Integrating Services, Platforms and Resources into a National Infrastructure Cluster for FAIR Language and Cultural Data

Giulia Pedonese, Daniele Melaccio, Michele Mallia, Monica Monachini, Francesca Frontini, Valeria Quochi, Anas Fahad Khan, Angelo Mario Del Grosso, Federico Boschetti, Riccardo Del Gratta

Consiglio Nazionale delle Ricerche,
Istituto di Linguistica Computazionale "A. Zampolli"
Via Moruzzi, 1, Pisa, Italy
name.surname@cnr.it

Abstract

In the context of evolving European and national policies for research infrastructure governance, this paper presents the contribution of CLARIN-IT, the Italian consortium of CLARIN (Common Language Resource and Technologies Infrastructure), to the construction of a national infrastructure supporting interoperability of language and cultural data through standard data models and shared vocabularies, sustainable trough integration with the Open Science Cloud framework. This was possible within the Humanities and cultural Heritage Italian Open Science Cloud (H2IOSC) project, funded by the European Union NextGenerationEU National Recovery and Resilience Plan (NRRP). This project involved the national nodes of four Research Infrastructures in the Social Sciences and Humanities (SSH): E-RIHS-IT, DARIAH-IT, OPERAS.it and CLARIN-IT and aimed at creating a federated cluster to provide access to advanced tools for intensive research on data.. In this framework, CLARIN-IT contributes to translating FAIR and Open Science principles into practice by integrating technical, methodological, and training dimensions. Its activities combine several coordinated components: workflows to align data to the FAIR principles; ontology-based metadata mediation to enhance semantic interoperability across infrastructures; the refactoring and exposure of services through a federated API gateway; and the implementation of a Linguistic Linked Open Data (LLOD) pilot for the validation, transformation, and publication of interoperable RDF datasets. In this framework a national training ecosystem — comprising a training platform and a repository for reusable training materials — was developed to support capacity building and the creation of FAIR-by-design learning materials. Finally, a permanent research observatory monitors community practices and needs, providing evidence-based insights for the continuous improvement of services and training provision. Together, these components demonstrate a coherent strategy for implementing FAIR principles and Open Science at the national level, while ensuring alignment with major European and national initiatives in the SSH data ecosystem.

Keywords: FAIR data; research infrastructures; CLARIN; Open Science; interoperability; metadata standards; training materials; linked data; language resources; policy coordination

1. Introduction and Policy Context

For at least the past two decades, the European Commission (EC) has been investing in research infrastructures (RIs) for the humanities, social and cultural heritage sciences, with policy priorities increasingly aligned with Open Science and FAIR principles[1] (European Commission: Directorate-General for Research and Innovation, 2025; Wilkinson et al., 2016), which has stimulated investments at the national level where we now witness the consolidation of several infrastructural initiatives linked to their European counterparts. More recently, the EC encourages RIs aggregation into clusters to further optimize resources, avoid duplication of efforts and investments, and foster interdisciplinary collaboration, as well as the reuse and sharing of data, tools, methodologies, protocols, and workflows (see i.e., the European Open Science Cloud, Almeida et al. (2017)). In parallel, similar initiatives are emerging at the national level, where nodes of major European RIs are beginning to operationalize the clustering promoted at the supranational level, e.g. HumaNum in France[2], Text+ in Ger-

[1] FAIR principles — Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016) — provide a framework for improving the management and reuse of research data. In the context of language and cultural resources, FAIR implementation involves several practical measures: the assignment of persistent identifiers and rich metadata to ensure discoverability; the adoption of standardized access protocols and authentication infrastructures; the use of shared semantic models and controlled vocabularies to enable interoperability; and the specification of clear licensing conditions to facilitate reuse. Research infrastructures play a crucial role in operationalizing these principles by providing tools, services, and governance frameworks that support FAIR data management across the entire research lifecycle.

[2] https://www.huma-num.fr/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1170–1178
11–16 May 2026.

many, SSHOC-CH in Switzerland[3], which bring together the national nodes of CLARIN, DARIAH and Operas; and CLARIAH consortia in various countries[4].

