Originally published in: Calzolari, Nicoletta/Béchet, Frédéric/Blache, Philippe [et al.] (Eds.): Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), May 11-16, 2020, Palais du Pharo, Marseille, France. - Paris: European Language Resources Association, 2020. Pp. 3406-3413. URL: *http://www.lrec-conf.org/proceedings/lrec2020/index.html#3406*

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3406–3413
Marseille, 11–16 May 2020

Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN

Franciska de Jong[1], Bente Maegaard[2], Darja Fišer[3], Dieter Van Uytvanck[4], Andreas Witt[5]
[1,2,3,4,5]CLARIN ERIC, [2]University of Copenhagen, [3]University of Ljubljana and Jožef Stefan Institute, [5]Leibniz Institute for the German Language and University of Mannheim
[1,4]Utrecht, The Netherlands, [2]Copenhagen, Denmark, [3]Ljubljana, Slovenia, [5]Mannheim, Germany;
[1]f.m.g.dejong@uu.nl, [2]bmaegaard@hum.ku.dk, [3]darja.fiser@ff.uni-lj.si, [4]dieter@clarin.eu, [5]witt@ids-mannheim.de

Abstract

CLARIN is a European Research Infrastructure providing access to language resources and technologies for researchers in the humanities and social sciences. It supports the use and study of language data in general and aims to increase the potential for comparative research of cultural and societal phenomena across the boundaries of languages and disciplines, all in line with the European agenda for Open Science. Data infrastructures such as CLARIN have recently embarked on the emerging frameworks for the federation of infrastructural services, such as the European Open Science Cloud and the integration of services resulting from multidisciplinary collaboration in federated services for the wider domain of the social sciences and humanities (SSH). In this paper we describe the interoperability requirements that arise through the existing ambitions and the emerging frameworks. The interoperability theme will be addressed at several levels, including organisation and ecosystem, design of workflow services, data curation, performance measurement and collaboration. For each level, some concrete outcomes are described.

Keywords: CLARIN, language resources, research infrastructure, Open Science, interoperability, multidisciplinarity, EOSC

1. Introduction

More and more data is becoming available in digital formats, and research infrastructures are expected to provide their communities with access to the available resources. Thematically focused research infrastructures typically do so through a combination of generic infrastructural service components such as e.g., persistent identifiers and bitstream preservation, and domain-specific or thematic components that take both the content and the context of the resources into account. Nowadays most publicly funded infrastructural initiatives adhere to the Open Science agenda and express the ambition to provide access to data collections in line with the FAIR principles for data management (Wilkinson et al., 2016) and thus aim at maximizing Findability, Accessibility, Interoperability and Reusability.[1]

As underlined by several important bodies such as the European Commission (Hodson et al., 2018), the added value in such agendas and the underlying values in the way they stimulate the relevant communities to work *towards* the realization of the objectives, rather than considering features such as interoperability or reusability as an absolute measure for acceptance by the ecosystem. Interoperability can only be pursued effectively if it is not just targeted at technical levels, such as syntactic interoperability of metadata formats, but as a set of goals embedded in a culture that is characterized by attention for all social, political, and organizational factors that impact system-to-system performance.

CLARIN (Common LAnguage Resources and technology Infrastructure) is one of the pan-European infrastructures. It is strongly rooted in the humanities and the field of Natural Language Processing (NLP) and has the mission to create and maintain an infrastructure to support the sharing, use and sustainable availability of language data and tools for research in the humanities and social sciences (SSH) and beyond.[2] Since its early days, the CLARIN consortium has aimed at building both a technical infrastructure and a sustainable organisation (Broeder et al., 2008; Hinrichs and Krauwer, 2014) while adhering to the interoperability paradigm at a range of levels.

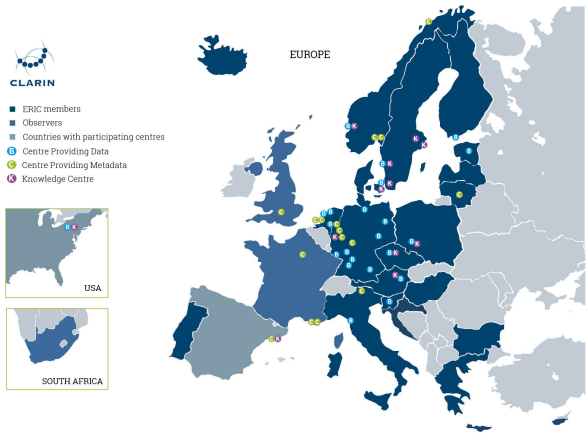

Figure 1: *Map of CLARIN members, observers, and participating centres by February 2020.*

