EMSO Eric Metadata Harmonization Efforts, International Conference on Marine Data and Information Systems - Proceedings Volume, p 105-107 https://dx.doi.org/10.13127/MISC/80/36

Mendelssohn, R. & Simons, R. (2008) ERDDAP - An Easier Way for Diverse Clients to Access Scientific Data From Diverse Sources, AGU Fall Meeting Abstracts. https://www.researchgate.net/publication/252989046_ERDDAP_-_An_Easier_Way_for_Diverse_Clients_to_Access_Scientific_Data_From_Diverse_Sources

Partescano, E., Brosich, A., Lipizer, M. et al. (2017) From heterogeneous marine sensors to sensor web: (near) real-time open data access adopting OGC sensor web enablement standards, Open geospatial data, softw. stand. 2, 22 https://doi.org/10.1186/s40965-017-0035-2

Tanhua, T.; Pouliquen, S.; Hausman, J.; O'Brien, K. M.; Bricher, P.; Bruin, T.; Buck, J. J.; Burger, E. F.; Carval, T.; Casey, K. S.; Diggs, S.; Giorgetti, A.; Glaves, H.; Harscoat, V.; Kinkade, D.; Muelbert, J. H.; Novellino, A.; Pfeil, B. G.; Pulsifer, P.; Van de Putte, A. P.; Robinson, E.; Shaap, D.; Smirnov, A.; Smith, N.; Snowden, D. P.; Spears, T.; Stall, S.; Tacoma, M.;

Thijsse, P.; Tronstad, S.; Vandenberghe, T.; Wengren, M.; Wyborn, L.; Zhao, Z., (2019) Ocean FAIR Data Services, Frontiers in Marine Science https://dx.doi.org/10.3389/fmars.2019.00440

Thomas, R., Lowry, R. K., and Kokkinaki, A., (2015) Moving Controlled Vocabularies into the Semantic Web,

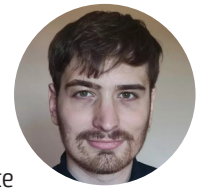
American Geophysical Union vol. 2015, Art. no. IN21D-1714, https://ui.adsabs.harvard.edu/abs/2015AGUFMIN21D1714T/abstract

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

Authors

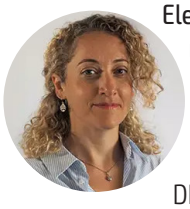
Alexia Cociancich acociancich@ogs.it

Alexia Cociancich has a master's degree in computer engineering and has worked as an IT consultant for 8 years in the field of industrial automation. Since 2019, she has been working at the National Institute of Oceanography and Experimental Geophysics (OGS) and is involved in database design and development for the National Oceanographic Data Centre (NODC).



Sebastian Plehan splehan@ogs.it

Sebastian Plehan is an experienced programmer with expertise in software development, database administration and IT support. He currently works at the OGS within the National Oceanographic Data Centre, where he contributes to the maintenance of services, and the development of new solutions for scientific data management projects. Passionate about new technologies, he is also studying at university to deepen his knowledge.



Elena Partescano epartescano@ogs.it

Elena Partescano is a data manager with expertise in managing, validating, analyzing, and organizing oceanographic data within relational databases. She supports real-time and delayed monitoring systems and works with SQL on Unix/Linux, metadata via XML and Mikado, XSLT, and GIS platforms. Elena also designs data entry interfaces, manages XML DB and XQuery, configures RESTful services, and ensures compliance with OGC standards for metadata catalogs and observations.

Alessandra Giorgetti agiorgetti@ogs.it

Alessandra Giorgetti is an expert in oceanographic data management. Senior technologist at OGS and deputy director of its Oceanography Section, she has led the Italian National Oceanographic Data Center since 2006, within UNESCO's IODE network. Giorgetti coordinates EMODnet Chemistry, promotes open science, and represents Italy in major European marine observation programs. In 2024, she was appointed to an EU expert group for harmonizing ocean observation standards.

DOI: 10.26314/GARR-Conf25-proceedings-18

Layout Parser, come creare un dataset di qualità per allenare l'Intelligenza Artificiale

Silvano Imboden, Gabriele Marconi, Simona Caraceni, Rossella Pansini, Fauzia Albertin, Antonella Guidazzoli

CINECA

Abstract. Negli ultimi anni la digitalizzazione del Patrimonio Culturale ha aperto nuove opportunità di accesso e studio, affrontando al contempo sfide significative, in particolare nella gestione e analisi dei dati. I progetti in corso in ambito europeo e nazionale riflettono una visione di accesso condiviso e collaborativo mirato all'arricchimento continuo delle conoscenze sul patrimonio culturale. La digitalizzazione dei quotidiani, ricchi di dati storici e sociali, richiede strumenti avanzati per l'analisi automatica, ottenuti grazie all'uso di dataset supervisionati per l'addestramento di modelli di AI. La piattaforma collaborativa sviluppata da CINECA, basata su Layout Parser, consentirà la classificazione automatica delle sezioni di un quotidiano. Il lavoro si focalizza sulla creazione di una ground truth per il fine-tuning del modello, con linee guida per l'annotazione di categorie come titoli, testi, pubblicità, fotografie e disegni

Keywords. Digital Humanities; Layout Analysis; Quotidiani; Intelligenza Artificiale

Introduzione (F. Albertin)

L'avvento di Biblioteche Digitali del Patrimonio Culturale a livello internazionale, europeo ed italiano ha aperto prospettive completamente nuove rendendo il Patrimonio accessibile a tutti e, al contempo, ha aperto nuove sfide.

Importanti progetti che riflettono la visione di un patrimonio culturale condiviso e collaborativo sono in corso, da iniziative europee come ECCCH [1], a iniziative nazionali come il Piano Nazionale di Digitalizzazione del Patrimonio Culturale (PND) [2]. In quest'ottica CINECA, partner del Ministero della Cultura – Digital Library, sta sviluppando applicativi innovativi per l'analisi di questo vasto patrimonio.

La prima sfida che si pone, infatti, è una elaborazione efficace ed efficiente dei dati, cruciale per trasformare questa mole di immagini in file clusterizzabili e analizzabili. In questo scenario, gli strumenti di AI sono risorse cruciali.

Una importante parte di questo patrimonio è rappresentata da quotidiani: non solo raccontano la nostra storia recente ma contengono importanti dati, come informazioni climatiche, statistiche demografiche e riguardanti annunci e campagne pubblicitarie. L'impiego di strumenti AI per l'estrazione e per la categorizzazione dei contenuti - distinguendo tra articoli, titoli e annunci pubblicitari - permetterebbe ai ricercatori di accedere facilmente ad un vasto dataset.

Conferenza GARR 2025 - Frontiere Digitali Selected Papers
Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

La piattaforma sviluppata, basata sull'applicativo Layout Parser [3][4], permette la classificazione automatica delle varie sezioni per poterle poi facilmente clusterizzare ed analizzare. In particolare, questa è stata sviluppata in un'ottica collaborativa per la creazione di una ground truth per il fine tuning dello strumento a partire dalle digitalizzazioni del quotidiano di Bologna, Il Resto del Carlino, degli anni 1939 e 1940.

Layout Parser: il modello AI (S. Imboden)

Layout Parser, introdotto da Microsoft Research Asia nel 2021 e rilasciato open-source su Github[5] e HuggingFace[6], è un framework open-source per l'analisi del layout documentale che implementa un approccio modulare ed estensibile per la segmentazione e il riconoscimento di elementi strutturali. L'architettura sfrutta modelli di deep learning basati su transformer, quali Detectron2 e Tesseract OCR, ottimizzati per il rilevamento di oggetti e il riconoscimento ottico dei caratteri. La pipeline di Layout Parser consente l'estrazione di primitive grafiche, quali regioni di testo, immagini e tabelle, mediante l'applicazione di modelli pre-addestrati o l'implementazione di modelli customizzati.
Il modello di partenza è stato specializzato in modi diversi per renderlo applicabile a diverse tipologie di documento. Tra queste, il riconoscimento di pubblicazioni scientifiche, trattati di matematica, o l'analisi di dati tabellari. La specializzazione riguardante i giornali invece è stata realizzata su un dataset con 16 milioni di immagini corredate della suddivisione in sezioni supervisionata manualmente relative al quotidiano Chronicling America [7].
Durante una fase preliminare di sperimentazione si è riscontrato come nell'analisi di giornali con una impaginazione a sei o sette colonne si ottenga buona precisione, ma questa decresca rapidamente man mano che ci si allontani da questo formato. Nel dataset da noi utilizzato si hanno impaginazioni che variano nel numero di colonne da due a nove, pertanto il modello non risulta adatto e si è quindi deciso di precedere con un fine-tuning mirato. Si è stimato che siano necessarie circa diecimila pagine annotate e supervisionate manualmente, un lavoro impegnativo ma importante nel panorama del "Digital Heritage" in Italia.

La piattaforma collaborativa (G. Marconi)

La piattaforma sviluppata è un'applicazione interattiva pensata per la visualizzazione e l'interazione con pagine di giornale digitalizzate.
L'architettura front-end è stata progettata utilizzando SvelteKit [8], un framework JavaScript moderno che consente di sviluppare applicazioni web.
Per la visualizzazione delle pagine di giornale digitalizzate, è stato integrato OpenSeadragon [9], una libreria JavaScript open-source ideale per il caricamento e la navigazione di immagini di grandi dimensioni, come le scansioni delle pagine di giornale, che possono essere esplorate a livelli di zoom elevati senza compromettere la qualità.
Per permettere agli utenti di creare la ground truth, è stata utilizzata D3.js [10], una libreria JavaScript per la manipolazione dei dati e la creazione di visualizzazioni dinamiche. In particolare, D3.js è stata impiegata per permettere agli utenti di creare e modificare aree

interattive sulle pagine di giornale consentendo agli utenti di spostarle, ridimensionarle o eliminarle a seconda delle necessità.
Il backend della webapp è stato sviluppato utilizzando PocketBase [11], una piattaforma di backend-as-a-service (BaaS) che fornisce un database integrato, consentendo di archiviare in modo efficiente e sicuro le informazioni relative alle pagine di giornale, le aree create dagli utenti e altre informazioni pertinenti. I dati sono organizzati in collezioni, che possono essere facilmente strutturate e consultate tramite API RESTful che supportano operazioni di lettura, scrittura, aggiornamento e cancellazione (CRUD).

L'interfaccia di lavoro di Layout Parser (S. Caraceni)

L'interfaccia di lavoro corrisponde ad una pagina di quotidiano su cui Layout Parser ha riconosciuto linee orizzontali e verticali e riquadri relativi a titoli e annunci (Fig. 1), e che risulta non accurata. Il passaggio successivo prevede la correzione delle linee individuate dal modello e l'identificazione delle regioni di testo e immagini. Layout Parser può indicare al massimo cinque tipologie di aree, corrispondenti a titoli, testo, fotografie, disegni e diagrammi, pubblicità (Fig. 2).