This contribution is situated within this broader context and presents the Italian efforts to establish a coordinated and functional aggregation of research infrastructures dedicated to the humanities, social sciences, cultural heritage, and open science. Specifically, it focuses on the role of CLARIN-IT[5] in implementing the Humanities and cultural Heritage Italian Open Science Cloud (H2IOSC)[6], a national federated model for a cross-disciplinary cluster of RIs designed to support, among other aspects, compliance with the FAIR principles, interoperabilty, data stewardship, resource curation, and training activities for the national target scholarly community (Degl'Innocenti et al., 2023). H2IOSC brings together services, platforms, and resources to help researchers manage, share, and reuse language and cultural data in interoperable ways. CLARIN-IT participates to the H2IOSC with other three national nodes of RIs in the SSH field: E-RIHS-IT. DARIAH-IT and OPERAS.it. Its main activities include service integration through a federated infrastructure, Linked Open Data publication, training, and community monitoring. More broadly, it acts as a national hub aligned with major European RIs and EOSC developments. CLARIN-IT focus in this context is on linguistic data management, standardization, and service integration, ensuring alignment with related initiatives under the European Open Science Cloud (EOSC).

The work towards the enrichment and integration of CLARIN tools and services for linguistic research into the national cluster is structured around four key pillars:

- alignment of data to the FAIR principles and the enhancement of interoperability and integrability of modular software services;
- implementation of pilot platforms that provide the community with fundamental services and functionalities;
- development of training environments and materials, with an emphasis on a FAIR-by-design approach;
- building a permanent observatory mapping the current landscape of use and the needs of the served scientific communities.

This paper reflects and expands on these strategic axes. Section 2 presents the integration of CLARIN's data and services into the technical federation and the "FAIRification"[7] Toolchain for assessing and enhancing the FAIRness[8] of linguistic and cultural resources. Section 3 introduces the Linguistic Linked Open Data pilot, which provides a dedicated platform for the modelling and publication of linguistic data according to the Linked Open Data paradigm. Section 4 describes the training strategy and tools developed for promoting and supporting an end-to-end FAIR lifecycle for learning objects. Finally, section 5 presents the permanent Research Observatory for monitoring communities and identifying user needs.

2. FAIRification, Interoperability and Service Publication

FAIRification workflows have been developed and implemented to enhance the interoperability and reusability of linguistic and cultural resources. Key actions have included:

- harmonization of metadata through ontology-based mediation and alignment with CIDOC CRM and SSHOCRo (in Section 2.1);
- exposure and refactoring of Services via API Integration Technologies (e.g., WSO2) and OpenAPI. Within the H2IOSC project, CLARIN-IT publishes its services through an API Gateway, adopting OpenAPI manifests for machine-readable documentation, versioning and policy enforcement (in Section 2.3)
- progressive alignment of CLARIN services to the H2IOSC catalogue and to its federated AAI for smooth accessibility (in Section 2.3);

These workflows ensure compliance with the FAIR principles while maintaining compatibility with European infrastructures such as the SSH Open Marketplace.

[3] https://sshoc.ch/

[4] e.g., https://clariah.nl/ in the Netherlands and https://www.clariah.es/ in Spain

[5] CLARIN-IT is the Italian node of CLARIN ERIC, a digital research infrastructure distributed across Europe (including centres in the USA and South Africa), offering data and tools to enhance research based on language resources. CLARIN is a European Research Infrastructure Consortium (ERIC) and its services are provided by national consortia in the countries that have joined the Infrastructure. Here is a map of CLARIN centres: https://centres.clarin.eu/map

[6] https://www.h2iosc.cnr.it/

[7] Which is the process of making data FAIR-compliant, see for example https://www.go-fair.org/fair-principles/fairification-process/

[8] Which is the adherence of a dataset description to FAIR principles, see for example: https://vocabs.sshopencloud.eu/browse/sshocterm/en/page/data_fairness_48

2.1. The Ontology Based mediation of Metadata

In parallel with the FAIRification workflows, an ontology-based mediation framework has been developed for enhancing metadata interoperability across H2IOSC ecosystem (Melaccio et al., 2025).

