

Advancing FAIR Language Data through Training and Capacity Building: the Swiss contribution to CLARIN

Joanna Blochowiak
CLARIN-CH

University of Zurich, Switzerland
joanna.blochowiak@uzh.ch

Cristina Grisot
CLARIN-CH

University of Zurich, Switzerland
cristina.grisot@uzh.ch

Abstract

The CLARIN-CH ecosystem is unique in its approach to language resources, as it brings together a diverse range of research infrastructures distributed across the various language regions of Switzerland. In a larger European context, we present in this paper the CLARIN-CH training program, which plays a central role in Switzerland's efforts to advance sustainable Open Research Data practices within the evolving landscape of Open Science. The program is structured around five principal axes of action: (i) assessing the needs of researchers across disciplines who work with language data, both within linguistics and beyond; (ii) collaborating with stakeholders, including researchers and training providers, to develop offerings tailored to these specific needs; (iii) organizing diverse training events using a variety of educational formats; (iv) co-developing Open Educational Resources aligned with the FAIR-by-design methodology; and (v) ensuring the publication and long-term accessibility of these resources via Zenodo for the CLARIN-CH community. In this contribution, we outline these core components of the training program and showcase three training initiatives delivered through multiple educational formats. The long-term goal of the CLARIN-CH training program is to contribute to capacity building among researchers working with language data, by promoting a spirit of lifelong learning and interdisciplinary collaboration. As such, Switzerland significantly contributes to CLARIN's efforts put into disseminating knowledge and building capacity about language resources and promote best practices aligned with FAIR and Open Research Data principles.

1 Introduction

The CLARIN-CH training program is a central component of Switzerland's efforts to promote sustainable Open Research Data (ORD) practices and to strengthen interdisciplinary collaboration among researchers working with language data. Rooted in the CLARIN-CH ecosystem, the program supports the development of a robust training infrastructure that addresses the evolving needs of a multilingual and multidisciplinary research community. In addition to researchers, the program explicitly encompasses data stewards and data professionals, who play a crucial role as both invaluable sources of expertise and as key beneficiaries of CLARIN-CH training activities. By fostering dialogue and knowledge exchange between researchers and data professionals, the program contributes to the development of shared, sustainable data practices. Its overarching goal is to enhance the capacity of all stakeholders involved in the research data lifecycle to engage with language data in alignment with FAIR principles (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al. 2016), thereby supporting a culture of open science across diverse fields including linguistics, medicine, law, journalism, theology, and beyond.

In this contribution, we describe the Swiss approach to training and capacity building and situate it in the larger context of CLARIN. The paper is structured as follows. Section 2 discusses the importance of training and capacity building for advancing FAIR compliance of language resources, as well for the sustainable development of research infrastructures. Section 3 describes existing training and capacity-building initiatives in CLARIN, both at the central level and at the country level. Section 4 presents on the Swiss approach and its five axes of action. Section 5 showcases three training initiatives provided in the framework of working groups, training sessions targeting linguists and activities targeting non-linguists. Section 6 concludes this contribution.

2 Why is training and capacity building important

Training and capacity building may target on the one hand FAIR and ORD data management practices, and the use of tools and research infrastructure components, on the other hand. Investing into training and capacity building is not a peripheral support activity, but a strategic prerequisite for their sustainable development of research infrastructures at both national and pan-European levels, as well as for empowering students, researchers, and data stewards to bridge the gap between theoretical knowledge and practical aspects of linguistic research data management, and ultimately to do better Open Science and data-driven research (see examples of training opportunities provided by CLARIN countries in section 3). For the rest of this section, we will focus in the importance of training and capacity building for the future of research infrastructures.

First, as research infrastructures increasingly operate as federated ecosystems rather than isolated centres or repositories, their effectiveness depends less on central technical capacity and more on the ability of research communities to produce data that are interoperable, reusable, and legally robust from the outset. In this context, training and capacity building enable technical infrastructures to scale and to federate effectively across countries and disciplines. While technical interoperability can be addressed through shared platforms and standards, the harmonisation of data management practices requires sustained capacity building, which enables the convergence of legal, semantic, and organisational practices supporting the vision of a distributed yet interoperable research ecosystem. In the domain of language resources and language technology, training is a decisive factor in ensuring that European research infrastructures remain interoperable, trustworthy, and strategically relevant in an increasingly AI-driven research landscape.

