FAIR Training Materials for Disciplinary Research Infrastructures: Metadata, Vocabularies, and Selective Harvesting in the H2IOSC Ecosystem

Daniele Melaccio[1,*], Giulia Pedonese[1], Francesca Frontini[1], Iulianna van der Lek[2] and Thalassia Kontino[2]

[1]CNR – Istituto di Linguistica Computazionale "A. Zampolli", Pisa, Italy
[2]CLARIN ERIC, Utrecht, Netherlands

Abstract

In a lifelong learning society and especially now in the era of Artificial Intelligence, high-quality, flexible learning resources are essential for acquiring the new skills required by the ever-evolving job market. Research infrastructures play a central role in providing users with specialised training and advanced digital tools, fostering innovation in their fields. CLARIN ERIC, the Common Language Resources and Technology Infrastructure, has consistently promoted training initiatives in linguistic and language technologies, especially through the CLARIN Learning Hub. The Italian national consortium CLARIN-IT has developed this approach within the NRRP project H2IOSC, aiming to create a cluster of research infrastructures in the field of Social Sciences and Humanities. In this context, CLARIN-IT led the work package dedicated to training and developed two platforms: an e-learning portal and a digital library for long-term deposit and curation of training materials as FAIR digital objects. The Skills4EOSC FAIR-by-Design methodology and the SSHOC vocabularies adopted in the H2IOSC project ensured metadata compatibility and interoperability of learning resources across domains. Moreover, the adopted standards will enable harvesting resources in the H2IOSC Marketplace and selective harvesting for domain-specific platforms such as the CLARIN Learning Resource Catalogue. This paper proposes an architecture for metadata mapping and harvesting of training resources to maximise their reusability across infrastructures.

Keywords

FAIR learning resources, metadata interoperability, CLARIN, H2IOSC, SSHOC, selective harvesting

1. Introduction

For its flexibility and immediate access, especially via e-learning systems, digital training is increasingly recognised as a key pillar of Open Science and Research Infrastructures (RIs), considered as networks of resources and tools used by research communities to perform specialised tasks in their scientific domains. In such a context, training initiatives are a central vehicle for capacity-building, bridging the gap between scientific communities and infrastructure services.[1] Moreover, training is integral to user involvement and outreach activities, ensuring that scholars acquire the new competencies required for the effective use of digital tools and resources. Yet the impact of training is hampered by the fragmentation of training resources and the lack of harmonised metadata that would enable easy discovery, reuse and cross-disciplinary uptake. Within the Italian National Recovery and Resilience Plan, the Humanities and cultural Heritage Italian Open Science Cloud (H2IOSC) project[1] addresses this challenge as a national federated infrastructure that integrates CLARIN, DARIAH, E-RIHS and

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.
†These authors contributed equally.
daniele.melaccio@cnr.it (D. Melaccio); giulia.pedonese@cnr.it (G. Pedonese); francesca.frontini@cnr.it (F. Frontini); i.vanderlek@uu.nl (I. v. d. Lek); t.kontino@uu.nl (T. Kontino)
https://www.ilc.cnr.it/ (D. Melaccio); https://www.ilc.cnr.it/ (G. Pedonese); https://www.ilc.cnr.it/ (F. Frontini); https://clarin.eu/ (I. v. d. Lek); https://clarin.eu/ (T. Kontino)

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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-08-20

OPERAS,[2] creating a shared environment[2] where educational resources from different communities can be aligned and jointly exploited. Against this backdrop, the paper proposes a metadata-centred, FAIR-by-Design model for disciplinary learning resources that can support interoperability across these infrastructures, enhance the visibility and reusability of resources, and contribute to a more coherent and sustainable training ecosystem in the Social Sciences and Humanities.

2. Background and Related Work

2.1. FAIR principles for learning resources

FAIR learning resources are educational assets such as slides, lesson plans, tutorials, datasets, code examples, videos, and exercises designed to be Findable, Accessible, Interoperable, and Reusable by both users and machines on the web according to the FAIR Guiding Principles for scientific data management and stewardship.[3] This means learning resources are not only open and accessible, but also described and structured so they are easily discoverable, understandable, and adaptable to suit the learning needs of other target audiences.

They are typically created by depositing the learning content in trusted repositories (institutional, thematic, or general-purpose), attaching standard and rich metadata, assigning persistent identifiers and open licences that clearly state reuse rights. Making learning resources FAIR is vital for Open Science, as it supports transparency and equity in knowledge transfer. This way, users can locate high-quality educational resources, build on them without redoing work, and update or translate them for new communities, thereby increasing the reach and long-term value of scientific education. For findability, learning resources should: have a persistent identifier (e.g., DOI, handle); include rich, standardised metadata sets integrating general information such as **title**, **authors**, **abstract**, **keywords**, **language**, **version**, **date** and didactic information like **learning outcome**, **target audience**, **topic**, **prerequisites** and **difficulty level**; and, finally, be indexed in open repositories, catalogues, and registries.

To maximise their impact, materials need to be legally and technically reusable. Key requirements for reusability include an open, clear license (such as, but not limited to, CC0 or CC-BY) describing the rights and the obligations of the re-user; provenance and attribution information either within the metadata schema or in supporting materials, including details about the authors, the date of creation and a correct citation of its sources; versioning and update history so users know what they are reusing; enough contextual documentation to enable adaptation from third parties (e.g. additional documents specifying setup instructions, expected outcomes, and assessment guidance); and editable source formats (e.g., .pptx/.odp alongside PDF).

To meet the interoperability requirements, materials should be developed and structured so they can work with other tools and resources. This usually involves: (1) using open, widely supported formats, e.g., Markdown, HTML, TXT, CSV, etc., rather than proprietary-only ones; (2) applying community vocabularies and metadata standards, e.g., Dublin Core, or discipline-specific schemas; and (3) referencing external resources through persistent identifiers and standard citation practices. Interoperable materials can be integrated into learning platforms, combined into larger curricula, and linked with FAIR datasets or software in the same research ecosystem.

We will provide insights into the metadata schema implemented in CLARIN-IT learning resources within the H2IOSC project in Section 3 and introduce two different typologies of learning resources in Section 4. These are called “training materials” and “courses”, corresponding to the intended audience (trainers and students, respectively).