Building upon the CLARIN Component Metadata Infrastructure (CMDI) and the CLARIN Concept Registry (CCR, Uytvanck et al. (2012); CLARIN ERIC (2023)), this approach extends metadata representation beyond syntactic conversion, introducing a formal semantic layer that ensures long-term stability and cross-domain integration. CMDI provides a flexible XML-based framework for the structured description of language resources, while the CCR anchors metadata elements in persistent and controlled vocabularies. In the ontology-based model, CMDI metadata are semantically mapped to the CIDOC Conceptual Reference Model (CIDOC CRM) and to the SSHOCro (Doerr, 2003; SSHOC Consortium, 2022), the reference ontology developed within the Social Sciences and Humanities Open Cloud. This mapping allows CLARIN metadata to be interoperable with those used in other research infrastructures fostering a unified semantic framework for the Humanities and Social Sciences. The mapping strategy combines top-down and bottom-up approaches. High-level CMDI descriptors such as Resource Type, Author, and Funding are aligned with CIDOC CRM entities (e.g., `E73 Information Object`, `E21 Person`, `E65 Creation`), while domain-specific extensions derived from SSHOCro refine the representation of linguistic datasets and tools. For example, CMDI profiles for corpora or lexica are modelled under `SHE1 Dataset`, with subclasses like `ParallelCorpus`, `ReferenceCorpus`, or `LexicalResource`. Similarly, services exposed through the CLARIN Language Resource Switchboard (Zinn, 2016) are represented as `LinguisticTool` subclasses (e.g., `CorpusQueryTool`, `PartOfSpeechTagger`, `NamedEntityRecognitionTool`), aligning practical tool taxonomies with formal semantic definitions. This ontology-first strategy bridges community practices and semantic technologies. It preserves CMDI's operational structure while enriching it with RDF/OWL-based representations, enabling integration into H2IOSC Common Semantic Framework and Linked Open Data environments. A representative case study —the *Musisque Deoque* corpus[9]— demonstrated the potential of this approach, showing how CMDI metadata can be linked to external authority files (e.g., ORCID, VIAF) and cultural heritage datasets while remaining interoperable within CLARIN.(VV., 2005) By embedding metadata within a shared ontology, CLARIN-IT not only enhances the FAIRness and discoverability of resources but also contributes to the broader semantic coherence of the H2IOSC federation.

Embedding metadata within a shared ontology, not only enhances the FAIRness and discoverability of resources but also contributes to the broader semantic coherence of the H2IOSC federation. Ontology-based mediation thus becomes a key enabler for semantic interoperability, reinforcing the alignment of linguistic data infrastructures with the European Open Science Cloud vision.

2.2. The FAIRification Tool

A key outcome of the H2IOSC initiative is the development of an automated FAIRification Tool designed to assess and enhance the FAIRness of linguistic and cultural resources.[10] The tool operationalizes the FAIR principles (Findable, Accessible, Interoperable, Reusable) through semi-automatic evaluation workflows and structured feedback mechanisms. Built upon the F-UJI automated FAIR data assessment service (Devaraju et al., 2020), it extends the FAIRsFAIR metrics (FAIRsFAIR Project Consortium, 2021) with domain-specific checks relevant to the humanities, language resources, and cultural heritage communities.

By offering interactive reports and detailed guidance on metadata completeness, licensing clarity, identifier persistence, and semantic interoperability, the tool helps data providers, researchers, and curators understand how to improve their resources in concrete and measurable ways. Through this process, it promotes the adoption of best practices in FAIR data management and fosters a shared culture of openness, transparency, and reproducibility. Beyond its technical functionality, the FAIRification Tool also serves as an awareness-raising and capacity-building instrument within the broader Open Science ecosystem.

From an architectural perspective, the FAIRification Tool integrates with the H2IOSC authentication service and interoperates with CLARIN repository systems such as ILC4CLARIN. It interacts with the metadata infrastructure, allowing the validation and improvement of datasets prior to publication, and supports multiple user roles — including data owners, curators, and administrators — within a controlled workflow for FAIR assessment and revision. Assessment workflows are containerized for scalability, while results are stored in a dedicated FAIRness index accessible via REST APIs and visual dashboards.

In practical terms, the tool supports a semi-automatic workflow in which datasets are first anal-

[9] www.mqdq.it

[10] FAIRification tool URL anonymized for peer review.

ysed through machine-actionable FAIR indicators and are then reviewed by curators, who can validate results, refine metadata, and trigger new assessment cycles. This makes the tool useful not only as an evaluation instrument, but also as a service for guiding data providers toward FAIR-by-design publication practices. Within H2IOSC, its role is therefore twofold: on the one hand, it provides an operational environment for assessing and improving the FAIR compliance of linguistic and cultural resources; on the other hand, it contributes to building a shared methodological framework for repository governance, metadata quality control, and cross-infrastructure interoperability.

The tool thereby bridges technical standardization with policy goals, supporting sustainable data stewardship, community engagement, and alignment with European initiatives under the EOSC and CLARIN ERIC frameworks. By embedding FAIR assessment within the broader H2IOSC service ecosystem, the FAIRification Tool also acts as a mediation layer between technical infrastructures and community-oriented data stewardship practices.