Second, language data are structurally complex, highly heterogeneous, multilingual, often multimodal, and frequently involve human subjects, which introduces significant legal and ethical constraints. The compliance to FAIR principles at onset, i.e. data creation and documentation, has a positive impact on making them usable for cross-project, cross-border, or long-term reuse. For this reason, systematic training and capacity building shifts FAIR and ORD practices upstream – at the level of researchers themselves, where they are most effective and impactful.

Finally, training and capacity building are also a key governance and sustainability mechanism. National and pan-European research infrastructures, such as CLARIN, but also other thematic infrastructure, such as DARIAH, SSHOC, as well as EOSC, rely on data quality, standardised metadata, and clear access conditions to operate efficient ingestion and service pipelines. Training empowers researchers for their scientific endeavours, but also data stewards and other data management professionals to act as co-producers of FAIR and ORD compliant resources. Consequently, this strengthens trust between national and European research infrastructures and their communities, and reinforces credibility within European research infrastructure consortia.

3 Training and capacity-building initiatives in CLARIN: centrally and in member countries

3.1. CLARIN Central Hub

The creation of the CLARIN Learning Hub¹ is the culmination of many efforts put by CLARIN into training for the last fifteen years. For CLARIN, training and capacity building is one of its core services rather than an auxiliary activity, which has been gradually developed as summer schools, tutorials, and workshops, national training initiatives, knowledge centers and competence networks, including coordination of training materials across member countries. In addition, CLARIN's participation in various EU-funded projects (such as Upskills²), its alignment with EOSC capacity-building agendas (see project skills4EOSC³), and increased emphasis on skills for FAIR data, reproducibility, and digital methods led to the expansion of the CLARIN Learning Hub, as it is today. The Hub aggregates training materials, tutorials, courses, and events produced across the CLARIN network and makes them discoverable and reusable for researchers, educators, and infrastructure professionals. Its underlying logic is federated

¹ <https://www.clarin.eu/content/learning-hub>

² <https://upskillsproject.eu/>

³ <https://www.skills4eosc.eu/>

and community-based: training content is created by distributed experts in national consortia and knowledge centers, then curated centrally to ensure visibility, reuse, and continuity across countries and disciplines.

In terms of thematic scope, the CLARIN training offer covers the full lifecycle of language data and technology use in research. Topics range from applied language technology, corpus linguistics, and computational methods to practical skills such as programming for NLP, data annotation, and standards and repositories (see van der Lenk & Fišer, 2023 *Introduction to Language Data: Standards and Repositories*⁴). There is also a strong emphasis on Open Science, Open Research Data, and FAIR practices, especially in sessions that explain how infrastructures enable reproducibility, sharing, and linking of language datasets to publications. This positioning shows that training is not only about technical tool adoption, but also about embedding good research data management practices into everyday research workflows.

A distinctive feature of CLARIN’s approach is its integration of training with infrastructure access, see for example the *CLARIN101 tutorials*⁵. Learning resources in these tutorials are tied directly to operational services such as federated search systems (the CLARIN Virtual Language Observatory and the CLARIN Federated Content Search), repositories maintained by CLARIN B-centers, virtual collections, and language processing tools such as the CLARIN Switchboard. Training therefore serves as an onboarding mechanism into the infrastructure ecosystem, helping users move from awareness to hands-on use.

The Learning Hub also supports methodological and pedagogical innovation. It promotes self-study materials, reusable tutorials designed following the FAIR-by-Design principle (Filiposka et al., 2025), ensuring that educational resources are findable, accessible, interoperable, and reusable. Training content thus can be embedded in university curricula, reflecting a long-term strategy to institutionalize infrastructure literacy within higher education. Recent initiatives around FAIR-by-design training materials further show that CLARIN is not only teaching FAIR data practices, but applying them to the training ecosystem itself, such as the *CLARIN Trainers’ Toolkit* (van der Lek, 2025). To facilitate the findability of training resources, *CLARIN’s Learning Resources Catalogue*⁶ is a non-exhaustive inventory of learning resources developed in the CLARIN community and allows users to search by topic, skill level, resource types, media type, license, publication and modification date. Each entry is described with consistent metadata, allowing learners and educators to find the materials corresponding to their needs, but also to understand how and by whom the materials were created.