CLARIN has always operated in line with the European agenda for Open Science[3] and it can be seen as an adopter of the FAIR data principles *avant la lettre* (de Jong et al., 2018). The focus on interoperability can be further illustrated along several dimensions, including the ambition to address the challenge of overcoming the obstacles stem-

[1]See also https://www.clarin.eu/fair

[2]See https://www.clarin.eu/content/vision-and-strategy

[3]See the 2016 Background note on Open Science available at https://ec.europa.eu/research/openscience/pdf/openaccess/background_note_open_access.pdf

ming from linguistic diversity that is to be faced by disciplines which use language materials as part of their scholarly workflows. With the growth of the number of participating countries, the number of languages for which data and analysis tools are integrated is steadily growing. This brings new requirements for the level of abstraction at which the resources are coded. But it also becomes more important to take into account the diversity of disciplines with an interest in the (re)use of the resources.

A salient feature of the CLARIN set-up is the federation of services. The building blocks of the CLARIN federation are the individual CLARIN centres. Centres have to conform to several technical and organisational principles and standards to ensure a seamless integration between the distributed resources available in the centres and the available services across the federation. To make sure that all criteria for syntactic and semantic interoperability are met, a two-step certification procedure is in place. After a centre has obtained certification by CoreTrustSeal,[4] the CLARIN Assessment Committee decides on the status of the centre. Interoperability is a crucial criterion, since federation can only lead to an effective common Research Infrastructure if the nodes are able to interact. By February 2020, 21 countries and over 50 centres were involved in CLARIN. See the map in Figure 1 for an overview.

Another perspective on interoperability comes into play with the more recently emerging European Open Science Cloud (EOSC). EOSC is envisaged to be an open and trusted environment for managing data from all research domains. EOSC will federate both existing and emerging data infrastructures and is meant to become the universal access channel to all registered data repositories and cloud-based services through which all European researchers will be able to access, use and reuse research outputs and data across disciplines.[5] In order for the CLARIN services to be suited for integration in EOSC a range of additional interoperability requirements on both technical and organisational level has to be faced.

The paper describes various aspects of interoperability in the CLARIN Research Infrastructure in more detail. Section 2. highlights the development and curation activities carried out by CLARIN in order to support scholarly workflows across disciplines and languages. Section 3. focuses on organisational and managerial issues. Some concluding remarks follow in Section 4.

2. Support for Scholarly Workflows

A crucial condition for the interoperability of language resources is their capability to interact or work together (Witt et al., 2009). Interoperability is prerequisite for any infrastructure and is even more challenging in case of a distributed data facility. In the design of the CLARIN infrastructure the idea that a seamless flow of data between web-based applications and services is crucial has therefore always been the guiding principle.

In line with its interoperability policy,[6] CLARIN has established the Virtual Language Observatory (VLO),[7] a registry of language resources in many languages based on the CMDI metadata standard (see Section 2.3. below). The VLO contains information about all language resources provided by the member countries, plus information from other registries that want to be visible through the VLO. In the following subsections CLARIN's implementation of interoperability geared towards infrastructural support for scholarly workflows will be described in more detail.

2.1. Interoperability of Tools

In an infrastructure for research data both static resources, i.e. the data, and tools to process the data, are in place. The tools of a distributed infrastructure must be accessible from different locations. Interoperability ensures that linguistic tools can be combined with language data in a common processing pipeline.

In CLARIN, web services have been put in place to encapsulate these tools and combine them in a common service-oriented architecture. The first CLARIN activity in this field led to the development of WebLicht (Hinrichs et al., 2010). The web-based linguistic chaining tool provides an environment that allows the processing of textually given resources in a pipeline architecture. WebLicht has been applied in different processing tasks and for resources of different languages (Schmidt et al., 2016; Çöltekin, 2015).

More recently, a tool has been developed (Zinn, 2016) to provide guidance on which service is recommended for which data, known as the Language Resource Switchboard.[8] The basic assumption behind the Switchboard is the focus on achieving a fairly basic but well-tested and robust level of interoperability, based on a lightweight and modular approach. It acts as a simple forwarding application that, based on the URL of an input file and a few simple parameters (language, mimetype, task), allows the user to select relevant NLP web applications that can analyze the input provided.