2.2. The role of disciplinary RIs

Disciplinar RIs play a key role in designing FAIR and sustainable learning resources by providing trusted platforms, shared standards, and community-driven curation. By aggregating and publishing

[2]CLARIN: https://www.clarin.eu/; DARIAH: https://www.dariah.eu/; E-RIHS: https://www.e-rihs.eu/; OPERAS: https://operas-eu.org/

domain-specific learning resources, they reduce duplication of effort, ensure disciplinary relevance, and help training content circulate across national and European nodes. In the Social Sciences and Humanities (SSH), several RIs have built mature training ecosystems that function as both repositories and communities of practice.

CLARIN - Common Language Resource and Technologies Infrastructure - supports training through a federated ecosystem that includes the CLARIN Learning Hub,[3] which offers open educational resources, complete training modules, and guidance for designing courses and workshops around language-data research. This environment is complemented by a Learning Resources Catalogue,[4] where materials are described with consistent metadata and released under Creative Commons licences, directly supporting reusability. At the national level, CLARIN nodes adapt and extend this approach to local communities. For example, CLARIN-CH (Switzerland) runs recurring training sessions and webinars that introduce researchers to the CLARIN-CH ecosystem[5] and its tools, showing how national nodes translate pan-European resources into targeted, community-specific learning pathways. Together, these layers ensure that training is discoverable across borders while staying responsive to disciplinary and linguistic contexts.

In digital humanities, DARIAH Campus[6] serves as a discovery and hosting platform for open learning resources, events, and curated external materials. It was developed to provide "trustworthy, reusable, open learning resources" for both learners and trainers, and is now a core DARIAH service. RIs that focus on scholarly communication and heritage science also have distinctive training priorities: OPERAS, the RI for open scholarly communication in SSH, explicitly aims to help communities adopt open and digital publication practices and to co-create services for discovery, dissemination, evaluation, and quality assurance. Its recent training framework emphasises modular, adaptable Open Science and scholarly-communication training aligned with FAIR and OER principles, reflecting the need to build capacity not just in tools, but also in open publishing workflows, licensing, and research assessment cultures.[4]

Similarly, E-RIHS, officially established as an ERIC in March 2025, coordinates advanced facilities and data for heritage science and provides dedicated training opportunities (including through its HS Academy) to build cross-disciplinary skills spanning humanities, experimental sciences, and digital methods. For E-RIHS, training needs are closely linked to enabling researchers to access, interpret, and reuse complex heritage-science data and instrumentation across distributed labs and platforms—again reinforcing the need for RI-level support for FAIR-aligned capacity building.

In short, disciplinary RIs act as amplifiers of FAIR training: they make materials visible and credible, align them with community standards, and sustain them through networks of national and thematic nodes.

2.3. The SSHOC cluster in EOSC

The Social Sciences and Humanities Open Cloud (SSHOC) is a Horizon 2020 cluster project that built the SSH contribution to the European Open Science Cloud (EOSC) by federating services, data, and communities from major SSH research infrastructures (e.g., CLARIN, DARIAH, CESSDA).[7] A central strand of SSHOC was capacity building: creating a shared training space and helping trainers and researchers navigate EOSC-ready SSH tools and datasets. In national alignment, initiatives such as the NRRP project H2IOSC extend this logic to the Italian context, connecting humanities and heritage communities to EOSC practices through modular training and two dedicated training platforms.[8]

A key SSHOC output in this direction was the SSH Training Discovery Toolkit (TDT),[9] an inventory

[3] https://www.clarin.eu/content/learning-hub
[4] https://www.clarin.eu/learning-resources-catalogue
[5] https://clarin-ch.ch/services/education-training/
[6] https://campus.dariah.eu/
[7] SSHOC: https://sshopencloud.eu/; EOSC: https://open-science-cloud.ec.europa.eu/; CESSDA: https://www.cessda.eu/
[8] SSHOC training: https://sshopencloud.eu/training; H2IOSC training infrastructure: https://www.h2iosc.cnr.it/training-infrastructure/; EOSC training resources: https://www.eosc-hub.eu/training-material
[9] https://sshopencloud.eu/ssh-training-discovery-toolkit

of SSH training materials[10] The TDT was designed to function as a bridge between distributed training providers and EOSC discovery, encouraging consistent descriptions, open licensing, and linking materials to the SSH Open Marketplace.[11] In parallel, EOSC-wide initiatives (such as the EOSC Knowledge Hub / EOSC Training Catalogue) have developed or adopted minimal metadata sets for training resources, informed by RDA work, to ensure that training can be catalogued, searched, and reused across domains.[5]

Together, SSHOC, H2IOSC, and EOSC training catalogues represent a layered approach: SSH-specific aggregation and curation, national adaptation and delivery, and cross-domain EOSC interoperability through shared metadata. However, some gaps were identified in earlier training efforts. For example, SSHOC deliverables reviewing pre-existing SSH training landscapes highlighted several structural weaknesses that motivated the cluster's interventions. First, learning resources were highly fragmented across many providers and websites, making them hard to locate systematically - hence the need for an inventory like the TDT.[6] Second, many resources lacked consistent or rich metadata, which limited findability and prevented reliable inclusion in cross-platform catalogues. Third, reuse was often constrained by unclear licensing, missing provenance, and non-editable formats, making it difficult to adapt materials even when they are found. Finally, gaps existed in coverage and accessibility: some disciplines, languages, and skill levels were well served, while others had few up-to-date learning resources, and pathways for beginners into EOSC services were not always clear.

Overall, the SSHOC cluster's training and metadata work can be read as a FAIR-oriented response to these gaps: aggregate what exists, describe it better, license it openly, connect it to EOSC discovery, and support national nodes (like H2IOSC) in turning shared resources into situated learning experiences.

3. Metadata Model for FAIR Learning Resources

High-quality metadata transforms training outputs into FAIR learning objects that can be portable across repositories, infrastructures, and national nodes. Building on FAIR-by-Design and EOSC/SSH training cataloguing work, three requirements are essential. Learning resources are produced and hosted on many platforms (RI catalogues, institutional repositories, Moodle instances, Zenodo communities, project sites). To be findable across this space, metadata must be semantically aligned: the same concept (e.g., "tutorial", "workshop", "trainer material") should be expressed with shared vocabularies or mapped between them. Cross-disciplinary FAIR initiatives emphasise agreement on common catalogue standards and metadata crosswalks so that aggregation services can interpret descriptions consistently.[12] In practice, this means using community schemas (like LRMI[13] or EOSC Training Profile models[14]) and, where local profiles differ, providing mappings that preserve meaning.

