The CLARINO Bergen Centre: Development and Deployment

Koenraad De Smedt, Gunn Inger Lyse, Rune Kyrkebø, Hemed Al Ruwehy,
Øyvind Liland Gjesdal, and Victoria Rosén
University of Bergen
Bergen, Norway
{desmedt|gunn.lyse|rune.kyrkjebo|hemed.ruwehy|oyvind.gjesdal|victoria}
@uib.no

Paul Meurer
Uni Research Computing
Bergen, Norway
paul.meurer@uni.no

Abstract

The CLARINO Bergen Centre (Norway) provides a language resource repository, corpus and treebank services and metadata management services. We explain the motivation for using the LINDAT repository software as a model and describe the cloning and adaptation of that software for the CLARINO Bergen Repository. We also describe how the other centre services addressing CLARIN goals have been integrated into the centre, focusing on the steps taken to adapt the INESS treebanking service to CLARIN standards.

1 Introduction

The CLARIN ERIC is a distributed research infrastructure, realized in the form of a network of centres which offer access to language data and tools and online services for search, analysis, visualization and other processing. The Norwegian research infrastructure project CLARINO is constructing a network of CLARIN centres in Norway. In this context, the CLARINO Bergen Centre was established through a cooperation between the University of Bergen (Norway) and Uni Research Computing (a research institute, also in Bergen). This centre was awarded CLARIN centre type B status in January 2016.[1]

Every CLARIN centre of this type is required to run a data repository in accordance with certain criteria for good practice and compatibility with the CLARIN infrastructure.[2] The first section of this paper exemplifies the value of sharing technical solutions within the CLARIN community by describing how the repository software from another CLARIN member was cloned and adapted for the CLARINO Bergen Repository.

In addition to providing a repository, a CLARIN centre may provide services for language data management, analysis, visualization, etc. In the CLARINO Bergen Centre these services include the INESS treebank management, annotation and search system (Meurer et al., 2013; Rosén et al., 2012), the Corpuscle corpus management and search system (Meurer, 2012a), and the COMEDI component metadata editor (Lyse et al., 2015). For such services, it is good practice to adopt the same criteria and standards as for the repository, in particular as regards authentication, metadata and persistent identifiers (PIDs). This adoption of CLARIN criteria and standards for the CLARINO Bergen Centre services is described in the second part of the paper, focusing in particular on INESS, as a useful example of how existing infrastructural systems can be integrated in the CLARIN ecosystem.

[1] http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-116

[2] http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-78

2 Repository development

2.1 Motivation

The University of Bergen Library (UBL), which participates in CLARINO, was assigned the task of implementing and running a repository, initially in order to manage the resources at the University of Bergen. In 2013, UBL decided to use the open software application DSpace,[3] as modified by the Institute of Formal and Applied Linguistics at the Charles University in Prague for their LINDAT repository (Mišutka et al., 2015).[4]

The motivations for this choice were the following. UBL had some previous experience with DSpace for the implementation of the Bergen Open Research Archive.[5] This experience showed that DSpace is a functional and stable platform which is open source and well maintained by an active user community. It provides long term storage and linking, suitable authentication mechanisms, handling of licenses for downloading of resources, PID support, and an OAI-PMH endpoint at which metadata can be harvested.

Moreover, LINDAT added certain features to make the DSpace software satisfy some essential CLARIN B centre requirements, such as support for CMDI metadata. They also added a method for license handling which enables the signing of licenses by users. The LINDAT software is in an open source software repository at GitHub.[6]

Furthermore, UBL attended the LINDAT presentation at the CLARIN meeting in June 2013 in Utrecht where the Prague group was willing to share their software and knowledge. Some strengths of the CLARIN community are the use of open source software and the mobility actions which can be used to get assistance across borders. For these reasons it was decided to proceed directly with implementing DSpace/LINDAT.[7]

2.2 Installation and adaptation

A mobility action funded by CLARIN enabled Jozef Mišutka to travel from Prague to Bergen in August 2013 in order to help set up the initial system. This mobility action was probably far more efficient than attempting to communicate at a distance. Indeed, within a few days, the first version of the installation was up and running.