Integrating an NLP web application with the Switchboard is fairly simple: basically the receiving application needs to accept a URL as an input parameter. The processing of the input and the rendering of the results are fully delegated to the NLP web application. While the simplicity of this approach does not allow for more advanced scenarios, like creating pipelines using several registered applications, processing large amounts of data or processing multilingual data, it does act as a low-threshold environment for testing real-world interoperability.

One particular feature of the Switchboard is the limited amount of metadata needed to process an incoming file. With its built-in ability to guess the format (mime-type) and content language of an incoming file,[9] it can also deal with many files that do not come with a rich metadata description. This makes it possible to call the Switchboard from many applications, even if they do not have access to detailed descriptions of a file, such as B2DROP.[10]

[4] See http://coretrustseal.org

[5] See https://www.eosc-portal.eu

[6] See https://www.clarin.eu/content/interoperability

[7] See http://vlo.clarin.eu

[8] See https://switchboard.clarin.eu

[9] Based on the Apache Tika library. Currently more complete libraries such as FITS are investigated.

[10] B2DROP is a generic service offered by EUDAT, based on

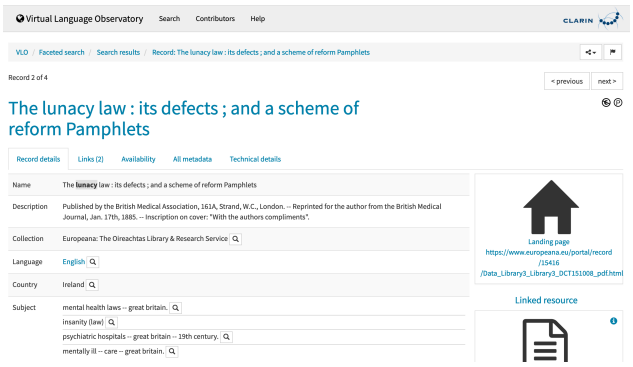

Figure 2: *The representation of a digitized pamphlet based on Europeana metadata in the VLO.*

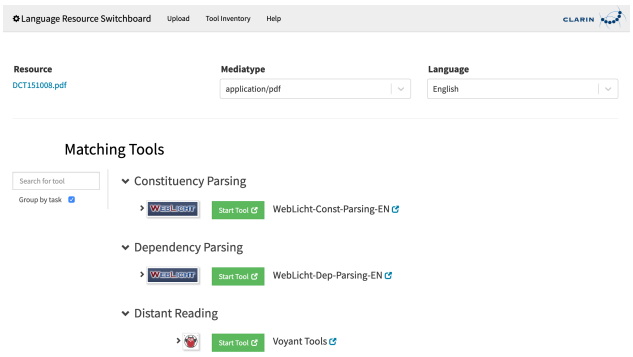

Figure 4: *The Language Resource Switchboard interface.*

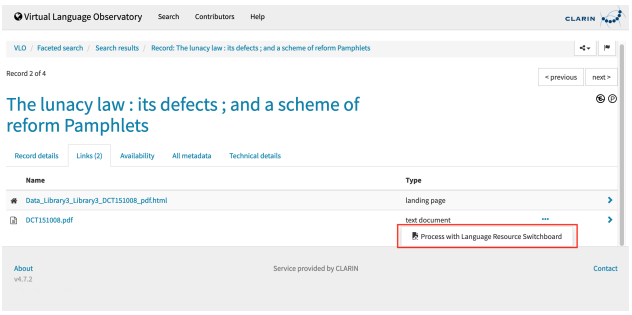

Figure 3: *Calling the Language Resource Switchboard for a specific file.*

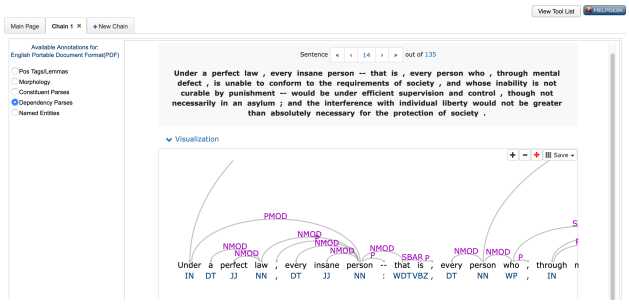

Figure 5: *The result of processing the selected file, in this case using the WebLicht dependency parser.*

Since the Switchboard is easy to connect to, both for the applications that need to call it and for the applications that it calls itself, a multiplicator effect arises – each registered file can be forwarded to each registered NLP tool, as long as the input parameters match. In that sense it can be seen as a good demonstration platform for the impact of interoperability.