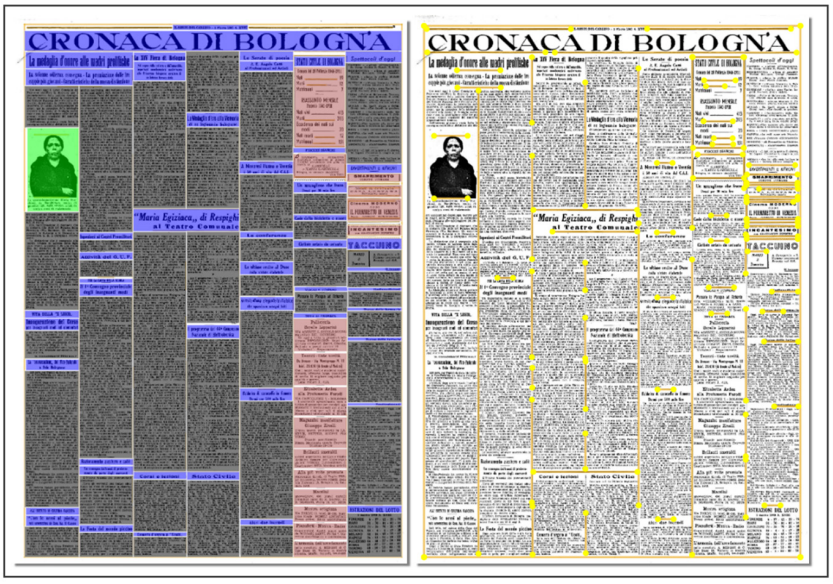


Fig. 1
Modalità di lavoro della piattaforma: a sinistra, tramite selezione delle sezioni: titoli - in blu; testo -in grigio; fotografie - in verde; annunci - in rosa; a destra, tramite modifica delle linee guida - evidenziate in giallo.

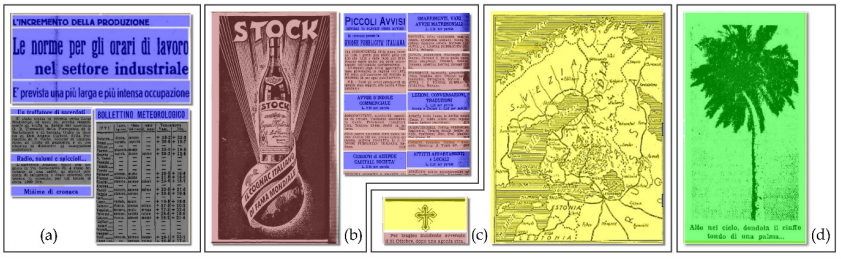


Fig. 2
Tipologie di testo identificabili dal sistema: (a) titoli e sottotitoli - in blu, testo e tabelle - in grigio; (b) avvisi ed inserzioni pubblicitarie - in rosa; (c) differenti tipologie di elementi grafici - in giallo; (d) fotografie - in verde.

Linee guida per l'annotazione (R. Pansini)

La costruzione della ground truth per il fine-tuning di Layout Parser ha comportato l'elaborazione di specifiche linee guida per l'annotazione.
Per garantire una comprensione completa del contesto, gli elementi sono stati considerati sia dal punto di vista tipografico, sia in relazione alle circostanze editoriali che hanno determinato la loro presenza nelle pagine del quotidiano.
I titoli di sezione, articolo e paragrafo, appartengono alla stessa categoria, ma sono distinti da una propria bounding box (Fig. 2(a)).
I testi sono raggruppati senza distinzione tra paragrafi e sottoparagrafi, ma con una separazione specifica per elementi particolari, come le tabelle (Fig. 2(a)).
Le inserzioni pubblicitarie comprendono tutti gli annunci pubblicati in seguito a una richiesta e a un pagamento da parte di privati. Questa categoria include sia le pubblicità testuali sia quelle contenenti elementi grafici, nonché gli annunci di nascite e decessi. Le inserzioni sono state annotate ciascuna con una propria bounding box, senza distinguere eventuali elementi grafici (disegni, scritte, fotografie, Fig. 2(b)).
Le fotografie sono incluse in una categoria separata insieme alle relative didascalie (Fig. 2(c)).
Infine, i disegni includono tutte le rappresentazioni grafiche, siano esse realizzate a mano o tramite altre tecnologie grafiche, insieme alle rispettive didascalie, quando presenti. Sono comprese in questa categoria anche le icone e i simboli utilizzati in alcuni tipi di annuncio, come i necrologi (Fig. 2(d)).

Bibliografia

[1] https://www.echoes-eccch.eu/
[2] https://digitallibrary.cultura.gov.it/pnrr-cultura/
[3] https://layout-parser.readthedocs.io/en/latest/
[4] Silvano Imboden, Giorgia Cardano, Corrado Consiglio, NEW PERSPECTIVES IN MANAGING HERITAGE DOCUMENTS, in A. Guidazzoli, M.C. Liguori (Eds.), AI, Cultural Heritage, and Art. Between Research and Creativity. Workshop proceedings – February 9-10, 2024. https://doi.org/10.1388/IIIWORKSHOPAIBC
[5] https://github.com/Layout-Parser
[6] https://huggingface.co/Eterna2/LayoutParser
[7] "The Newspaper Navigator Visual Content Recognition Model" https://github.com/LibraryOfCongress/newspaper-navigator
[8] https://svelte.dev/docs/kit/introduction
[9] https://openseadragon.github.io
[10] https://d3js.org
[11] https://pocketbase.io

Autori

Fauzia Albertin è una fisica specializzata nella diagnostica scientifica per i Beni Culturali, con competenze che spaziano dalla tomografia a raggi X all'analisi chimica dei

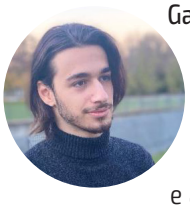
materiali.
Attualmente lavora presso CINECA (Bologna, Italia), dove partecipa al Programma Nazionale di Digitalizzazione (PND) a cura del Ministero della Cultura (MiC), dedicato alla digitalizzazione del Patrimonio Culturale italiano e all'impiego di tecnologie avanzate, come l'intelligenza artificiale, per l'analisi e l'arricchimento dei dati.
Ha preso parte a importanti iniziative europee nel campo della digitalizzazione come Time Machine Europe, e ha svolto attività di ricerca post-dottorato in diverse istituzioni italiane ed europee. Dal 2013 al 2017 ha ricoperto il ruolo di Lead Scientist del progetto Virtual X-ray Reading presso l'EPFL (Lausanne, Svizzera), dove ha lavorato allo sviluppo di una tecnica innovativa di tomografia a raggi X per la digitalizzazione di manoscritti antichi.

Antonella Guidazzoli è responsabile del VisiT Lab del dipartimento di High Performance Compuitng di Cineca. dove coordina attività di ricerca su computer grafica, virtual heritage, tecnologie immersive e Intelligenza Artificiale applicata ai beni culturali. Laureata in Ingegneria Elettronica (1988) e in Storia con lode (2007) all'Università di Bologna, unisce competenze tecnico-scientifiche e umanistiche per innovare la fruizione e la conservazione del patrimonio culturale. Ha realizzato progetti pluripremiati come Apa l'Etrusco ed Experience Etruria, contribuendo alla creazione di mostre, musei virtuali e narrazioni interattive. Partecipa regolarmente a conferenze internazionali di rilievo (come SIGGRAPH e Digital Heritage) e cura il workshop biennale su AI, Patrimonio, Arte e Scienza. È una convinta sostenitrice dell'uso dell'Intelligenza Artificiale e dei digital twin in una prospettiva di collaborazione uomo-macchina, promuovendo approcci etici e creativi. Appassionata di tecnologie emergenti, è anche una divulgatrice entusiasta del quantum computing.

Simona Caraceni si occupa di comunicazione e multimedialità con i nuovi media e dal '94 si occupa di applicazioni dei nuovi media e delle nuove tecnologie nella comunicazione e nell'arte. Ha insegnato presso le Università di Bolzano, Milano, Firenze, Macerata e Bologna. Giornalista freelance, è editorialista e editorialista per Artribune http://www.artribune.com e per altre testate giornalistiche. Ha scritto un libro sui musei virtuali nel 2012 e numerosi articoli scientifici. Dottore di ricerca presso il Planetary Collegium dell'Università di Plymouth (Regno Unito), coordinatrice del patrimonio virtuale presso Cineca, la sua attività di ricerca riguarda musei virtuali, interfacce uomo-macchina, multimedialità e multimodalità, e nuovi linguaggi mediatici della comunicazione nel campo del patrimonio. Nell'ambito della ricerca applicata si occupa delle opportunità della comunicazione online, delle interfacce multimediali e multimodali, della realtà aumentata, del patrimonio comunicativo e dell'e-learning con l'Università di Bologna, il Comune di Bologna, la Regione Emilia Romagna e Cineca. Membro del Consiglio Direttivo di AVICOM dal 2009 al 2013, eletta Vicepresidente di AVICOM nel 2013 e Segretario Generale nel 2015, ha fondato e coordinato la commissione italiana "Audiovisivi e nuove tecnologie" per ICOM-Italia dal 2007 al 2017.

Silvano Imboden, laureato in Scienze dell'Informazione, dal 2000 lavora presso il dipartimento HPC del CINECA di Bologna. Le sue principali competenze riguardano la computer grafica, la visualizzazione scientifica, la progettazione e lo sviluppo di applicazioni e framework. Nel corso degli anni ha partecipato a numerosi progetti finanziati, affrontando

temi quali: analisi forense, rendering in tempo reale e offline, produzioni video, archeologia virtuale, elaborazione di dati sismici e interfacce utente/interazione. Più recentemente si sta occupando dell'utilizzo di tecniche di intelligenza artificiale per il trattamento di dati nel campo dei Beni Culturali Digitali.



Gabriele Marconi è un HPC Scientific Application Specialist, attualmente impegnato in CINECA in progetti multidisciplinari. Si occupa dello sviluppo di applicazioni web, in particolare per la creazione di interfacce dedicate al fine-tuning e all'interazione con modelli di intelligenza artificiale. Parallelamente, è coinvolto in progetti di grafica 3D, sia come modellatore che come sviluppatore. Collabora inoltre a iniziative dedicate alla gestione e alla valorizzazione del patrimonio culturale, in cui le tecnologie digitali giocano un ruolo centrale. È laureato in Ingegneria Informatica presso l'Università di Bologna (Alma Mater), dove ha svolto una tesi sulla realizzazione di ambienti virtuali per la simulazione tridimensionale.