The next step consisted of local modifications and configurations, which mainly affected the routines for authentication and persistent identifiers (PIDs), the explanatory textual parts, and the graphical profile.

Federated single sign-on was installed by UBL in cooperation with the IT-department at the University of Bergen, with helpful guidance from the Shibboleth setup guide by Sander Maijers, published by CLARIN ERIC.[8]

A Handle[9] configuration was set up by the UBL in order to assign PIDs to resources. There was no need to develop a PID server, since it comes as a built-in feature in DSpace. UBL bought the handle prefix and configured the PID server as described in their documentation. From there, the installation was repeated several times for upgrading, and further local customizations were made, mostly to the user interface.

A graphical profile for CLARINO was designed by Talan Memmott (University of Bergen), with results as shown in Figure 1. The color scheme and logos were designed to be partly compatible with those of CLARIN but differ slightly so as to express the CLARINO branding.

The explanatory texts, such as the description of the site, terms of service, policies and submission lifecycle, were adapted to the CLARINO context, but remained largely similar to those of LINDAT as they reflect general CLARIN policies.

[3] `http://www.dspace.org/`

[4] `http://lindat.mff.cuni.cz/`

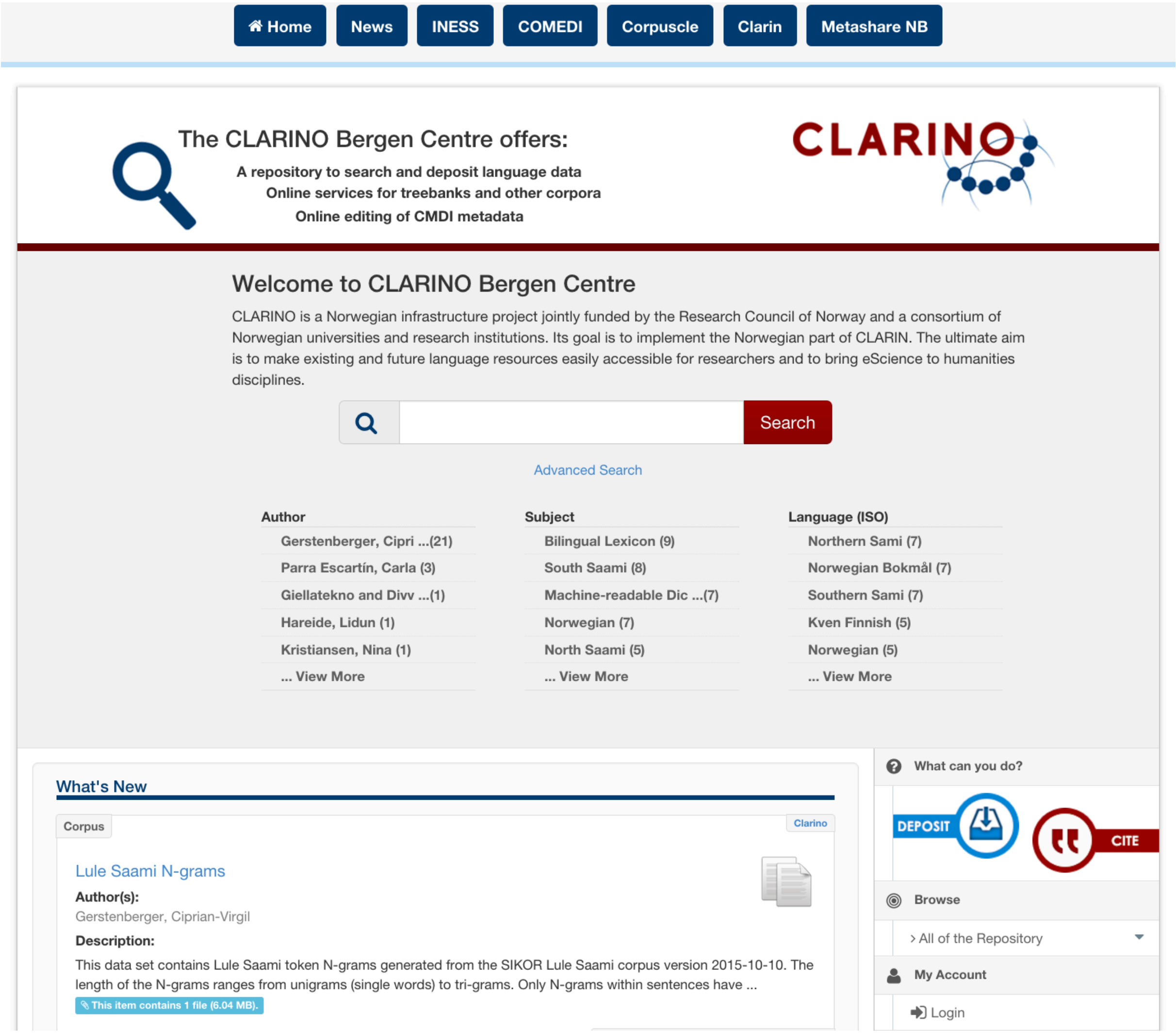
[5] `http://bora.uib.no`

[6] `https://github.com/ufal/lindat-dspace`

[7] In the future we might look at how FEDORA is implemented both in the CLARIN community and elsewhere to build repository infrastructure.

[8] Sander Maijers: Your own Shibboleth Service Provider within the Service Provider Federation. `https://cdn.rawgit.com/clarin-eric/SPF-tutorial/master/Shib_SP_tutorial.html`

[9] `http://handle.net`



The CLARINO Bergen Centre offers:

A repository to search and deposit language data
Online services for treebanks and other corpora
Online editing of CMDI metadata

Welcome to CLARINO Bergen Centre

CLARINO is a Norwegian infrastructure project jointly funded by the Research Council of Norway and a consortium of Norwegian universities and research institutions. Its goal is to implement the Norwegian part of CLARIN. The ultimate aim is to make existing and future language resources easily accessible for researchers and to bring eScience to humanities disciplines.