FAIR training metadata should not only be human-readable, but also machine-actionable so that catalogues and EOSC aggregators can automatically harvest, index, and recommend materials. The specifications of the EOSC Training Catalogue rely on minimal, structured metadata that can be exchanged via open APIs. GO FAIR's "Metadata for Machines" approach highlights that metadata templates should be structured for automated processing and reuse from the outset.[15] Concretely, machine-actionability implies: persistent identifiers; metadata exposed via standard protocols (e.g., OAI-PMH, JSON-LD, schema.org markup); controlled fields instead of free text where possible; and explicit links to related objects (versions, prerequisites, datasets, tools).

Generic descriptors such as **title author**, and **license** are necessary but not sufficient for SSH training. FAIRness improves when metadata uses discipline-specific concept schemas that reflect how a community describes methods, objects, and learner needs. LRMI - built on Dublin Core and

[10]The term *digital learning resources* is often used broadly for slides, courses, and tutorials. In this paper, we use *learning resources* as a general label and distinguish between training materials and courses/programmes in Section 4.
[11]https://eosc.eu/use-case/ssh-open-marketplace
[12]https://www.fairsfair.eu/fair-semantics-interoperability-and-services-0
[13]https://www.dublincore.org/specifications/lrmi/
[14]https://wiki.eoscfuture.eu/display/PUBLIC/EOSC+Training+Resource+Profile+-+Data+Model
[15]https://www.go-fair.org/today/making-fair-metadata/

schema.org - provides education-specific properties (**learning resource type**, **intended audience**, **educational level**, **time required**) and encourages controlled vocabularies for these fields. Domain-specific vocabularies (e.g., CLARIN/DARIAH/CESSDA terms, heritage-science ontologies) can then be layered into LRMI/EOSC profiles to ensure that a Digital Humanities tutorial or a corpus-linguistics lesson is described in the language its users actually search for, while still interoperating with EOSC-wide discovery.

In a federated setting such as H2IOSC, where learning resources are produced and described within different research infrastructures, metadata interoperability is not only a matter of shared fields and formats, but also of consistent interpretation. Even when common metadata profiles are adopted, concepts such as resource type, contributor role, or relations between learning resources and other research assets may be understood and implemented differently across infrastructures. Ensuring conceptual coherence across these heterogeneous descriptions is therefore a key requirement for cross-infrastructure discovery and reuse. These requirements make training materials portable educational assets: consistently understood across infrastructures, automatically harvestable, and richly described in the disciplinary terms that enable meaningful reuse.

In H2IOSC, learning resources are created and adapted following the guidelines of the FAIR-by-Design Methodology[7] developed in the Skills4EOSC project,[16] and include the application of a specific metadata set from the Research Data Alliance (RDA) community. This set is composed of 14 labels among descriptive information (title, abstract, authors, language, keywords and version date), access information (URL, URL type, access cost, license) and educational information (target group, learning resource type, learning outcome, expertise level). This set was adapted for CLARIN resources, adding more educational information (such as prerequisites and expertise level), and was used as the starting point for both the FAIRification of existing educational resources and the design of new ones, as well as the technical implementation of the H2IOSC training platforms, which will be detailed in Section 4.

To better align with the fit SSH and RI practices, the schema is extended using the CLARIN adaptation, adding fields such as *contributors*, *workload* expressed in ECTS, and *persistent identifiers* (PIDs).[8] These additions strengthen provenance, support citation of learning objects as scholarly outputs, and make the Library's records more informative for both trainers and aggregators.

3.1. FAIR principles in practice

Applying FAIR principles to learning resources means making a few concrete design choices from the start, so that materials remain traceable, citable, and safely reusable over time.

Every training object (slide deck, tutorial, notebook, video, exercise set, or full course) should have a persistent identifier (PID) such as a DOI, handle, or stable repository URL. A PID guarantees a durable reference even if files move or platforms change. It also enables correct citation, automated harvesting by training catalogues, and reliable linking between related objects (e.g., a lesson pointing to a dataset or tool). PIDs are typically assigned when materials are deposited in trusted repositories (institutional repositories, RI catalogues, or general platforms like Zenodo) and should be displayed prominently within the material and its landing page.

Training resources evolve: errors are fixed, tools change, examples and exercises become outdated, and content may be translated or localised to national contexts. FAIR practices, therefore, require explicit versioning (e.g., v1.0, v1.1, v2.0) and a visible change log. Each version should remain accessible, with a clear "latest version" pointer, so users know exactly what they are reusing and citing. Alongside versioning, materials need provenance metadata: who created them, when, under which project or RI, and what sources or dependencies they build on (datasets, software, images, third-party lessons). Provenance supports trust, reproducibility, and proper attribution, and helps future trainers assess whether a resource still fits their context.

FAIR reuse is impossible without legal clarity. Training materials should carry an explicit open license that states reuse conditions in plain terms and in machine-readable form. Creative Commons licences

[16] https://www.skills4eosc.eu/

are the default in Open Science training; CC BY (attribution) is often recommended because it maximises reuse while ensuring credit, and CC BY-SA is useful when you want adaptations to stay open. Whatever the choice, the license must be attached to both the landing page metadata and the files themselves (e.g., a slide footer or README). Reuse also benefits from noting any third-party content with different rights (images, excerpts, datasets), so downstream users can adapt the material confidently without legal risk.

3.2. Controlled vocabularies

Applying FAIR principles also entails implementing controlled vocabularies to ensure semantic interoperability. Controlled vocabularies replace ambiguous free-text tags with stable, shared concepts that machines and humans can interpret the same way. In the SSH ecosystem, this work is increasingly coordinated through RI and cluster-level services, enabling a consistent description of training materials across EOSC and RI catalogues. SSHOC promoted a common approach to describing SSH training resources through its Training Discovery Toolkit and the wider SSH Open Marketplace ecosystem. The Toolkit functions as a cross-disciplinary inventory of training materials, and its discoverability depends on using harmonised labels for topics, learning resource types, and disciplines (see above - SSHOC Discovery Toolkit). SSHOC also provides a dedicated SSH Vocabulary Service (Vocabs) for collaboratively maintaining and publishing SKOS vocabularies to support interoperability among SSH datasets, tools, and training resources.[17].