Rossella Pansini è archeologa e digital humanist, attualmente impegnata in CINECA in progetti che riguardano la gestione e valorizzazione dei beni culturali, tra cui lo sviluppo dell'ecosistema IPaC (Infrastruttura per il Patrimonio Culturale), promosso dal Ministero della Cultura – Digital Library. In passato è stata assegnista di ricerca e docente universitaria a contratto e ha partecipato a progetti nazionali e internazionali dedicati all'analisi e alla valorizzazione di siti archeologici e documenti d'archivio. I suoi interessi principali riguardano l'applicazione delle tecnologie allo studio dei beni culturali, con particolare attenzione all'IA, al GIS, alla fotogrammetria e alla ricostruzione 3D.

DOI: 10.26314/GARR-Conf25-proceedings-19

BIOBANCA VIRTUALE WOAH

Domenico Nilo Mazza

IZSLER: Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna "Bruno Ubertini"

Abstract. L'IZSLER ha sviluppato un progetto per realizzare un network di biobanche e condivisione di materiale biologico di qualità per la ricerca: WOAH-VB, WOAH VirtualBiobank. Il progetto prevede fasi: 1 lo sviluppo del sistema in un gruppo di laboratori per verificarne l'operatività 2 la sua diffusione presso i laboratori WOAH e realtà terze interessate. Un utente potrà connettersi al portale WOAH-VB e chiedere la disponibilità di materiali; la richiesta passerà alle biobanche collegate per presentare le risposte al richiedente. Il portale farà inoltre da repository della documentazione in materia.

Keywords. Biobanche Veterinarie, Biobanche Virtuali, Virtual Biobank, WOAH

Introduzione.

L'Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (nel seguito indicato come "IZSLER") ha sviluppato negli anni un proprio sistema di conservazione dei campioni biologici in ambito veterinario (Biobanca) perseguendo alcuni scopi ben precisi (dal sito http://www.ibvr.org/AboutUs.aspx):

1. Accedere rapidamente e facilmente ai materiali biologici della biobanca, con informazioni dettagliate.
2. Ampliare la varietà e il numero di campioni conservati, rendendoli disponibili sia all'interno di IZSLER che alla comunità scientifica.
3. Produrre materiali con processi standardizzati, con livelli di qualità elevati e conservati in ambienti controllati.

Il progetto ha via via coinvolto altri Istituti Zooprofilattici Sperimentali (IIZZSS) in Italia che hanno nel tempo adottato la stessa soluzione all'interno di un progetto di network di valenza internazionale volto a mettere a fattore comune e rendere disponibile all'intera comunità scientifica un vasto ed articolato patrimonio di materiale biologico veterinario utile e necessario alle attività di ricerca in tale campo.

Il progetto Vitual Biobank WOAH

Da questa particolare ed importante esperienza è nato nel 2019 il progetto volto a mettere in rete a livello internazionale il maggior numero di biobanche nel mondo, a partire da quelle detenute dai laboratori di riferimento della World Organisation for Animal Health

Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

(WOAH), nell'ottica di realizzare una rete mondiale per la condivisione di materiale biologico di elevata qualità ed interesse per la ricerca scientifica in ambito veterinario e non solo.
Dall'esperienza specifica di IZSLER, arricchita dalle esperienze di altri soggetti internazionalmente competenti in materia, e dalle esigenze di WOAH per la realizzazione di una rete di tali caratteristiche è nato il progetto WOAH-VB, WOAH Virtual Biobank.
Il progetto si articola in due fasi principali, così riassumibili:

1. nella prima è stato sviluppato, collaudato e sta per essere messo in esercizio il sistema nelle sue diverse componenti, in un numero ristretto di laboratori di riferimento WOAH, corrispondente alla rete italiana delle biobanche veterinarie gestite dai vari IIZZSS, allo scopo di testare l'operatività del sistema e verificare la sua rispondenza alle esigenze del WOAH, dei suoi laboratori e degli utenti; la durata di tale fase sperimentale è prevista in circa tre anni;
2. nella seconda fase si procederà alla diffusione del sistema presso altri laboratori WOAH ed all'integrazione di quelle realtà che pur non adottando completamente il sistema oggetto della presente soluzione, intendono comunque integrarsi alla WOAH-VB.

Quello implementato è un sistema a due livelli e tre componenti che permette di mettere in rete le biobanche dei laboratori di riferimento WOAH e di altri laboratori, allo scopo di realizzare un sistema virtuale integrato al servizio dei ricercatori.
Lo schema di massima prevede che un utente registrato possa connettersi al portale WOAH -VB per richiedere la disponibilità di determinati materiali; il portale inoltrerà la richiesta alle singole biobanche collegate attendendo da loro per un tempo definito le relative risposte, che saranno poi riepilogate e presentate all'utente.
Il portale fungerà inoltre da repository documentale di tutta la documentazione inerente il proprio contesto operativo, la normativa e le linee guida in materia, documentazione che sarà in parte pubblica e liberamente accessibile.
La figura seguente riassume l'architettura del sistema e le sue componenti principali:

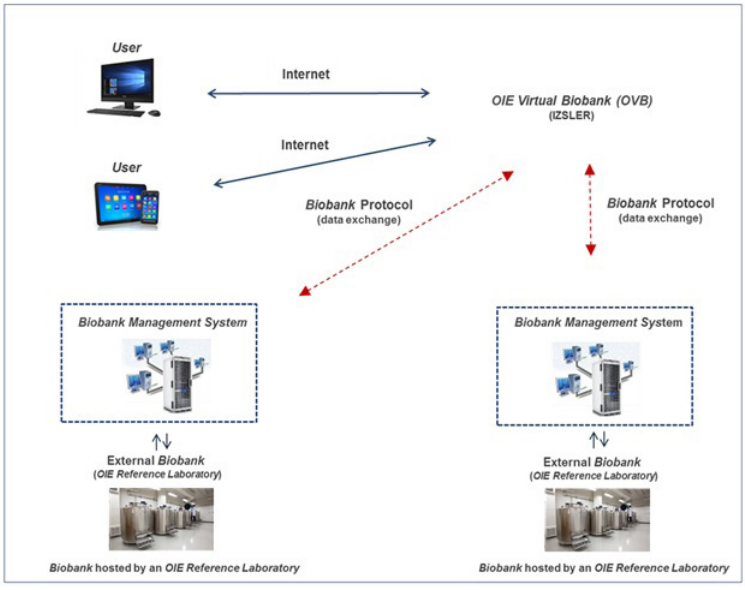


Fig. 1
Architettura del sistema

Le componenti del sistema sono le seguenti:

- il Portale WOAH Virtual Biobank (WOAH-VB): sistema pubblico posto in cloud che mette in rete le biobanche collegate e rappresenta il punto di accesso unico al sistema per gli utenti che necessitano di ricercare determinati materiali biologici;
- il Biobank Management System (BMS) che gestisce localmente la biobanca ed i materiali in essa conservati, controlla il processo di movimentazione interna/esterna dei campioni e si occupa di comunicare con il WOAH-VB per la messa in rete dei propri dati;
- il Protocollo che definisce ed implementa le regole di comunicazione fra il portale WOAH-VB ed i differenti BMS collegati garantendo la piena apertura a sistemi di gestione delle biobanche differenti dal BMS che volessero connettersi al sistema.

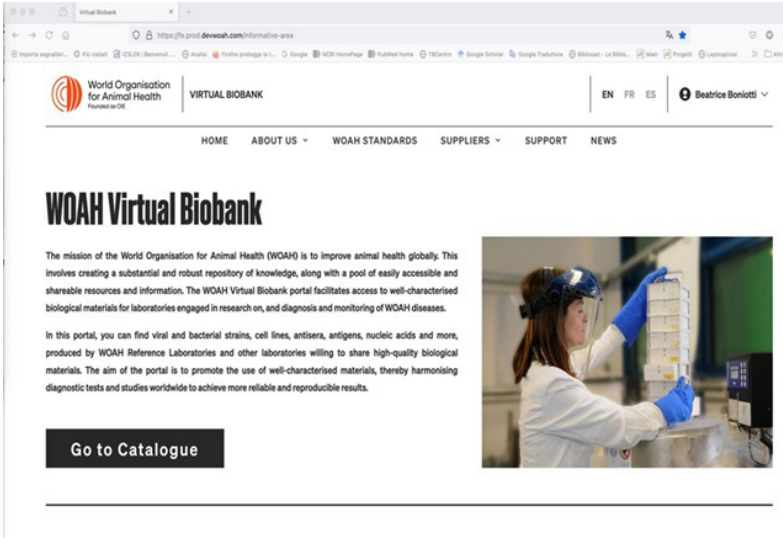


Fig. 2
La home Page del portale

Fra le principali funzion implementate sono previste:

- catalogo dei materiali
- funzioni di ricerca avanzate
- richieste di materiali
- mini-biobanca locale per i piccoli laboratori
- documentazione e linee guida
- protocolli ed agreement

Fra i punti di attenzione del progetto si possono evidenziare:

- portale ad elevata scalabilità ed affidabilità
- BMS operativo su più architetture
- apertura a sistemi terzi
- integrazione di entità minori

- standardizzazione informazioni sui materiali
- ampia possibilità di documentazione materiali
- gestione dell'intero ciclo operativo

Il cronoprogramma iniziale prevedeva la messa in produzione nel corso del 2022, ma la pandemia da Covid19 ha fermato il progetto per il forte coinvolgimento dell'IZSLER nel supporto all'attività di laboratorio degli Enti Sanitari.
La pianificazione aggiornata prevede la piena entrata a regime del sistema per la rete italiana entro la fine del 2025, per poi procedere con l'estensione ai laboratori WOAH nel mondo a partire dal 2026.

L'intero progetto verrà rilasciato in modalità Open Source secondo la GNU General Public License della Free Software Foundation

Conclusioni

Il progetto si pone l'obiettivo di realizzare un ambiente aperto e multipiattaforma per la condivisione certificata di materiale biologico: la sua struttura è modulare ed estensibile, e non ne limita l'utilizzo all'ambito veterinario.
Il sistema ha permesso di realizzare uno strumento di condivisione in rete in ambito scientifico sfruttando le tecnologie dell'informazione per estenderne la fruibilità ad una platea la più ampia possibile consentendo la partecipazione del maggior numero di laboratori certificati.
Un particolare ringraziamento va alla Dr.ssa Beatrice Boniotti, responsabile WOAH del progetto, ed alla Dr.ssa Anna Mor, coordinatrice della Biobanca IZSLER.

