Gaps on the linguistic eco-system, project goals and achievements (October 2024)

In the context of the requirements of Open Science and of FAIR principles, on the one hand, and of that of more challenging data sets (such as, sensitive or with copyright issues), on the other hand, we identified several gaps regarding the current situation in Switzerland and we are addressing them in this ORD project:

Gap (1) The provision and the usability of the individual corpus platforms across Switzerland: their drawback is that they do not satisfy the Interoperability principle.

Gap (2) The status of several linguistic corpora: most of them do not satisfy all four FAIR principles.

In order to address these first two gaps this project is moving forward the implementation of the LiRI Corpus Platform at the national level following the FAIR principles. This will significatively improve discovery, access, integration, usability and reusability of corpora according to FAIR principles, as well as simplify re-formatting, assembling, harmonizing and standardizing the data and metadata.

- The LiRI Corpus Platform LCP is a software system for handling and querying corpora of different kinds: text, audio and video (see the modules **catchphrase**, **soundscript** and **videoscope**).
- Numerous tests with importing and quering various corpora have been carried out on the test version. Now, the public version is available, so users can query corpora directly from their browser, and upload their own corpora using a command-line interface.
- The platform is going to be presented during the first LCP Day on November 1st. The program will consist of project presentations and practical software demonstrations. We welcome participants on-site and on Zoom (on-site participation is limited).

Gap (3) The lack of meaningful metadata and of infrastructures to manage a diversity of annotations of linguistic data, this being linked to the Findable and the Reusable principles.

Regarding this third gap, the implementation of VIAN-DH (presently known as the **videoscope**) has been used as a test case for modelling complex annotation schemes that LiRI Corpus Platform has to offer. Data processed within videoscope is complex interactional data consisting of verbal, paraverbal and non-verbal annotation levels. The project is collecting further complex annotation models via the CLARIN-CH and NCCR partners as well as via other clients LiRI is already working with. It is then evaluated whether the LCP@LiRI can also map these. This process goes hand in hand with the development of best practices to show researchers how to deal with complex annotations, but also to showcase the resulting opportunities.

- Complex annotation representation is constantly being tackled during LCP implementation process and lead to a new approved proposal (Technologieplattform UZH) to represent even more complex annotations (audio/video) than in UpLORD planned.
- Up to now, data converters from (TEI-)XML, generic XML and CWB format to CoNLL-U+ have been implemented.

Gap (4) The proper management of sensitive data, informed consent, copyright and intellectual property issues, which are necessary so that the sets of data adhere to the FAIR requirements.

Gap (5) The metadata schema on SWISSU-base which must be adapted and amended to satisfy the Findable principle.

Gap (6) Uploading workflows on SWISSUbase.

Regarding gaps from 4-6, the already implemented modular linguistic metadata schema of LaRS@SWISSUbase will be specified and adapted to the needs that are not included for the moment. To upgrade the usability of SWISSUbase, it is also possible to provide a workflow via an API. In addition, easy workflows between LCP and videoscope will be implemented.

- Version control workflows and launch of API for data upload on SWISSUbase are implemented
- Data Service Unit@UZH established as single point of entry for linguistic data to be deposited on SWISSUbase
    - Principles of data curation for linguistics
- Webinar on the use LaRS/SWISSUbase services (February 2023), see also: Webinar: The Language Repository of Switzerland, powered by SWISSUbase (youtube.com)
- New SWISSUbase Info Website have been launched: info.swissubase.ch. Exploring the new Info Website it is possible to learn more about the SWISSUbase Consortium and partner Data Service Units; their mission, platform and services; and discover the extensive Help Centre to support researchers in preparing, sharing and accessing data

Gap (7) the quality control and data curation in SWISSUbase to satisfy the Reusable principles,

In order to address this gap, the project set up national CLARIN-CH working groups whose role is to develop specific metadata for various types of linguistic data (e.g., sociolinguistics data, experimental psycholinguistics, neurolinguistics data, conversational analysis data, lexicography data, computational data, acquisitional data, multimodal data) that satisfy the FAIR principles.

- To develop specific metadata, a mixed approach was adopted: (1) a CLARIN-CH WG on Learner corpora, (2) individual discussions with researchers (e.g. for sign language data) and (3) the reuse of CLARIN CMDI metadata schema.
- The mapping of existing metadata schema on SWISSUbase to CLARIN CMDI profiles has been done.
- Adapting the metadata schema according to the needs and requirements of the linguistic community:
    - Challenge of balancing individual and more general needs/requirements: one additional focus that has emerged as very important is on functionality of SWISSUbase as a catalogue for data and also for services (linkage functionality only).
- A survey to collect information about recommended/standard data formats was created. After an initial pre-test with data experts working at LiRI, SWISSUbase and NCCR Evolving Language (April 2024), the survey circulated in the CLARIN-CH community (May-June 2024). A detailed explanation of the survey results and a summary with format recommendations are available.

Gap (8) The setting of ethical standards for best practices, good habits and frames of mind in linguistic data management and collaboration, as well as the lack of training for the target scientific communities to adopt these standards.

Regarding gap 8, this project is developing showcases, best practices and documentation for data management according for FAIR principles (such as, planning and writing data management plans, using sustainable file formats, setting solid file-naming conventions, finding data storage backup schemes, creating informative metadata and documentation) promoting ORD practices in the CLARIN-CH, the NCCR "Evolving Language" and other linguistic communities. Special attention is given to find solutions that overcome disciplinary boundaries.

- Existing LaRS@SWISSUbase guides:
  - For Researchers: Prepare your data for deposit in SWISSUbase
  - User guide for Linguistics studies and data
  - Linguistics Metadata guide
- Documentation on CLARIN-CH Documentation platform
  Research data lifecycle
  - Data Management Planning
  - Collecting data
  - Processing data
  - Sharing data
  - Archiving data
  - Reusing data
  - FAIR data
  - Copyright
  - Licenses
  - Data protection
  - Data access and security
  - Metadata standards
  - Standard data formats

- A national event, CLARIN-CH Day, on the theme "ORD: Challenges and Opportunitie"s took place in September 9, 2024, with presentations of data pitches and exchanges between researchers and experts.
- 9 webinars on the topics of data protection (legal and technical perspectives), data anonymization, intellectual property rights in linguistic data, legal aspects of sharing and reusing linguistic data organized in the frame of the CLARIN-CH WG on Management of Sensitive and Personal data, Ethical and Legal issues for linguistic data.