An Infrastructural Approach to Terminology Work: The Case of Research Infrastructures

Tanja Wissik

Austrian Centre for Digital Humanities and Cultural Heritage
Austrian Academy of Sciences
Vienna, Austria
`tanja.wissik@oeaw.ac.at`

Abstract

This study explores the role of research infrastructures, in particular the role of CLARIN and DARIAH, with regard to terminology work in institutional settings (academic and non-academic) by analyzing a body of qualitative interview data, collected in 2023 across Europe. The contribution also discusses how research infrastructures (RIs) could reach out to new non-academic communities e.g., in the public sector and to reevaluation of existing terminology infrastructure models and include research infrastructures.

1   Introduction

It is not new to use the term *infrastructure* for the organisation of terminological collaboration and terminological activities (Pilke et al. 2021, 101). Already Galinski (1998) described an infrastructural approach to terminology, dividing terminological infrastructures into horizontal and vertical infrastructures. The horizontal infrastructure includes five elements: "terminology (planning) policy, terminology creation centres, terminology information and documentation centres, terminology associations and corporate cooperation groups led by the private sector" (Galinski, 1998). The vertical infrastructure concerned the different ways of carrying out terminological activities within different domains (Galinski, 1998). An adapted version of this infrastructural model (Galinski & Giraldo, 2023) is used in a cooperation project between Austria and Mongolia[1]. This model includes five horizontal layers "(1) Individual users and data creators (as well as their groups and networks), (2) Intermediaries and service providers to individual users/groups, (3) Organizations/networks of data creators and curators, (4) Terminology infrastructure coordination authority/ies, and (5) Policies and high-level authorities. At all levels six vertical functions and activities occur: (A) Organizational aspects, (B) Services, (C) Education and R &D, (D) ICT systems and tools, (E) Support and promotion, and (F) Legal & technical regulations." (Galinski & Lušicky, 2024). In these existing infrastructural models for terminology, research infrastructures are not explicitly included, however there are several connecting points between terminology work and research infrastructures (e.g., Andersen & Gammeltoft, 2022; Wissik & Declerck, 2020; Wissik, 2022). Stakeholders in terminology work can be on the one hand data providers and on the other hand they can be users of data, tools and services provided by RIs and benefit from the knowledge sharing infrastructure to exchange knowledge and promote collaboration.

However, there is little insight into the role and use of such research infrastructures, in particular CLARIN and DARIAH, within the community of stakeholders involved in terminology work, especially in institutional settings, besides some case studies (e.g., Andersen & Gammeltoft, 2022). So, this paper

---

[1] The name of the project is "Terminology planning strategy and terminology infrastructure for Mongolia to support scientific and educational development and innovation".

wants to close this gap and explores the role of research infrastructures with regard to terminology work in institutional settings (academic and non-academic) based on qualitative interview data. The paper is structured as follows: after an introduction, the research method is described, and the results are discussed. The contribution focuses on resources (e.g., corpora) and repositories as possible links between research infrastructure and the community of stakeholders involved in terminology work, mentioning also other possible areas of cooperation such as training materials and tools.

2 Background

2.1 Research Infrastructures in the Humanities

Among the first Research Infrastructures (RIs) to support research in the Arts and Humanities were libraries, museums and archives (Moulin et al., 2011). In today's digital age a number of RIs have emerged at European and national level to support digital research. These RIs offer technical infrastructures in a more stable and sustainable way than research projects that run only short period. As technical infrastructure they provide resources, tools and services to the scientific community in order to support top-level research activities. Furthermore, RIs provide a social infrastructure for collaboration and knowledge exchange and they act as promoter for the use of common methods and standards. RIs also play a crucial role in training and educating future generations of researches and research engineers (Wissik & Declerck, 2020).

In the Social Sciences and Humanities (SSH) the European Strategy Forum on Research Infrastructures (ESFRI) recognizes five large European research infrastructures and six research infrastructure projects (ESFRI 2021). RIs can be generic or domain specific. As generic research infrastructures, for this paper, we understand research infrastructures that can be used by researchers from a variety of research fields within the SSH. In the Humanities, e.g., CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities) count as generic research infrastructures (Doel & Maes, 2012) and for example EHRI (European Holocaust Research Infrastructure) can be seen as a domain specific research infrastructure (Wissik & Declerck, 2020).

In the following we will describe the two generic RIs CLARIN and DARIAH that are most relevant for the field of terminology research and practice.

CLARIN stands for Common Language Resources and Technology Infrastructure) and it was established as an ERIC in 2012 with "the mission to create and maintain a digital infrastructure to support the sharing, use and sustainability of language data (in written, spoken, or multimodal form) available through repositories from all over Europe, in support of research in the humanities and social sciences and beyond" (de Jong et al, 2022). Currently, CLARIN currently has 26 member countries that operate a distributed network of data, service and knowledge centres. Through its network, CLARIN provides access to digital language resources, services and expertise to scholars, researchers, and students from different disciplines in the SSH.