Autore

Domenico Nilo Mazza danilomazza64@outlook.it danilo.mazza@izsler.it
Informatico laureato a Pisa, ha lavorato come progettista in differenti realtà dell'ICT quali Olivetti, Cap Gemini, TIM operando in ambito industriale, finanziario, logistico e trasporti. Lavora in sanità dal 2005 come Responsabile dei Sistemi Informativi prima in strutture private e dal 2009 come Dirigente in strutture pubbliche italiane. Attualmente è responsabile dell'ICT dell'IZSLER, del quale è RTD, dove si occupa sia di progetti di digitalizzazione infrastrutturale ed applicativa a supporto dell'attività sanitaria e della ricerca. Segue inoltre vari gruppi di lavoro in ambito ICT

DOI: 10.26314/GARR-Conf25-proceedings-20

L'Archivio videoteatrale di Giacomo Verde. Il progetto I_PAD

Anna Maria Monteverdi

Università Statale di Milano, Dipartimento Beni culturali e ambientali

Abstract. Il progetto I_PAD, vincitore del bando Prin 2022, ha permesso di digitalizzare e conservare a lungo termine, il patrimonio archivistico di Giacomo Verde (1956-2020), videomaker e tecno performer. Per la prima volta fotografie e video degli anni Ottanta sono usciti da un vecchio archivio analogico per mostrare l'esplosione di creatività di un artista che ha esplorato in quarant'anni di attività, diversi linguaggi, dal teatro al video alle installazioni, "frantumando generi", come lui amava dire. Pioniere della videoarte del videoteatro, è stato tra i primi in Italia a dare vita a opere di computer graphics, animazione e net art creando in Italia il primo spettacolo di narrazione interattivo, Storie Mandaliche (1998-2001). La sua arte, è sempre stata finalizzata a una comunicazione diretta, autentica, a un'idea di impegno sociale e politico. Verde ha esplorato l'uso intelligente e alternativo della televisione e in generale, delle tecnologie. Verde mescolava follia, gioco, a una radicale critica al sistema della cultura ufficiale. L'ICT della Statale di Milano ha affiancato il team di ricerca per preservare i materiali dell'artista grazie al sistema innovativo di archiviazione digitale ARKIVE

Keywords. Giacomo Verde, Videoteatro, Arkive, I_PAD, archivi videoteatrali

Introduzione

Il progetto I_PAD (Italian Performance Archive in Digital)[1], vincitore del bando PRIN 2022 - Progetti di ricerca di rilevanza nazionale - finanziato dal MUR, riunisce studiosi di due Università italiane (Università di Milano e Università Link di Roma) e del CNR-ISTI (Pisa); tutti i soggetti sono stati coinvolti nelle diverse fasi del lavoro. Il focus riguarda il restauro, la digitalizzazione e la disseminazione dei materiali dell'archivio video dell'artista Giacomo Verde, già considerato dal Ministero dei Beni Culturali, di interesse nazionale. L'obiettivo del progetto è dimostrare l'importanza di documentare, salvaguardare, promuovere e diffondere la memoria del videoteatro e della videoarte italiana degli anni '80 e '90 a partire dall'Archivio di Verde. Il progetto si articola in due percorsi interconnessi e paralleli: uno teorico, in cui sono state selezionate, visionate e analizzate le più significative opere video teatrali, e l'altro di natura pratica, in cui ci si è focalizzati sul loro recupero e digitalizzazione, prevedendo una conservazione attraverso specifici repository on line, dopo una catalogazione, descrizione analitica e metadatazione del contenuto.

La conservazione dei dati implica una riflessione sul metodo per renderli accessibili e riutilizzabili in futuro: l'archivio video, ben organizzato ma fragile e non di facile accesso,

1 Per maggiori dettagli sul progetto I_PAD: Monteverdi, A. M. (2024), Giacomo Verde's Archive. Mimesis Journal, 13(2), 55–63. https://doi.org/10.13135/2389-6086/10107

Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

necessitava di un restauro tecnico, una digitalizzazione e successivamente di un nuovo modello di archiviazione per la conservazione a lungo termine. Alla fine del processo conservativo è stato creato un ambiente VR per la fruizione virtuale. I_PAD ha aderito anche al progetto Dataverse della Statale che prevede un deposito Fair dei data set della ricerca. @DataUNIMI è l'unico repository universitario in Italia ad aver ottenuto il Core TrustSeal, certificazione internazionale per l'affidabilità del repository e l'elevata qualità dei dati della ricerca depositati, imponendo altissimi standard di qualità e di controllo nella gestione e pubblicazione dei dati della ricerca.
È possibile consultare il progetto nella sua completezza dal sito: ipadprin.isti.cnr.it

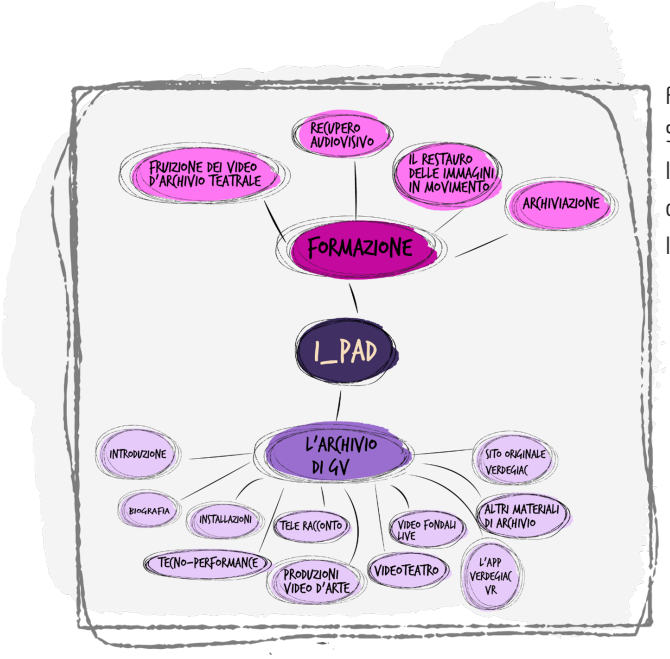


Fig. 1
Schema del progetto I_PAD. Grafica di Francesca Pardini (CNR/ISTI)

Giacomo Verde

Giacomo Verde (1956-2020) è stato uno dei protagonisti della videoarte e del videoteatro in Italia. In oltre quarant'anni di attività ha realizzato centinaia di installazioni elettroniche e opere video a carattere teatrale, autoprodotte o commissionate da grandi Fondazioni, presentandole nelle più importanti rassegne di videoarte italiana, promuovendo sistematicamente la sua personale estetica della "bassa definizione". Teknoartista e videoartivista, comincia a occuparsi di teatro e arti visive negli anni Settanta. A partire dagli anni Ottanta realizza oper'azioni collegate all'utilizzo creativo della tecnologia "povera": opere di videoarte, tecno-performances, spettacoli teatrali, installazioni. Si cita a questo proposito il 1° Video Totem Est-Etica Antica-t-Astr-Fica, esposto nel 1986 alle Gallerie Civiche di Arte Moderna di Palazzo dei Diamanti, a Ferrara, che segna l'inizio di questa tipologia di produzioni. Sempre negli anni Ottanta, Verde fonda il gruppo di teatro-musica Bandamagnetica con cui comincia a fare capolino in diverse trasmissioni televisive siglate RAI. Nel 1989, vince il concorso "Le scritture del visibile" al Pow - Progetto Opera Videoteatro – di Narni,

con lo storyboard Stati d'animo ispirato al trittico di Boccioni, all'origine dell'opera video omonima realizzata l'anno seguente in computer grafica. Agli anni Novanta si deve la nascita del Progetto Tele-Racconto - performance teatrale che coniuga narrazione, microteatro e macro ripresa in diretta – la cui tecnica sarà utilizzata anche nei video-fondali-live presentati in seguito, a partire dal 1998, in concerti, recital di poesia e spettacoli teatrali in giro per il mondo. Agli anni Novanta risalgono anche le prime installazioni interattive, inaugurate da Degli Avi (1992) e l'inizio delle collaborazioni internazionali, come nel caso della Van Gogh TV di Amburgo (siamo sempre nel 1992), con cui realizza il progetto di tv interattiva Piazza Virtuale per Documenta IX di Kassel. Le collaborazioni si moltiplicano anche in Italia e nascono nuove oper'azioni e anche dei personaggi virtuali, a cui dà, ad esempio, "vita" con un cyberglove, come nel progetto Euclide di Stefano Roveda (1994). Tra gli anni Novanta e Duemila, Giacomo Verde è stato tra i primi italiani a realizzare opere di arte interattiva e net-art, a intraprendere un uso creativo del cellulare, a includere nelle proprie produzioni l'impiego del QR code, creando connessioni tra i diversi generi artistici.

Restauro dell'Archivio video: l'intervento di UniMi e UniLink

L'Archivio Giacomo Verde è esemplificativo dello sviluppo tecnologico del video (prima analogico poi digitale) perché tutti i formati sono rappresentati al suo interno. Data la grande quantità di copie in VHS è stato necessario capire il formato sorgente e rintracciare il Master (o al limite, la duplicazione DUB o SUB Master) per effettuare la digitalizzazione a partire dai materiali originari di miglior qualità. Prima di intervenire con la digitalizzazione bisognava ripristinare le migliori condizioni possibili per le videocassette. Il Team di I_PAD ha selezionato decine di cassette Vhs, Betacam e U-Matic dell'Archivio, purtroppo già in cattivo stato di conservazione; successivamente sono state portate allo studio professionale Plurimedia di Treviso di Gabriele Coassin che aveva conosciuto Giacomo Verde negli anni Novanta e con il quale aveva collaborato a diverse produzioni artistiche. Le problematiche riscontrate erano diverse: dallo Sticky Shed Syndrome o sindrome da nastro colloso, al degrado delle immagini in forma di segmenti di linea (drop out) o jitter di quadro (disallineamento). Gli interventi avevano l'obiettivo di ripristinare le condizioni originarie attraverso metodi diversi: apertura delle cassette e pulizia con lubrificante al teflon per portare via muffa e polvere, riavvolgimento del nastro giocando con la trazione del videoregistratore, "cottura" ad alte temperature, uso dell'apparecchio TBC (Time Base Corrector) per il riallineamento.