Author	Subject	Language (ISO)
Gerstenberger, Cipri ...(21)	Bilingual Lexicon (9)	Northern Sami (7)
Parra Escartín, Carla (3)	South Saami (8)	Norwegian Bokmål (7)
Giellatekno and Divv ...(1)	Machine-readable Dic ...(7)	Southern Sami (7)
Hareide, Lidun (1)	Norwegian (7)	Kven Finnish (5)
Kristiansen, Nina (1)	North Saami (5)	Norwegian (5)
... View More	... View More	... View More

What's New

Corpus — Clarino

Lule Saami N-grams

Author(s):
Gerstenberger, Ciprian-Virgil

Description:
This data set contains Lule Saami token N-grams generated from the SIKOR Lule Saami corpus version 2015-10-10. The length of the N-grams ranges from unigrams (single words) to tri-grams. Only N-grams within sentences have ...

This item contains 1 file (6.04 MB).

What can you do?

Browse
› All of the Repository

My Account
Login

Figure 1: Main page of the CLARINO Bergen Centre website.

To accentuate the prominent role of the repository, it was decided that the main URL for the centre[10] would be an immediate entry to the repository. This is different from the front page of LINDAT, which requires an extra click to get to the repository. The main page of the centre, shown in Figure 1, also has a highly visible top menu with links to the other parts of CLARINO: the CLARINO news blog, the INESS treebanking infrastructure, the COMEDI component metadata editor, and the Corpuscle advanced corpus management and search system. There are also links to the CLARIN ERIC website and the Norwegian META-SHARE node.

2.3 Metadata handling in the repository

The Component Metadata Initiative (CMDI) (Broeder et al., 2010) has led to a standard with the benefit of modularity through reusable components and standard profiles. The basic building blocks of CMDI are *components* which may consist of sets of *elements* and other components. CMDI is flexible in that the user can choose any set of components that together constitute a CMDI *profile*. At the same time, the reuse of components offers a degree of stability, since equal components may then appear in a number of individual metadata profiles. Existing CMDI profiles and components are stored in the Component Registry of the Component Metadata Infrastructure[11] as XML files. In the CLARIN infrastructure, the CMDI format is generally recommended and B centres are required to deliver harvestable CMDI metadata.

