Franciska de Jong

The scope of CLARIN: From language as linguistic data to language as social and cultural data

Abstract

CLARIN is a European Research Infrastructure providing access to language resources and technologies for researchers in the humanities, the social sciences and beyond. It supports the study and use of language data in general and aims to increase the potential for comparative research of cultural and societal phenomena across the boundaries of languages.

This paper outlines how the design and implementation of CLARIN are compliant with the emerging Open Science framework and supports findability, accessibility, interoperability and re-usability of data. The paper also explains how the CLARIN research infrastructure contributes to the potential of digital language resources for research of cultural and societal phenomena across the boundaries of languages and modalities, and in particular for comparative studies in a multidisciplinary context.

1. Introduction

CLARIN (Common Language Resources and Technology Infrastructure) is the European research infrastructure that provides access to language resources and technologies for researchers in the humanities, the social sciences and beyond.[1] CLARIN gives easy and sustainable access to (i) digital language data (in written, spoken, video or multimodal form) and (ii) advanced tools to discover, explore, exploit, annotate, analyse or combine them, wherever they are located. For this purpose, CLARIN has set up a distributed network of trusted data and service centres.[2] The access service is offered through a single sign-on environment. CLARIN also serves a network of technical services as an ecosystem for knowledge sharing.

CLARIN was established as an ERIC (European Research Infrastructure Consortium) in 2012, and since 2016 it is a so-called ESFRI Landmark.[3] Between 2012 and 2017 the CLARIN membership has grown from nine member countries to a consortium of more than twenty countries (19 members, 2 observers).[4]

---

1 See the CLARIN website for detailed documentation and background information: www.clarin.eu.

2 For an up-to-date overview, see https://www.clarin.eu/content/clarin-centres.

3 For details about the status of ERICs as legal entity and their role in the context of ESFRI: https://ec.europa.eu/research/infrastructures/index.cfm?pg=eric.

4 For an up-to-date overview of participating countries and third parties, see https://www.clarin.eu/content/participating-consortia.

The CLARIN infrastructure is widely seen as a matured entrance point for researchers that study language from a linguistic point of view. The recent surge of data that is available through online platforms of various kinds has given rise to new research agendas that call for a focus on content and context of language materials rather than on the linguistic layers alone. There is increased potential for comparative language-based research of cultural and societal phenomena across the boundaries of languages and modalities, and this comes with a demand for the renewal of methodological frameworks and new models of multidisciplinary collaboration. CLARIN is strongly rooted in a research community that has a deep understanding of how linguistic diversity is both an asset and a challenge. This is enabling CLARIN to contribute to the emerging paradigms for studying social and cultural phenomena based on language data.

In Section 2 it will be explained how the mission and vision of CLARIN relates to the Open Science agenda at large that is stimulating the implementation of models for the wider accessibility and reuse of research data. In Section 3 the potential of CLARIN to support multidisciplinary research based on language data will be illustrated.

2. CLARIN and Open Science

An inherent goal of CLARIN is to integrate and contribute to Europe's Open Science policies. With the growing availability of digital language data ('digital born' or digitized analogue resources), the possibilities beyond the mere archiving and viewing of language resources have become significant. Increasing the potential for re-use and repurposing of digital data has become part of the objectives that the CLARIN community has embraced with the aim to contribute the generation of new knowledge and scholarly impact.

Supporting the re-use and repurposing of language data determines a good deal of the principles that have been adopted in the design of CLARIN's infrastructure from the very beginning. Their implications and value largely coincide with the so-called FAIR data principles (Wilkinson et al. 2016) that are now widely promoted as part of the Open Science paradigm. In the following paragraphs we describe the data architecture of CLARIN in a nutshell and the compliance with the FAIR data framework, or in other words: the way in which CLARIN data sets are Findable, Accessible, Interoperable, and Re-usable.

2.1 CLARIN and FAIR

As a distributed infrastructure, CLARIN exists of a network of centres (nodes). A CLARIN centre typically provides a data repository where language resources are stored and made available for researchers in a sustainable manner. Additionally,

many centres also provide tools (web applications, web services or stand-alone applications) for the processing of language data.

