

UPSKILLS Two Years on: Teaching about Language Resources and FAIR Data Principles with CLARIN

Iulianna van der Lek
CLARIN ERIC
Utrecht, The Netherlands
i.vanderlek@uu.nl

Maja Miličević Petrović
Department of Interpreting and Translation
University of Bologna
Forlì, Italy
maja.milicevic2@unibo.it

Silvia Bernardini
Department of Interpreting and Translation
University of Bologna
Forlì, Italy
silvia.bernardini@unibo.it

Adriano Ferraresi
Department of Interpreting and Translation
University of Bologna
Forlì, Italy
adriano.ferraresi@unibo.it

Olga Arsić
Department of Interpreting and Translation
University of Bologna
Forlì, Italy
olga.arsic2@unibo.it

Abstract

The ongoing technological developments have a major impact on language-related research and professional activities. This highlights an increasing demand for new skills and profiles, as well as a greater focus on making language data and resources open and FAIR. In this paper, we present long-term efforts that address these issues, starting from the Erasmus+ project UPSKILLS (2020–2023). We specifically focus on follow-up activities at the University of Bologna in collaboration with CLARIN ERIC, proposing a conceptual framework that combines progressive competence development across all study levels with infrastructure-based pedagogy to support discipline-specific Open Science education. Using CLARIN as a learning environment, lecturers can modernize curricula, improve students' FAIR data literacy, and improve graduate employability.

1 Introduction

The rapid advances in language technologies require adaptation not only on the part of individual researchers and language professionals, but also through initiatives of higher education institutions offering language- and linguistics-related degree courses. One of the recent projects with related objectives was the Erasmus+ strategic partnerships UPSKILLS: *UPgrading the SKills of Linguistics and Language Students*.¹ This three-year project (2020–2023) sought to identify and address the gaps and mismatches in the skills of linguistics and language students with respect to the demands of the current job market and the broader research landscape. The project's results received the Erasmus+ good practice label² and were ranked as the second runner-up in the 2023 competition for the best Digital Humanities training materials.³ Two years later, we revisit the results by focusing specifically on the impact the project has had on one of the consortium partners, the Department of Interpreting and Translation (DIT) of the University of Bologna (UNIBO). We expand on the original UPSKILLS project results by exploring in particular how the CLARIN research infrastructure can be integrated into teaching practices to raise awareness among students about the basic principles of FAIR research data management, as well as the vast body of knowledge, expertise, data and tools available in open access.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹<https://upskillsproject.eu>

²<https://erasmus-plus.ec.europa.eu/projects/search/details/2020-1-MT01-KA203-074246>

³<http://dhawards.org/dhawards2023/results/>

The paper is structured as follows. Section 2 presents the UPSKILLS project, outlining its main results, the spin-off materials hosted at UNIBO, as well as the changes that DIT implemented in its study programmes. Section 3 describes the CLARIN-UNIBO collaboration and the topics covered in the CLARIN guest lectures at DIT. Section 4 presents the conceptual framework and a matrix of proposed learning outcomes and potential activities using CLARIN's research infrastructures, which could be embedded in any language-related research programme. Section 5 concludes with reflections on how to ensure the long-term impact of the project and a call to the CLARIN community to help improve the proposed framework.

2 From UPSKILLS to UpskillIng@UNIBO

2.1 A summary of UPSKILLS outputs

In the UPSKILLS project, researchers from eight partner institutions across Europe worked toward developing a new curriculum component and a new teaching approach targeting the skills needed for industry and academic research among (primarily) undergraduate students of language-related subjects. The deliverables comprise: (1) a needs analysis, based on several components and summarized through a new target profile of *language data and project specialist*, skilled in the disciplinary domain, but also in technology, research and data and project management (Miličević Petrović et al., 2021), (2) a series of guidelines and best practices on research-based teaching and the use of research infrastructures (Simonović et al., 2023; van der Lek et al., 2023), (3) a set of Moodle-based materials for online and blended teaching and learning in the areas identified as in need of improvement (Gledić et al., 2023),⁴ and (4) educational games developed for, or adapted within the project (Arsić et al., 2023).⁵

UNIBO and CLARIN collaborated on the development of guidelines and teaching materials, which focused on language resources, namely the *Processing Text and Corpora* course (Bernardini et al., 2023)⁶, the course *Introduction to language data: Standards and repositories* (van der Lek & Fišer, 2023)⁷ and the *Guidelines for Integrating Research Infrastructures into Teaching* (van der Lek et al., 2023). While parts of the materials produced in the project, included those by UNIBO and CLARIN, were piloted at its end, during the UPSKILLS Summer School in August 2023,⁸ more extensive inclusion of the content into UNIBO curricula took place in the years that followed.