This effect can be illustrated by looking into the connection between the metadata provided by Europeana[11] on the one hand, and the Virtual Language Observatory on the other — as a portal to explore this metadata and as a caller of the Switchboard. After identifying a relevant language resource with the VLO (e.g. a scanned pamphlet as a PDF that includes the OCRed text, see Figure 2), one can call the Switchboard (see Figure 3 and 4) to perform a specific NLP task (e.g. dependency parsing, see Figure 5). Eventually the selected target application is called and presents the results, as illustrated in Figure 5.

2.2. Interoperability of Data Sets

In addition to top-down development efforts to ensure interoperability of CLARIN services, several bottom-up initiatives which assess interoperability issues with the CLARIN resources from the perspective of the end user have been carried out as well (Odijk, 2014; Lušicky and Wissik, 2017; Sanders, 2017). They all identify a clear need for more comprehensive metadata on the provenance and annotation of the resources, standard formatting, uniform concordancing and text analytics options that enable not only the use of corpora to tackle research questions the corpus was compiled for but also allow comparisons across corpora and use of corpora across disciplines and methodological approaches as well as for cross-disciplinary and trans-national comparative research. This is a much more challenging task and surveys conducted in the framework of the CLARIN Resource Families (Fišer et al., 2018) have identified several issues suggesting room for improvement.

First and foremost, some of the deposited resources or their most recent versions cannot be identified by the uninitiated researchers through the VLO due to lacking, idiosyncratic or vernacular names, keywords or description fields. The second type of issues is the incomplete documentation for many of the corpora that can range from basic fields such as corpus size, period (i.e., when the source texts were originally written or recorded), linguistic annotation, or license information, but can also highlight the lack of awareness for cross-disciplinary needs, such as information on the directionality of translations in parallel corpora which is not very important for machine translation researchers but is crucial for translation studies researchers (Cartoni et al., 2013, e.g.). The third type of issues is the different granularity of data annotation which can significantly hinder or even prevent comparative research, such as the uneven availability of speaker information (name, sex, role, party affiliation), or the use of different tagsets for the same annotation layers (e.g., morphosyntactic tagging) across parliamentary corpora. A major bottleneck for interoperability

Nextcloud. For more details see https://eudat.eu/services/b2drop

[11] See https://www.clarin.eu/europeana

for a lot of SSH researchers is also the fact that corpora of the same type for different languages are made available through different concordancers that not only offer different analytical functionalities (e.g., keyword lists) but also use different statistics for the same functionalities (e.g., collocations), which are not always explicitly documented, making it difficult to compare the results across corpora. Mechanisms to improve the identified issues with metadata and documentation of legacy data sets are already in place via GitHub. For new data sets, guidelines, best practices and awareness raising activities are planned that can help prevent avoidable interoperability gaps.

2.3. Metadata Curation

As metadata is an important cornerstone of an interoperable and FAIR data infrastructure, metadata quality is also a key factor for successful use, reuse and repurposing of language resources. As described in Section 2., CLARIN makes use of Component Metadata (CMDI), harvesting metadata records from the distributed centre repositories and bringing the results together in the Virtual Language Observatory search portal (Van Uytvanck et al., 2012).
This CLARIN-specific process of gathering and processing metadata has been going on for almost a decade, and over time several categories of issues with metadata quality have come up. Below, we provide a non-exhaustive list as illustration:

- Descriptions might be incorrect, incomplete or completely missing.
- Some fields might not be using a controlled vocabulary, leading to inconsistencies.
- All kind of technical and formal issues, like metadata that does not validate according to the XML schema.

To deal with these issues, CLARIN has been following several routes towards achieving a higher metadata quality and as a consequence increased interoperability:

- encouraging the use of controlled environments for authoring metadata,
- providing best practices documents,[12]
- performing automated and regular quality checks.