I_PAD e Arkive

La conservazione dei dati implica una riflessione sul metodo per renderli accessibili e riutilizzabili per lungo tempo, garantendo:

1. Interpretabilità del contenuto
2. Leggibilità (tecnica) del file
3. Integrità del file e del contenuto

I_PAD è tra i primi progetti di archiviazione selezionati in tutto l'Ateneo di Milano per

essere inserito in Arkive[2], un'infrastruttura ideata per la conservazione a lungo termine dei dataset della ricerca universitaria.
Le caratteristiche di Arkive si possono riassumere in:

– Grande quantità di storage
– Scalabilità
– Interoperabilità
– Automazione dei processi e delle operazioni di data curation
– Policy per la cura e la gestione dei dati
– Affidabilità, certificazione dell'archivio

Arkive ha un'infrastruttura modulare basata sullo standard di riferimento per gli archivi digitali OAIS. Il sistema di storage è basato su tecnologia cloud Swarm Datacore (S3) su server di proprietà e su rete di UNIMI. Il "motore" della gestione dei dati è Archivematica che permette di analizzare, normalizzare, impacchettare i dati e spostarli nel deposito. iRODS è il componente che realizza la virtualizzazione del file-system permettendo di modificare ogni parte del sistema lasciando inalterata la struttura dell'archivio[3].

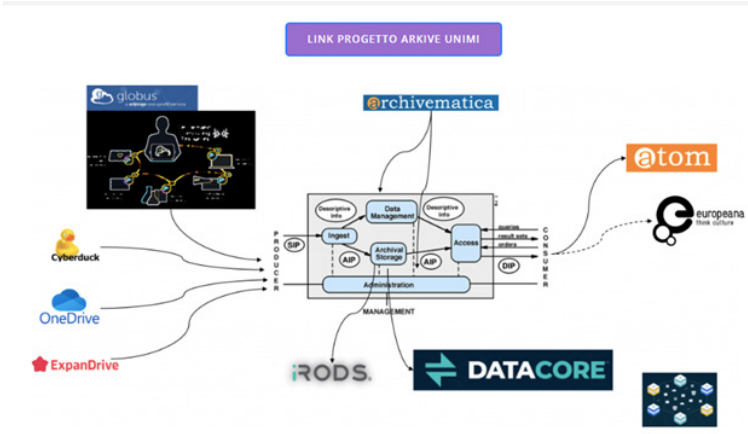


Fig. 2
Schema del funzionamento di Arkive (su gentile concessione di ICT UniMi)

Arkive per I_PAD

Per I_PAD è stato sviluppato grazie all'ICT della Statale, un workflow che segue diverse fasi cruciali: dalla generazione di un'impronta digitale dei file alla loro verifica, all'arricchimento con metadati e, quando possibile, all'estrazione del testo. Al termine, ogni dato è confezionato in un pacchetto di archiviazione (AIP) che viene inserito nel sistema Arkive. In alcune circostanze viene creato anche un pacchetto di distribuzione (DIP), così da permettere il riuso dei dati in applicazioni esterne. Una volta completata la procedura di ingestion, gli AIP vengono conservati nella sezione centrale dell'archivio, inaccessibile

2 Il Gruppo di progetto di ARKIVE è composto da: Giorgio Bagnato; Michele Sciarabba; Federica Zanardini e Matteo Zoppi.
3 Su Arkive cfr: F. Zanardini Progetto Arkive: un'infrastruttura per l'archiviazione a lungo termine dei dati della ricerca dell'Università degli Studi di Milano, GARR Conference proceedings, 2022 (on line)

dall'esterno. I proprietari dei dati possono accedere e scaricare le informazioni che li riguardano. La finalità è garantire al fondo digitale IPAD una conservazione che sia a lungo termine e che consenta, in futuro, di adattare i materiali a nuove tecnologie e modalità di fruizione. La creazione di dati aggregati è l'elemento innovativo del lavoro: non è stata effettuata una semplice digitalizzazione e organizzazione dei file, ma si è creato una serie di connessioni significative tra i materiali, affinché l'utente finale possa accedervi seguendo percorsi specifici e suggeriti.

Contenuto dell'Archivio Verde in Arkive: l'intervento di UniLink

L'archivio Verde su Arkive -denominato IPAD – GIAC -è suddiviso in 11 collezioni, ciascuna rappresentante una tipologia artistica. Ogni collezione contiene opere specifiche, ognuna racchiusa in un pacchetto. Ogni pacchetto include una cartella DATA, contenenti i dati multimediali alla massima risoluzione, ed una cartella METADATA che contiene un file json con tutte le informazioni dettagliate: il titolo dell'opera, la tipologia dell'oggetto (ad esempio: audiovisivo, sonoro, fotografico, testuale) e i rispettivi numeri, insieme a tutte le tracce digitali e i metadati della fonte.

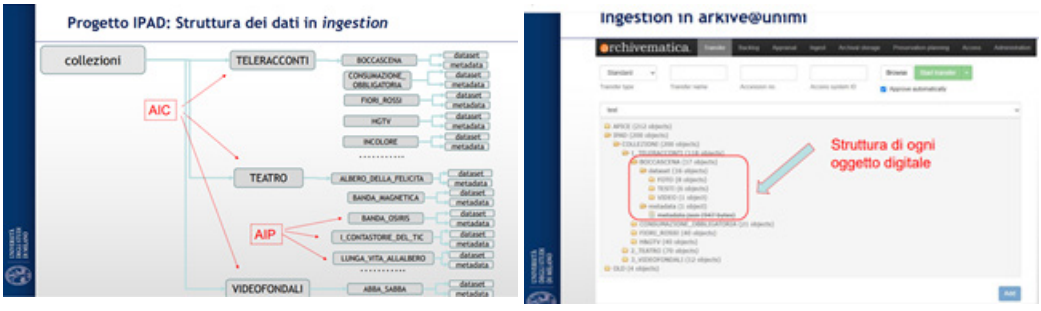


Fig. 3
Schemi di suddivisione della struttura dati del progetto I_PAD per Arkive

L'Archivio Verde in VR: l'intervento del CNR/ISTI

Alcune delle installazioni video o interattive documentate nell'Archivio Verde, il cui valore è nell'esperienza diretta e quindi non efficacemente descrivibili da documentazione visiva, sono state ricreate in Virtual Reality. Questa metodologia ha permesso di dare nuova vita ad installazioni che, a causa dell'obsolescenza dei materiali HW/SW impiegati, non sono più riproponibili al pubblico. L'approccio utilizzato dal team del CNR coordinato da Massimo Marini rappresenta un esempio di metodologia potenzialmente applicabile a tutte le installazioni di arte elettronica basate su una tecnologia ormai obsoleta. Uno degli obiettivi principali del progetto I_PAD è quello di riportare in vita un piccolo numero installazioni, riproducendole in un ambiente virtuale. In questo modo, è possibile offrire un'esperienza simile a quella originale in un contesto immersivo, facilmente fruibile grazie ai visori VR portatili di nuova generazione.

L'applicazione ha la sua scena base collocata in un piccolo teatro dove è allestita una scenografia cyberpunk che rimanda agli immaginari tipici degli anni più attivi di Verde (1990-2005). Sul palco, interagendo con un mandala che richiama all'opera teatrale Storie Man-

daliche, si può accedere a quattro scene diverse, che rappresentano altrettante categorie di opere di Verde:

– Interno Neve (installazione interattiva)
– Bit (personaggio virtuale)
– Videoloop/Videofondali (azione videoperformativa)
– Video (selezione video dall'archivio)

Fig. 4
Frame dagli ambienti immersivi

Interno neve Questa scena ripropone in realtà mista l'installazione omonima di VerdeGiac. Originariamente, il pubblico interagiva con quattro enormi fiocchi di neve proiettati su teli in tulle sensorizzati. In VerdeGiac VR, grazie alla modalità passthrough dei visori XR, gli stessi fiocchi compaiono nella stanza reale dell'utente, integrandosi visivamente con l'ambiente circostante. Il risultato è un'esperienza immersiva e poetica, dove la materia digitale fluttua nello spazio fisico.

Bit In questa scena torna Bit, il pupazzo virtuale ideato da Verde e già protagonista di numerose sue opere. L'interazione assume qui una forma ludica: il giocatore guida Bit su una scala composta da scatole etichettate con parole tratte dal vocabolario di Verde, cercando di comporre una sua frase caratteristica. L'estetica e la meccanica di gioco richiamano il celebre videogame Qbert, offrendo un omaggio ironico e affettuoso al linguaggio verbale e visivo dell'artista.

Videoloop Questa scena rende omaggio a una delle tecniche più distintive di Verde: il loop video generato da retroazione visiva. Verde era solito inquadrare un monitor con una microcamera, inviando il segnale simultaneamente al proiettore e di nuovo al monitor stesso. Interponendo oggetti tra camera e schermo, otteneva ripetizioni infinite e distorsioni geometriche e cromatiche, influenzate dalla posizione della camera e dai suoi parametri (esposizione, tinta, ecc.). In VerdeGiac VR, questa tecnica analogica è stata fedelmente ricostruita attraverso un modulo software originale, che simula anche la leggera latenza di uno o due frame, fondamentale per ricreare l'effetto autentico del loop.

Videoinstallazioni Sfruttando le capacità dei nuovi visori Meta di mappare lo spazio circostante, questa scena permette all'app di "appendere" virtualmente degli schermi alle

pareti dell'ambiente reale dell'utente, visibile grazie al passthrough. Al centro della stanza compare la TV modificata da Verde, uno dei suoi oggetti iconici. Sia sugli schermi virtuali che sulla TV vengono proposti alcuni dei video più significativi dell'artista, trasformando lo spazio fisico in una videoinstallazione dinamica e personale

Autore

Anna Maria Monteverdi anna.monteverdi@unimi.it

DOI: 10.26314/GARR-Conf25-proceedings-21

LibRA: A Tool for Researcher Metrics Management

Chiara Rebuffi[1], Roberto Cavanna[2], Paolo Uva[2]

[1]Scientific Direction, IRCCS Istituto Giannina Gaslini, Genoa, Italy, [2]Clinical Bioinformatics, IRCCS Istituto Giannina Gaslini, Genoa, Italy

Abstract. Research conducted at the Gaslini Pediatric Hospital is primarily funded by the Italian Ministry of Health based on bibliometric indicators of scientific performance. The analysis of these metrics can be facilitated by software that automates bibliometric processes and calculations for use by administrative staff in research evaluation. To this end we developed LibRA (Library for Research Assessment), a Python-based application which leverages application programming interface (API) technology to query relevant databases (e.g., Scopus – Elsevier) and retrieve performance indicators in real time

Keywords. Research evaluation, bibliometric analysis, APIs, citation databases

1. Introduction

The Giannina Gaslini Institute is a pediatric hospital that integrates healthcare and assistance with research activities, aiming to advance biomedical knowledge. This is largely made possible thanks to funding from the Italian Ministry of Health (MoH). Research and clinical activities undergo a competitive assessment among the IRCCS based on the scientific performance for the allocation of Ricerca Corrente funds, as well as additional funding opportunities such as Ricerca Finalizzata, Piano Nazionale di Ripresa e Resilienza (PNRR), and thematic networks.