2.2 Reforming the UNIBO curriculum

Towards the end of the UPSKILLS project and upon its conclusion, the UNIBO team developed additional materials, partly adapted to meet the needs of the local target community, students of intercultural communication, translation and interpreting in Italy. The needs analysis was complemented by a survey that gives students an initial indication of the UPSKILLS subprofile they might wish to pursue, i.e. language data analyst, language data scientist, language data manager, language project manager, depending on their inclination towards data analysis versus organizational tasks. Furthermore, the original UPSKILLS corpus of job advertisements, a central component in the needs analysis, was accompanied by an updated 2023 edition (with a 2026 edition in preparation), providing more up-to-date information, especially on AI-related job positions.⁹ The resources converged in a website that helps students discover a suitable profile and leads them to relevant learning materials.¹⁰

At the same time, DIT started partially reforming its degree programmes in line with the UPSKILLS findings. At the MA level, the degree course in Specialized Translation gained a new curriculum in Trans-

⁴All courses can be viewed directly on Moodle (upon free registration) at <https://upskills.fil.bg.ac.rs>, and they can be downloaded from the CLARIN.SI repository at <http://hdl.handle.net/11356/1865>. Descriptions of individual courses, with access and download links, can be found at https://upskillsproject.eu/deliverables/io3/upskills_learning_materials/

⁵<https://upskillsproject.eu/deliverables/io4/>

⁶https://upskillsproject.eu/project/text_processing/

⁷<https://upskillsproject.eu/deliverables/io2/research-infrastructures/>, available in HTML format at <https://www.clarin.eu/content/introduction-language-data-standards-and-repositories-0>; see also the Italian translation and adaptation by van der Lek et al., 2024 at <https://h2iosc-training-platform.ilc4clarin.ilc.cnr.it>.

⁸<https://upskillsproject.eu/events/summer-school/>

⁹<https://bellatrix.sslmit.unibo.it/noske/upskills/#open>

¹⁰<https://upskilling.dipintra.it>

lation and Technology, heavily focused on research skills, and incorporating courses such as *Profession-Based Research*, *Machine Translation*, *Natural Language Processing*, and *Language Data Analysis*. Later on, a module called *Language, Technology, Research* was introduced as an option in the BA course in *Languages and Technologies for Intercultural Communication*, with courses in *Computational Thinking* (I year), *Text Processing* (II year) and *Researching Language* (III year). Finally, the principles of research-based teaching have been implemented, most notably in the MA course in *Corpus Linguistics*, jointly with an evaluation approach based on peer assessment.

In addition, a MOOC titled *Processing Texts and Corpora: An Introduction* (Bernardini et al., 2025) has recently been created and made available on the UNIBO Open Knowledge platform,¹¹ as part of the project *Edvance – Digital Education Hub per la Cultura Digitale Avanzata*.¹² This course revisits a number of topics addressed in its UPSKILLS predecessor *Processing Texts and Corpora*, elaborating on CLARIN-related topics. Its unit “Corpus sourcing” focuses on CLARIN ERIC: it explains how to locate corpora via CLARIN’s Virtual Language Observatory (VLO)¹³ and how to select tools via the CLARIN Switchboard,¹⁴ and helps learners assess suitability for reuse based on metadata, licensing, and documentation.¹⁵ The original UPSKILLS outputs and later additions are both shown in Figure 1.