To enable automated checks, Ostojic et al. (2017) developed the so-called Curation Module.[13] This application checks all metadata records harvested and checks a broad range of criteria (schema validity, the occurrence of fields like language and availability, etc.). Based on these checks, an overall quality score is calculated.
It should be noted that these scores should not be seen as a final outcome, but merely as an indicator for potential improvements. They are mainly used as a trigger to approach metadata providers with a request to look into particular issues that have been detected. Based on the scores, a dialogue is initiated with the aim to enable the user to curate the metadata at the source if this can be done with reasonable efforts.
In certain cases, the original metadata cannot be curated. Then the VLO can perform some post-hoc curation steps, e.g., by mapping the value of a field to an equivalent entry of a controlled vocabulary.
A very specific part of metadata curation is checking the validity of links. Such links can take the form of a uniform resource identifier or a persistent identifier – e.g., a handle[14] or a digital object identifier.[15] In any case, practice has shown that many links cannot be resolved without issues: in November 2019, these amounted to around 10% of the 5.2 million links in the VLO. Having this information at hand is key to detecting metadata and data issues proactively. Therefore the Curation Module includes a specific component that regularly crawls all the links encountered and stores the result of accessing these links into a database. Link checkers exist for other infrastructures as well: DataCite has one since the end of 2018 (Dasler, 2018) and Europeana is piloting similar approaches (personal communication, August 2019). From a global perspective it would make sense to federate the respective databases with results to avoid redundant checks and to increase data consistency. This is a theme that has some potential to be developed within the European Open Science Cloud.

2.4. Encoding standards

In line with CLARIN's strategic priority to increase interoperability among data sets and tools, CLARIN has recently developed instruments to coordinate standardisation of formats of specific data types across the infrastructure which are applicable to both legacy and newly developed resources. As the first example parliamentary corpora have been selected because they are already available for most CLARIN countries,[16] and have great potential for research in multidisciplinary and multilingual settings. In addition, this data type also has several advantages compared to many others:

- There are no copyright or data protection issues with source data.
- Source data is typically readily available in digital form.
- Text and speaker metadata are readily available.

Nevertheless, parliamentary data come with its own set of issues:

- Parliamentary proceedings are subject to country-specific (and often not transparent) procedures.
- Digital sources of parliamentary data are distributed in many different formats and are structured quite differently.
- Parliamentary corpora are mostly compiled by computational linguists who are often not aware how proceedings are produced and how they are used by SSH

[12] See the CMDI best practices guide and Eckart et al. (2017)
[13] See https://www.clarin.eu/curation
[14] See https://www.handle.net
[15] See https://www.doi.org
[16] See http://clarin.eu/resource-families/parliamentary-corpora

researchers beyond computational and corpus linguistics.

As part of the initiative known as ParlaFormat,[17] a proposal has been developed for a common standard for the encoding of parliamentary metadata, including the speakers and political parties, the structure of the corpus, the encoding of the speeches and notes, linguistic annotation and multimedia (Erjavec and Pančur, 2019). The proposal is available at GitHub[18] and comprises the ODD specification,[19] the derived HTML guidelines and XML schemas, and example documents. As next steps, a shared task is planned in order to test the proposed common standard, as well as the development of conversion scripts from and to other standards that are commonly used for parliamentary data, such as RDF[20] or Akoma Ntoso.[21] Consolidated parliamentary corpora will substantially boost cross-disciplinary and trans-national research agendas, will stimulate application and further development of interoperable NLP tools, and will enable linking parliamentary corpora with background documents (e.g., legislation), external knowledge sources (e.g., taxonomies) and other related research data sets (e.g., party manifestos, campaign speeches, social media, broadcasts).

2.5. Enabling Replication

The increased awareness of the importance of reproducible science and the need to preserve the underlying data to enable reuse and replication has resulted in initiatives such as the Research Data Alliance and the adoption of the concept of FAIR data. (See Section 1. Introduction).

In the field of Language Resources and NLP, this theme is especially addressed in the 4REAL workshops (Branco et al., 2016; Branco et al., 2018) and in the subsequent REPROLANG shared task. The latter is centered around the concept of a "replication paper": participants adopt a previously described NLP task and try to replicate it, describing their results in a new paper. Next to such a publication, the tools and the data are also published in the form of a Docker container that can relatively easily be executed by anyone else. The outcomes of this shared task will be presented at the LREC2020 conference.

Often, the replication of NLP tasks require considerable amounts of computing power, especially in the case of machine learning based on deep learning techniques. This is where the expertise of research infrastructures like CLARIN ERIC about containerised deployments[22] is connected to the processing capabilities offered through the European Open Science Cloud.