Adopting SSHOC vocabularies in training metadata, therefore, means: tagging content with shared topic terms (e.g., specific Open Science or method areas); selecting standardised resource types (tutorial, exercise, webinar, lesson plan, etc.); associating materials with recognised disciplinary categories, so that catalogues can align results from many providers and offer a faceted search that works across infrastructures. While SSHOC offers an umbrella layer, real semantic precision comes from integrating RI-specific knowledge organisation systems: CLARIN CCR (Concept Registry) and CMDI components. CLARIN maintains a Concept Registry of well-defined data categories and a Component Registry used to build metadata profiles (CMDI) from reusable blocks.[18] Linking training materials to CCR concepts (e.g., language resource types, linguistic annotation categories, tool functions) allows CLARIN-based training to be indexed in ways that match how its users search and how CLARIN services harvest metadata.

DARIAH taxonomies (e.g., TaDiRAH). DARIAH curates digital-humanities vocabularies such as TaDiRAH, a SKOS taxonomy for digital research activities, designed to enhance the discoverability and interoperability of DH resources.[19] Using TaDiRAH (and related DARIAH vocabularies) in training metadata makes DH learning objects searchable by method and activity (“annotating”, “visualising”, “modelling”, etc.), not only by generic keywords. OPERAS vocabularies. OPERAS coordinates open scholarly communication services for SSH within EOSC, and its service ecosystem (including discovery and publishing workflows) relies on shared terminology to describe publication types, dissemination practices, and scholarly-communication functions. https://operas-eu.org/services/ Aligning training metadata with OPERAS-relevant vocabularies ensures that materials about open access books, diamond OA models, peer-review innovations, or multilingual dissemination are tagged with concepts understood consistently across OPERAS services and communities.

4. Training Materials and Courses: a Conceptual Distinction

A conceptual distinction made explicit in the Skills4EOSC FAIR-by-Design methodology and echoed in H2IOSC training publications and platforms is between (a) reusable training materials and (b) courses or programmes. The FAIR-by-Design approach treats training outputs as FAIR digital objects, often called

[17] https://vocabs.sshopencloud.eu/browse/en/
[18] https://www.clarin.eu/content/clarin-concept-registry
[19] https://vocabs.dariah.eu/tadirah/en/

learning objects: small to medium-granularity resources (a slide deck, a tutorial, a worked example, a notebook, a short video, an exercise set) that are designed to be published, versioned, cited, and reused by educators in new contexts. H2IOSC similarly frames its Training Library and Training Environment (details in Sections 4-5) as spaces where modular, FAIR learning objects can be created and shared, enabling lecturers and trainers to pick them up, adapt them, and assemble them into local teaching.

By contrast, courses/programmes are learner-facing pathways: structured sequences of activities with pacing, facilitation, and assessment (e.g., a semester course, a summer school, a MOOC, a workshop series). These may use learning objects, but their primary purpose is to guide learners through a curriculum, not to maximise the modular reuse of each component. This distinction matters because FAIRness is partly about who the object is for and how it will be shared across the web. A learning object is optimised for train-the-trainer reuse: clear objectives, editable formats, instructor notes, and licensing that enables adaptation. A course is optimised for an end-user learning journey: coherence, sequencing, learner support, and often platform-specific delivery. Both can be FAIR, but the design targets differ, and so do the metadata signals that make them discoverable and reusable. Skills4EOSC explicitly asks designers to define purpose, learning objectives, and target audience early in the workflow, because the same content can serve different roles depending on context.

H2IOSC publications emphasise the same point: the infrastructure is meant to foster collaboration between teachers and students, and to support the circulation of reusable materials among educators while still serving learners through assembled courses.[9] Importantly, "teacher" and "student" here are pedagogical roles, not personal attributes: a single person can be a trainer in one setting and a learner in another. Treating the role as contextual avoids stereotyping while still allowing the resource to be properly indexed and reused. Once you separate learning objects from courses, specific metadata fields from the recommended set developed by Research Data Alliance become more relevant:

- **Learning Resource Type** – Is the item a learning object (slide deck, lesson, exercise, tutorial, dataset-in-teaching, instructor kit) or a course/programme (workshop, MOOC, syllabus, curriculum)? Skills4EOSC templates and H2IOSC cataloguing both rely on this to support correct discovery and assembly.
- **Audience** – Who is the primary user? For reusable training materials, the intended audience is educators/trainers (for reuse in teaching from a train-the-trainer perspective), even if learners may also read them. For courses, the intended audience is learners, sometimes with a specified level (BA/MA/PhD, beginner/advanced, discipline). Declaring this up front improves findability, prevents misapplication, and specifies the role of the resource in a learning scenario: instructor material, learner material, or both? The FAIR-by-Design methodology adapted in H2IOSC emphasised contextual clarity to enable reuse. In H2IOSC's modular environment, marking pedagogical role helps trainers rapidly filter for "ready-to-teach" units versus student-only handouts or activities.
- Although educational information is key to both end users, some metadata fields become more relevant from either a student's or a teacher's perspective, depending on the type of digital object being described. For example, in the case of training materials, access information, such as **license**, is key to determining reusability potential and adaptation in different fields, whereas the information about access cost is fundamental for a student wanting to access a course.

5. The H2IOSC Training Library

The H2IOSC Training Library is a core component of the H2IOSC Training Infrastructure: a repository specifically designed for the deposit, long-term preservation, and reuse of FAIR training materials in the Social Sciences and Humanities. Rather than functioning as a learner-facing platform, the Library is conceived primarily from the train-the-trainer perspective, allowing educators to publish modular learning objects that others may discover, cite, adapt, and integrate into courses. The Training Library addresses a recurring gap in earlier SSH training landscapes: materials were often scattered, poorly

described, and difficult to reuse. Therefore, the repository is framed as a FAIR-by-design solution that treats learning materials as digital scholarly objects, aligning training with EOSC principles and RI ecosystems. The platform provides a single discovery point for reusable SSH teaching resources curated within H2IOSC, while remaining open to contributions from the broader Italian and international community.[20]

The platform relies on a shared methodology for producing and converting materials into FAIR learning objects. Existing training content was reviewed and, when necessary, converted into the most open and editable formats available to maximise future reuse (e.g., by releasing source files alongside PDFs). Publication in the Library follows guided deposit practices: creators upload the resource, add rich metadata, choose licenses, and link the object to relevant topics and disciplines, so it is interoperable with H2IOSC and EOSC discovery services.