Overall, CLARIN’s training model is infrastructure-centric, distributed, and lifecycle-oriented. It addresses technical competencies (tools, NLP, corpora), data stewardship practices (FAIR, ORD, metadata), and research integration (linking data, services, and publications). This makes training a strategic instrument for strengthening the adoption, sustainability, and cross-border interoperability of language resources and technologies across the European research landscape.

3.2. Overview of CLARIN member countries

In parallel to the centralized efforts made by CLARIN ERIC, CLARIN Knowledge Centers and national initiatives provide domain-specific guidance, educational materials, and expert support, reinforcing the link between skills development and sustainable infrastructure uptake. The CLARIN Learning Hub⁷ provides a centralized entry point for the training and capacity building initiatives from CLARIN countries, allowing users to find their way (see non-exhaustive overview in Table 1). Training materials may be available in English, but also in national languages.

National consortia	Open Learning and Training Resource
CLARIN-IT ⁸	materials for the Italian community and beyond; H2IOSC Training Environment

⁴ <https://www.clarin.eu/content/introduction-language-data-standards-and-repositories-0>

⁵ <https://www.clarin.eu/content/intro-clarin-tutorial>

⁶ <https://www.clarin.eu/learning-resources-catalogue>

⁷ <https://www.clarin.eu/content/guidelines-and-best-practices>

⁸ <https://clarin-it.it/>

CLARIAH-NL ⁹	materials for Linguistics and Media Studies
CLARIAH-VL ¹⁰	materials on national services, tools and resources, Digital Humanities
CLaDA-BG ¹¹	materials for Bulgarian Language Technology for Digital Humanities
CLARIN-DK ¹²	reports, presentations, tutorials related to the language technologies hosted in the Danish infrastructure, and topics such as FAIR and data management topics
CLARIN:EL ¹³	user manual and video tutorials to help users explore and use the CLARIN:EL research infrastructure
CLARIN-LV ¹⁴	videos and tutorials
CLARIN-PL ¹⁵	practical guides, video tutorials, and workshop materials on how Polish applications and infrastructure services are used in research practice
CLARINO ¹⁶	extensive documentation and tutorials on tools and services integrated in the Norwegian infrastructure.
CLARIN-SI ¹⁷	training events and workshops aimed to teach university students, linguists, language teachers and DH scholars how to use the CLASSLA web corpora and NoSketch Engine concordance in South Slavic language research
FIN-CLARIN ¹⁸	guidelines, introduction videos, courses and training materials related to topics related to text and speech processing and analysis, language technology and data management topics
LINDAT/CLARIAH-CZ ¹⁹	slides, videos and tutorials on how to use the LINDAT repositories, concordancers and other tools and services, including best practice guidelines
Portulan CLARIN ²⁰	workbench of tools and services with embedded documentation
CLARIN-CH ²¹	education and training sessions, webinars, workshops
CLARIN-UK ²²	workshops and tutorials on tools and services, and maintains a <u>collection</u> of datasets used in teaching

Table 1 Overview of Open Learning and Training Resources within CLARIN consortia (from CLARIN Learning Hub, made by Iulianna van der Lek, CLARIN Training Officer)

An outstanding example for its complexity, modularity and capacity to serve several languages is the CLASSLA-Express, a workshop series organized under the CLASSLA Knowledge Centre²³ within the Slovenian CLARIN.SI infrastructure. CLASSLA focuses on South Slavic languages and serves as a regional hub for corpus resources, tools, and expertise. CLASSLA-Express was conceived as a flexible and mobile training format that delivers *hands-on, practice-oriented instruction* directly to researchers, students, and other stakeholders working with language data across several South Slavic countries, including Croatia, Serbia, North Macedonia, Bulgaria, and Slovenia. The training program covers a wide methodological spectrum, ranging from introductory corpus-linguistic techniques, such as concordancing, frequency analysis, and corpus exploration using the CLASSLA-web platform, to more advanced topics, including the application, evaluation, and critical assessment of large language models in corpus-