2.3. Exposure and Refactoring of CLARIN Services

Within H2IOSC, CLARIN-IT is progressively refactoring and exposing its service portfolio through a federated API gateway based on the WSO2 middleware platform[11] and machine-readable OpenAPI manifests[12]. This transition marks a shift from a set of heterogeneous and project-specific endpoints to a coherent, standards-based service layer that enables consistent access, authentication, and monitoring across the federation. Through this harmonisation process, services become more easily discoverable, reusable, and interoperable. Moreover, fostering their integration into shared federated workflows for linguistic and cultural data management is currently under development to further exploit this harmonisation. The exposed services cover a broad spectrum of operations supporting the processing, annotation, and publication of language resources. They include natural language processing and text enrichment pipelines, tools for Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR), and services for the creation and management of digital editions. Other services focus on semantic data publication and vocabulary management, enabling Linked Data integration and the alignment of terminologies across repositories. Each service is documented through an OpenAPI manifest, providing clear definitions of inputs, outputs, and usage policies while ensuring transparency, persistence, and alignment with FAIR-by-design principles. From a governance perspective, the adoption of a federated API gateway enforces common access policies, usage monitoring, and authentication mechanisms, that support both human and machine interaction. This refactoring effort promotes sustainability and long-term interoperability within the H2IOSC federation, while aligning the CLARIN service architecture with European Open Science frameworks and the emerging Common European Language Data Space. Several of these services have also been made available to researchers and institutions through demonstration and trial phases within national and transnational access calls, to the purpose of strengthening the link between infrastructure development and community engagement.

[11] https://apim.docs.wso2.com/

[12] https://spec.openapis.org/oas/v3.1.0/

3. The Linguistic Linked Open Data pilot

Key pillar of CLARIN-IT contribution to H2IOSC towards enhancing sharing and interoperability, the Linguistic Linked Open Data (LLOD) Pilot provides a dedicated platform for the publication, validation, and long-term hosting of language and culture-related resources as Linked Open Data[13] thus supporting their integration within the broader LLOD ecosystem and thus facilitating their reuse. By combining technical, organizational, and training activities it aims at facilitating and supporting the publication of Open Linked Data in the humanities, as well as promoting the making of linguistic data FAIR-by-design through a suite of interconnected services. As a side-effect, it also demonstrates how a domain-specific infrastructure can operationalize the FAIR principles through concrete tools and workflows. As such, the LLOD Pilot extends the FAIRification Tool by offering a concrete environment in which datasets can be assessed, transformed or enriched, and published as interoperable and interconnected RDF graphs. The main objectives of the pilot are to: (i) support the validation and publication of LLOD, incentivizing the adoption of standardized Linked Data models such as SKOS and OntoLex-Lemon; (ii) promote semantic interoperability; and (iii) foster capacity building and awareness of the advantages

[13] Linked Data is a paradigm for publishing and connecting structured data on the Web using standard technologies such as URIs, RDF, and HTTP. It enables datasets to be described in a machine-readable way and linked to other datasets, allowing information from different sources to be integrated and queried across domains. Linked Open Data (LOD) extends this paradigm by requiring that the data be openly accessible and reusable under open licenses. By combining semantic web standards with open access principles, LOD promotes interoperability, transparency, and the reuse of data across communities and disciplines.

of LLOD among researchers and data providers.

The architecture of the LLOD Pilot integrates two fundamental components designed to support the full FAIR lifecycle of linguistic data. The first is a dedicated triplestore environment for storing and querying RDF resources, enabling long-term preservation and access via public SPARQL endpoints. The second is a vocabulary management platform based on SKOSMOS, which facilitates the publication and visualization of controlled vocabularies and other conceptual reference resources encoded in SKOS.(Mallia et al., 2025) A custom dashboard further supports resource validation, transformation, and publication. An integrated authentication mechanism, based on H2IOSC identity and access management system, ensures secure access for data providers and curators[14]. An illustrative example of this workflow concerns the transformation of tabular lexical datasets into LOD representations. This was the case of the several terminologies from the Osservatorio di terminologie e politiche linguistiche in Milan, which were FAIR-ified and published according to the Semantic Web principles within the framework of an early adopters call; just two examples out of many: MADIN-TERM (Zanola and Osservatorio di terminologie e politiche linguistiche, 2025) and the Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume (Zanola et al., 2023).