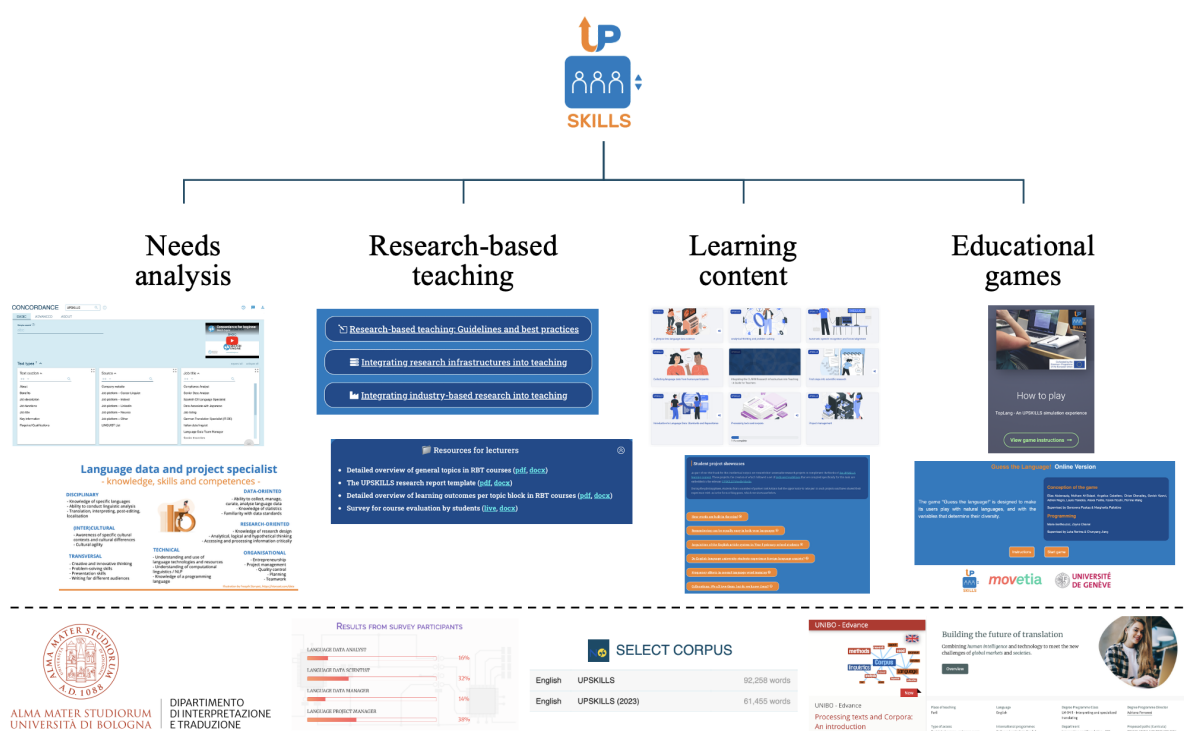


Figure 1: The UPSKILLS and additional UNIBO outputs.

3 Embedding FAIR competences through CLARIN

3.1 Focus on FAIRness

Open Science, defined as “an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone” (UNESCO, 2021, p. 7), includes as one of its cornerstones the **FAIR data principles** (*Findability, Accessibility, Interoperability, and Reusability*). As stated by Wilkinson et al. (2016), these foundational principles apply to data in the conventional sense, as well as the tools and workflows that led to that data.

¹¹<https://book.unibo.it/>

¹²<https://www.edvance.it/>

¹³<https://vlo.clarin.eu/?0>

¹⁴<https://switchboard.clarin.eu/>

¹⁵For another Italian experience, see Pedonese et al. (2025) on the use of UPSKILLS materials at the University of Ferrara.

Although Open Science and FAIR data principles are increasingly embedded in research policy and funding requirements, their integration into university curricula remains largely fragmented. Several initiatives¹⁶ acknowledged a lack of awareness among lecturers, students and researchers of research data management (RDM) practices, including knowledge of principles to make digital resources FAIR. Consequently, the *FAIR Competence Framework for Higher Education* (Demchenko et al., 2021) proposed a set of core competencies for FAIR data education that universities can use to integrate relevant content into their curricula. Students, teachers, and researchers from all disciplines are encouraged to acquire fundamental skills for Open Science, including the ability to effectively interact with **federated research infrastructures** and Open Science tools for collaborative research. Furthermore, research on RDM education consistently highlights the importance of embedding data skills within disciplinary teaching rather than offering generic, de-contextualized training (Cox et al., 2017). Effective approaches emphasize alignment with research data lifecycle, use of authentic data and tools, and progression from awareness to independent practice.

Since Open Science originated largely in the natural sciences, its application to the humanities has required adaptation to disciplinary epistemologies, data types and ethical concerns (Borgman, 2015). Data in the humanities are often heterogeneous, multilingual, and context-dependent, raising challenges for standardization and reuse (Edmond, 2020). In language-related disciplines specifically (Translation Studies, Corpus Linguistics, Language Technology, and others), researchers and students routinely work with complex language resources such as corpora, parallel texts, linguistically or otherwise annotated data, or terminology databases, where FAIR principles have proven highly relevant, and metadata quality, documentation, and interoperability directly affect reuse and reproducibility (Krauwer & Hinrichs, 2014). Yet, explicit training in research data management, licensing, metadata and data publication is often limited in this area.

Research infrastructures such as CLARIN are commonly framed as enablers of advanced research, but work in digital humanities and NLP pedagogy argues that infrastructures can also function as learning environments (De Smedt et al., 2024). Exposing learners to real-world data, standards, repositories, and infrastructures, supports experiential learning and fosters data literacy grounded in disciplinary practice. This paper proposes a progressive, infrastructure-based teaching framework for Open Science, FAIR data and research data management in language-related disciplines based on the materials CLARIN produced in the UPSKILLS project, namely the course *Introduction to Language Data: Standards and Repositories* (van der Lek & Fišer, 2023) and *Best Practice Guidelines for Integrating Research Infrastructures into Teaching* (van der Lek et al., 2023). The materials demonstrate how instructors can use the CLARIN open data repositories, language resources, and integrated NLP tools to promote the added value of Open Science and guide students in applying FAIR data principles in language research.¹⁷

3.2 Lectures on CLARIN resources at UNIBO

Throughout the academic year 2024/2025, CLARIN-related topics were introduced in the DIT curriculum both as in-class guest lectures and embedded in an e-learning course. In March 2025, a total of three 90-minute guest lectures were delivered in two courses within DIT's degree programmes at UNIBO, with the objective of familiarizing the students with the CLARIN research infrastructure and raising awareness about basic concepts of Open Science, FAIR data principles and research data management. After the guest lectures, the materials were adapted for the MOOC *Processing Texts and Corpora: An Introduction* (Bernardini et al., 2025). The content of all courses was based on the materials CLARIN developed in UPSKILLS (van der Lek & Fišer, 2023).