⁹ <https://clariah.nl/>

¹⁰ <https://clariahvl.hypotheses.org/training-materials>

¹¹ <https://clada-bg.eu/en/dissemination/videos.html>

¹² <https://info.clarin.dk/en/publications/presentations-posters-etc./>

¹³ <https://www.clarin.gr/en/support/user-manuals>

¹⁴ <https://www.clarin.lv/en-us/publications>

¹⁵ <https://clarin-pl.eu/>

¹⁶ <https://clarin.w.uib.no/>

¹⁷ <https://www.clarin.si/info/about/>

¹⁸ <https://www.kielipankki.fi/language-bank/>

¹⁹ <https://www.lindat.cz/#education>

²⁰ <https://portulanclarin.net/>

²¹ <https://clarin-ch.ch/services/education-training/>

²² <https://www.clarin.ac.uk/oxford-text-archive>

²³ <https://www.clarin.si/info/k-centre/>

based research. In addition to technical skills, the workshops place emphasis on methodological transparency and the informed reuse of shared language resources, aligning with broader CLARIN goals related to open and reusable research data. A defining feature of CLASSLA-Express is its deliberate decentralized design. Rather than concentrating training activities in a single institutional setting, the program is structured to travel to different locations and host institutions, which lowers participation barriers and extending its reach to geographically dispersed and linguistically diverse research communities. This approach enables close engagement with local research contexts while fostering cross-border collaboration and contributing to the development of a shared regional community of practice around language data.

Taken together, these initiatives illustrate a continual emphasis on bridging technical tool literacy with domain-specific research needs, promoting FAIR and ORD awareness. This broader CLARIN context situates the CLARIN-CH training program within a rich ecosystem of capacity-building efforts that address language data challenges in multifaceted research environments.

4 Axes of action of the CLARIN-CH training approach

The CLARIN-CH training approach is created within the collaborative and community-driven approach and is structured around five principal axes of action, each contributing to the development of a sustainable training infrastructure that supports researchers working with language data across Switzerland's multilingual and interdisciplinary research landscape.

First, a key priority is the systematic **evaluation of the needs of researchers** who work with language data across various disciplines, both within linguistics (see Grisot, 2024) and beyond (see Blochowiak et al., under review). This includes researchers in fields such as medicine, law, journalism, theology, and business, where language data plays an increasingly central role. By conducting targeted needs assessments, such as interviews and surveys, the program aims to gain a comprehensive understanding of the existing competencies, practices, and gaps in knowledge related to ORD and the use of Natural Language Processing techniques. For instance, among the main gaps pointed out by linguists in the 2023 CLARIN-CH Survey (see Grisot, 2024) were numerous, such as data management and legal support, access to large, diverse and multimodal corpora, access to NLP tools and reusable components, access to functional annotation tools, and training in corpus building, coding and data analysis and statistical methods. Outside linguistics, for instance in the area of Film Studies, the main problem is that prohibitive copyright restrictions on audiovisual material establishes a culture of data silos, where datasets are stored on private drives because legal pathways for sharing remain too opaque (see Blochowiak et al., under review). This evidence-based approach ensures that the training offer is built on a clear understanding of actual research needs, grounded in the day-to-day challenges faced by researchers.

Second, the program places strong emphasis on **collaboration with key stakeholders** across the Swiss research ecosystem. This includes engaging both researchers and training providers in co-developing educational content and formats that are tailored to specific disciplinary and linguistic contexts. Through these partnerships, CLARIN-CH fosters a participatory and community-driven model of training design, ensuring that the resulting materials and activities are relevant, adaptable, and aligned with institutional and national priorities related to Open Science and FAIR data practices.

Third, the training program is committed to **organizing a diverse portfolio of training events**, delivered in a variety of educational formats to maximize accessibility and impact. These include interactive working groups, thematic webinars, and practical workshops held both onsite and online. This multimodal approach allows the program to reach a broad and varied audience, from early-career researchers to experienced data stewards, while accommodating different learning preferences and institutional contexts. Importantly, many of these events are designed to foster interdisciplinary exchange and knowledge-sharing, creating spaces where participants can learn from one another and explore how tools and methods can be applied to their specific datasets, but also across disciplinary boundaries.