The other relevant generic infrastructure in the humanities is DARIAH, which stands for Digital Research Infrastructure for the Arts and Humanities, and was established in 2014 as an ERIC with the mission "to empower research communities with digital methods to create, connect and share knowledge about culture and society." (DARIAH, 2025) Currently DARIAH has 22 member countries and 17 co-operating partners in non-member countries. DARIAH represents a network of "people, expertise, information, knowledge, content, methods, tools and technologies." (DARIAH, 2025a).

In this section, the two generic RIs CLARIN and DARIAH were introduced. In the next section we will discuss some offerings from those RIs, that could be of interest for stakeholders working in the area of terminology.

2.2 Research Infrastructures and Terminology

As already mentioned before, there are several connecting points between terminology work and research infrastructures (e.g., Andersen & Gammeltoft, 2022; Wissik & Declerck, 2020; Wissik, 2022). In the following we will look into some of these connecting points by discussing services offered by RIs.

Stakeholders in the area of terminology on the one hand produce language data as a result of the terminology workflow. The data can be in the form of terminology databases, glossaries or terminological dictionaries. Furthermore, they also might create specialized corpora, focusing on a specific domain or subdomain. On the other hand, they often consult already existing language resources, such as terminology databases, glossaries, dictionaries or corpora in the process as well. Therefore, we will look into ways of accessing existing language resources as well as services for depositing language resources that are provided by the before mentioned RIs.

In CLARIN, there are several ways of accessing the language resources and tools that are available within the infrastructure: One way of accessing is via the Virtual Language Observatory (VLO), a search catalogue to find language resources via different facets (Van Uytvanck et al., 2012; Goosen & Eckart, 2014); Another way of accessing is via the so-called CLARIN Resource Families, a manually curated overview of language resources and tools organized by type (e.g., glossaries, corpus query tools) (Fišer et al., 2018). Another option of accessing is directly via the data repositories hosted by the CLARIN B-Centres network.

Besides exploring and searching for available resources via their metadata, it is also possible to search for patterns within the language resources directly online. One option is the Federated Content Search (FCS)[2], which allows the search in different language resource, hosted at different centres, at the same time. The FCS was originally created for searching in full-texts with optional annotation layers. Recently a new LexFCS extension (Eckart et al., 2023) was designed to make also lexical resources such as dictionaries, word lists or semantic wordnets, searchable via the FCS. A first implementation of this extension is included in the Text+ FCS Aggregator (Körner et al., 2024).

Furthermore, it is also possible to search in individual corpora via provided concordance tools by dedicated CLARIN B-Centres (e.g., CLARIN.si or LINDAT/CLARIAH-CZ).

Another service, that is provided by the CLARIN infrastructure is the depositing service for language resources. This service is offered by CLARIN B-Centres, who maintain certified data[3] repositories in order to archive and disseminate language resources. Depositing language resources is one essential step in making the language resources findable, accessible, interoperable and reusable.

Since DARIAH is serving a broader community, there are very diverse tools and services listed in the DARIAH Tools and Services Catalogue. In particular, the Vocabs services, hosted and maintained by the ACDH-CH might be of interest for the terminology community. The Vocab services contain on one hand a controlled vocabulary repository and on the other hand an editor in order to collaboratively create and maintain controlled vocabularies.

Furthermore, the SSH Open Market place (König et al. 2024), a discovery portal maintained by the three RIs DARIAH, CLARIN and CESSDA, can be of interest for the terminology community, on one hand to search for tools and services but also to publish and share their own tools and services with the broader SSH community. When searching for the keyword "terminology" in the SSH Open Market place, 36

---

[2] https://www.clarin.eu/content/content-search

[3] A Clarin B-Centre needs to have the Core Trust Seal and must fulfill the requirements set by the CLARIN Technical Centre Assessment Committee. More information available here https://www.clarin.eu/content/clarin-b-centre-assessment

resources were listed (21 data sets, 10 tools and services, 4 publications and 1 training material. However, no workflow was listed[4]).

Besides a technical infrastructure, RIs also provide expertise and a space for knowledge exchange.

CLARIN maintains a knowledge infrastructure, comprising a network of CLARIN Knowledge Centres, so-called CLARIN K-Centres, to make expertise and knowledge available in a structured way. When searching with the keyword "terminology" through the CLARIN K-Centre inventory, there are 7 CLARIN K-Centres listed with expertise in the area of terminology: CLARIN-ELEXIS Knowledge Centre for Lexicography, Czech CLARIN Knowledge Centre for Corpus Linguistics, CLARIN Knowledge Centre for Dutch, CLARIN K-Centre for Natural Language Processing in Greece, CLARIN Knowledge Centre for Polish Language Technology, CLARIN Knowledge Centre for The Languages of Sweden and CLARIN K-Centre for Terminology Resources and Translation Corpora.