The undergraduate course in *Language, Technology, Research II: Text Processing* aimed to introduce BA-level students to basic text processing methods and teach them how to design and build compara-

¹⁶In addition to the UPSKILLS projects, especially the EOSC Working Group on Skills & Training (<https://eosc.eu/opportunity-area-exp/oa5-skills-training-rewards-recognition-upscaling/>) and their publication *Digital Skills for FAIR and Open Science* (<https://data.europa.eu/doi/10.2777/59065>).

¹⁷These materials were built on the handbook *How to be FAIR with your data* (Engelhardt et al., 2022), comprising ready-made lesson plans and guidelines for teachers, and *The Open Handbook of Linguistics Data Management* (Berez-Kroeker et al., 2022), a useful toolkit for linguists.

ble and parallel corpora. After being introduced to character encodings, file formats, and principles of corpus design and construction, the students had to build corpora of various kinds: scientific articles, tasting notes on beverages other than wine, and manuals of household appliances. Before the students started with data collection, they were introduced to data discovery through two sessions organized by a CLARIN guest lecturer who showed how to locate already digitized text collections available in public repositories, at the same time gradually introducing the students to basic concepts in linguistic data management and FAIR principles. The sessions were attended by 10 students with different language combinations, and three lecturers with a background in translation and interpreting studies. The topics described below were covered in two sessions organized at the undergraduate level, and one at MA level.

3.2.1 A gentle introduction to project and research data management

The opening session introduced shared best practices in project and research data management, with a particular focus on teamwork for collecting and building resources such as corpora. As students were working in pairs to collect texts from the web to build their corpora, some user-friendly cloud-based tools were shown for collaboration and task management, e.g. Trello.¹⁸ Additionally, the concept of data management plan (DMP)¹⁹ was introduced to show how it can be used to plan for each step in the project: data collection, documentation and metadata, ethics and legal compliance, storage and backup, data sharing and how to split responsibility and resources during the project.

This approach was adopted because, while students typically gain practical project management experience only during internships, university lectures can also introduce good project and (research) data management practices as early as the undergraduate level. This early academic exposure, which often includes joint work when collecting and building resources such as corpora, allows students to refine essential soft skills in planning and managing complex data workflows throughout their degree, workflows essential for the professional profile of the “language data and project specialist”.

3.2.2 Locating linguistic data in public repositories

UNIBO students also learned how to find research data, starting from their institutional repository, and using registries such as re3data (REgistry of REsearch Data REpositories) to locate domain-specific repositories used in linguistics, such as CLARIN, TROLLing (Tromsø Repository of Language and Linguistics), and OLAC (Open Open Language Archives Community).²⁰

Specifically, the CLARIN representative demonstrated how to discover language data in CLARIN’s Virtual Language Observatory and Resource Families. By browsing the VLO, a search portal containing over a million records, students learned that CLARIN helps researchers make their published language resources FAIR, e.g. the resources are described with consistent metadata (through CMDI – Component Metadata Infrastructure) and are assigned a persistent identifier (a Handle) ensuring that the resource can be reliably retrieved and cited in their work even if the physical location of the data changes. After locating a suitable digital text collection they could potentially download and use to build their own corpus, the students first needed to review the metadata to identify whether the resource allowed public, academic or restricted access. In the case of restricted access, students got to test the CLARIN federated login system, i.e. the single-sign on,²¹ which allowed them to log in to the repository hosting the respective resource with their institutional credentials.

3.2.3 FAIR assessment of corpora

The students learned about the FAIR data principles through a hands-on activity that involved evaluating the available corpora from a FAIR perspective. They were split into groups and asked to pick a corpus from a proposed list and assess its FAIRness based on a checklist inspired by Frey et al. (2019). The students went through each stage of FAIR, checking whether the corpus and its metadata were easily accessible for use, whether the corpus was assigned a unique and persistent identifier (PID), whether

¹⁸<https://trello.com>

¹⁹<https://www.openaire.eu/what-is-a-data-management-plan>

²⁰<https://www.re3data.org>, <https://site.uit.no/trolling/>, <http://www.language-archives.org/>

²¹<https://www.clarin.eu/content/identity-provider-technical-details>

it was developed using common standards and formats, was properly documented, and was assigned a clear and accessible license usage (e.g., a Creative Commons License).

3.2.4 Corpora depositing and sharing

To make students aware of the depositing workflow, a brief demonstration of the Italian ILC4CLARIN repository²² was given. The repository is hosted at the Institute for Computational Linguistics of the National Research Council in Pisa. Scholars in Italy can use the repository to search for linguistic data and tools, as well as deposit their data to safely store them, ensuring that it is also findable and properly credited if reused by other scholars in research or teaching. To be able to deposit data, the data owner is prompted to log in using their institutional credentials. Once logged in, the depositing workflow is relatively straightforward. For example, corpus data needs to be provided with metadata in standard formats recommended by the CLARIN community²³. The depositor needs to fill in all the required metadata, after which the submission is curated by the repository manager. Once published, the corpus metadata is freely accessible not only via the Italian repository but also via the Virtual Language Observatory.