Approximately 10,000 researchers work in these institutes whose career progression is regulated by a merit-based, competitive system aligned with the standards of public healthcare research institutions.

Performance evaluation is therefore conducted at multiple levels — including the institute, research groups and individual researchers. The analysis of the parameters on which this evaluation is based plays a crucial role in the allocation of funding.

This evaluation increasingly relies on bibliometric indicators, such as the h-index and citation index which are calculated using proprietary databases selected by the MoH. Among these are databases provided by Elsevier, such as Scopus (for citation data) and SciVal (for citation analysis) (Research Intelligence, 2019). The relevant regulations specify not only the tools to be used for bibliometric analysis, but also the methodology to be followed and the thresholds to be met. For example, the use of unique author identifiers (e.g. Scopus ID) is mandatory for retrieving scientific publications. To complete the literature collection PubMed APIs were also used.

These assessments are managed by the staff of the Scientific Directions, who are involved

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

in the preliminary eligibility checks for funding calls, in monitoring the validated results communicated by the MoH, and in executing the internal redistribution of funds.
These processes require significant staff involvement and are highly time-consuming, as the data often undergo multiple normalization steps to ensure fairness across different research areas, researcher seniority, and other variables. Moreover, these evaluations typically rely on large-scale data extraction and manipulation which, in the absence of adequate technological support, may introduce a considerable margin of error.
To address these challenges and support the transition towards a data-driven research hospital, a web application for automating bibliometric data collection has been proposed as a pilot initiative. The hosting of LibRA on the GARR Virtual Data Center is currently under evaluation.

2. Methods

2.1 System overview

LibRA is composed of two web interfaces - Administrators and Researchers - built using Streamlit (https://streamlit.io), a Python-based open-source framework for creating interactive data applications. Streamlit provides the web interface, while Pandas, Psycopg2, and Requests python libraries are used for data handling, database interaction, and API calls, respectively. Configuration settings, such as API keys and parameters, are stored securely in a secrets.toml file. Authentication for the Researcher web application is implemented using basic session handling within Streamlit, while the Administrator interface is protected through internal network access and/or external authentication mechanisms, depending on deployment context.
The Administrator and Researcher web applications share a common backend and are connected to a PostgreSQL relational database and external bibliometric data sources via public Scopus (https://dev.elsevier.com/api_docs.html) and PubMed APIs.
An overview of the architecture is provided in Figure 1. At the core of the system is a central server that coordinates data retrieval, processing, and interaction between components.
The core workflow involves:

- Importing demographic data via Excel file
- Fetching publication metadata and citation metrics from Scopus and biomedical literature from PubMed through APIs
- Computing bibliometric metrics in Python across different timeframes. The system also detects first, last, and corresponding authorship roles to identify the contribution in each publication
- Storing metrics and raw metadata in a PostgreSQL database for reuse and comparison
- Displaying and editing information via the Admin or Researcher interface to explore and filter data, view metrics in tabular form, and export reports and publication lists in Excel format

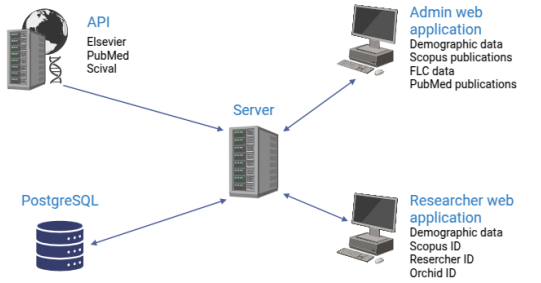


Fig. 1
Overview of the software architecture

3. User interface

LibRA has two web interfaces for managing and viewing bibliometric data.

3.1 Admin web application

The Admin interface is used to manage demographic data, update publication records, compute bibliometric indicators and authorship roles (first, last, corresponding) from Scopus and aggregate metrics across departments or researcher age groups. It is composed of 8 pages:

- Home. Includes several subsections:
  - Personal Data Import – Import of researcher demographic data via Excel file. Each row includes name, surname, department, Scopus ID, ORCID and ResearcherID
  - Register – Retrieval of publications via Scopus API. For each researcher, h-index, number of publications and total number of citations over three time frames (entire career, last 10 and 5 years) are computed
  - FLC from Scopus – Retrieval of First, Last, and Corresponding authors (FLC)
  - PubMed – Retrieval of publications available on PubMed
- Demographic Data. List of researchers. Can be sorted, filtered by year and exported to Excel
- Researcher Details. Metrics and PubMed publications not yet listed in Scopus. Also enables updates to demographic data
- Register. Table of researcher bibliometric metrics sorted by h-index
- Statistics. Individual and aggregated metrics (median h-index by Research Unit, including all members or under 40)
- Requests. Interface for approval of external requests to join the researcher list
- Scopus. Publications from Scopus, grouped by year and researcher, with a secondary table for missing PubMed entries to highlight gaps
- PubMed. Publications from PubMed

3.2 Researcher web application

This allows researchers to update their own demographic information, and access to the list of publications.

- Researcher Dashboard (login required). Publications (reference, Scopus Author ID, DOI, PubMed ID, citation count, and authorship role) and metrics (h-index, number of publications, citation count, and the number of publications where the author is listed as FLC). Allows updating of demographic data
- Public Request Form (no login). Researchers not in the database may request inclusion, subject to Administrator review

4. Use case

The software has been under testing for approximately one year at the Giannina Gaslini Institute. In early 2025, it was updated with the list of researchers from the Workflow della Ricerca, a web-based platform of the MoH, which includes 460 individuals working at our Institute with a valid Scopus ID. Metrics were also aggregated for each of the 68 RU. The most recent update carried out in June 2025 - four months after the previous one - required less than 5 minutes and showed changes in the number of publications or related metrics for 426 out of 460 researchers (93%).

5. Conclusions

Some limitations emerged that suggest areas for improving LibRA's performance. In particular, data collection was hindered by insufficient metadata in the source database, resulting in occasional inaccuracies in author attribution.

For example, it is not currently possible to identify shared author positions in articles because this information is not tracked by Scopus. We plan to identify possible solutions to these biases in future releases.

At the same time, outreach initiatives for researchers have been launched: our Code of Research Integrity promotes regular updating of unique identifiers to prevent information loss during data extraction due to duplicate author profiles.

Finally, we plan to add novel functionalities based on complex performance analyses from SciVal.

In our Institute LibRA has proven to significantly reduce the time required for data collection and to minimize the risk of human error. We believe that it could also be useful to other IRCCSs, because to our knowledge no tool calculates all the bibliometric indicators provided by LibRA as required by MoH (e.g. Moschini and Molinari, 2022).

6. Data Availability

The source code along with documentation and Excel template for demographic data import is available at https://github.com/igg-bioinfo/igg-biblio (Administrators) and https://github.com/igg-bioinfo/igg-researcher (Researchers).

References

Moschini U., Molinari E. (2022). Designing ecosystems to enable recognition and adoption of Open Science measures [Internet]. Genoa Open Access Week. Available from: https://openscience.unige.it/sites/openscience.unige.it/files/2022-12/07.1_GenOAweek2022_

Moschini.pdf
Research Intelligence (2019). Research Metrics Guidebook [Internet]. Elsevier. Available from: https://brand.elsevier.com/share/nSegnbJj4TG2qKKGxrYE

Authors

Chiara Rebuffi chiararebuffi@gaslini.org

Graduated in Archival Sciences, Documentation, and Library Science, she has been working for about 15 years in Libraries and Scientific documentation services for biomedical research institutions, currently works at the IRCCS Gaslini institute in Genoa. She is involved in the evaluation of research and researchers, both at national and international levels, and regularly uses tools, such as citation databases, to assess performance.

Roberto Cavanna robertocavanna@gaslini.org

Software developer with over twenty years of experience, including the last ten dedicated to scientific research at the IRCCS Giannina Gaslini Institute. Specialized in designing and maintaining web applications for clinical registries and research trials, he combines deep technical expertise with a strong understanding of research workflows and data integrity.

Paolo Uva paolouva@gaslini.org

Coordinator of the Bioinformatics Unit at the IRCCS Giannina Gaslini Institute in Genoa. The activities of his research group focuses on the development and application of methods for the analysis of clinical and omics data generated through high-throughput technologies. The group also works on the digitalization and automation of procedures to ensure compliance with FAIR principles, supporting the institute's transition toward a data-oriented model.

DOI: 10.26314/GARR-Conf25-proceedings-22

The 'encrypted cable': FPGA implementation of secure communication based on cryptographic algorithms

Antonio Mastrandrea[1], Paolo Palazzari[2], Pasquale Tommasino[1]

[1]Dept. of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome, [2]ENEA, ICT-HPC Division, Casaccia Research Center, Rome

Abstract. The "Encrypted Cable", a secure and high-speed communication system for public communication networks, is proposed as a solution for encrypted data exchange between nodes in a smart grid. Encryption and decryption are performed on an FPGA using the QP-Dyn cryptographic algorithm. The proposed system achieves an intrinsic throughput of 750 MB/s and has been successfully tested in encrypted transmissions between ENEA sites in Casaccia and Portici.

Keywords. Smart grid, Cyber security, Cryptography, FPGA

Introduction

In modern smart grids, both the energy distribution network and its associated data communication network are extensive and highly branched. Ensuring secure data transmission and management is essential—from individual devices, such as smart meters installed in apartments, to entire buildings and neighborhoods [Fouda 2011]. These smart meters, capable of bidirectional data exchange, may be located in private residences or connected to charging stations for electric vehicles [Cheng 2024]. Encrypting transmitted data is crucial to prevent unauthorized access and safeguard user privacy.

Various types of encryption algorithms exist, typically categorized into three main classes: symmetric, asymmetric, and those based on hash functions [Alenezi 2020]. These algorithms serve to protect data from theft, tampering, and unauthorized processing by entities responsible for managing smart meter data. Encryption requirements vary across the hierarchy of the network architecture. While individual smart meters—often constrained in terms of computational resources—operate within private networks, higher-level systems must handle significantly larger volumes of data. Although these upper levels generally have greater processing power, they often depend on public networks, which are inherently less secure.