Fourth, CLARIN-CH works closely with training providers and domain experts to **develop high-quality Open Access Educational Resources** that are aligned with the FAIR-by-design methodology developed in the skills4EOSC project. These resources cover each stage of the language data life cycle from data collection, processing, and annotation to ethical considerations, legal compliance, sharing, and long-term archiving. Special attention is given to discipline-specific needs and to demonstrating the

added value of tools such as the LiRI Corpus Platform²⁴ (University of Zurich) and Swiss-AL²⁵ (Zurich University of Applied Sciences ZHAW) for researchers working outside linguistics. The materials are designed to be modular, reusable, and adaptable for both self-directed learning and formal teaching contexts.

Fifth and finally, the training program ensures the **publication and long-term accessibility of all educational resources** via Zenodo²⁶. By using this open and trusted repository, CLARIN-CH guarantees that training materials are freely available to the broader research community and can be easily cited, shared, and integrated into other teaching or training initiatives. This axis of action reflects the program's strong commitment to sustainability and transparency, and its role in advancing open science practices not only in Switzerland, but also within the wider CLARIN and European research infrastructures.

5 Educational program

Guided by these core principles, CLARIN-CH is developing a training infrastructure designed to meet the needs of researchers working with language data within Switzerland's linguistically and disciplinarily diverse academic environment. In this section, we highlight three initiatives led by the CLARIN-CH team, each utilizing distinct educational formats and addressing the needs of different target audiences involved in language data research.

5.1 Working Groups

As part of the CLARIN-CH training program, several Working Groups (WG) have been created. In what follows we describe two of them: the first is dedicated to sensitive and personal data management (Section 5.1.1) and the second focuses on management of language research data (Section 5.1.2).

5.1.1 The CLARIN-CH sensitive and personal data management Working Group

The first CLARIN-CH Working Group is dedicated to sensitive and personal data management: *Working Group on the management of sensitive and personal data, ethical and legal issues for linguistic data*²⁷. Launched in September 2023, this WG aims to serve as a national platform for disseminating information, promoting best practices, and offering tailored guidance to researchers across disciplines who engage with language data, especially in multilingual and cross-disciplinary settings. The WG began its activities with a kick-off event that brought together experts and researchers to discuss critical issues related to data collection, protection, and long-term preservation. Particular attention was given to the specificities of various types of linguistic data, such as multimodal, historical, experimental, sociolinguistic, and social media-based data, as well as data collected from diverse demographic groups. The event featured three keynote presentations (recorded and made available on the CLARIN-CH Documentation Platform²⁸) alongside an open call for questions from the community. These questions, often discipline-specific, highlighted widespread concerns around the management of sensitive and personal data, as well as copyright and licensing issues, particularly in relation to digital and online sources.

In response to these expressed needs, the WG designed and launched a monthly webinar series for the 2024 spring and autumn semesters. These lunchtime talks feature a 30-minute expert presentation followed by a 30-minute open discussion. The topics covered include both legal and technical aspects of working with sensitive data, as well as practical case studies, such as anonymization techniques for textual, video, and audio data, and legal considerations for cross-border research or social media data use. Each session is documented through recordings and presentation slides, which are made publicly available via the CLARIN-CH Documentation Platform. In addition to the webinars, the WG produces concrete outputs designed to support researchers in practice. These include: (i) a growing FAQ section compiling answers to community questions; (ii) slides and video recordings from the keynote talks and webinars; (iii) personalized advice (when needed) and guidance provided by the WG members or through collaboration with legal and institutional experts.