Despite the availability of cloud and container technology that potentially simplifies replication, it is clear that in practice replication is not always so straightforward. Documenting scientific workflows in an executable fashion demands a considerable amount of additional effort. Nevertheless, the expectation is that this additional work is rewarded with suitable academic credits for the researchers taking up the gauntlet.

3. Interoperability at the organisational level

As articulated above interoperability plays a role at a range of levels. In this section we will address how the dynamics coming from multidisciplinary collaboration in the context of the European Strategy Forum for Research Infrastructures (ESFRI)[23] affects the organisational strategy of CLARIN.

3.1. Interoperability across Research Infrastructures

Recently, several policy instruments have been introduced to stimulate research infrastructures (RIs) to reinforce collaboration beyond the thematic domain in which they have their primary communities of use. This collaboration in so-called 'clusters' is assumed to contribute to stimulate multidisciplinary research and the potential to effectively address societal challenges and the potential for innovation. For the social sciences and the humanities (SSH) this has led to an increased cooperation between the five thematic research infrastructures that are listed in the ESFRI Roadmap 2018[24] as mature research infrastructures (so-called 'Landmarks'[25]) and have ERIC status.[26]

With funding from the H2020 programme,[27] these ERICs have joined forces with a number of other parties, including LIBER,[28] to work towards the integration of part of their service offer into what has become known as the Social Sciences and Humanities Open Cloud (SSHOC)[29] and the European Open Science Cloud. SSHOC is one of five cluster projects which are preparing a thematic or domain-specific contribution to EOSC.[30] It will leverage and interconnect existing and new infrastructural facilities offered by the project partners to foster synergies across disciplines and expedite interdisciplinary research and collaboration. Not surprisingly the interoperability issues related to the

[17] See https://www.clarin.eu/event/2019/parlaformat-workshop

[18] See https://github.com/clarin-eric/parla-clarin

[19] ODD stands for: One Document Does it all; see also https://wiki.tei-c.org/index.php/ODD and Romary and Riondet (2018)

[20] RDF stands for: Resource Description Framework, a standard model for the interchange of web data; see https://www.w3.org/RDF

[21] Akoma Ntoso is an XML standard for parliamentary, legislative and judiciary documents. See also http://www.akomantoso.org

[22] See https://www.clarin.eu/reproducibility

[23] See https://www.esfri.eu/about

[24] ESFRI regularly issues a strategy report on the European landscape of research infrastructures and the ESFRI vision of the evolution of Research Infrastructures in Europe. The most recent edition can be found at http://roadmap2018.esfri.eu

[25] Apart from CLARIN, these research infrastructures are CESSDA, DARIAH, ESS and SHARE.

[26] 'ERIC' is short for European Research Infrastructure Consortium. For an overview of all ERICs established (including the five RIs in the SSH cluster) and the links to their websites, see the information pages of ERIC Forum on the ERIC Landscape.

[27] The SSHOC project was proposed in response to H2020 call INFRAEOSC-04-2018.

[28] See http://www.libereurope.eu

[29] See https://sshopencloud.eu

[30] See the announcement on the EOSC Portal (link).

planned pilot studies in the SSHOC work plan all relate in one way or another to the infrastructural challenges that are shared among the consortium members:

- the distributed character of the data infrastructure involved,
- the multilinguality of (a part of) the datasets,
- the need for arranging secure access to sensitive data.

The SSHOC project aims to realize an open platform for researchers from social sciences and the humanities where data, tools, and training materials are available and accessible. The SSHOC consortium's work plan covers the entire data life cycle, from data creation and curation to optimal data reuse. In order to test the functionality of the SSH Open Cloud, a series of pilot projects is planned that will focus on the support of multidisciplinary collaboration for the following research areas: migration and mobility, election studies as well as culture and heritage. In addition, training and other outreach activities are undertaken. SSHOC will also develop a governance model suited to the services that are rooted in social sciences and the humanities, and which will be offered through a common SSHOC platform or the emerging EOSC platform.
For CLARIN, the clustering in SSHOC is yet another context in which the ongoing effort towards optimizing the service we offer for comparative research based on the CLARIN Resource Families (Fišer et al., 2018) can be coordinated. In particular for the work on better integration of parliamentary data and the support for the development of methodologies for working with heterogeneous data sets, the SSHOC project has a big potential for impact, as it is calling for collaboration with political scientists and linguists, the coupling of parliamentary corpora with other political research data sets, such as the party manifestos,[31] and the integration of textual data and quantitative data from polls.
Another boost that SSHOC is envisaged to bring is expected for the services that CLARIN offers for researchers working with (recordings of) interview data. Interviews represent a data type that can be complementary to what is available from surveys conducted by social scientists. The first workshop has been organized during the DH2019 conference in Utrecht, focusing on support for the variety of scholarly approaches and research paradigms in creating and processing interviews. Barriers for the uptake of available tools for e.g., transcription in existing workflows may stem from not overlapping terminology, lack of digital skills, limited experience with handling personal data and disjoint publication cultures, but as the workshop demonstrated, a lot can be gained from training sessions that take the diversity into account.[32]