1. Encryption algorithm choice

This research focuses on the development and hardware implementation of cryptographic algorithms optimized for deployment on programmable logic devices, particularly targeting the upper levels of the smart grid network hierarchy. A preliminary analysis was con-

Conferenza GARR 2025 - Frontiere Digitali Selected Papers
Bari, 13-15 maggio 2025 - ISBN: 978-88-946629-4-8 - Creative Commons CC-BY

ducted to identify algorithms offering the best performance for this context.
In [Abood 2017], symmetric and asymmetric encryption algorithms suitable for secure data transmission in smart grids were compared in terms of encryption/decryption times and the estimated time required to break the encryption for short plaintexts. Among them, the asymmetric RSA algorithm with a 1024-bit key [Rivest 1978] was found to be the slowest, while the Advanced Encryption Standard (AES) [Daemen 2002] demonstrated both the highest security and fastest execution. Additionally, in [Alenezi 2020], common symmetric algorithms were evaluated based on throughput, encryption time, and CPU usage for varying text sizes. AES, RC4, and RC6 yielded the best performance overall; however, only AES showed high resistance to known cryptographic attacks, as noted in [George 2023].
Based on performance and security benchmarks, AES emerges as the most suitable candidate for data encryption in smart grids. However, secure communication in this context often requires additional cryptographic operations, given the hierarchical architecture comprising at least three levels: from smart meters (lowest level), to one or more data aggregation centers, up to the highest level (e.g., the power operator, PO), which is also responsible for key management and distribution. Asymmetric encryption is often employed at this top level to ensure secure key distribution. For instance, the protocol described in [Uludag 2015] uses the Diffie-Hellman algorithm for key exchange, AES-256 for data encryption, and SHA-256 for digital signatures. Consequently, while AES is optimal for primary encryption tasks, it must often be supported by auxiliary cryptographic mechanisms to ensure complete security.
In addition to AES, the QP-Dyn algorithm [Abundo 1992][Accardi 2011] has also been considered. QP-Dyn is a symmetric encryption algorithm based on the chaotic behavior of a class of deterministic dynamical systems known as Anosov systems. These systems produce very long periodic orbits that pass standard randomness tests, despite not being fully chaotic, as their orbits cannot originate from irrational initial points.
A comparative study [Italiano 2009] evaluating encryption algorithms for mobile applications showed that QP-Dyn generates longer secret keys more quickly than conventional algorithms. Specifically, when comparing its stream cipher version to AES in Cipher Feedback mode (AES-CFB), QP-Dyn—with a 279-bit key—outperformed AES-CFB with a 256-bit key for input sizes greater than 256 bytes. Although AES in block cipher mode (its standard configuration) remains faster overall, the performance gap is minimal (<60 ms) for blocks smaller than 512 bytes.

2. Hardware implementation and transmission tests

Secure communication between two users—referred to as the "encrypted cable"—was implemented using the VITIS environment [Amd 2023], enabling encrypted data exchange via the AXI-Stream interface. The data (e.g., from a smart meter) are streamed from memory to a QP-Dyn encoder. The encrypted data is then sent back to memory, with overall latency primarily determined by memory access times.
The QP-Dyn algorithm is implemented using dynamical systems modeled as a matrix M

of size d × d. Starting from an initial state $S_0 = [S_{0,1} \dots S_{0,d}]$, the system generates an orbit S_0, S_1, ... through the recurrence relation

$$S_{i+1,j} = (\Sigma^{d}_{k=1} M_{j,k}\, S_{i,k}) \bmod p \qquad j=1,2,\dots,d \qquad (1)$$

The parameter p is typically a large number. In this implementation, two independent dynamical systems M_A and M_B with d = 4 were used. Secret keys K_A and K_B were derived at each iteration i from the states $S_{i,A}$ and $S_{i,B}$, respectively, using a Key Generating Function (KGF), and combined via an XOR operation. KGF constructs the encryption key by concatenating the words derived from $S_{i,j}$ (j=1,2,...,d) removing all leading zeros up to and including the first 1.

The modulo p operations are efficiently handled using Barrett's algorithm [Barrett 1986] which avoids division operations.

Figure 1 shows a schematic of the QP-Dyn architecture. The "Dynamic System" block implements equation (1); on even cycles, it computes the evolution of system M_A, while M_B is updated on odd cycles.

The XOR operation on KGF outputs generates the final key

$$K_i = K_{i,A} \text{ XOR } K_{i,B}$$

which is used to encode the input word via a bitwise XOR.

The initial states $S_{0,A}$ and $S_{0,B}$, which are parameters of the encryption algorithm, are loaded during the first two clock cycles via the multiplexer input selected by the condition i≤1.

Once synthesized on an AMD ALVEO U280 board, the "encrypted cable" was deployed between two nodes of the ENEA network: an encryption node at the Casaccia site and a decryption node at the Portici site. The design was synthesized with a target clock frequency of 100 MHz.

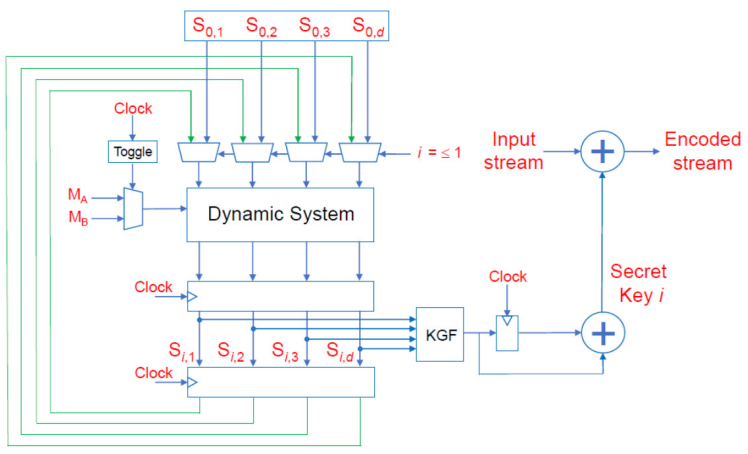


Fig. 1
QP-Dyn algorithm implemented with two dynamic systems MA and MB

The synthesis results are summarized in Figure 2, demonstrating the low hardware resource utilization of the QP-Dyn implementation.
The system was validated through a series of tests: first on a single node hosting both transmitter and receiver on two U280 boards, and then in a real-world deployment between the ENEA Casaccia and Portici sites. The first test evaluated the intrinsic throughput of the "encrypted cable", considering only encryption, decryption, and memory transfers, and yielded a performance of 750 MB/s.

Name	LUT	LUTAsMem	REG	BRAM	URAM	DSP
Platform	194876 [14.95%]	23758 [3.95%]	278667 [10.69%]	330 [16.37%]	0 [0.00%]	10 [0.11%]
User Budget	1108804 [100.00%]	578092 [100.00%]	2328693 [100.00%]	1686 [100.00%]	960 [100.00%]	9014 [100.00%]
Used Resources	8840 [0.80%]	644 [0.11%]	6351 [0.27%]	4 [0.24%]	0 [0.00%]	124 [1.60%]
Unused Resources	1099964 [99.20%]	577448 [99.89%]	2322342 [99.73%]	1682 [99.76%]	960 [100.00%]	8890 [98.40%]
Memory2Stream	1308 [0.12%]	277 [0.05%]	1510 [0.05%]	2 [0.12%]	0 [0.00%]	0 [0.00%]
Memory2Stream_1	1308 [0.12%]	277 [0.05%]	1510 [0.05%]	2 [0.12%]	0 [0.00%]	0 [0.00%]
Stream2Memory	1194 [0.11%]	367 [0.06%]	1913 [0.06%]	2 [0.12%]	0 [0.00%]	0 [0.00%]
Stream2Memory_1	1194 [0.11%]	367 [0.06%]	1913 [0.06%]	2 [0.12%]	0 [0.00%]	0 [0.00%]
krnl_qp_dyn	6338 [0.57%]	0 [0.00%]	2928 [0.13%]	0 [0.00%]	0 [0.00%]	124 [1.38%]
krnl_qp_dyn_1	6338 [0.57%]	0 [0.00%]	2928 [0.13%]	0 [0.00%]	0 [0.00%]	124 [1.38%]

Fig. 2 Synthesis results on the ALVEO U280 board

The inter-site test, on the other hand, reported a throughput of 60 MB/s, limited by the available bandwidth of the communication channel between the two locations.

3. Conclusions

We presented an FPGA implementation of the QP-Dyn encryption algorithm, which generates pseudo-random numbers using a dynamical system. These numbers are used as ciphering keys for point-to-point encrypted communication within the context of energy smart grids.

References

Abood, O. G. et al. (2017). Investigation of cryptography algorithms used for security and privacy protection in smart grid. In 2017 Nineteenth International Middle East Power Systems Conference (MEPCON) (pp. 644-649). IEEE.
Abundo, M. et al. (1992). Hyperbolic automorphisms of tori and pseudo-random sequences. Calcolo, 29, 213-240.
Accardi, L. et al. (2011). The Qp-Dyn Algorithms (Vol. 8, pp. 1-15). Singapore: World Scientific Publishing Co. Pte. Ltd.
Alenezi, M. N. et al. (2020). Symmetric encryption algorithms: Review and evaluation study. International Journal of Communication Networks and Information Security, 12(2), 256-272.
Amd: UG 1399 - Vitis High-Level Synthesis User Guide. 2023.
Cheng, R. et al. (2024). LLRA: A Lightweight Leakage-Resilient Authentication Key Exchange Scheme for Smart Meters. IEEE Transactions on Smart Grid.
Daemen, J., & Rijmen, V. (2002). The design of Rijndael (Vol. 2). New York: Springer-verlag.
Fouda, M. M et al. (2011). A lightweight message authentication scheme for smart grid communications. IEEE Transactions on Smart grid, 2(4), 675-685.
George, D. J., & Thomas, T. (2023). A Comparative Study of Symmetric Key Algorithms. International Journal of Computer Science and Mobile Computing, Vol.12 Issue.6, June-

2023, pg. 71-75.
Guan, D. J. (2003). Montgomery algorithm for modular multiplication. Departement of Computer Science, National Sun Yar-Sen University, Taiwan.
Italiano, G. F. et al. (2009, August). Benchmarking for the QP cryptographic suite.
MontgomeryBarrett, P. L. (19865). Modular multiplication without trial division. Mathematics of computation, 44(170), 519-521. Implementing the Rivest Shamir and Adleman Public Key Encryption Algorithm on a Standard Digital Signal Processor. Advances in Cryptology – CRYPTO' 86. Lecture Notes in Computer Science. Vol. 263. pp. 311–323.
Rivest, R. L. et al. (1978). A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2), 120-126.
Uludag, S. et al. (2015). Secure and scalable data collection with time minimization in the smart grid. IEEE Transactions on Smart Grid, 7(1), 43-54.

