

Language Research Data Management Node (L-RDM Node) – Concept

Introduction

Language data represents a rapidly growing and highly diverse research domain in Switzerland. It is central not only to linguistics and digital humanities, but also to disciplines such as law, medicine, history, and the social sciences. At the same time, language data introduces specific research data management (RDM) challenges due to multilingualism, sensitive personal content (e.g., voice recordings), heterogeneous formats (text, audio, video), copyright and licensing constraints, and ethical issues related to language preservation and vulnerable communities.

The proposed **Language Research Data Management Node (L-RDM Node)** addresses this need by strengthening RDM support capacity in the Swiss language research community. It aims to provide practical, scalable resources and training for research support professionals and data stewards, aligned with academic, funder, and international RDM standards. The node will build on both newly developed training materials and the aggregation of existing resources.

The L-RDM Node is closely connected to and developed in collaboration with **CLARIN-CH**, the Swiss national node of the pan-European CLARIN infrastructure. It builds on the experience and network of the existing CLARIN-CH RDM Working Group while extending its scope toward research support services and data stewardship communities.

In addition, the node addresses the timely and fast-evolving topic of **Artificial Intelligence (AI) and Large Language Models (LLMs)** in language data management, raising awareness and providing guidance on opportunities and risks of AI-assisted workflows such as transcription, annotation, metadata generation, and data organization.

Objectives

The L-RDM Node aims to strengthen RDM practices across the Swiss language research community by leveraging the expertise of institutional data stewards, libraries, and research support services. It promotes consistent and FAIR-aligned approaches to managing language research data across institutions and disciplines.

Specifically, the node provides a forum for knowledge exchange, expert input, and peer-to-peer support on topics including:

- best practices for managing multilingual and multimodal language datasets (text, audio, video)

- data management planning (DMPs) for language research, including discipline-specific requirements
- FAIR and reusable language data workflows (documentation, metadata, standards)
- legal and ethical aspects of language data (copyright, licensing, privacy, voice data, consent)
- preservation and sharing of language resources, including endangered language materials
- practical use of AI and LLMs for language data workflows (e.g., transcription tools such as OpenAI Whisper or Otter.ai, automated annotation, metadata enrichment)
- responsible use of AI tools, including transparency, quality control, and reproducibility

Beyond training and exchange, the node will contribute to sustainable resource development by producing reusable guidelines, templates, and training materials for SRDSN members.

Community

The L-RDM Node is aimed at SRDSN members working with or supporting research involving language data, including:

- data stewards and RDM professionals at Swiss HEIs with language-related research, particularly within the CLARIN-CH consortium (UZH, ZHAW, UNIBAS, UNIFR, USI, UNIGE, UNIL, UNIBE, UNINE)
- research support staff working with language and humanities data
- library professionals supporting digital humanities and linguistic research
- SRDSN members with interest in AI-assisted data management and language data workflows

Node membership is open to all interested SRDSN members who agree to the SRDSN Code of Conduct.

The node aims to foster a growing network of research professionals working with language data across disciplines, and to strengthen connections between institutional support services and national initiatives.

Activities

The L-RDM Node will support SRDSN members through training, resource development, and structured community exchange. Planned activities include:

- organizing training workshops and events on RDM practices for language research
- providing DMP-focused training (general principles and hands-on language-specific examples)

- raising awareness about AI and LLM tools for language data management, including benefits, limitations, and risks
- sharing and curating reusable training materials, guidelines, and examples
- fostering strategic collaboration among institutional and national partners
- presenting outcomes, updates, and challenges at SRDSN meetings
- contributing to SRDSN communication channels and knowledge-sharing platforms

Activities are open to SRDSN members, with topics tailored primarily to research support staff and data stewardship communities.

Operational Mechanisms

The L-RDM Node is coordinated by a node lead consisting of a chair and supporting members. The lead team is responsible for organizing activities, maintaining collaboration with SRDSN structures, and ensuring continuity and sustainability of node outputs.

The node lead is responsible for:

- setting deliverables and yearly activity plans
- organizing training events and workshops
- coordinating collaborations with CLARIN-CH and other relevant partners
- managing the node's community exchange mechanisms
- supporting onboarding of new members (in coordination with SRDSN processes)
- producing annual reporting to the SRDSN Steering Committee
- presenting node updates and outcomes at SRDSN meetings
- ensuring documentation and archiving of node resources
- ensuring compliance with SRDSN Code of Conduct

Timeline

The L-RDM Node is planned as a long-term initiative. Its activities and relevance will be reviewed through annual reporting and SRDSN community feedback. In case of reduced activity or demand, the node's scope and timeline will be adjusted accordingly.