The Library is a searchable and citable catalogue of learning objects where users can explore materials through a simple search interface and discipline/topic filters; each item has a dedicated landing page. Crucially, the Library supports versioning and visibility of updates (e.g., "new version" flags and timestamps on records), helping trainers select current materials while preserving traceability to earlier editions. Each record also provides standard citation support, including export formats such as BibTeX, reinforcing the treatment of training outputs as scholarly contributions.

Consistent with the FAIR-by-design approach, the Library depends on structured, machine-actionable metadata. These shared standards for describing training materials allow them to be harvested and reused across infrastructures. Records include fields such as learning resource type, discipline, intended audience, and topical keywords, and are aligned with SSHOC/H2IOSC vocabulary practices to ensure semantic interoperability.

The Training Library is paired with the H2IOSC Training Environment [21], a virtual learning platform for interactive courses. In Pedonese et al.'s model, learning objects deposited in the Library can be assembled into learner-oriented courses delivered through the Training Environment, creating a full lifecycle from reusable educator materials to structured training programmes. This division of labour mirrors the conceptual distinction outlined earlier: the Library optimises reuse and discoverability for educators, while the Environment optimises pedagogy and learning paths for students.

5.1. Architecture

The metadata architecture of the H2IOSC Training Library is designed to make learning objects FAIR-by-Design and interoperable with EOSC/SSH discovery services. The schema combines a shared core from the Research Data Alliance with RI-specific enrichment, and a controlled-vocabulary layer aligned to SSHOC.

The Library organises content through six nested containers (**learning path** → **course** → **section** → **unit** → **module** → **learning object**), so trainers can deposit and download materials at different granularities. The corresponding hierarchical structure is illustrated in Annex A (see Fig. 1).

Only two nodes are required at the time of the deposit: the learning path (the root container with contextual documents like README, facilitation guide, license, citation file, syllabus, and release notes) and the learning object (the leaf node that hosts the actual reusable materials). Each container is described using the same metadata grid, and lower levels can inherit metadata from their parents, reducing duplication while keeping consistency across the hierarchy. This setup mirrors the FAIR-by-design distinction between reusable educator materials and learner-facing courses, while keeping everything within a single coherent metadata framework.

At every level, H2IOSC uses the RDA Minimal Metadata Set for Learning Resources as the backbone, ensuring cross-platform interoperability and EOSC compatibility. The RDA profile provides the essential descriptive fields needed for discovery and reuse (e.g., title, description, keywords, learning outcomes,

[20] The H2IOSC Training Library is available at this link: https://h2iosc-training-library.ilc4clarin.ilc.cnr.it/it/home

[21] The H2IOSC Training Environment is available at this link, and accessible to academic users in Europe via CLARIN Federated Login: https://h2iosc-training-platform.ilc4clarin.ilc.cnr.it/login

resource type, audience, language, dates), and is already widely recognised in EOSC training cataloguing. An example of the metadata grid adopted in the deposit workflow is provided in Annex A (see Fig. 2).

A key interoperability layer is the use of controlled vocabularies drawn from the SSHOC Marketplace for the metadata fields that require standardised values (notably topics, learning resource types, and disciplinary areas). Pedonese et al. note that H2IOSC will reuse SSHOC vocabularies and provide Italian translations where needed, preserving semantic alignment while supporting multilingual discovery. This alignment serves two purposes: it guarantees that Library records can be harvested and indexed consistently alongside other SSHOC/EOSC training resources; and it allows H2IOSC materials to participate in shared faceted search and cross-RI workflows within the H2IOSC Marketplace[10] and the wider SSH Open Marketplace ecosystem.

Last but not least, the Library's metadata is designed to flow upward into the H2IOSC Marketplace. The Marketplace serves as a discovery layer with slightly coarser granularity, but its metadata remains compatible because both systems share the same conceptual model and vocabulary backbone. This enables a single search interface to return mixed object types (courses, events, and reusable learning objects) while keeping them semantically distinguishable.

5.2. Content types

The H2IOSC Training Library is designed to host a broad range of FAIR learning objects that reflect how SSH training is actually produced and reused. Core content types include: tutorials (step-by-step guides focused on practical skills), lectures (conceptual or methodological presentations, often slide-based), Jupyter notebooks (executable, literate programming resources that combine narrative, code, and outputs), and demonstrators (hands-on examples or mini-workflows that showcase tools, datasets, or methods in action). A key requirement of the Library is support for multimodality: learning objects may combine text, images, audio/video, code, and data within the same training package. This allows trainers to publish not only static materials (e.g., PDFs or slide decks) but also interactive and media-rich resources—such as notebooks paired with datasets, screencasts linked to exercises, or demonstrators including video walkthroughs—while keeping all components together under a single, citable, FAIR record.

6. Selective Harvesting across Distributed Infrastructures

6.1. Need for selective harvesting

As SSH training resources proliferate across projects and infrastructures, a growing challenge is how to harvest metadata without flooding catalogues. The SSHOC Training Discovery Toolkit already aggregates hundreds of items from dozens of sources, and its value comes from curation and targeted discovery rather than exhaustive ingestion. If every small update, duplicate slide deck, or highly local course instance were automatically harvested into EOSC-level catalogues, users would face catalogue overload: search results become noisy, quality signals weaken, and truly reusable materials are harder to find. Selective harvesting (ingesting only those resources that meet clear relevance and FAIRness thresholds) is therefore necessary to keep discoverable resources usable and trustworthy.

A complementary reason is the need to provide domain-specific views on the same underlying training corpus. Different communities search differently: a linguist looking for corpus annotation workflows, a digital humanist searching by method, or a heritage scientist needing instrument-oriented modules will not benefit from a single flat list. Disciplinary and national infrastructures thus act as filters and lenses, exposing subsets of training materials that are meaningful for their users while still contributing to broader EOSC discovery through shared metadata. H2IOSC, for example, positions its Training Library as a curated repository of reusable educator-oriented learning objects, rather than an indiscriminate mirror of all training activity; this makes its catalogue navigable and aligned with SSHOC vocabularies for cross-platform harvesting. SSHOC plays a similar role at the cluster level

by selecting and describing items in ways that support trainers across SSH, avoiding duplication and emphasising reusability.