3.2. Measuring Progress – Key Performance Indicators

There is growing consensus that it is important to assess the value and performance of research infrastructures, and therefore performance monitoring systems have been proposed. Key Performance Indicators (KPIs) are widely discussed and used for describing the extent to which a research infrastructure is achieving its objectives in quantitative terms. KPIs are not only considered useful for measuring progress, but also for collecting feedback on the strategy of a research infrastructure. Consequently, in 2018 CLARIN ERIC started work on a framework for KPIs that would help the CLARIN community to describe the progress in developing and operating the research infrastructure for language resources in quantitative terms. (Obviously, until then progress was also measured annually, but not based on formal KPI framework.)
As mentioned by Kolar et al. (2019), in order to ensure effectiveness and feasibility of implementation, KPIs should be:

- Relevant – i.e. closely linked to the objectives to be achieved;
- Accepted – e.g., by staff and stakeholders;
- Credible for non-experts, unambiguous and easy to interpret;
- Easy to monitor – e.g., data collection should be possible at low cost;
- Robust – e.g., against manipulation.

In line with this recommendation the point of departure for the CLARIN KPI framework has been the objectives and activities as listed in Article 2 of the CLARIN ERIC Statutes.[33] For each objective in the statutes a KPI and a method for measurement have been specified. Eventually 12 KPIs have been identified as suitable for measuring the performance related to the objectives in its statutes. Out of these, one KPI is related to the use and promotion of standards and mappings that are essential for interoperability. Also some of the other KPIs have a direct or indirect relation to the interoperability issues addressed in this paper, such as the number of certified centres, and the number of formalized collaborations with other research infrastructures and organisations from the cultural heritage domain. As not all performance can be measured quantitatively, progress is always also described through narratives, which are considered critically complementary to the metrics adopted, especially when it comes to measuring the performance related to interoperability, and establishing and supporting multidisciplinary collaboration.
Apart from this focused initiative from CLARIN itself, for some time ESFRI has been working on a framework for cross-infrastructure KPIs to be used for the monitoring of all research infrastructures under the ESFRI umbrella. Obviously, this is a further challenge compared to a single-infrastructure KPI system, because the envisaged KPI system needs to cut across disciplines, from physics and energy to social sciences and humanities, across single-sited and distributed infrastructures, and across physical and virtual infrastructures. Therefore the ESFRI Monitoring group could not depart from actual objectives that were shared

[31] See https://manifestoproject.wzb.eu

[32] See https://oralhistory.eu/workshops/dh2019 for more details on the SSHOC oral history workshop held at DH2019.

[33] See the CLARIN ERIC statutes, HDL:11372/DOC-143

across infrastructures (there are some, but not many), but rather had to go for slightly more general objectives that could be agreed upon by all or many infrastructures, such as the aim to enable and support scientific excellence. Eventually ESFRI came up with 20 KPIs which, to the extent that they make sense for a specific infrastructure, may be used by all ESFRI infrastructures[34]. The emergence of a common KPI framework illustrates how the CLARIN policy is influenced by ESFRI's aim to cover all infrastructures in a discipline-independent manner. At the same time it shows that a cross-disciplinary approach cannot stand alone; focused KPIs which measure the crucial features for each individual infrastructure, such as interoperability in the case of CLARIN, remain of crucial importance.

4. Concluding Remarks

With the increasingly multidisciplinary contexts in which data-driven research, machine learning and artificial intelligence are becoming commonplace, there is a strong need for novel frameworks for integrated processing of multiple heterogeneous data types on the one hand and for deepening insights in the validity of analysis outcomes and interpretability of the models on the other.