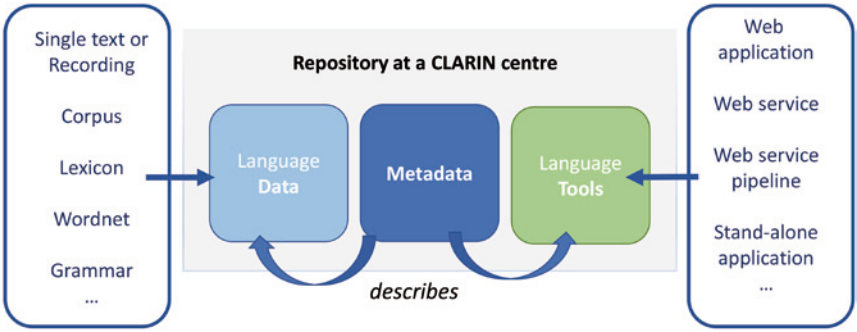

Fig. 1: Repository perspective: metadata generation for the description of data and tools

Use and re-use of registered language resources is only possible when they can be easily and reliably found by researchers. This requires that they are stored with persistent identifiers, which secure the citeability of data, rich open metadata, which can be indexed and made searchable, and links from the metadata to the data identifier. CLARIN requires CMDI metadata (Goosen et al. 2015) for the description of data and tools in repositories. The so-called Virtual Language Observatory (VLO) is the portal through which all the harvested metadata records can be searched for (Van Uytvanck et al. 2012). Figure 2 is a graphical representation of the way in which CLARIN harvests metadata from the associated centres.[5]

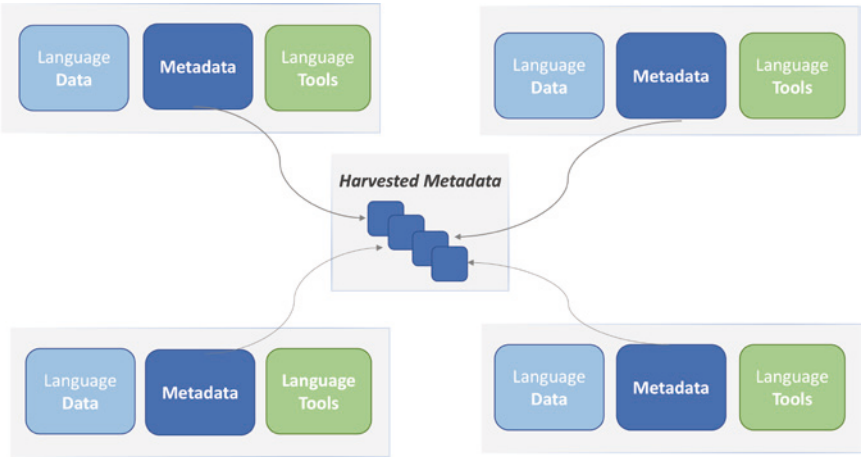

Fig. 2: Metadata harvesting for data and tools distributed over multiple registries

[5] See De Jong et al. (2018) for a more detailed description of the architecture.

With more than 1,600,000 entries in the VLO, CLARIN supports access to a wealth of language materials, including cultural heritage data, in many languages. Apart from data *findability*, data *accessibility* is a prerequisite for reuse. This comes with the need for a standardized communication protocol, and the option for easy authentication of researchers that want to access the data (in CLARIN supported through a federated login service), and authorisation when needed.

To ensure *interoperability* of data CLARIN has implemented a formal, shared and broadly applicable framework, for the linking between metadata and data, and recommendations for the use of standard data formats. Semantic interoperability can be offered though flexibility of metadata and open vocabularies for the description of language resources.

In accordance with the requirements for *re-usability* of data, a clear licence and provenance information is provided. The bottom-up structure of the centres structure definitely brings along close ties with such community-based practices and standards.

2.2 CLARIN as a platform for data analysis services

CLARIN comprises not only central discovery services that give access to the data on offer in the distributed collection of resources, but the associated centres also provide access to repositories and services for curation, analysis, modeling and knowledge sharing.