²⁴ <https://www.liri.uzh.ch/en/services/LiRI-Corpus-Platform-LCP.html>

²⁵ <https://www.zhaw.ch/en/linguistics/research/swiss-al>

²⁶ <https://zenodo.org/communities/clarin-ch/records?q=&l=list&p=1&s=10>

²⁷ <https://clarin-ch.ch/about/working-groups/sensitive-personal-data/>

²⁸ <https://clarin-ch.ch/documentation-platform/>

5.1.2 The CLARIN-CH Working Group on research data management for language data

The second CLARIN-CH Working Group focuses on research data management practices for linguistic research: *Working Group on Research Data Management (RDM) for language data*²⁹. Launched in October 2025, this WG aims to strengthen RDM practices across the Swiss linguistic research community by leveraging the expertise of institutional data stewards and research support services. The WG strives to serve as a collaborative platform for developing resources, training, and support that improve awareness, capacity, and consistency in managing language data, while laying the foundation for a future CLARIN-CH Training Program aligned with national and international RDM standards. The WG began its activities with a kick-off workshop collocated with the CLARIN-CH Day 2025 at the University of Lausanne, bringing together data stewards, research support professionals, and linguists to discuss the challenges and opportunities of implementing FAIR principles for language data. Particular attention was given to the specificities of language data management across various research contexts, including multilingual corpora, endangered language documentation, experimental linguistic data, and computational linguistics resources. The discussions highlighted critical needs around Data Management Plans (DMPs), institutional alignment on RDM practices, and the integration of domain-specific requirements within broader Open Research Data frameworks.

To address these needs, the WG designed an action plan targeting specifically the DMPs, as they stood out being of particular interest to researchers as well as to data professionals. As a first step, the WG convenors started to work on a two-part DMP workshop series: Part I addresses the general principles and importance of DMPs for SNSF-funded projects, emphasizing early-stage planning and integration into the research lifecycle; Part II provides hands-on guidance with concrete examples tailored to language data management scenarios. Each session will be documented, and materials will be made available through the CLARIN-CH Documentation Platform to support broader adoption across institutions. In addition to the training activities, the aim of the WG is to produce concrete outputs designed to support researchers and data professionals in practice. These include: (i) best practice guidelines for FAIR-compliant language data management; (ii) DMP templates and examples specific to linguistic research; (iii) a comparative review of DMP tools and templates used across Swiss higher education institutions; (iv) training materials and resources accessible through a centralized repository; (v) personalized consultation and guidance provided by WG members in collaboration with institutional data stewardship teams. The WG fosters strategic collaboration with institutional and national partners, including major Swiss universities, the Swiss Academy of Humanities and Social Sciences, swissrn.org, and doctoral schools, ensuring that outputs are broadly relevant and sustainable. This collaborative approach positions the WG as a key contributor to professionalizing RDM support for language data across Switzerland and establishing CLARIN-CH as the national reference point for FAIR and ORD-compliant practices in linguistic research.

5.2 Training sessions for linguists

As part of its commitment to fostering sustainable and inclusive practices in working with language data, the CLARIN-CH training program organized a series of online training sessions in Spring 2025, titled *Exploring Swiss Language Resources and Tools*³⁰. This initiative aimed to introduce researchers, both newcomers and experienced users, to the rich ecosystem of infrastructures, tools, and services developed within CLARIN-CH with a precise aim to offer practical, hands-on guidance on how to effectively leverage these resources for research purposes. The training series was designed to support the growing community of researchers relying on language data across a wide range of disciplines. Participants were introduced to a variety of topics, including corpus querying and analysis, media data processing, multimodal data modeling, and best practices for data sharing and reuse. The sessions were developed with the support of two nationally co-funded projects: *Upgrading the Linguistic ORD Ecosystem (UpLORD)*³¹ and *Moving towards a FAIR-Compliant Ecosystem of Federated Infrastructure for Language Data (FAIR-FI-LD)*³².

²⁹ <https://clarin-ch.ch/about/working-groups/wg-research-data-management/>

³⁰ <https://clarin-ch.ch/training-sessions-2025/>

³¹ <https://www.liri.uzh.ch/en/projects/Completed-Projects/UpLORD.html>

³² <https://www.liri.uzh.ch/en/projects/Completed-Projects/FAIR-FI-LD.html>

The series was delivered collaboratively by key institutions and stakeholders in the CLARIN-CH network, including the CLARIN-CH Coordination Office, the Linguistic Research Infrastructure (LiRI) at the University of Zurich, the ZHAW Digital Discourse Lab, the Language Repository of Switzerland (LaRS@SWISSUbase), and Università della Svizzera italiana (USI). Each session featured expert contributions from representatives of these institutions which enabled the dissemination of up-to-date knowledge and the showcasing of current best practices. The sessions were held biweekly and offered participants a structured journey through the CLARIN-CH ecosystem and its tools for working with language data. A total of 98 participants from Switzerland and beyond took part in the online training sessions. For lack of resources, the training sessions were carried out only in English. In the future, depending on the available financial resources, we plan to create training materials in all Swiss national languages to ensure equitable access.