For multidisciplinary collaborative initiatives in SSH there is a big potential for impact from use cases based on applying mixed methods to linguistic content that is interlinked with survey and other kinds of structured data. With increased interoperability of data and analytical techniques, the promise of impact on scientific excellence and innovation may be brought closer to realization.

Investment in well-organized support for the integrated processing and interpretation of heterogeneous data is therefore crucial, as well as establishing adequate models for multidisciplinary collaboration in which the potential obstacles for conceptual interoperability are taken into account. With the NLP processing of data made available through Europeana, we presented an example of how the CLARIN infrastructure contributes with small but concrete implementation steps to improved interoperability in practice.

Acknowledgements

The work reported in this paper has been supported by the countries participating in CLARIN ERIC, and it has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777536 for project EOSC-hub, and grant agreement No. 823782 for project SSHOC. Steven Krauwer participated in the development of the CLARIN framework for KPIs as described in Section 3.; Twan Goosen made the conversion of Europeana metadata as described in Section 2. possible.

Bibliographical References

Branco, A., Calzolari, N., and Choukri, K. (2016). Proceedings of the workshop on research results reproducibility and resources citation in science and technology of language.

Branco, A., Calzolari, N., and Choukri, K. (2018). LREC 2018 workshop proceedings: 4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language, 12 May 2018, Miyazaki, Japan. European Language Resources Association.

Broeder, D., Nathan, D., Strömqvist, S., and Van Veenendaal, R. (2008). Building a federation of language resource repositories: the DAM-LR Project and its continuation within CLARIN. In *Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. ELRA.

Cartoni, B., Zufferey, S., and Meyer, T. (2013). Using the europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics*, 27:23–42.

Çöltekin, Ç. (2015). Turkish NLP web services in the WebLicht environment. In *Proceedings of the CLARIN Annual Conference*.

Dasler, R. (2018). Link checker is here. DOI:10.5438/vywf-6s91.

de Jong, F., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, D. (2018). CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3259 – 3264. ELRA.

Eckart, T., Goosen, T., Haaf, S., Hedeland, H., Ohren, O., Van Uytvanck, D., and Windhouwer, M. (2017). Component Metadata Infrastructure Best Practices for CLARIN. Budapest, Hungary.

Erjavec, T. and Pančur, A. (2019). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings.

Fišer, D., Lenardič, J., and Erjavec, T. (2018). CLARIN's key resource families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1525–1531.

Hinrichs, E., Hinrichs, M., and Zastrow, T. (2010). WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29, Uppsala, Sweden, July. Association for Computational Linguistics.

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., and Wittenburg, P. (2018). Turning FAIR into reality. Final report and action plan from the EC expert group on FAIR data. DOI:10.2777/1524.

Kolar, J., Harrison, A., and Gliksohn, F. (2019). Key performance indicators of Research Infrastructures. http://www.accelerate2020.eu/key-performance-indicators-of-research-infrastructures/.

Lušicky, V. and Wissik, T. (2017). Discovering resources in the vlo: A pilot study with students of translation studies. In *Selected papers from the CLARIN Annual Confer-*

[34]The final report of the ESFRI Working Group on the Monitoring of Research Infrastructures was published in December 2019.

ence 2016, number 136, pages 63–75, Aix-en-Provence, France. Linköping University Electronic Press.

Odijk, J. (2014). Discovering resources in CLARIN: Problems and suggestions for solutions.

Ostojic, D., Sugimoto, G., and Ďurčo, M. (2017). The Curation Module and Statistical Analysis on VLO Metadata Quality. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, France*, pages 90–101. Linköping University Electronic Press.

Romary, L. and Riondet, C. (2018). EAD ODD: a solution for project-specific EAD schemes. *Archival Science*, 18:165 –184.

Sanders, W. (2017). Focus group on user involvement. HDL:11372/DOC-139.

Schmidt, T., Hedeland, H., and Jettka, D. (2016). Conversion and Annotation Web Services for Spoken Language Data in CLARIN.

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: the virtual language observatory. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA.

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. *Scientific Data*, 3(160018).

Witt, A., Heid, U., Sasaki, F., and Sérasset, G. (2009). Multilingual language resources and interoperability. *Language Resources and Evaluation*, 43(1):1 – 14.

Zinn, C. (2016). The CLARIN Language Resource Switchboard. In *Abstracts of the CLARIN Annual Conference 2016*, Aix-en-Provence, France.