The handling of metadata in CMDI format represents a challenge for a DSpace repository, since CMDI fields are hierarchically structured while DSpace internal metadata fields represent a flat structure. Moreover, filling out metadata in DSpace does not handle arbitrary CMDI profiles from the component registry. The LINDAT extensions, however, facilitate the use of CMDI metadata in the repository. A CMDI metadata file can either be imported by the data depositor as part of the upload process, or it can be imported afterwards by the repository administrator. Once uploaded, the CMDI metadata are harvestable at the repository's metadata harvesting endpoint. The CMDI metadata are also available for the user by a click on the *CMDI* button in the citation box on the *View Item* page. The repository thus handles CMDI format metadata both for upload and export, while at the same time operating with the ordinary set of DSpace metadata fields. On export from DSpace the harvester module checks if there is an XML file present in the metadata bundle for the item. If so, this entire file is exported as metadata, instead of using the contents of the DSpace internal metadata fields.

We see some technical challenges with this situation and ideally we wish to relate to only one metadata set for each repository item. We encourage the use of the COMEDI metadata editor which is the most flexible for the production of CMDI metadata files. To create new metadata, CLARINO has developed the metadata editor COMEDI, which is now also a part of the CLARINO Bergen Centre services (Lyse et al., 2015). COMEDI is a web-based editor for CMDI-conformant metadata which supports the creation of new CMDI metadata files, cloning of existing files and upload and modification of existing metadata. A metadata file in COMEDI can be exported as a CMDI XML file and can be harvested with OAI-PMH. Currently, the workflow is to export an XML file from COMEDI which is then manually uploaded to DSpace either as part of the first upload of the item, or as a later import. A tighter technical integration between COMEDI and the repository should be achievable. Since COMEDI offers an OAI-PMH endpoint, a possible solution might be that COMEDI uses the DSpace REST API to post files to the repository.

3 Integration of treebank services

3.1 Infrastructural initiatives for treebanks

The rich annotation in treebanks makes them good sources for empirical research on syntax and the lexicon, for work on parsers, and to some extent also for exploring semantics, for information extraction, and for other 'deep' processing. The accessibility of treebanks is therefore important for several target researcher groups in CLARIN.

Although hundreds of treebanks exist which are potentially useful for research, their effective exploitation has until recently often been impeded by practical issues regarding distribution, access, metadata,

[10] `http://clarino.uib.no`

[11] `http://catalog.clarin.eu/ds/ComponentRegistry/`

licensing and use. Search within treebanks has often required downloading the data to one's own computer as well as downloading and installing standalone tools. Furthermore, query languages and tools are often specific to certain annotations and formats. Such limitations are typical of the kinds of problems that CLARIN in general wants to see solved.

In recent years, some treebanking efforts linked to CLARIN projects have started to address these issues. For instance, whereas the standalone tool Dact[12] already provides a user-friendly alternative to the earlier Alpino treebank tools (van Noord et al., 2013), online search alternatives for these tools have also become available, such as the example-based Gretel (Vandeghinste and Augustinus, 2014) and PaQu,[13] which is especially handy for relations between word pairs.

Access to the Czech treebanking resources and services has also considerably evolved. The distribution of treebanks in the LINDAT repository (based on DSpace) has become well integrated in the overall CLARIN architecture by the consistent use of CMDI metadata (Broeder et al., 2010), PIDs, federated authentication and license handling. The current LINDAT infrastructure offers a wide selection of dependency and constituency treebanks for different languages which can be individually searched and visualized through its online service PML Tree Query.[14]

Taking another approach at CLARIN integration, the TüNDRA[15] web tool for treebank research is accessible online in WebLicht (Martens, 2013). It provides federated authentication, browsing, search and visualization for TüBA treebanks (Telljohann et al., 2012) and some other treebanks with constituency or dependency annotation. WebLicht also offers a service for building one's own parsebank (De Kok et al., 2014).

3.2 INESS as a relevant infrastructure for CLARIN

INESS (Infrastructure for the Exploration of Syntax and Semantics) is similar to the efforts described above in its goal of making treebanks more accessible, but it handles a wider range of treebank types and online services. INESS hosts treebanks of many current annotation types and formats. This includes structural representations in Lexical Functional Grammar (LFG) and Head-Driven Phrase Structure Grammar (HPSG), besides constituency annotation and three current flavors of dependency annotation. It also handles parallel treebanks, even those having different annotation types on each side.