Although a formal evaluation of the activities was not conducted, the interactions during the sessions showed that the students appreciated being introduced to general and linguistic-specific repositories, which they would otherwise have not been able to discover through simple Google searches.

3.2.5 Emerging career paths

In the master's course in *Profession-Based Research*, part of DIT's mentioned Translation and Technology curriculum, a CLARIN guest lecture was delivered to ten students and two lecturers. This session introduced students with backgrounds in modern languages and translation to the CLARIN research infrastructure, highlighting potential career paths, such as data stewards, communication specialists, and training officers supporting researchers doing language research.

It was highlighted that graduates of language and linguistic programmes with strong language data management skills are well suited to take on data stewardship roles across all sectors where their disciplinary expertise is required to enhance the findability and reusability of language data through high-quality metadata, and manage the legal and ethical constraints governing access and reuse.

A summary of activities promoted by CLARIN and UNIBO is provided in Figure 2; the figure also shows some of the key elements of the next section, which introduces recommendations that extend beyond any single university or degree course.

4 Learning Objectives and Curriculum Design

4.1 Experience-based recommendations on how to teach about language resources

The CLARIN teaching experience at UNIBO confirmed the need for a more systematic approach for integrating CLARIN into university curricula. The authors of the present paper propose a conceptual framework for doing so based on three interrelated pedagogical principles:

1. **Situated Open Science learning**, which implies that Open Science and FAIR data principles are embedded in concrete research and data practices as early as possible. Students at all levels should work with real corpora and public data repositories, making openness and reusability tangible.
2. **Progressive competence development across educational levels**, which aligns with competence-based approaches to Open Science education (European Commission, 2017; Whyte et al., 2024); specific progressive learning objectives are summarized in Table 1 below, based on a learning trajectory that gradually moves from recognition to original contribution.
3. **Infrastructure-based pedagogy**, which supports end-to-end learning across the research data life-cycle, based on CLARIN's ecosystem of services for language data discovery, standards, tools and repositories.

²²<https://ilc4clarin.ilc.cnr.it/en/>

²³<https://standards.clarin.eu/sis/views/view-centre.xq?id=ILC4CLARIN>



Figure 2: Joint CLARIN-UNIBO activities.

In other words, our approach puts students at the centre, with hands-on activities based on field-specific data and tools being given more importance than frontal teaching and generic examples, and with FAIR-related topics being presented jointly with other topics related to text processing.

Using also the learning objectives designed for the UPSKILLS course *Introduction to Language Data: Standards and Repositories* as a topical basis for the conceptual framework, Table 1 lists the main topics related to infrastructure and shows the progressive development of competence using level-specific keywords. A more elaborate version of the table is available online as a Google sheet,²⁴ i.e. in the form of a matrix, and it has been discussed with fellow researchers and teachers during the CLARIN Annual Conference in Vienna, 2025. The subsections below elaborate on these learning objectives and translate them into more specific recommendations that can be implemented in curriculum design.

Topic	Learning objectives		
	BA	MA	PhD
Open Science and FAIR Data Principles	Recognize	Use	Apply
Research Infrastructures and Data Repositories			
Finding and (Re)Using Language Resources	Define	Assess	Solve
Analyzing Language Resources			
Building and Managing Language Resources	Explain	Analyze	Contribute
Sharing and Archiving Language Resources			

Table 1: A summary of language resource topics and learning objectives.

4.2 Bachelor's level: Awareness, discovery and responsible use

At the undergraduate level, teaching should raise awareness and build foundational literacy in Open Science, FAIR data, and language resources, with a strong emphasis on discoverability, reuse and basic standards rather than data creation.

²⁴<https://tinyurl.com/yyhckrtj>

Recommended approach for teachers

- Introduce Open Science concepts early, focusing on Open Access, Open Data, FAIR and CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics) principles, and why they matter in language-related research.
- Use CLARIN as an entry point to research infrastructures, helping students understand what a research infrastructure is and how it supports research communities.
- Emphasize finding and reusing existing language resources rather than producing new ones.

Key topics and learning outcomes

Students should be able to:

- Explain Open Science, Open Access, Open Data and FAIR and CARE principles.
- Understand what research data and language resources are, and distinguish between major resource types (written corpora, speech corpora, multimodal corpora, etc.).
- Discover language resources using CLARIN tools, such as the Virtual Language Observatory and Content Search.
- Recognize and compare licence conditions and basic legal/ethical constraints for reuse.
- Gain introductory exposure to annotation methods, corpus analysis tools, and common standards from written and speech corpora.
- Understand the benefits and challenges of sharing and archiving data, and the role of CLARIN centres and repositories.