In the case of DARIAH, a lot of expertise can be found in the DARIAH Working Groups. The goal of the DARIAH Working groups is "to consolidate infrastructure and scholarship in certain areas of research and to create or reinforce the network of expertise inside of DARIAH" (Scharnhorst et al. 2019, 9). At the moment there is no WG specifically targeting the terminology community, but there are several working groups where there are connecting points: for example, with the WG "Lexical Resources"[5] because they are dealing with all kinds of digital resources not only dictionaries, also lexicons, thesauri, word lists etc. and they are not only working with semasiological methods and standards but also with onomasiological methods and standards. Another connection point can be seen with the WG "Multilingual DH", that want to enhance digitally-enabled research in under-resourced languages by adapting or developing tools for those languages. Furthermore, the Research Data Management WG might be of interest for the terminology community, regarding best practices in data management, the use of data repositories, FAIR data principles etc.

2.3   Related Research

There was some qualitative research done on interinstitutional cooperation in terminology work by Chiocchetti & Ralli (2013). They conducted expert interviews with 17 terminologists in institutional settings. However, when the interviews were conducted between 2011 and 2012, research infrastructures in the Humanities were fairly new, so the topic of research infrastructure was not discussed in their contribution. Budin (2015) provided a theoretical discussion on the role of research infrastructure, particularly CLARIN, in the context of computational terminology, without evidence from empirical data. For the ELRC White Paper (ELRC, 2022) on "AI for a multilingual Europe – Why Language Data Matters" a survey was conducted to get insights into the current use and importance of language technologies and into common European practices with respect to translation, data management and sharing in public administrations and SMEs. 73 people responded to the survey: their answers were the basis for the ELRC White Paper. The national CLARIN consortia were mentioned several times in the ELRC White Paper in the country report section.

Furthermore, there are recent initiatives, besides the already known catalogues and registry (e.g., Virtual Observatory, ELRC Share, the ELRA Universal Catalogue), to collect terminology data as shown by the TeresIA project (Maroto, 2024). In this project a survey for terminology providers was set up to collect information on the data and metadata in order to integrate these terminology resources into a newly created meta-search portal[6].

In this section we have discussed those parts of the technical and social infrastructure of RIs, that might be relevant for the stakeholders in terminology work and we have reviewed some related research.

---

[4] Here the search performed on January 29, 2025 can be found https://marketplace.sshopencloud.eu/search?q=Terminology
[5] https://www.dariah.eu/activities/working-groups/lexical-resources/
[6] https://proyectoteresia.org/colaborar

However, until now there was little insights, which role RIs, especially CLARIN and DARIAH, have within the community of practice.

3   Method

The present contribution is part of a larger study, carried out in 2023, exploring the role and impact of new technologies and new paradigms, such as open data, on terminology work performed in institutional settings and how workflows, tasks and roles are influenced consequently.

To gain insight in this area, 15 semi-structured expert interviews (Meuser & Nagel, 1991: 443) were conducted with individuals involved in terminology workflows in different institutional settings in different roles to better understand terminology workflows in the digital age (Wissik, 2024).

Experts that were eligible for this study had to work in an institution[7] that (1) performs practical terminology work and (2) maintains a publicly accessible terminology database or terminological dictionaries. Moreover, the interviewees had to be regularly involved in practical terminology workflows. An initial selection of candidates was identified through relevant terminology networks. Additionally, relevant international organizations were added to the list. From this initial selection of candidates, 15 experts consented to participate in the study (Wissik, 2025). One limitation of the study is the small sample size. However, the final sample covered academic and non-academic institutions: 4 regional/state administrations in Europe, 2 European institutions, 2 international organizations, and 7 academic/research institutions in Europe (universities or academy-based terminology institutes and language centres). There were both, institutions from CLARIN and/or DARIAH member countries and non-member countries (see Table 1). Moreover, the sample covered institutions with different sizes of publicly accessible terminology database, ranging from large databases with over 400,000 entries to smaller ones with fewer than 10,000 entries and all major types of terminology work: systematic, ad hoc, translation-oriented/text-based, preparatory for standardization (e.g., terminology planning), proactive, and a posteriori terminology work. Furthermore, different roles within the terminology workflow were represented including directors of terminology units, terminologists, terminology managers/coordinators, technology managers, developers/IT experts, and members of standardizing committees (Wissik, 2025).[8] So, even though given the small sample size, it reflects most of the scenarios of institutional terminology work. And experiments with qualitative interview data have shown, that more interview does not necessarily mean more information and that already with 12 interviews data saturation was reached (Guest et al. 2006)

The questions for the expert interviews were based on a previous study on terminology workflows by Chiocchetti and Ralli (2012, Attachment A) and were modified for this study (see question list on Zenodo).[9]

Table 1. Interview participant profiles (adapted from Wissik, 2025)

---

[7]Experts from commercial settings or freelancer were not included in this study.
[8]For more details on the interview participants see Wissik (forthcoming).