Training is also a key component of this pilot, and is aimed at promoting the adoption of Linked Data standards and supporting format conversion through hands-on examples and reproducible workflows. Specifically, Jupyter notebooks and online training materials guide users through the creation, transformation, and publication of linguistic LOD datasets. Indeed, many linguistic and cultural datasets originally are in tabular formats, which, while convenient for human and processing use, lack the semantic structure required for interoperability and reuse. Within the LLOD Pilot, a dedicated workflow has been developed which supports, via Jupyter notebooks, the transformation of these datasets into interoperable RDF representations, compliant with SKOS and OntoLex-Lemon models. This includes: (a) direct harvesting of metadata from the certified institutional or disciplinary repositories offering APIs (e.g., the ILC4CLARIN repository); (b) mapping of the structure and semantics of the original tabular data onto well-known Linked Data models, such as Ontolex, and shared vocabularies, such as Dublin Core, by means of human-readable templates based on YARRRML or RML; (c) serialization into turtle or rdf formats suitable for publication within the LLOD infrastructure.

Quality checks verify URI encoding, language tagging, and label harmonization before deployment. The resulting RDF resources can be published in the national triplestore and visualized via SKOSMOS interfaces, enabling access through both human-readable and machine-actionable endpoints.

Beyond its technical function, the workflow also serves as a training instrument, as each stage of the conversion process is documented and reproducible through Jupyter notebooks, allowing users to learn, replicate, and adapt the methodology for their own data. This way, the conversion methodology not only ensures compliance with FAIR and LOD standards, but also contributes to the development of digital skills and sustainable data practices across the user communities.

4. FAIR Learning Materials: Training Platform and Library

The training infrastructure developed in H2IOSC provides access to digital learning materials across language technology, digital humanities, and heritage science. In particular, the repository called Training Library (from now on the Library) ensures that these materials are curated, described, and published as FAIR Learning Objects. The Library adopts a shared metadata schema, supports multilingual documentation, interoperability with existing CLARIN repositories, and integration into broader educational ecosystems. These components enable researchers and educators to search, reuse, and adapt materials while ensuring long-term preservation and FAIR compliance.

Training is crucial to the functioning of research infrastructures, as it bridges the gap between scientific communities and what infrastructures have to offer. It is closely linked to dissemination activities that promote the use of products and services by researchers, students and citizens. For this reason, training initiatives have always played a prominent role through the organisation of events such as seminars, workshops and summer schools, the provision of scholarships and mobility grants, and the creation of digital archives through which users can access educational material, indexed according to shared standards as digital objects in accordance with FAIR principles.

In line with the European landscape, the training in H2IOSCis aimed primarily at Italian user communities, with the aim of equipping them with specific interdisciplinary skills and competences in the SSH sector and of training them on the resources that disciplinary infrastructures can offer at national and international level. Part of this objective is to train new professionals who can, in turn, train future generations on how to integrate infrastructural services into the research methods and practices with

[14] https://pllod.clarin-it.it/login

a train-the-trainers approach. This requires modular and easily reusable teaching materials that teachers can integrate into their training courses as needed.

The training infrastructure consists of two separate platforms: the Training Environment, a learning management system for delivering synchronous and asynchronous courses that will enable H2IOSC and the participating research infrastructures to offer direct training to both their staff and user communities; and the Training Library for storing, sharing, citing and reusing training materials as licensed FAIR digital objects with a version management system, in order to create a community of trainers who share and reuse their respective materials within academic and training programmes. This training infrastructure is a fundamental part of the H2IOSC Marketplace, i.e. the platform that will offer a unified access point to resources and services related to the humanities and social sciences at the national level, designed on the model of the Social Sciences and Humanities Marketplace. These technical and functional requirements have been translated into specification documents useful for the implementation of both training platforms.

The standard referred to is the methodology for developing FAIR-by-Design teaching materials (Filiposka et al., 2023) devised by the Skills4EOSC project[15]. This framework ensures that teaching materials are aligned with FAIR principles to increase their impact and reusability within the scientific community and adopts a six-step backward design process: preparation, discovery, design, development, publication and quality control. In accordance with these recommendations, the minimum modular unit of reference has been identified as the learning object, defined in Skills4EOSC as a package of a lesson, an activity and an assessment with a single learning objective and a concrete learning outcome.