Activities

- Hands-on discovery exercises in the VLO, Content Search and Resource Families of corpora and short reflective assignments on FAIRness and licensing

These suggestions are in line with the *Minimum Viable Skillsets* framework developed in the Skills4EOSC project²⁵ in the domain of Open Science and FAIR data principles and practices (White and Green, 2022, see also Whyte et al., 2023), according to which students at this level need to have foundational digital research skills, understand why it is important for society that research data is open and FAIR, and have good organization and documentation skills. The notion of a research infrastructure can also be introduced at this level and the CLARIN infrastructure can be used to show examples of published digitized text collections that are available and can be explored with free tools and services. By browsing the resources in the VLO, students can learn that published data and tools are described with consistent metadata, carry licences with different possibilities of reuse (public, academic, restricted), and can be located and cited using a persistent identifier (e.g., Handle or DOI). Teachers can also use the VLO to show students how to search for text collections in plain text format, download them and upload them to their preferred corpus-building tool. For instance, the VLO provides direct access to Voyant Tools²⁶ via the CLARIN Switchboard. As a free, open-source, and web-based platform, Voyant Tools can be introduced at the undergraduate level to gently introduce students with a humanities background to basic quantitative text analysis and visualization. For more advanced text analysis, the CLARIN resource families of corpora²⁷ can be used to show students how they can explore different types of corpora through the integrated concordancers (NoSketch Engine, Korp, or KonText). Finally, students can be introduced to the benefits and challenges of sharing and archiving language resources, as well as to the basic taxonomy of repositories (generic, discipline-specific, institutional) and to the idea that “responsible openness” may require additional safeguards when working with sensitive data, in line with CARE principles.

²⁵<https://www.skills4eosc.eu/>

²⁶<https://voyant-tools.org/>

²⁷<https://www.clarin.eu/resource-families>

4.3 Master's Level: Applying FAIR principles and managing data

At the MA level, teaching should shift towards active data management, evaluation, and informed use, with students learning to apply FAIR principles in practice and work more independently using infrastructure services and available tools for language data processing and analysis.

Recommended approach for teachers

- Frame CLARIN as a practical research environment, not just a catalogue.
- Integrate FAIR assessment, standards and workflows into research methods.
- Encourage students to critically evaluate data quality, repositories and tools.
- Direct students to CLARIN Knowledge Centres (K-centres) for specialized expertise on language-specific corpora, methods and tools.

Key topics and learning outcomes

Students should be able to:

- Identify data sources on the web relevant for collecting open data for research projects in language/linguistics (e.g. repositories of research data, or data from governments and industry).
- Assess how FAIR existing language resources are, including repository practices.
- Distinguish between research data, metadata and documentation, and identify relevant data sources.
- Evaluate the quality and suitability of open data for specific research questions.
- Identify legal and ethical issues in language data collection and reuse.
- Understand the research data lifecycle, including planning, documentation, storage and sharing.
- Apply recommended community standards, file formats, and best practices for building and managing language resources.
- Use tools for language resource management, annotation, querying and basic programming-based analysis.
- Understand data publication, citation, and reuse, and how CLARIN repositories implement FAIR principles.

Activities

- Start by designing simple assignments that follow a full research cycle: discover data in the VLO, analyse data with Switchboard tools, export and interpret results, document and share outputs via a general-purpose repository, such as Zenodo. Such assignments will teach students to think in terms of “workflows”.
- Embed CLARIN resources into project-based assignments, e.g. write a simple data management plan before the start of the project, prepare a small dataset for potential deposit in a linguistic repository hosted by a CLARIN national consortium.

In line with the *Minimum Viable Skillsets*, at the MA level, students are expected to be able to integrate Open Science skills in their project workflows and assess FAIR concepts in their respective field of study, understand the data life cycle when conducting research. A useful starting point is to clarify foundational terminology by distinguishing *research data*, *datasets* and *metadata*, and by discussing how metadata quality affects findability and reuse. The research infrastructure can be used to encourage students to explore available resources and services, starting from a research question, and to apply the FAIR guiding principles to assess whether the published resources are suitable for reuse. For example, the CLARIN resource families of corpora can be used for interdisciplinary and comparative research; the Switchboard can be used to incorporate natural language processing (NLP) tools in the classroom and teach students basic NLP methods, such as text segmentation, tokenization, part-of-speech tagging,

syntactic parsing, and named entity recognition. In parallel, students can practice locating and accessing resources via persistent identifiers assigned by repositories, and become familiar with common standards and file formats used in CLARIN-oriented workflows. To strengthen good scholarly practice, they can be taught established guidelines for citing research data, as well as practical routines for sharing, reusing and citing language resources. After getting acquainted with basic NLP methods, students can be taught more advanced digital skills, such as programming in Python and R.

4.4 Doctoral level: Advanced research data management, publishing and stewardship

At PhD level, teaching should support researchers as data producers and stewards, focusing on more advanced research data management, legal/ethical responsibility, and data publication within their institutional repository or a domain-specific repository, e.g. a CLARIN national centre.

Recommended approach

- Treat Open Science and FAIR data as integral to doctoral research quality.
- Align training with the RDM requirements of funders and institutions.
- Use CLARIN to demonstrate end-to-end data stewardship, from planning to publication.