[9]List of questions is available on Zenodo with the following link https://doi.org/10.5281/zenodo.11144968

Interview Number	Role	Type of Institution	CLARIN Member	DARIAH Member	Less-resourced language
INT 1	Developer / IT expert	Research / Academic	no	yes	yes
INT 2	Technology manager	Research / Academic	no	yes	yes
INT 3	Head Terminology Unit/ Terminologist	Research / Academic	yes	yes	yes
INT 4	Member of Terminology Committee	Research / Academic	no	yes	yes
INT 5	Head of Terminology and Legal Translation Unit, Deputy Director for Development	Administration (national level)	yes	cooperating partner	yes
INT 6	Terminologist	Administration (regional level)	yes	yes	yes
INT 7	Terminologist	Research / Academic	yes (at the time of the interview not yet full member, K-Centre)	yes (at the time of the interview not yet full member)	Yes
INT 8	Head Unit Project Management / Terminologist	Administration (regional level)	yes (at the time of the interview not yet full member, K-Centre)	yes (at the time of the interview not yet full member)	yes
INT 9	Terminologist	Research / Academic	yes (at the time of the interview not yet full member, K-Centre)	yes (at the time of the interview not yet full member)	yes
INT 10	Terminology Coordinator / Terminology Manager	Intergovernmental Organisation	no	no	yes
INT 11	Head Unit Terminology / Terminologist	Intergovernmental Organisation	no	no	no
INT 12	Technology Manager	Intergovernmental Organisation	no	no	no
INT 13	Head Unit Terminology / Terminologist	Intergovernmental Organisation	no	no	yes
INT 14	Terminology Coordinator / Terminology Manager	Research / Academic	yes	Cooperating Partner	yes
INT 15	Head Unit Terminology / Terminologist	Administration (regional level)	Observer	yes	no

Most interviews were conducted in English, two interviews were conducted in German. The transcribed and anonymized interviews were analyzed by using a thematic qualitative text analysis (Kuckartz, 2014). The data was encoded with CATMA (Gius et al., 2023), an open-source annotation tool that allows to create your own categories to annotate the data (Wissik, 2025).

This study explored the role of RIs in particular the role of CLARIN and DARIAH with regard to terminology work in institutional settings. The relevant questions asked in this context where (1) which material/resources do the participants use for their terminology work and (2) if the participants use or create corpora when doing terminology work and if they publish the corpora, they have created. Furthermore, (3) if they deposit their terminological data in data repositories and (4) if they collaborate with Research Infrastructures, in particular CLARIN and/or DARIAH.

4 Results

When analysing the interview data, several aspects regarding the actual role of RIs in terminology work and potential connection points emerged. In this contribution we will focus on the following aspects: use of corpora and other language resources when compiling terminological resources, use of data repositories for the created language resources and collaboration or engagement with RIs. All of these aspects will be described and discussed below and illustrated with examples from the interview data.

4.1 Corpora and other language resources

When compiling terminological resources, documentation is collected to extract the terms to be included into the final resource. Therefore, using digital corpora in terminology work is not new (Bowker, 1996; Pearson, 1998). When doing ad hoc terminology work, for example, answering requests from query services (Žagar Karer & Fajfar, 2023), terminologists usually resort to already existing corpora:

> "We use already compiled corpora, for example general language corpora, to check how a specific term is used or some domain corpora just to check whether they [terms] are used and frequencies, when we deal with several term candidates, which one is more represented which one is less, that kind of things." (INT 7)

Besides corpora also other language resources are used for researching and verifying terms such as already existing dictionaries:

> "[...] we search through dictionaries, [...] depends on the type of the problem and in specialized texts and in corpora. We have a lot of corpora in [Name of Country], we have corpora of academic texts, and general corpora and all sorts of specialized corpora, so we can check in different kind of corpora to see the situation in language." (INT 3)

When doing systematic terminology work for a specific domain in order to create for example, a specialized dictionary or to enrich a terminology database with a new domain, terminologists also create specialized corpora from scratch: "In the beginning when we started compiling […] dictionaries we always prepare a specialized corpus of the texts that the experts give them [the texts] to us and then we the terminologist, prepare wordlist" (INT 3).