Examples are support for full-text federated content search or for aggregating resources into virtual collections. Another important service is the so-called Language Resources Switchboard (LRS): a single point of access where researchers can find analysis tools that are suited for the processing of the data resource they have selected. LRS sorts the tools in terms of the tasks they perform, and presents a task-oriented list to the user and can integrate the tools into a web-based pipeline of analysis services (Zinn 2016). See Figure 3.

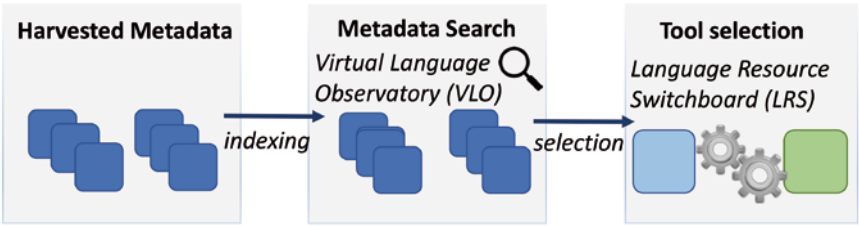

Fig. 3: Central processing of metadata, selection of suitable tools

Among the functionalities that are suited for the kind of contentoriented analysis that is typically useful for the collections of textual or spoken data which are relatively big in size are: text classification, topic clustering, named entity extraction,

sentiment analysis, etc. In most cases this category of tools typically is built from of a combination of language-specific algorithms and language specific resources such as vocabularies and part-of-speech taggers that are available as distributed web services. See Figure 4 for a schematic overview of the chaining of analysis steps supporting specific research workflows.

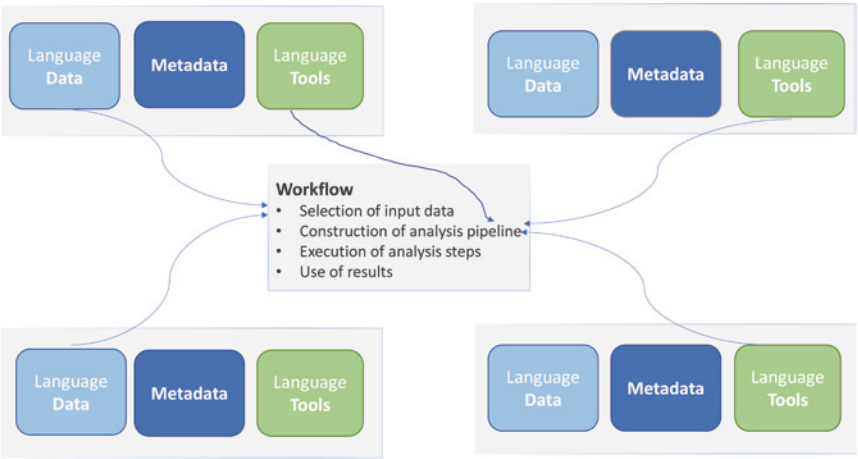

Fig. 4: Illustration of a stepwise workflow based on CLARIN web services for distributed data and tools

3. CLARIN and data analysis

The data that CLARIN is collecting covers languages from all locations, regions and periods. The following non-exhaustive list of examples can illustrate the topical spread and the variety in terms of scholarly origin (in arbitrary order):

- Parliamentary records
- Literary texts
- Social Media data
- Historical letters
- Oral History data
- Disciplinary libraries
- Institutional archival data
- Broadcast archives
- Newspaper archives

The diversity in language resources that are available across Europe and beyond underlines that language is a rich carrier of cultural content, a significant reflection of societal dynamics, and a central part of the identity of individuals and groups.

As language is more and more recognized as social and cultural data (Nguyen et al. 2017), CLARIN is developing activities that contribute to the visibility of the scholarly fields involved as pillars of data science.