Finally, selective harvesting is increasingly relevant also for cross-cluster and cross-domain initiatives. The Cultural Heritage Cloud ecosystem, developed through projects like ECHOES–ECCCH, aims to build community-led capacity-building and curricula for heritage professionals. https://www.echoes-eccch.eu/ Because heritage science training intersects with SSH but also with conservation, museums, and technical labs, it will require tailored entry points rather than full ingestion into generic SSH catalogues. Selective harvesting enables ECHOES–ECCCH to surface training that is both relevant and technically ready for reuse, while allowing EOSC and SSH catalogues to link to domain-focused collections rather than duplicating them.

6.2. Technical approach

Selective harvesting in the H2IOSC ecosystem relies on a lightweight yet robust technical stack that can move training metadata across catalogues while preserving semantic precision and avoiding overload. The approach combines standard harvesting protocols, controlled-vocabulary filtering, and tight alignment with existing SSH infrastructures. The H2IOSC Training Library exposes its records through open machine interfaces, enabling other services to harvest and index them without manual intervention. Two complementary channels are used: OAI-PMH endpoints support batch harvesting in a widely adopted, repository-to-catalogue standard. This supports incremental updates (only new/changed records), ensures long-term compatibility with library and EOSC harvesting tools, and keeps the workflow simple for both sides.

REST/JSON APIs enable more flexible, on-demand access to records, allowing downstream platforms to query by parameters (e.g., topic, resource type, audience, language) and build custom discovery features. Using both routes ensures that H2IOSC materials can be harvested by "traditional" repository aggregators and by modern EOSC-style marketplaces. To prevent catalogue overload and to provide meaningful disciplinary description, the H2IOSC Library implements filtering rules directly on metadata fields populated with controlled vocabularies. Typical filters include: topics (from SSHOC-aligned topic vocabularies); learning resource types (tutorial, lecture, notebook, demonstrator, course, etc.); disciplines/communities (SSHOC disciplinary terms, with mappings to RI taxonomies); intended audience / pedagogical role (trainer-oriented vs learner-oriented objects).

Because these values come from shared vocabularies, a harvester can reliably interpret them and apply consistent inclusion/exclusion rules across providers. This turns selective harvesting into a transparent, machine-actionable curation step rather than an ad-hoc human decision. As far as the H2IOSC Training library is concerned, the harvested metadata is designed to flow into two discovery layers with different but compatible purposes: The H2IOSC Marketplace serves as the national SSH discovery front end. The Library's schema and vocabularies are aligned with the Marketplace model, so learning objects can be surfaced there without loss of meaning. This enables unified search across H2IOSC assets (training materials, tools, datasets, events) while still keeping resource types clearly separated and facetable.

For training on language resources, alignment with CLARIN is essential. By keeping H2IOSC learning-resource metadata compatible with CLARIN's educational descriptors and vocabulary practices, H2IOSC materials can be cross-referenced or selectively harvested into the CLARIN Learning Hub or related catalogues. This supports a two-way enrichment loop: H2IOSC gains visibility in a pan-European RI training space, while CLARIN users gain access to Italian SSH training objects that match their disciplinary filters. Within the H2IOSC ecosystem, selective harvesting and discovery are supported not only by metadata exchange mechanisms but also by a shared conceptual alignment between catalogues. This alignment is realised in practice through the data model of the H2IOSC Marketplace, which acts as the technical implementation of the Common Semantic Framework adopted in the project. The Marketplace data model translates the shared conceptual layer into a pragmatic structure for cataloguing and orchestration, enabling resources described in different systems to be aggregated and explored within a unified discovery environment while preserving their original metadata representations. In this

way, semantic alignment is not confined to an abstract modelling layer, but informs concrete discovery and orchestration services within H2IOSC.

6.3. Interoperability challenges

Even with shared standards, interoperability remains hard in practice because learning resources live in a highly heterogeneous repository landscape. SSH learning objects may be stored in institutional repositories, RI catalogues, project websites, GitHub/GitLab, Zenodo communities, or learning platforms such as Moodle. Each environment exposes metadata differently (or not at all), uses different protocols, and applies local interpretations of fields like "resource type" or "audience." This heterogeneity means that harvesting pipelines must cope with uneven metadata quality, incompatible schemas, and varying support for persistent identifiers, licenses, or versioning.

A second challenge is metadata granularity. Some providers describe training at a coarse level (a whole course or workshop as one record), while others provide fine-grained records for every learning object (slides, exercises, notebooks, datasets-in-teaching). When these are harvested into a common catalogue, the mismatch can produce duplication, gaps, or misleading search results. For example, a user may retrieve many small components without the context of the parent course, or conversely, only see a course record without visibility of reusable sub-items. Addressing this requires explicit modelling of part–whole relations (learning object belongs to module/course), inheritance-aware schemas like the H2IOSC hierarchy, and selective harvesting rules that decide which granularity levels are most useful to expose in each discovery layer.

7. Discussion: Advantages of an Ontology-Mediated Approach

Building on the interoperability challenges outlined above, this discussion reflects on the ontology-mediated approach developed within H2IOSC for the integration of CLARIN resources and considers its advantages as a complementary semantic layer for aligning heterogeneous metadata schemas in federated discovery and training environments.

The H2IOSC ecosystem for training materials is grounded in shared metadata profiles, controlled vocabularies, and standard harvesting mechanisms that ensure interoperability across platforms and research infrastructures. Within such a federated setting, however, interoperability is not only a matter of shared fields or exchange formats, but also of consistent interpretation of concepts across infrastructures that originate from different disciplinary traditions and cataloguing practices. The ontology-mediated approach adopted in H2IOSC addresses this issue by introducing a shared conceptual reference model that enables semantic alignment without replacing existing metadata schemas.

A first advantage of the ontology-based approach concerns the alignment of heterogeneous metadata schemas through a shared conceptual reference model. In the CLARIN–H2IOSC integration, CMDI metadata—anchored in the CLARIN Concept Registry (CCR)[11]—were mapped to an ontology-based framework grounded in CIDOC CRM (ISO 21127) and its SSH-specific extension SSHOCro.[12] [13]. This mapping strategy was explicitly designed to be non-invasive: CMDI profiles remain fully operational within CLARIN, while the ontology provides an additional semantic layer that enables cross-infrastructure interpretation and alignment [14].