Another resource that was mentioned, especially in the translation context, were parallel corpora: "[a]s part of our work in terminology we have built up parallel corpora" (INT2). A special type of parallel corpora, translation memories were also mentioned when asked for the use of corpora: "No [we don't use corpora for systematic terminology work], we use our databases primarily and translation memories for that but corpora we use when answering these consultations were there more language problems, morphology problems and those cases." (INT5)

Other interview partners have stated, that they tried to create their own corpora, but they encountered various obstacles, such as little data for specific less-resourced language or missing specific IT support, because the institution has only a generic IT department and not a specific IT department that supports only language technologies (INT 7):

> "We made some experiments in creating our own corpora, but we did not get good results. Because we have little data on [Name of less-resourced language]. Corpus was compiled [in] [Name of well-resourced language]. You really need a large spectrum to cover all the areas to have a balanced corpus and we never ended with having balanced corpora. There were always some areas that were more present. Not all the subject fields we wanted were represented. And because we did not have an IT department, that only work for us, for our need, it did not work for us." (INT 7)

The snippets from the interview show, that in institutional terminology work on the one hand already existing corpora are consulted, as well as other already existing language resources such as dictionaries and on the other hand specialized corpora are created from scratch. When asked about publishing those corpora, some interview partners mentioned, that they publish their corpora on their own website, depending on legal constrains:

> "We do publish a lot of our corpora. We have a national corpus project as well […] and on our own website we have the contemporary corpus of [Name of less-resourced language] which is about 100 million worth of contemporary [Name of less-resourced language] published over the last decade or so and we have a parallel corpus of legislative materials from EU legislation and national legislation and they are both freely available on our website. Some of the material is to download, some of them are there only to search, again depending on the copyright restrictions." (INT 4).

However, most interview participants create the corpora for internal use only. They do not publish them or deposit them in a repository: "We usually keep it as a working material. It's more like really like a stage in preparing a dictionary, it is not annotated with POS [part of speech]. I would take us too much time for this." (INT3). Sometimes they are also shared through a corpus management platform but they are not published:

> "[W]e store them [corpora] in SketchEngine, it's collaborative so you can share the corpus with other people in the organization or outside the organization so that's very useful and we typically leave it there. I mean we don't export them we don't. Sometimes we will use them but we don't publish them or we don't, you know, otherwise store them except for sketch engine where we have a license and some storage." (INT 13).

4.2 Data repositories

Data repositories are a way to archive language resources for the long-term and make the language resources (e.g., corpora, terminological resources) available to the community in a reliable way. Usually, data repositories assign persistent identifiers and therefore the data can be cited easily. Through the available metadata it is also possible to search for the language resources efficiently.

Regarding the use of repositories, Wissik (2024) analysed the use of data repositories for terminological data in general in the context of sustainability and the findings showed, that it was not a very common practice among the interviewees to store terminological data in a data repository. However, most of them had multiple other access points to their data and alternative data backup strategies.

For this contribution we only analysed the use of data repositories that are related to RIs. Only one interview participant reported, that their terminological data is stored in a national CLARIN repository (INT 3) and one interview participant reported, that they had recently talked with a national CLARIN representative also about the possibility of archiving the data in a CLARIN repository in the future (INT 14). For publishing their corpora, none of the interview participants used data repositories (see also section 4.1).

4.3 Collaboration or engagement with Research Infrastructures such as CLARIN or DARIAH

A part of the interview was also dedicated to participation in networks and collaborations with other institutions with special focus on CLARIN and/or DARIAH. When looking at the interview profiles in Table 1, it shows, that out of the 15 interview partners, 11 interview partners were from a country that are part of CLARIN and/or DARIAH or at least cooperating partners.

All the interviewees mentioned that they are active in different networks and associations regarding terminology, or specific languages etc., as institutions or as individuals. Several of the non-academic institutions mentioned collaborations with universities. Regarding the explicit collaboration with CLARIN and DARIAH, most academic institutions were aware of both RIs, and some had direct links. Besides using for example, the repositories in the CLARIN infrastructure, also activities in committees were mentioned: "I think one of my colleagues is member in CLARIN, she is active member in some

committee" (INT 3). However, most of the units responsible for terminology had no direct links: "I think [Name of University] possibly has some DARIAH links, but not our unit." (INT 4). Furthermore, some interview partners mentioned, that they are planning to collaborate in the future: "And regarding CLARIN and DARIAH we are not collaborating with this research infrastructures at the moment but we are considering such collaboration in the future." (INT 9). Several interview participants, especially those from non-academic institutions, were not familiar with CLARIN and DARIAH (e.g., INT 5, INT 6, INT 8).