In line with FAIR principles and the Open-by-design principles, the materials produced for these sessions were developed as Open Educational Resources (OER) and published on the CLARIN-CH Zenodo Community, ensuring long-term accessibility and reusability, as shown in Table 2.

Training session	KPIs
CLARIN-CH Training Session#1: Introduction to the CLARIN-CH ecosystem of infrastructures in the era of Open Research Data	67 views, 69 downloads
CLARIN-CH Training Session#2: LiRI Corpus Platform: Data Model and Query	64 views, 76 downloads
CLARIN-CH Training Session#3 Querying and analysing Swiss media data with Swissdox@LiRI	44 views, 53 downloads
CLARIN-CH Training Session#4: LiRI Corpus Platform: Data Analysis Using Text, Audio, and Video Corpora	54 views, 54 downloads
CLARIN-CH Training Session#5: Modelling of Multi-Modal Data in LiRI Corpus Platform and Beyond	54 views, 57 downloads
CLARIN-CH Training Session#6: Data analysis with Swiss-AL for textual data	47 views, 23 downloads
CLARIN-CH Training Session#7: Publishing Data on the Language Repository of Switzerland (LaRS)	60 views, 53 downloads
CLARIN-CH Training Session#8: Reusing data for teaching and secondary analyses	49 views, 14 downloads

Table 2 CLARIN-CH Training Series and Zenodo KPIs (status on March 31, 2026)

Collectively, the series provided comprehensive, practice-oriented training designed to strengthen the capacity of researchers working with linguistic data in line with FAIR principles and ORD standards.

5.3 Emerging initiative for non-linguists

As part of its ongoing efforts to enrich its educational offering, CLARIN-CH has begun laying the groundwork for developing training programs tailored to non-linguistic disciplines. As language data continues to gain prominence across a broad spectrum of research disciplines beyond linguistics, including medicine, history, theology, business communication, translation and interpretation, and language education, new challenges have emerged regarding the management of such data in line with ORD principles. Researchers in these fields often use language data in ways that diverge significantly from traditional linguistic approaches. For instance, while a linguist may examine syntactic structures, a medical researcher might analyze communication strategies to improve patient care. More specifically, corpus-based studies have become a valuable tool for analyzing the specialized language of these domains, since they are crucial for addressing such domain-related research questions effectively. In recent years, however, conducting corpus-based research has become increasingly regulated, driven by growing positive pressure to define requirements and implement ORD practices, as pointed out by disciplinary experts interviewed in Blochowiak et al. (under review).

A survey conducted by CLARIN-CH within the Swiss community in 2023 highlighted an acute need for structured training in ORD practices (Grisot, 2024). Researchers expressed a strong desire for guidance in technical, legal, and ethical aspects of working with language data. While past initiatives, such

as webinars, workshops, and training programs, have made strides toward addressing these needs, they have often been limited in scope, sporadic in implementation, and focused primarily on linguistics. Researchers from other disciplines who use language data have been left underserved, due to their unique approaches and specific research objectives. To address this gap, the CLARIN-CH ecosystem has launched a new initiative called *Language Data Across Disciplines* aimed at strengthening ORD competencies across the broader research community in Switzerland. The LaDaD initiative is currently in its conceptualization phase and includes the following fundamental disciplines: medicine, history and political sciences, theology, film studies, financial and business communication, translation, terminology and interpretation, and language education. It has three main objectives. The first objective is to develop a comprehensive understanding of the competencies needed to manage language data in these fields at each step of the data life cycle. A needs assessment, conducted through a series of guided interviews as well as a national survey, will identify specific challenges faced by researchers in adhering to ORD practices. The second objective is to promote the development of a collaborative research community gathering researchers and data stewards from different fields working with language. The community will seek to exchange knowledge, address common challenges, and establish a network of researchers and data stewards skilled in managing language data in compliance with FAIR principles. As a result, the community will provide actionable recommendations for composing a comprehensive compendium of ORD best practices. The third objective is the development of training sessions to address gaps in ORD knowledge and skills. These resources will include tutorials and practical guides that showcase best practices for working with language data tailored to the disciplinary needs of researchers working with language data. By promoting a deeper understanding of ORD practices and creating reusable training resources, the project aims to contribute to the sustainable advancement of research data management of language data in the spirit of open science beyond linguistics.