The metadata describing each individual learning object has been identified in the Minimal Metadata Set for Learning Resources (Hoebelheinrich et al., 2022) proposed by the Research Data Alliance (RDA)[16], which establishes 14 fields divided into three categories of information (descriptive, access, educational) and is currently being evaluated in projects such as OpenPlato, SSHOC Training Discovery Toolkit and NI4OS Training Platform. It is a flexible scheme, open to possible additions depending on needs related, for example, to different teaching perspectives (formal, professional and informal), and is aimed at maximising the searchability of data without overburdening the descriptor system, ensuring compliance and reusability of existing material. For H2IOSC training, we have adopted the model (van der Lek et al., 2025) already modified by CLARIN, which has invested in training by promoting standard practices and creating a community of trainers through the CLARIN Trainer's Network[17]. This model modified the metadata set by adding specific fields, such as contributors, workload in ECTS, Persistent Identifier (PID), version date and standard citation, thus making it even more flexible (Melaccio et al., 2026).

Recognising that the use of data by third parties can raise ethical and legal issues, we have developed a course[18] that explores the application of the GDPR to linguistic data within the Italian national context. The course is available in modules that can be freely downloaded from the training library and applies specifically to spoken linguistic data, though it offers a method that can be adapted to other contexts. (Draxler et al., 2025)

The Training Environment[19] provides an interactive environment for e-learning courses. From the student's point of view, the platform allows them to create an account with a personal profile that includes the courses taken and certifications obtained, the possibility to spontaneously enrol in open or closed courses (in the latter case, by entering an access code provided by the teacher), and the opportunity to interact with other users through the integrated chat and forum features. The courses already available include introductory courses to language data management and standards, teachings on oral archives, and on Linguistic Linked Open Data.

The Training Library [20] is a repository that allows the storage of modular teaching material organised according to the methodology promoted by Skills4EOSC, adapted to the needs of H2IOSC (Section 3). In accordance with the guidelines adopted, the materials for the courses, converted into `.md` (Markdown) files and organised according to a hierarchy that allows users to find and download even just part of the entire course, depending on their needs. The platform also supports search and discovery features for users interested in reusing educational materials: all the content is open for consultation and available for download upon user authentication through a federated login system. The search interface allows simple natural language searches, which sort results based on keywords entered in the keyword section of the metadata. A faceted system allows users to fur-

[15] https://www.skills4eosc.eu/about

[16] https://www.rd-alliance.org/

[17] https://www.clarin.eu/trainers-network

[18] https://h2iosc-training-library.ilc4clarin.ilc.cnr.it/it/ricerca/archive/150

[19] https://h2iosc-training-platform.ilc4clarin.ilc.cnr.it/login

[20] https://h2iosc-training-library.ilc4clarin.ilc.cnr.it/it/home

ther narrow their search based on a wider selection of metadata, such as publication date, language, author name, format and main topic.

5. The permanent observatory: Collecting habits and needs for monitoring communities

A central pillar of the national infrastructure is the Observatory, a dynamic environment designed to map communities, collect requirements, and monitor trends across disciplines. Building upon earlier landscaping activities in the humanities and heritage domains, the Observatory now evolves into an active coordination and feedback mechanism that guides the development of interoperable digital services. This observatory model enables continuous interaction between data providers and users, fostering community-driven governance of language and cultural data infrastructures.

The H2IOSC Permanent Observatory is one of the flagship components of the Italian Open Science Cloud for the Humanities and Cultural Heritage. Conceived as both a scientific instrument and an operational environment, it monitors, analyses, and valorises the ecosystem of research infrastructures, services, and community practices that converge within the H2IOSC federation. Acting as a data-driven interface complementary to the H2IOSC Marketplace (Sichera et al., 2024, 2025) , the Observatory supports evidence-based decision-making on how data, tools, and users interact within the broader Open Science environment.

The Observatory follows a mixed-methods approach that integrates quantitative surveys, qualitative interviews, and semi-automatic data collection protocols to map communities and their evolving needs. The resulting insights are made accessible through an interactive dashboard that provides continuously updated visualisations and indicators on data usage, interoperability levels, and FAIR compliance. In this way, the Observatory functions as a reflexive mechanism that informs the strategic evolution of the H2IOSC federation and supports long-term policy alignment.

Beyond its analytical function, the Observatory contributes to fostering a culture of Open Science and FAIR data stewardship. The public dashboard also encourages community participation, allowing researchers and infrastructure managers to explore correlations between community needs, available services, and FAIRification progress. This interaction creates a feedback loop that helps prioritise the integration of new datasets and tools into the H2IOSC Marketplace and guides the design of training activities promoting FAIR and Open Science practices.