5   Discussion

Regarding the use of corpora, most of the interviewees reported, that they were using already existing corpora, especially in the context of ad hoc terminology, but also other resources such as dictionaries. In these cases, also corpora from CLARIN national consortia were mentioned and used. However, not all interviewees were aware of the available variety of resources through CLARIN and DARIAH. The possible ways of accessing these resources, via the VLO via the CLARIN Resource Families and via the SSH Open Market Place could be promoted further to this target group. Especially Corpora of Academic Texts, Legal Corpora and Parallel Corpora, Dictionaries and Glossaries might be of particular interest for the terminology community. As described before, in the interview data, translation memories were mentioned as a resource, and not so often parallel corpora. In the CLARIN Resource Family for Parallel Corpora, also Translation Memories are listed, but it is not mentioned in the title. If it is not too long, also Translation Memories could be integrated into the title, to show that Translation Memories are included and that this CLARIN Resource Family is among the hit list, when searching for Translation Memories. Besides the resources themselves, also the possibilities of the Federated Content Search and the use of online concordance tools could be beneficial for terminology community, especially when checking for example the use and frequencies of several term candidates.

Furthermore, consideration should be given to adapt the LexFCS extension to terminology resources or to design and implement a new extension for the FCS in order to search through distributed (multilingual) terminology data, which is already available in the CLARIN infrastructure.

Regarding the creation of their own corpora, several participants mentioned, that for systematic terminology work, they are compiling their own corpora. When ask about publishing, only few participants reported, that they are publishing their own corpora, mainly on their own websites. None of the participants published their corpora in a data repository. Most of the interview partners do not publish their corpora, because they only see them as internal working material towards the final product, the terminological resource. Furthermore, often these corpora are not annotated with part of speech, because it would be too time consuming to add them, and therefore the corpora are also considered by the creators less valuable for others and are therefore not shared.

It can be seen, that most institutions, creating terminological resources have also other valuable language resources, that are not yet shared with a wider community but could be of interest for the CLARIN and also DARIAH community. So, RIs could promote in this community the value of sharing and reusing language resources, and that language resources do not always need to be annotate with for example Part of Speech Tagging, in order to be valuable for certain user groups.

During the interviews also challenges when creating their own corpora were mentioned. For example, the lack of not enough IT support. In these cases, CLARIN could provide training and training materials to terminologists, how to create and maintain corpora, so that terminologists are capable to do it on their own without or with minimal support from IT units.

Another topic that was discussed in the interviews were data repositories. The use of depositing services offered by RIs for publishing terminological data, i.e., to publish a terminological dictionary or the data export of a terminology database in a data repository, was not a very common practice at the time of the interviews. In fact, only one participant mentioned, that they deposit their data, where copyright allows it, in a national CLARIN repository. This is in line with the findings in Wissik (2024: 110) that the use of data repositories, in general, is not a widespread practice in this community. So, in this respect, awareness raising regarding the use of data repositories, and the benefits of it, e.g., sustainability of the

data, adhering to the FAIR data principles, additional dissemination channel for the data in this community would be needed, in both academic and non-academic settings. This could be done together by the RIs through dedicated CLARIN K-Centres, and through dedicated DARIAH WGs such as DARIAH WG on "Lexical Resources" and "Research Data Management".

A part of the interview was also dedicated to networks and collaborations with special focus on CLARIN and/or DARIAH. All of the interview participants were very active in relevant networks and associations regarding terminology or specific languages, regardless if they were academic or non-academic institutions. Most non-academic institutions also mentioned, that they have collaborations with academic institutions such as universities. However, most participants from non-academic institutions did not have collaborations with RIs and some of them were not familiar with CLARIN and DARIAH, even though their institutions were located in a member country or the country had at least a K-Centre (in case of CLARIN). These results are not so surprising, as the priority within the RIs so far was to reach out to academic users and to broaden the academic user base. However, recently RIs started to engage with non-academic communities as well, such as the public sector (e.g., Lyding et al., 2022). In the case of CLARIN, with the help of dedicated K-Centres, training materials could be created, that specifically target the terminology community. Furthermore, an ambassador from the terminology community within the public sector could be used to get engage with others in the public sector. By recruiting ambassadors also from the public sector, the already existing and successful CLARIN ambassadors programme could be used to reach out to the public sector. Furthermore, CLARIN and DARIAH could engage with this community via terminology or language associations, where they are members. Furthermore, terminology communities that already benefit from RIs like in Norway (Andersen & Gammeltoft, 2022) could be used as success stories to highlight the benefits of RIs in terminology work. Another finding of the analysis was, that interviewees that were aware of CLARIN and/or DARIAH, reported, that the specific academic unit involved in terminology work was not having links with these RIs. It is clear, that institutions such as universities are complex but it could be worthwhile to investigate, how to best target relevant users within an institution that is already part of an RI consortium.