3.1 Support for comparative research

This diversity of data types and languages covered makes CLARIN highly relevant for the study of national and regional languages and cultures. Actually, it is the combination of data and tools available for multiple languages that makes CLARIN a key enabler of comparative studies across regions, periods, languages and cultures. In particular the data and tools offered for study of phenomena that are characteristic for the culture of Europe based on language data can be boosted by CLARIN. Research into topics such as language variation, migration patterns, intellectual history, language acquisition, or parliamentary discourse, obviously can benefit from well structured overviews of data that could reinforce the comparative perspective. Therefore, CLARIN has recently started targeted actions that promote the findability and visibility of specific data types and families of resources (Fišer et al. 2018). This is likely to increase the potential for uptake of CLARIN in the so-called digital humanities (De Smedt et al. 2018).

3.2 Validation: Key for uptake

Sustainable scientific and societal impact of the tools offered can only be expected if the validity of analysis results can be explained and assessed by relevant researcher communities. This requires that language data is not studied in isolation, but that whenever possible, contextual information that may be available can also be taken into account. As a consequence, language data will sometimes be integrated in a larger collection of heterogenous data types, including for example, numerical data or geo-information.

Attention for the validity of analysis results also comes with the need to document and explain the performance levels that can be expected from the analysis tools, and thereby of the suitability of certain tools for specific scenarios of use. Scenario-based testing is particularly relevant for the uptake of CLARIN functionality in the context of multidisciplinary collaboration where methodological frameworks rooting from different scholarly traditions have to be combined. As CLARIN is not only a technical infrastructure but also an ecosystem for the development and sharing of knowledge and expertise the conditions are in place to successfully pave the way to the stimulate collaboration across communities of use in the social sciences and humanities and thereby the uptake of the infrastructure for the study of language as social and cultural data.

References

Arppe, A./Bruun, S./Koskenniemi, K./Lindén, K./Oksanen, V./Westerlund, H. (eds.) (2011): *A report including model licensing templates and authorization and authentication scheme. CLARIN Deliverable D7S-2.1.* Utrecht: CLARIN ERIC.

Fišer, D./Lenardič, J./Erjavec, T. (2018): Meet CLARIN's key resource families. In: *Proceedings of the Eleventh Language Resources and Evaluation Conference* (*LREC*)*, 7-12 May 2018, Miyazaki, Japan*.

Goosen, T./Windhouwer, M./Ohren, O./Herold, A./Eckart, T./Ďurčo, M./Schonefeld, O. (2015): CMDI 1. 2: Improvements in the CLARIN component metadata infrastructure. In: *Selected papers from the CLARIN 2014 Conference, October 24–25, Soesterberg, The Netherlands*. (= Linköping Electronic Conference Proceedings 116). Linköping: Linköping University Electronic Press, 36-53.

De Jong, F.M.G./Maegaard, B./De Smedt, K./Fišer, D./Van Uytvanck, D. (2018): CLARIN: Towards FAIR and responsible data science in the area of language. In: *Proceedings of the Eleventh Language Resources and Evaluation Conference* (*LREC*)*, 7-12 May 2018, Miyazaki, Japan*.

Nguyen, D./Doğruöz, A.S./Rosé, C.P./De Jong, F.M.G. (2017): Computational sociolinguistics: A survey. In: *Computational Linguistics* 42, 3, 537-593.

De Smedt, K./De Jong, F.M.G./Maegaard, B./Fišer, D./Van Uytvanck, D. (2018): Towards an Open Science infrastructure for the digital humanities: The case of CLARIN. In: Proceedings of the Digital Humanities Nordic Conference 3rd Conference (DHN 2018), 7-9 March 2018, Helsinki.

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. et al. (2016): The FAIR guiding principles for scientific data management and stewardship. In: Scientific Data 3 (article 160018).

Zinn, C. (2016): The CLARIN language resource switchboard. In: Proceedings of the CLARIN Annual Conference October 26-28, 2016, Aix-en-Provence, France. Utrecht: CLARIN ERIC.

Bibliographical information

This text was first published in the book:
Gerhard Stickel (ed.) (2018): National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences. [299 pages.]

The electronic PDF version of the text is accessible through the EFNIL website at:
http://www.efnil.org