Read in continuity with the metadata-centred approach adopted for FAIR learning resources in H2IOSC, this experience illustrates how different descriptive traditions—such as CMDI, Dublin Core–derived profiles, and discipline-specific schemas—can coexist within a shared semantic framework. Rather than enforcing a single metadata schema, ontology-mediated mapping supports interoperability by making conceptual correspondences explicit, while preserving the autonomy of local cataloguing practices. A second advantage lies in the support for cross-domain querying and unified discovery. In the CLARIN case, datasets and tools were aligned within the same semantic space, even though they are exposed through different catalogues and services, such as the Virtual Language Observatory, the Resource Families classification, and the Switchboard. These catalogues offer different, complementary perspectives on the same ecosystem: some focus on resource typology, others on functional

compatibility or user-oriented discovery paths. Ontology mediation makes it possible to interpret these perspectives coherently by grounding them in a shared conceptual layer, ensuring that comparable concepts retain a consistent meaning across discovery contexts.

This perspective is fully consistent with the layered discovery approach described earlier in this paper, where metadata profiles and controlled vocabularies enable selective harvesting and faceted search across infrastructures. Ontology mediation does not replace catalogue-level discovery mechanisms, but rather contributes to their semantic consistency, ensuring that comparable concepts are interpreted in a uniform way across federated environments such as H2IOSC. Importantly, the Common Semantic Framework adopted in H2IOSC is not confined to an abstract modelling layer. As discussed in [14], the data model of the H2IOSC Marketplace represents a pragmatic technical implementation of this shared conceptual framework. The Marketplace translates the ontology-based alignment into an operational structure for cataloguing, discovery, and orchestration, enabling resources described in different systems to be aggregated and explored within a unified environment while preserving their original metadata representations. In this sense, semantic mediation informs concrete infrastructure services, bridging conceptual modelling and technical interoperability.

Finally, ontology-mediated interoperability contributes to semantic consistency and long-term sustainability. Metadata schemas and cataloguing practices naturally evolve in response to new requirements, standards, and community needs. By anchoring metadata descriptions to a stable, well-documented reference ontology—such as CIDOC CRM and SSHOCro—it is possible to preserve the meaning of descriptions even as local schemas are adapted or extended[12] [13].

The CLARIN–H2IOSC experience shows that an ontology-first strategy can support long-term interoperability without constraining existing workflows or descriptive practices[14]. In this sense, ontology mediation can be understood as a stabilising layer that complements FAIR-by-Design metadata approaches and controlled vocabularies, helping federated infrastructures maintain coherence and interpretability over time.

8. Conclusion and Future Work

This article has argued that making SSH learning resources FAIR requires treating them as first-class research objects, supported by disciplinary infrastructures and by interoperable metadata practices. We presented a conceptual distinction between reusable learning objects for educators and learner-facing courses, and showed how this distinction translates into concrete metadata requirements (resource type, intended audience, pedagogical role). We also discussed the role of controlled vocabularies in enabling semantic interoperability across SSH infrastructures, and the need for selective harvesting to avoid catalogue overload while still supporting cross-platform discovery. Finally, we outlined a technical approach based on OAI-PMH/API harvesting, vocabulary-driven filters, and alignment between national and European training discovery layers.

Future work will focus on two complementary directions. First, we will strengthen integration with adjacent SSH and cultural heritage ecosystems, notably by connecting the H2IOSC Training Library and Marketplace to the emerging ECHOES–ECCCH training landscape and by improving interoperability pathways with DARIAH Campus. This will require sustained work on vocabulary mappings, multilingual term alignment, and shared harvesting profiles so that domain-specific views can coexist with EOSC-wide discovery without duplication or loss of meaning. Second, we will deepen the linking of training resources to the wider research ecosystem by systematically connecting learning objects and courses to the services and datasets they teach. This includes persistent, machine-actionable links from training records to tools, workflows, corpora, collections, and data publications, enabling users to move seamlessly from “how to” materials to executable environments and reusable data. Strengthening these links will improve both reuse (trainers can rapidly assemble coherent programmes) and impact (learners can directly apply skills using EOSC/RI services and datasets), completing a FAIR-by-design lifecycle from infrastructure to training to research practice.

Acknowledgements

H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union – NextGenerationEU – NRRP M4C2 - Project code IR0000029 - CUP B63C22000730005. We acknowledge the collaboration of CLARIN-IT, DARIAH-IT, OPERAS-IT, and E-RIHS.it.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Monachini, F. Frontini, Infrastrutture digitali per le scienze umane e sociali, in: Digital Humanities, Carocci, Roma, 2023, p. 256 : 263. URL: https://www.carocci.it/prodotto/digital-humanities.

[2] E. Degl'Innocenti, M. Monachini, A. Bucciero, E. Pasini, B. Fanini, F. Frontini, H2IOSC: Humanities and Heritage Open Science Cloud, in: E. Carbé, G. L. Piccolo, A. Valenti, F. Stella (Eds.), La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD, AIUCD, Siena, Italy, 2023, pp. 63–64. URL: https://iris.unive.it/retrieve/0f226d38-e332-418b-9b14-d5558d1a0d9d/AIUCD2023.pdf.

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. d. S. Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. oen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. v. Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. v. d. Lei, E. v. Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. URL: http://www.nature.com/articles/sdata201618. doi:10.1038/sdata.2016.18.

[4] K. Avanço, Framework for the development of open science and scholarly communication training (OPERAS-PLUS Deliverable 6.3), ND (2025). URL: https://zenodo.org/records/16970400. doi:10.5281/zenodo.16970400, publisher: Zenodo.

[5] L. Eric, A. C. Vallo, L. Eric, S. Venkataraman, P. Principe, D9.2 / EOSC Knowledge Hub (Training Catalogue and Platform) (2022). URL: https://eoscfuture.eu/wp-content/uploads/2023/10/EOSC-Future-WP9-LIFEWATCH-D9.2-EOSC-Knowledge-Hub-2022-09-30.pdf.

[6] R. Braukmann, E. Leenarts, S. Saldner, V. Keck, J. Wehmeyer, A. Albuerne, K. Illmayer, M. Durco, V. Kalaitzi, T. Yankelevich, D. Fiser, A. Cvek, A. Vlad, D6.11 SSHOC Trainer Toolkit (final), ND (2022). doi:10.5281/zenodo.6564283, publisher: Zenodo.

[7] S. Filiposka, D. Green, A. Mishev, V. Kjorveziroski, A. Corleto, E. Napolitano, G. Paolini, S. Di Giorgio, J. Janik, L. Schirru, A. Gingold, C. Hadrossek, I. Souyioultzoglou, C. Leister, G. Pavone, S. Sharma, E. M. Mendez Rodriguez, E. Lazzeri, D2.2 Methodology for FAIR-by-Design Training Materials (2023). URL: https://zenodo.org/records/8305540. doi:10.5281/zenodo.8305540, publisher: Zenodo.