Key topics and learning outcomes

Doctoral candidates should be able to:

- Design and implement a research data management plan (DMP) for language-centric research.
- Identify high-level workflows in the creation and processing of language resources.
- Critically evaluate NLP and computational tools in terms of reproducibility and reusability.
- Address complex legal and ethical issues, including handling sensitive personal data and obtaining consent.
- Prepare language resources for deposit in a FAIR-compliant repository.
- Deposit, publish and cite datasets using CLARIN repositories and persistent identifiers.
- Create and share virtual collections to support transparency and reproducibility.

Activities

- Integrate CLARIN training into doctoral seminars and supervision, linking dataset publication directly to dissertation outputs and Open Science practices.

Students at this level are usually required to follow general Open Science and research data management courses, and are expected to be able to apply OS practices and FAIR data principles in their research (e.g., research data storage and archiving, handling legal and ethical issues surrounding the research data, selecting appropriate licences to protect potentially sensitive data). The infrastructure can be used to teach students how to find a suitable CLARIN repository to deposit language data and resources they may have created in their PhD project (e.g., corpora, audio and video recordings, annotations, grammars), and make them available to the community in a reliable manner, following established good practices in research ethics. If resources contain sensitive data, they can be protected by an institutional login, data encryption, and specific terms and conditions for data protection. Finally, doctoral students can take responsibility for preparing resources for depositing and archiving (packaging, metadata completeness, documentation, format choices), addressing legal and ethical issues in data sharing (e.g., personal data and consent, access restrictions), and depositing and publishing resources in a repository under conditions that balance openness with the protection of sensitive data.

5 Concluding remarks

Drawing on the UPSKILLS project and subsequent pedagogical activities at the University of Bologna, this paper proposed an integrated approach that combines progressive competence development with infrastructure-based pedagogy to foster sustainable and discipline-specific Open Science education. The conceptual framework described in Table 1 can help curriculum designers transition from abstract policy-driven instruction to situated learning using authentic language data from the CLARIN repositories and professional workflows. This framework refines the learning objectives of the *Introduction to Language Data: Standards and Repositories* course, employing a graduated matrix that moves from foundational recognition of FAIR principles at the Bachelor's level to integrated assessment at the Master's level and full professional application and data stewardship at the Doctoral level. Such a model directly addresses the data-related skills gap, supporting a smoother transition from student to researcher, and at the same time preparing students for emerging professional profiles, such as the *language data and project specialist* and *data steward*.

Although the proposed framework represents a significant step towards infrastructure-based pedagogy, it is not without limitations. A primary objective for our future research is to conduct a robust empirical evaluation of the framework's effectiveness in diverse classroom settings to validate its impact on student learning. To address this, we intend to refine our proposal through collaborative engagement with the broader academic community, inviting lecturers and trainers within the CLARIN network to contribute their expertise and feedback on the matrix of learning objectives. Another path worth exploring is the systematic deployment of the interactive research tracking tool, developed by consortium partners at the University of Zurich, which could enable lecturers to monitor student progress and provide targeted feedback during complex data-driven projects at all study levels. By integrating these practical tracking mechanisms with the resources available in the CLARIN Learning Hub, we aim to ensure the framework remains aligned with the evolving needs of the *language data and project specialist* profile and the broader European Open Science landscape.

References

- Aršić, O., Assimakopoulos, S., Bernardini, S., Ferraresi, A., & Vella, M. (2023). *Enhancing learning with branching scenario simulations: Guidelines to using H5P content blocks* (tech. rep.). Zenodo. <https://doi.org/10.5281/zenodo.8353727>
- Berez-Kroeker, A., McDonnell, B., Koller, E., & Collister, L. (Eds.). (2022). *The Open Handbook of Linguistic Data Management*. The MIT Press. <https://doi.org/10.7551/mitpress/12200.001.0001>
- Bernardini, S., Ferraresi, A., & Tedesco, N. (2023). Processing texts and corpora. https://upskillsproject.eu/project/text_processing/
- Bernardini, S., Polizzi, D., & Tedesco, N. (2025). Processing texts and corpora: An introduction [UniBOOK – UniBO Open Knowledge platform]. https://apps-book.unibo.it/learning/course/course-v1:UNIBO+DEH07+aa_24-25/home
- Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press. <http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=3339930>
- Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, 68(9), 2182–2200. <https://doi.org/10.1002/asi.23781>
- De Smedt, K., van der Lek, I., van den Heuvel, H., Balvet, A., Janssen, M., Cinková, S., Sanz, A., Assimakopoulos, S., & ten Bosch, L. (2024). CLARIN in training and education. In K. Lindén, T. Kontino, & J. Niemi (Eds.), *Selected papers from the clarin annual conference 2023*. <https://doi.org/10.3384/ecp210011>
- Demchenko, Y., Stoy, L., Engelhardt, C., & Gaillard, V. (2021). *D7.3 FAIR competence framework for higher education (Data stewardship professional competence framework)* (tech. rep.). Zenodo. <https://doi.org/10.5281/zenodo.5361917>