More theoretically, it would be beneficial to integrate research infrastructures such as European Research Infrastructure Consortia (ERICs) and also similar initiatives such as the European Digital Infrastructure Consortia (EDIC)[10] into the terminology infrastructure models proposed by Galinski (1998) or Galinski & Giraldo (2023).

6 Concluding remarks

This contribution has discussed the potential role of Research Infrastructures such as CLARIN and DAIRAH in practical terminology work and explored the actual role of those RIs in practical terminology work in institutional settings by analysing 15 recorded expert interviews with a qualitative approach. To sum up, the case study has shown, that the role of RIs in terminology work in institutional settings has potential but is still expandable. Due to the small size of the sample, it is difficult to genialize the results. However, since the 15 expert interviews covered most current terminology approaches and most common scenarios where terminology work is done in institutional settings, the study contributes to our understanding of the current relation between Research Infrastructures and the terminology community, especially in institutional settings.

Furthermore, the contribution has discussed possible measures how CLARIN and DARIAH could engage more with the terminology community, for example by involving more dedicated CLARIN K-Centres and DARIAH WGs in awareness raising and training measures. The contribution also discussed measures on how to specifically target potential new user groups in the public sector. One suggestion was to expand the CLARIN ambassador programme by recruiting also ambassadors from the public sector. So, for examples, stakeholders in terminology work for example in public administration could act as ambassadors to engage with non-academic communities who could benefit from the data, services

---

[10]One example of an EDIC is the Alliance for Language Technologies EDIC (ALT-EDIC). More information can be found here https://alt-edic.eu/about-us/.

and knowledge provided by RIs. In this contribution, we have only discussed language resources and depositing services and the sharing of expertise as possible connecting points between the terminology community and RIs. However, also other connecting points could be discussed such as digital tools. Several tools to manage, edit and visualize data play an important role in terminology work (e.g., terminology management systems, corpus management tools, term extraction tools) which could be also a possible area of interaction with research infrastructures. Another area of possible interaction could be the use of AI and LLMs in terminology work especially for less-resourced languages: There the CLARIN K-Centre for Terminology Resources and Translation Corpora and the CLARIN K-Centre for LLMs4SSH could collaborate.

Moreover, a reevaluation of existing terminology infrastructure models and integration of research infrastructures and similar constructs into the terminology infrastructure models is recommended.

Acknowledgement

References

Andersen, G, & Gammeltoft, P. (2022). The Role of CLARIN in Advancing Terminology: The Case of Termportalen – the National Terminology Portal for Norway. In Fišer, D. and Witt, A. (eds.) CLARIN: The Infrastructure for Language Resources, Berlin, Boston: De Gruyter, 249–274.

CLARIN: The Infrastructure for Language Resources, ed. by Darja Fišer and Andreas Witt, 249–274. Berlin, Boston: De Gruyter.

Budin, G. (2015). Digital Humanities, Language Industry, and Multilingualism: Global Networking and Innovation in Collaborative Methods." In: CIUTI-Forum 2014: Pooling Academic Excellence with Entrepreneurship for New Partnerships. 423 – 448, Lausanne.

Bowker, L. (1996). A Corpus-Based Approach to Terminography. *Terminology, 3*(2), 27–52.

Chiocchetti, E. & Ralli, N. (2013). Let's do it together: Instances of cooperation in terminology work: Roles, tools, needs and difficulties. In Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria, 515 – 524.

DARIAH (2025). Mission & Vision. In DARIAH Website. https://www.dariah.eu/about/mission-vision/

DARIAH (2025a). DARIAH in a Nutshell. In: DARIAH Website. https://www.dariah.eu/about/dariah-in-nutshell/

De Jong, F. Van Uytvanck, D., Frontini, F., Van den Bosch, A., Fišer, D. and Witt, A. "Language Matters". In Fišer, D. and Witt, A. (eds.) CLARIN: The Infrastructure for Language Resources, Berlin, Boston: De Gruyter, 31–58.

Doel, W. van den and Maes, K. 2012. Social Sciences and Humanities: Essential Fields for European Research and in Horizon 2020. League of European Research. Accessed January, 17 2025 https://www.leru.org/ files/Social-Sciences-and-Humanities-Essential-Fields-for-EuropeanResearch-and-Horizon-2020-Full-paper.pdf

ELRC (2022). AI for a Multilingual Europe. Why Language Matters. ELRC White Paper. Saarbrücken: Research Center for Artificial Intelligence (DFKI).

Fišer, D., Lenardič, J., Erjavec, T. (2018), CLARIN's Key Resource Families. In N. Calzolari (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan.