INESS currently provides access to more than 200 treebanks in 48 languages. Since the average user will not be interested in all of these, INESS offers treebank selection based on user choices, which currently include language, collection, annotation type, and linguality (monolingual or parallel).

In order to offer more uniform exploration, the online search tool INESS-Search has a readable, compact and expressive query language (Meurer, 2012b) which shares important syntax features across annotation frameworks. Thus, notations for categories and words, operators for dominance and precedence, etc. are the same, regardless of the grammatical approach or type of annotation, to the largest extent possible. It also allows simultaneous search in several treebanks selected by the user, in other words, virtual treebank collections can be defined as search domains.

Similarly, INESS offers visualization of common structural representations in any type of treebanks and has user dependent options for visualization preferences (e.g. tree or arc diagrams for dependencies). INESS also supports online parsing, interactive parse selection and parsebank construction with LFG grammars and discriminant disambiguation. Briefly, INESS has evolved into a virtual laboratory for treebank management and exploration (Meurer et al., 2013; Rosén et al., 2012, inter alia).

Initially, the INESS treebanking infrastructure was not linked to CLARIN. Work on INESS began in 2010, two years before the start of the CLARINO project. The INESS project was originally a freestanding specialized research infrastructure with two main goals: (1) the construction of NorGramBank, a large Norwegian parsebank (i.e. treebank obtained by parsing), and (2) making treebanks more accessible. While the former will not be discussed here, the latter goal is relevant for a wide CLARIN audience.

[12] `http://rug-compling.github.io/dact/`

[13] `http://zardoz.service.rug.nl:8067/info.html`; see also Jan Odijk, *Linguistic research with CLARIN*, this volume.

[14] `http://lindat.mff.cuni.cz/services/pmltq/`

[15] `http://weblicht.sfs.uni-tuebingen.de/Tundra/`

INESS initially did not comply to basic CLARIN standards as regards metadata, PIDs, licensing and authentication. One of the objectives of the CLARINO project has therefore been to integrate INESS into the world of CLARIN. The remainder of this section describes how INESS has adopted CLARIN standards, how it addresses needs of the CLARIN user community, and how INESS together with LINDAT have formed a K-centre in the CLARIN Knowledge Sharing Infrastructure (KSI).

3.3 Metadata and PIDs

All CLARINO services aim to ensure CMDI compatibility, either by creating metadata in CMDI from scratch, or by converting to CMDI from their pre-existing internal format. Initially, INESS used META-SHARE (Losnegaard et al., 2013) to create metadata for the resources it makes available. The transition to CMDI metadata implies partly a conversion from earlier metadata, particularly from META-SHARE, and partly the creation of new metadata.

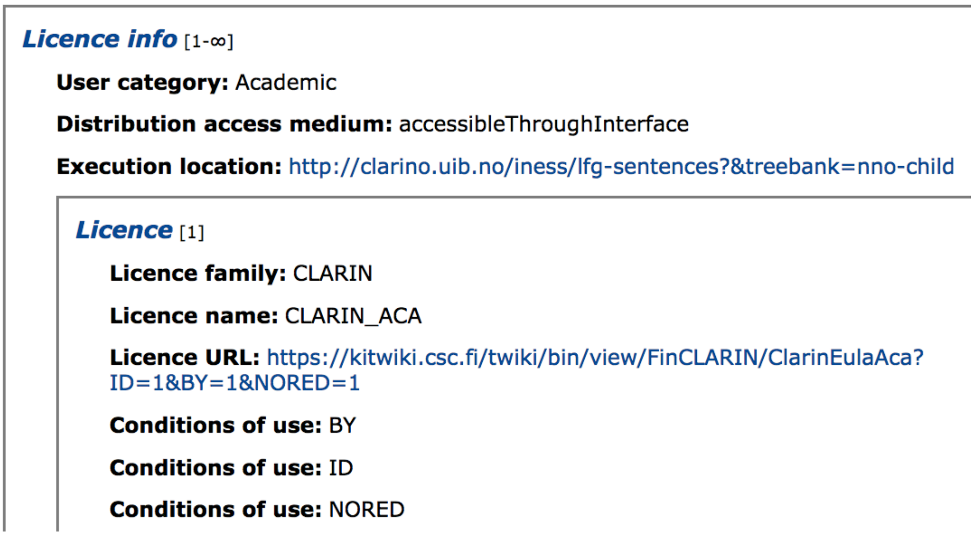
While supporting the principle of reusing existing components and profiles, the CLARINO working group on metadata nevertheless found that it was difficult to identify satisfactory existing profiles and components for treebanks and other resources in the Component Registry. In particular, since INESS and several other national partners had already described many resources using the META-SHARE framework, it was natural to start from the CMDI profiles derived from the corresponding META-SHARE schemata. It was found, however, that they did not have sufficient descriptive coverage for all the main resource types expected to be present in the CLARINO consortium. To ensure homogeneity in CLARINO, it was therefore decided to create a set of profiles, to be recommended by CLARINO, to accommodate all expected main resource types in the CLARINO consortium. In a national effort by the Norwegian metadata working group, improved CMDI profiles and components were developed, including the *corpusProfile* which is being applied to all treebanks available in INESS and all corpora available in Corpuscle.