6 Conclusion

In this paper, we discussed how CLARIN ERIC and other national CLARIN consortia invest into training and capacity building with the aim to empower users, such as students and researchers, to do better Open Science and data-driven research by bridging technical tool literacy with domain-specific research needs and raising FAIR- and ORD-compliance awareness. In alignment with these efforts, the CLARIN-CH training program stands as a key pillar in fostering sustainable and inclusive ORD practices for researchers and data stewards working with language data in Switzerland. Through its multidimensional approach based on five axes of action, ranging from targeted needs assessment to the development of FAIR-aligned educational resources, it supports a growing, interdisciplinary community engaged in Open Science. By continuing to expand its initiatives and reach beyond linguistics, CLARIN-CH is laying the foundation for a robust and collaborative research environment across disciplines.

When viewed in the broader context of training initiatives developed across the CLARIN network, such as decentralized workshop series or federated learning hubs in other countries, the CLARIN-CH program distinguishes itself through its strong emphasis on systematic needs assessment, close integration of data stewards and data professionals, and its explicit focus on disciplines beyond linguistics. By combining international best practices with solutions tailored to Switzerland's multilingual and interdisciplinary research landscape, CLARIN-CH contributes to the diversification and strengthening of the European CLARIN training ecosystem. The CLARIN-CH training program is currently largely dependent on project-based funding, which can limit the continuity and long-term planning of training activities. As a result, ensuring a comprehensive and consistently updated training offer across disciplines and also across Switzerland's linguistic regions remains challenging, particularly beyond the duration and scope of specific funded initiatives.

By continuing to expand its initiatives and fostering collaboration across disciplinary, institutional, and professional boundaries, CLARIN-CH is laying the foundation for a robust and sustainable research environment in which language data can be managed, shared, and reused responsibly across domains. In doing so, the program not only addresses current training needs, but also supports long-term capacity building aligned with FAIR principles and the evolving practices of Open Science.

References

- Blochowiak, J., Fischer, M., Krasselt, J., Liste Lamas, E., Vukovic, T. & Grisot, C. (under review). Going beyond linguistics: assessing the status quo and exploring the needs of disciplines using language data in the Swiss context. Submitted to CLARIN2026.
- Grisot, C. (2024). Report on 2023 CLARIN-CH User Survey. Zenodo. <https://doi.org/10.5281/zenodo.19458087>
- Grisot, C., Körner, E., Eckart, T., Osenova, P., Piasecki, M., Lennes, M., & van der Lek, I. (2025, October 17). CLARIN101 - Introduction to CLARIN. CLARIN Annual Conference, Vienna. Zenodo. <https://doi.org/10.5281/zenodo.17379981>
- Filiposka, S., Green, D., Mishev, A., Kjorveziroski, V., Corleto, A., Napolitano, E., Paolini, G., Di Giorgio, S., Janik, J., Schirru, L., Gingold, A., Hadrossek, C., Souyiultzoglou, I., Leister, C., & Pavone, G. (2025). FAIR-by-Design Methodology for learning materials. Zenodo. <https://doi.org/10.5281/zenodo.14711525>
- van der Lek, I., Fišer, D. (2023). Introduction to Language Data: Standards and Repositories. In UPSKILLS Learning Content. https://upskillsproject.eu/project/standards_repositories/. CC BY 4.0.
- van der Lek, I. (2025). CLARIN Trainers' Toolkit (1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.15029514>
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9. <https://doi.org/10.1038/sdata.2016.18>.