Technically, the Observatory is implemented as a modular web application integrated with the H2IOSC Marketplace. The two systems exchange information bidirectionally: the Marketplace provides access to digital assets, while the Observatory analyses their use and evolution over time. This synergy enables real-time monitoring of resource ingestion, service uptake, and user engagement, supporting adaptive management of the federated ecosystem. Together, these components constitute a dynamic model for continuous improvement and sustainable coordination across research and cultural data infrastructures.

The Observatory's analyses have also highlighted significant training needs within the research and cultural heritage communities. Survey results show that, while awareness of FAIR and Open Science principles is steadily increasing (cf. Poljak Bilić and Posavec 2024), many researchers still face challenges in their practical application—particularly in metadata creation, licensing, and interoperability. These findings are being used to guide the development of targeted educational programmes within the H2IOSC Training Environment and the H2IOSC Training Library, described in Section 4. By aligning training provision with empirically identified community needs, the infrastructure ensures that capacity-building activities reinforce the broader objectives of FAIRification and sustainable data stewardship.

6. Conclusions and Outlook

The infrastructural undertaking described in this paper is embedded in the broader European Open Science landscape. It contributes to ongoing efforts toward effective interoperability of metadata, tools, and standards. Through its links with other research infrastructures in the Social Sciences and Humanities, as well as in Cultural Heritage, it strengthens cross-domain and cross-disciplinary collaborations and fosters open scholarly publishing also of data and software. The integration of observatory functions, FAIRification workflows, and training services represents a scalable model for both national and supra-national coordination and orchestration of efforts across scientific communities. The ongoing collaboration between research infrastructures demonstrates how Open Science principles can effectively translate into concrete policies and sustainable services.

Future work will focus on the consolidation of the observatory and the continuous FAIRification of both research outputs and training materials. Future developments may also align with other initiatives, such as the Common European Language Data Space (LDS) (Rehm et al., 2024) and ALT-EDIC[21], which mainly target industrial sec-

[21] https://www.alt-edic.eu/

tors, ensuring that national assets are interoperable within the wider European data ecosystems, or data spaces. Our work also connects to the ECCCH-ECHOES project[22] in fostering cooperation with European cultural heritage infrastructures and supporting multilingual access to data and services. In this perspective, the efforts undertaken by CLARIN-IT within H2IOSC, as described in this paper, place it as an interesting model and key actor in the emerging European ecosystem for FAIR and reusable research and training data.

7. Acknowledgements

H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

This work has also been partially supported by LLMs4EU “Large Language Models for the European Union” project, funded by the European Union through the Digital Europe Programme (DIGITAL-2024-AI-B-06-LANGUAGE - GA 101198470).

8. Bibliographical References

Ariovaldo Veiga de Almeida, Maria Manuel Borges, and Licinio Roque. 2017. The european open science cloud: A new challenge for europe. In *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM 2017, New York, NY, USA. Association for Computing Machinery.

CLARIN ERIC. 2023. Clarin concept registry (ccr). https://www.clarin.eu/content/clarin-concept-registry. Accessed: 2025-10-24.

Emiliano Degl’Innocenti, Monica Monachini, Alberto Bucciero, Enrico Pasini, Bruno Fanini, and Francesca Frontini. 2023. H2iosc: Humanities and heritage open science cloud. In *La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella*, page 63–64.

Anusuriya Devaraju, Robert Huber, Peter Herterich, Andreas Petzold, Peter Wittenburg, Hendrik Diener, Ulrich Schwardmann, and Ulrike Wößner. 2020. F-uji: An automated fair data assessment tool. In *Proceedings of the 17th International Conference on Digital Preservation (iPRES 2020)*, Beijing, China. International Conference on Digital Preservation. Developed under the FAIRsFAIR project to automate the evaluation of FAIR data compliance.

Martin Doerr. 2003. The CIDOC conceptual reference model: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92.

European Commission: Directorate General for Research and Innovation. 2025. The european strategy on research and technology infrastructures. https://data.europa.eu/doi/10.2777/9553939. Accessed: 2025-10-22.

FAIRsFAIR Project Consortium. 2021. Fairsfair data object assessment metrics v0.8. Zenodo. Deliverable D5.4. Defines indicators and metrics for automated FAIR data assessment.