[8] I. van der Lek, F. Frontini, D. Fiser, A. König, Making the CLARIN Training Materials FAIR-By-Design, 2025. URL: https://zenodo.org/records/14699078. doi:10.5281/zenodo.14699078.

[9] G. Pedonese, F. Frontini, L. Francalanci, Designing a Repository for FAIR Learning Objects: H2IOSC Training Library, in: Proceedings of the CLARIN Annual Conference 2025, 2025. doi:https://cris.brighton.ac.uk/ws/portalfiles/portal/60194662/CLARIN2025_ConferenceProceedings.pdf#page=141.

[10] P. Sichera, M. Monachini, V. Quochi, N. Giampietro, V. Fabiani, R. Ottaviani, R. B. Luzietti, Synergies between CLARIN-IT and OPERAS-IT within H2IOSC: Monitoring Communities and Orchestrating Digital Services, in: CLARIN Annual Conference Proceedings 2025, 2025, pp. 26–30. doi:10.5281/zenodo.17357825.
[11] I. Schuurman, M. Windhouwer, CLARIN concept registry: The new semantic registry (2015).
[12] C. Bekiari, A. Kritsotaki, E. Tsouloucha, M. Theodoridou, D4.20 SSHOCro (final version) (2022-04-27). URL: https://zenodo.org/records/6771757, publisher: Zenodo.
[13] C. Bekiari, G. Bruseker, E. Canning, M. Doerr, P. Michon, C.-E. Ore, S. Stead, A. Velios, Cidoc conceptual reference model, version 7.1.3, 2024. URL: https://cidoc-crm.org/Version/version-7.1.3, produced by the CIDOC CRM Special Interest Group. Volume A: Definition of the CIDOC Conceptual Reference Model.
[14] D. Melaccio, F. Boschetti, M. Monachini, Interfacing CLARIN with H2IOSC: Metadata Interoperability through Ontology-based Mediation, in: Proceedings of the CLARIN Annual Conference 2025, 2025. doi:10.5281/zenodo.17357825.

A. Supporting Figures

This annex provides supplementary figures supporting the description of the H2IOSC Training Library architecture and metadata model.

A.1. Training Library hierarchy

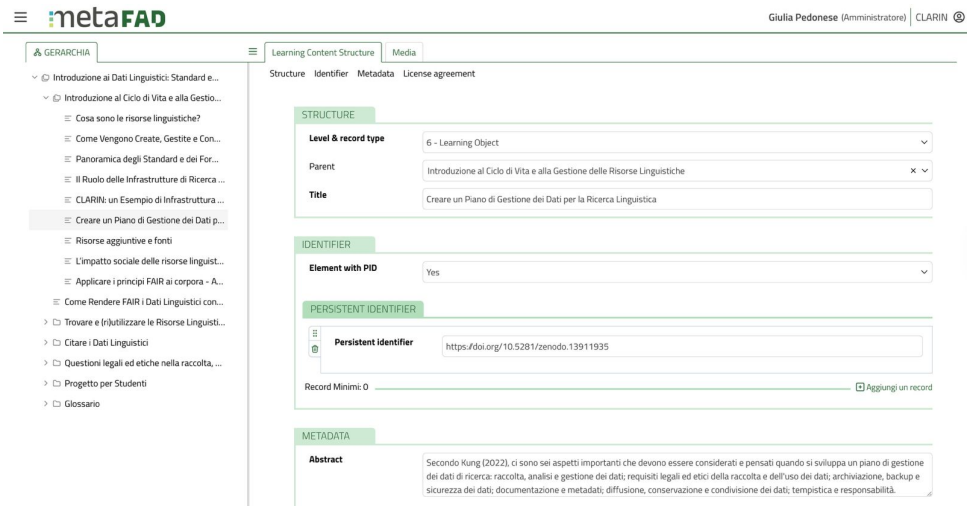

Figure 1: Structure of the Training Materials hierarchy in the H2IOSC Training Library.

A.2. Metadata grid adopted in the deposit workflow

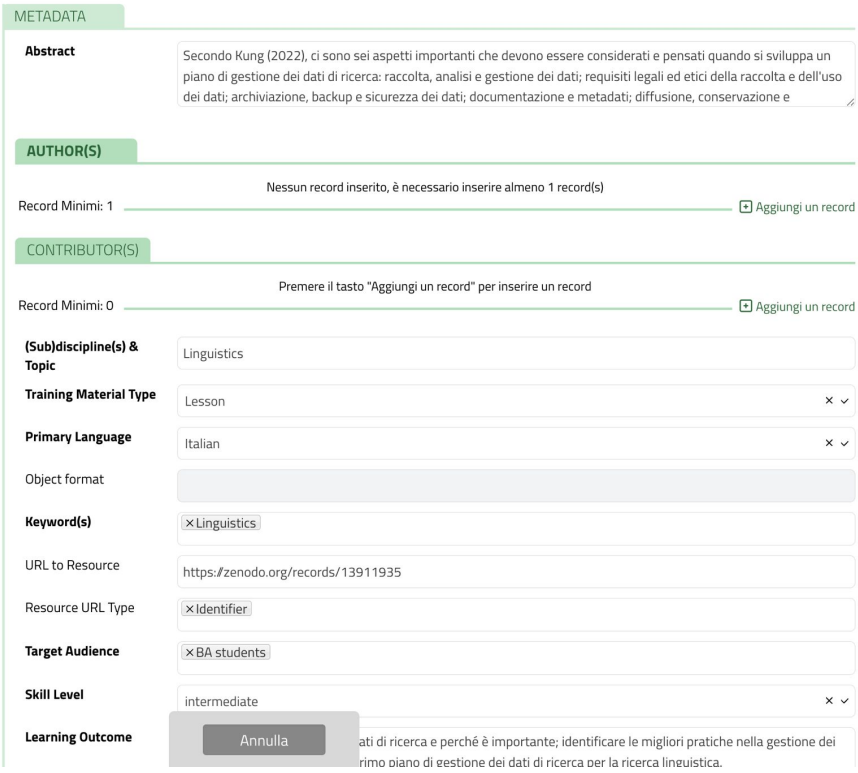

Figure 2: Example of the metadata grid adopted in the Training Library deposit workflow.