Authors

Antonio Mastrandrea antonio.mastrandrea@uniroma1.it

Antonio Mastrandrea received the master's(Laurea) degree (cum laude) in electronics engineering and the Ph.D. degree, from the Sapienza University of Rome, Rome, Italy, in 2010 and 2014,respectively. He is a Research Assistant with the Department of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome. His current research interests include digital system-on-chip architectures and nano-CMOS circuits oriented to high-speed computation.

Paolo Palazzari paolo.palazzari@enea.it

Paolo Palazzari (M.Eng. 1989, Ph.D. 1994) has been a researcher at ENEA since 1996. He founded ENEA's first spin-off, Ylichron srl, developing the HCE High-Level Synthesis tool. From 2010 to 2018 he was detached to PLDA Italia as CTO, contributing to the development of the QuickPlay High-Level Synthesis flow. Since 2018 he has returned to ENEA as a senior researcher, focusing on FPGA algorithm development using High-Level Synthesis tools.

Pasquale Tommasino pasquale.tommasino@uniroma1.it

Pasquale Tommasino is a researcher at the Department of Information, Electronics and Telecommunications Engineering of the University "La Sapienza" of Rome. His research activity focuses mainly on the design of integrated circuits for communication applications in the microwave and millimetre wave field, in its different aspects of methodology, modelling and development of circuit topologies, and also on issues concerning spectral analysis and processing of radar and communication signals.

DOI: 10.26314/GARR-Conf25-proceedings-23

CLIC (Cloud In Cresco): towards HPC/HPDA-as-a-Service

Marco Faltelli, Alessandro Peloso, Francesco Iannone, Matteo Fois, Massimo Celino and Giovanni Ponti

ENEA, Lungotevere Thaon di Revel, 76 – 00196 Rome, Italy

Abstract. . The demand for High-Performance Computing (HPC) services is rapidly growing, driven by the advent of new accelerators and increasingly complex tasks, with artificial intelligence (AI) being a major focus. Despite this, HPC cluster users continue to rely on bare-metal resources for their computations. In this position paper, we argue that the time has come for HPC to transition to virtualized services. Virtualization has demonstrated its scalability, reliability, and efficiency while also abstracting the underlying hardware's intricate details. Additionally, virtualized environments offer significant flexibility, allowing for easy adjustment to varying workloads and security through isolation, ensuring that different virtual machines can operate securely on the same physical hardware. We present a PoC architecture that shall be deployed in our virtualized HPC cluster based on OpenStack (as part of the IPCEI-CIS project), showing the future challenges and issues that may emerge

Keywords. HPC, cloud, GPU, IAAS

1. Introduction

The IPCEI-CIS (Next Generation Cloud Infrastructure and Services) project aims at creating a multi-provider, European-sovereign cloud-edge continuum between different partners, such as telco operators, industry operators, and R&D centers. ENEA is part of IPCEI-CIS through the DataCLEEN project: DataCLEEN's goal is to create a highly scalable and reliable cloud infrastructure based on HPC technologies.

ENEA is already a major HPC provider in Italy through the CRESCO infrastructure, with CRESCO8 being launched in April 2025; DataCLEEN will enable the creation of the first cloud HPC cluster called CLIC (CLoud In Cresco), which is expected for the end of 2025.

We believe that creating a cloud HPC environment brings significant benefits to the community and poses significant challenges in engineering and maintaining such an infrastructure.

In this paper, we first describe the motivations behind DataCLEEN's virtualized HPC infrastructure. Then, we describe our proof-of-concept architecture for the future CLIC supercomputer, focusing on the challenges and some preliminary results, which show very close performances compared to bare-metal supercomputers.

2. Why cloud HPC

Adopting VMs in HPC clusters brings a series of benefits to both users and maintainers, which are described here.

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

2.1 The users' side

Unified access to resources: HPC suppliers usually provide users with isolated clusters with different login portals, requiring them to choose a cluster a-priori based on their needs. Virtualization offers a unified login portal, with hypervisor controllers selecting the best cluster for each workload.

Integrated data handling: A unique infrastructure for all the clusters brings the significant benefit of having a unique data storage platform that is geographically distributed and redundant among the clusters. In this way, new clusters and storage can be added modularly.

Specialized environments: HPC is traditionally related to topics like science materials, fusion power, bioscience, and weather forecasts. Lately, AI/ML has gained a lot of interest because of the increasing hardware requirements of LLMs. Each of these different fields requires specific software and packages, and it is hard to provide a catalog that can satisfy all possible applications. Through virtualization, users can create and replicate VMs equipped with specific software for their needs. For example, AI users can create VM images equipped with CUDA, PyTorch, Numba, and LLM-specific libraries.

Failure resilience: HPC jobs often involve many nodes and long durations, increasing the chances of hardware failure. Virtualization minimizes the consequences of failures through periodic snapshots and live migration of instances.

2.2 The administrators' side

Enhanced security: virtualization implies isolation both at a software level and at the network level, as well as advanced monitoring and logging features.

Easier maintenance: through multiple availability zones and migration capabilities, administrators can operate in isolated parts of the cluster without impacting the users.

Better resiliency: in case of any fault (hardware, network, software), VMs can be migrated to a different availability zone, thus giving the cluster a high degree of resiliency to failures.

3. A primer on CLIC

3.1 Infrastructure

Our PoC infrastructure is based on two ENEA HPC clusters: 20 nodes of the dismissed CRESCO4 cluster and 11 nodes from CRESCO5F. The CRESCO5F nodes feature 64 AMD EPYC 7313 CPU cores, 256GB RAM, and Infiniband Connect-X6 100Gbps NICs. Seven of these nodes are equipped with NVIDIA GPUs, including A100 and H100 models.

OpenStack is the main component of our infrastructure and it orchestrates our cluster's different computing, storage, and network resources through an Infrastructure-As-A-Service (IAAS) paradigm. OpenStack is composed of different services that can be added to our infrastructure modularly; each service is responsible for a certain feature in OpenStack. OpenStack relies on Canonical frameworks like juju and MAAS: the former permits us to manage OpenStack and its dependencies, while the latter discovers and provisions the available resources.

3.2 Challenges

3.2.1 Performance

The first requirement for CLIC is that performances should be equal (or very close) to bare-metal clusters. In other words, virtualization should not add a significant overhead. We underline that this is not a trivial question to be answered, as in HPC it is of paramount importance to effectively virtualize resources such as GPUs and Infiniband devices. As we show in Section 3.3, preliminary results highlight very similar performances between virtualized and bare-metal HPC hardware.

3.2.2 Licensing

Another challenge (at the moment of writing this paper) is GPU virtualization, which can be exclusively done on NVIDIA GPUs. Virtual GPUs need a license from NVIDIA, which must be bought separately. Otherwise, their performance is degraded twenty minutes after the VM startup, and after 24 hours, CUDA stops working. This is a substantial limitation in adopting vGPUs in HPC scenarios, especially since most of these clusters are used for non-profit, scientific research reasons. At the time of writing, AMD is planning to introduce virtual GPUs in the next year, which could be a viable alternative for GPU virtualization.

3.3 First results

Here, we showcase two tests focusing on fundamental HPC performances. The first one highlights the performances of a licensed and virtualized NVIDIA A100 GPU when running common HPC benchmarks, such as HPL and HPCG. We can see from Table 1 that the performance of the virtualized hardware is very close to that of the bare-metal one. Table 2 focuses on the communication overheads that the virtualization may cause. Surprisingly, the performances of the virtualized infrastructure perfectly match the bare-metal ones, with zero overhead. We believe these two tests make a first case for virtualized HPC clusters. We plan to do further performance tests in the near future.

	vGPU	Bare-metal
HPL 21.4 (TFLOPS)	9.73	9.85
HPCG 24.09 (GFLOPS)	228	230

Tab. 1: HPC tests on NVIDIA A100

	VM	Bare-metal
OFED perf (read)	96 Gb/sec	96 Gb/sec
OFED perf (write)	96 Gb/sec	96 Gb/sec
MPI host-to-host	88 Gb/sec	88 Gb/sec

Tab. 2: HPC communication tests with ConnectX-6 100Gbps Infiniband NICs

4. Future Plans

Our future plans include using the architecture described here for CLIC, the next ENEA HPC cluster, which will be distributed among four different ENEA research centers: the main cluster is going to be hosted in Frascati, while the other three will be hosted in

Casaccia, Portici, and Brindisi research centers. We foresee that such architecture will create innovative solutions for HPC and will foster collaboration with other partners of the IPCEI-CIS project.

Acknowledgement

This work was funded by the Italian Ministry MIMIT in the frame of the European initiative IPCEI CIS, under the PNRR/NextGenerationEU, as part of the project DataCLEEN (CUP: I38H23000710006).

Authors

Marco Faltelli marco.faltelli@enea.it
Marco Faltelli is a Researcher in the ENEA High Performance Computing Lab. In June 2023 he obtained his PhD in Computer Science under the supervision of Prof. Giuseppe Bianchi and Prof. Francesco Quaglia. He is a Microsoft Research PhD Fellowship recipient. His research activities combine computer architecture, HPC, and computer networks to design high-performing, scalable architectures for both network hardware and software solutions.

Alessandro Peloso alessandro.peloso@enea.it
Alessandro Peloso is a software architect in the ENEA TERIN department. His interests combine operative systems, cloud computing

Francesco Iannone francesco.iannone@enea.it
Francesco Iannone holds a degree in Physics. Currently he is researcher directory at ENEA specialising in High-Performance Computing (HPC). He works on the development and optimisation of scientific applications. His expertise includes parallel computing, numerical modelings, and support for HPC infrastructures within national and European projects.

Matteo Fois matteo.fois@enea.it
Matteo Fois holds a master degree in Applied Mathematics and a master course in HPC from the University of Rome Sapienza. He works as a researcher in ENEA's HPC laboratory of Portici, managing the compute infrastructure and collaborating with different research groups to help the development of parallel codes and simulations on ENEA's clusters.

Massimo Celino massimo.celino@enea.it
He studied Physics at the University of Rome Sapienza and holds a PhD from the University "L. Pasteur" of Strasbourg (France). He promotes and manag-es national and European projects in the field of HPC and cloud. He is cur-rently the PI of the ENEA DataCLEEN project as part of the IPCEI-CIS cloud initiative. He is the author of more than 100 scientific papers in international journals.

Giovanni Ponti giovanni.ponti@enea.it
He is Director of the ENEA ICT Division and member of the GARR board of di-rectors. PhD in Computer Engineering in 2010 (University of Calabria), dur-ing which he focused on data modeling and innovative data mining algo-rithms. Senior ENEA researcher for 15 years with research activities in HPC, Data Science, Cloud Computing, Big Data, Data Analytics and AI. In these con-texts, he is the author of prestigious scientific publications (more than 100 with referee), member of the conference program committee and reviewer for important journals and conferences.