Sonja Filiposka, Dominique Green, Anastas Mishev, Vojdan Kjorveziroski, Andrea Corleto, Eleonora Napolitano, Gabriella Paolini, et al. 2023. D2.2 methodology for FAIR-by-design training materials.

Nancy J Hoebelheinrich, Katarzyna Biernacka, Michelle Brazas, Leyla Jael Castro, Nicola Fiore, Margareta Hellström, Emma Lazzeri, et al. 2022. Recommendations for a minimal metadata set to aid harmonised discovery of learning resources.

Michele Mallia, Fahad Khan, Silvia Calvi, and Klara Dankova. 2025. Methodology for converting and publish tabular data into skos resources via python notebooks. In *CLARIN Annual Conference Proceedings*, page 113.

Daniele Melaccio, Federico Boschetti, and Monica Monachini. 2025. Interfacing CLARIN with H2IOSC: Metadata Interoperability through Ontology-based Mediation. In *Proceedings of the CLARIN Annual Conference 2025*.

Daniele Melaccio, Giulia Pedonese, Iulianna Van der Lek, Thalassia Kontino, and Francesca Frontini. 2026. Fair training materials for disciplinary research infrastructures: Metadata, vocabularies, and selective harvesting in the h2iosc ecosystem. In *Proceedings of the 22nd Conference on Information and Research Science Connecting to Digital and Library Science (IRCDL 2026)*, Modena, Italy. To appear.

[22] https://www.echoes-eccch.eu/

Ljiljana Poljak Bilić and Kristina Posavec. 2024. Fairness of research data in the european humanities landscape. *Publications*, 12(1).

Georg Rehm, Stelios Piperidis, Khalid Choukri, Andrejs Vasiljevs, Katrin Marheinecke, Victoria Arranz, Aivars Bērziņš, Miltos Deligiannis, Dimitris Galanis, Maria Giagkou, Katerina Gkirtzou, Dimitris Gkoumas, Annika Grützner-Zahn, Athanasia Kolovou, Penny Labropoulou, Andis Lagzdiņš, Elena Leitner, Valérie Mapelli, Hélène Mazo, Simon Ostermann, Stefania Racioppa, Mickaël Rigault, and Leon Voukoutis. 2024. Common European language data space. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3579–3586, Torino, Italia. ELRA and ICCL.

Pietro Sichera, Cristina Marras, and Enrico Pasini. 2024. Orchestrazione API per workflow applicativi nell'ambito delle digital humanities.

Pietro Sichera, Monica Monachini, Valeria Quochi, Nicola Giampietro, Vittoria Fabiani, Roberta Ottaviani, and Roberta Bianca Luzietti. 2025. Synergies between CLARIN-IT and OPERAS-IT within H2IOSC: Monitoring communities and orchestrating digital services. In *Proceedings of the CLARIN Annual Conference 2025*, Vienna, Austria. CLARIN ERIC.

SSHOC Consortium. 2022. Sshoc reference ontology (sshocro). Accessed: 2025-10-24.

Dieter Van Uytvanck, Andreas Witt, and Daan Broeder. 2012. Cmdi: A component metadata infrastructure for CLARIN. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. ELRA. Describes CMDI's component-based metadata framework.

Iulianna van der Lek, Francesca Frontini, Darja Fišer, and Alexander König. 2025. Making the CLARIN training materials FAIR-by-design.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1).

Claus Zinn. 2016. The CLARIN language resource switchboard. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. ELRA.

9. Language Resource References

Christoph Draxler and van den Heuvel, Henk and van der Lek, Iulianna and Francesca Frontini and Giulia Pedonese. 2025. *Introduzione alla Gestione dei Dati Orali*. H2IOSC Training Library. PID https://h2iosc-training-library.ilc4clarin.ilc.cnr.it/it/ricerca/archive/150.

AA. VV. 2005. *Musisque Deoque (MQDQ)*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli". PID http://hdl.handle.net/20.500.11752/OPEN-555.

Zanola, Maria Teresa and Dankova, Klara and Grimaldi, Claudio and Serpente, Anna. 2023. *Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli". PID http://hdl.handle.net/20.500.11752/OPEN-987.

Zanola, Maria Teresa and Osservatorio di terminologie e politiche linguistiche, Università Cattolica del Sacro Cuore. 2025. *MADIN-TERM : Agri-food terminology*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli". PID http://hdl.handle.net/20.500.11752/OPEN-2142.