Galinski, C. (1998). Terminology infrastructures and the terminology market in Europe. *TRANS Internet-Zeitschrift für Kulturwissenschaften, 0*. https://www.inst.at/trans/0Nr/galinski.htm

Galinski, C. & Giraldo, S. (2023). Mongolian Infrastructure Layers. Presentation at the Project Meeting (Project Nr. KOEF 09/20), October 2023, Vienna.

Galinski, C. & Lusicky, V. (2024). Empowering Knowledge Societies Using Terminological Approaches – How Top-Down Policies Can Meet Bottom-Up Application. Book of Abstracts of the International Conference "Terminology in Scientific and Technological Development (TSTD) – 2024, September 13, 2024, Ulaanbaatar, Mongolia, 7–8.

Goosen, T., & Eckart, T. (2014). Virtual Language Observatory 3.0: What's New? Proceedings of the CLARIN Annual Conference 2014, Soesterberg, Netherlands.

Gius, E. et al. (2023). *CATMA 7 (Version 7.0)*. Zenodo. DOI: 10.5281/zenodo.1470118.

Guest, G., Bunce, A. & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability". *Field Methods* 18 (1), 59–82.

König, A., Barbot, L., Grisot, C., Michael Kurzmeier, Gray, E. J. (2024). SSH Open Market Place and CLARIN. In Lindén, K., Kontino, T. Niem, J. (eds.). Selected Papers from the CLARIN Annual Conference 2023.

Körner, E., Eckart, T., Helfer, F. & Uwe Kretschmer, U. (2024). Federated Content Search: Advancing the Common Search Infrastructure. In Kontion, T. & Vandeghinste, V. (eds). CLARIN Annual Conference Proceedings 2024, 15 – 17 October 2024 Barcelona, Spain, 41–45.

Kuchartz, U. (2014). *Qualitative Text Analysis: A Guide to Methods, Practice and Using Software*. London: Sage.

Lyding, V., Stemle, E., König, A. (2022). Collaborating on Language Resource Infrastructures with Non-ResearchPartners: Practicalities and Challenges. In Monica Monachini and Maria Eskevich (eds.). Selected Papers from the CLARIN Annual Conference 2021, 88–100.

Maroto, N. (2024). TeresIA. Spanish Access Portal to Terminologies and Artificial Intelligence Services. In Proceedings of the 3rd International Conference on Multilingual Digital Terminology Today (MDTT 2024), Granada, Spain, June 27-28, 2024.

Meuser, M., & Ulrike Nagel, U. (1991). "Experteninterviews – vielfach erprobt, wenig bedacht." In Gerz, D. & Karaimer, K. (eds.) Qualitative-empirische Sozialforschung. Konzepte, Methoden, 441–471. Opladen: Westdeutscher Verlag.

Moulin, C., Nyhan, J., Ciula, A., Kelleher, M., Mittler, E., Tadić, M., Ågren, M., Bozzi, A., Kuutma, K., (2011) Research Infrastructures in the Digital Humanities. ESF Science Policy Briefing 42. Strasbourg: European Science Foundation (ESF).

Pearson, J. (1998). *Terms in Context*. Amsterdam: Benjamins.

Pilke, N., Nissilä, N. & Landqvist H. (2021). Organising terminology work in Sweden from the 1940s onwards. Participatory expert roles in networks. *Terminology Terminology, 27(1),* 80–109.

Scharnhorst, A. Morselli, F. Admiral, F. (2019). DARIAH Working Groups Policy Statement 2019, v05. DARIAH ERIC. https://www.dariah.eu/wp-content/uploads/2019/09/DARIAH-Working-Groups-Policy-Statement_v5.pdf

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. In N. Calzolari (Ed.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, May 23rd-25th, 2012, 1029–1034. European Language Resources Association (ELRA).

Wissik, T. (2025). Impact of automatic term extraction on terminology work: a qualitative interview study in institutional settings. *Terminology 31(1)*, 110–135.

Wissik, T. (2024). Dimensions of sustainability in terminology practices in institutional settings. *Terminology Science & Research*, 27, 93–116.

Wissik, T. (2022). Research Infrastructures and Lexicography: An European Perspective. *Mongolian Terminology Studies,* 133–43. Ulaanbaatar: Mongolian Academy of Sciences.

Wissik, T. & T. Declerck, T. (2020). Using an Infrastructure for Lexicography in the Field of Terminology. *Terminologie & Ontologie: Théories et Applications Actes de la conférence TOTh 2019*, 365–379. Chambéry: Presses, 365–379.

Žagar Karer, M., & T. Fajfar, T.(2023). Terminological problems of terminology users: Analysis of questions in terminological counselling service on the Terminologišče website. Terminology 29(2): 78–102.