- Edmond, J. (Ed.). (2020). *Digital Technology and the Practices of Humanities Research*. Open Book Publishers. <https://books.openedition.org/obp/11899>
- Engelhardt, C., Barthauer, R., Biernacka, K., Coffey, A., Cornet, R., Danciu, A., Demchenko, Y., Downes, S., Erdmann, C., Garbuglia, F., Germer, K., Helbig, K., Hellström, M., Hettne, K., Hibbert, D., Jetten, M., Karimova, Y., Hansen, K. K., Kuusniemi, M. E., ... Zhou, B. (2022). *How to be FAIR with your data: A teaching and training handbook for higher education institutions*. Universitätsverlag Göttingen. <https://doi.org/10.17875/gup2022-1915>
- European Commission. (2017). *Providing researchers with the skills and competencies they need to practise Open Science*. Publications Office of the European Union. <https://data.europa.eu/doi/10.2777/121253>
- Frey, J.-C., König, A., & Stemle, E. W. (2019). How FAIR are CMC corpora? *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)*, 25–30. <https://cmccorpora19.sciencesconf.org/resource/page/id/15>
- Gledić, J., Đukanović, M., Budimirović, J., Soldatić, N., Miličević Petrović, M., Bernardini, S., Ferraresi, A., Tedesco, N., van der Lek, I., Fišer, D., Puskas, G., Pallottino, M., Berthouzoz, M., Kraš, T., Podboj, M., Simonović, M., Samardžić, T., van der Plas, L., Tanti, M., ... Vella, M. (2023). Upskills teaching and learning content [Slovenian language resource repository CLARIN.SI]. <http://hdl.handle.net/11356/1865>
- Krauwer, S., & Hinrichs, E. (2014). The CLARIN research infrastructure: Resources and tools for e-humanities scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1525–1531. <https://dSPACE.library.uu.nl/handle/1874/307981>
- Miličević Petrović, M., Bernardini, S., Ferraresi, A., Aragrande, G., & Barrón-Cedeño, A. (2021). *Language data and project specialist: A new modular profile for graduates in language-related disciplines* (tech. rep.). Zenodo. <https://doi.org/10.5281/zenodo.5030929>
- Pedonese, G., Frontini, F., del Fante, D., & Federici, E. (2025). Adapting UPSKILLS learning modules to the university curricula: Best practices and lessons learnt from the H2IOSC training experience at the University of Ferrara. *Selected papers from the CLARIN Annual Conference 2024*, 37–47. <https://doi.org/10.3384/ecp216.04>
- Simonović, M., Arsenijević, B., van der Lek, I., Assimakopoulos, S., ten Bosch, L., Fišer, D., Kraš, T., Marty, P., Miličević Petrović, M., Milosavljević, S., Tanti, M., van der Plas, L., Pallottino, M., Puskas, G., & Samardžić, T. (2023). *Research-based teaching: Guidelines and best practices* (tech. rep.). Zenodo. <https://doi.org/10.5281/zenodo.8176220>
- UNESCO. (2021). UNESCO Recommendation on Open Science. <https://doi.org/10.54677/MNMH8546>
- van der Lek, I., Fišer, D., Frontini, F., & Pedonese, G. (2024). Introduzione ai dati linguistici: Standard e archivi digitali [Zenodo]. <https://doi.org/10.5281/zenodo.13911935>
- van der Lek, I., & Fišer, D. (2023). Introduction to language data: Standards and repositories. https://upskillsproject.eu/project/standards_repositories/
- van der Lek, I., Fišer, D., Samardžić, T., Simonović, M., Assimakopoulos, S., Bernardini, S., Miličević Petrović, M., & Puskas, G. (2023). *Integrating research infrastructures into teaching: Recommendations and best practices (version 2)* (tech. rep.). Zenodo. <https://doi.org/10.5281/zenodo.8114407>
- White, A., & Green, D. (2022). Minimum Viable Skillsets: A briefing for competence centres. <https://doi.org/10.5281/zenodo.14865952>
- Whyte, A., Green, D., Avanço, K., Di Giorgio, S., Gingold, A., Horton, L., Koteska, B., Kyprianou, K., Prnjat, O., Rauste, P., Schirru, L., Sowinski, C., Torres Ramos, G., van Leersum, N., & Lazzeri, E. (2024, June). Minimum Viable Skillsets - Catalogue of Open Science Career Profiles. <https://doi.org/10.5281/ZENODO.11469300>
- Whyte, A., Green, D., Avanço, K., Di Giorgio, S., Gingold, A., Horton, L., Koteska, B., Kyprianou, K., Prnjat, O., Rauste, P., Schirru, L., Sowinski, C., Torres Ramos, G., van Leersum, N., Sharma, C., Méndez, E., & Lazzeri, E. (2023). *D2.1 catalogue of open science career profiles - minimum viable skillsets* (tech. rep.). Zenodo. <https://doi.org/10.5281/zenodo.8101903>

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>