Following the principles of reusing existing components and profiles, the existing META-SHARE profiles in the Component Registry were reused whenever possible. However, it became apparent that quite a few of the original META-SHARE components needed modifications. Even simple general changes had repercussions for many META-SHARE components.

As an example, META-SHARE has individual components for the different *actor roles* that persons and organizations may have in relation to a resource (with individual components such as *creatorPerson*, *creatorOrganization* for any role such as IPR holder, funder, validator, annotator, etc.). The existing selection of *actor role* components did not always cover the descriptive needs seen in CLARINO. For instance, the META-SHARE *author* component did not seem to accommodate the author's year of death, which may be quite relevant for historical texts, which form the source of several treebanks relevant to e.g. philology studies. CLARINO therefore decided to collapse the different META-SHARE components for person and organization roles into a generic component *actorInfo*, applicable for any role, independently of whether the *actor* is a person or an organization, and where a sufficient number of elements are available (including e.g. year of death). Replacing all role-specific components with the generic component *actorInfo* meant that a considerable number of original META-SHARE components had to be replaced by new CLARINO components, even if each of the changes was a minor one.

Another significant departure from META-SHARE concerns the license component in the metadata. This component should promote searchability according to user rights. For the CLARIN indexing service (such as the VLO), all licenses are classified into three main usage categories so that users of search services can easily filter their search. The usage category labels, informally designated as 'laundry tags', are PUB (public), ACA (academic) and RES (restricted) (Oksanen et al., 2010). CLARIN also has a classification scheme for describing license conditions with a standardized set of conditions of use, since different license families differ in how explicitly each condition is formulated. Since META-SHARE has not integrated the CLARIN licensing scheme, CLARINO modified the license component to include the CLARIN *User category* and *conditions of use*, as exemplified in the COMEDI view in Figure 2.

The CLARINO working group on metadata has developed the following set of profiles and components based on our estimated needs. These are in a test phase and will be published in the Component Registry in 2016. Further profiles will be added as needed.



Licence info [1-∞]

User category: Academic

Distribution access medium: accessibleThroughInterface

Execution location: http://clarino.uib.no/iness/lfg-sentences?&treebank=nno-child

Licence [1]

Licence family: CLARIN

Licence name: CLARIN_ACA

Licence URL: https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinEulaAca?ID=1&BY=1&NORED=1

Conditions of use: BY

Conditions of use: ID

Conditions of use: NORED

Figure 2: An example of a CLARINO license component describing a resource licensed under a CLARIN ACA license.

- corpusProfile[16] – for corpora of all types and modalities (including treebanks)
- lexicalProfile[17] – for lexical resources
- teiProfile[18] – in collaboration with Clarin-DK, extending the current TEI profile
- toolProfile[19] – for tools and services

In addition to these general profiles, CLARINO has also developed specific profiles to assist repositories aiming to become CLARINO-compatible with regard to metadata. To this end, a profile *electronicSingleObjects*[20] has been developed in collaboration with the ELMCIP knowledge base on electronic literature[21]. Similarly, a *dataverseProfile* has been created[22] in collaboration with the Tromsø Repository of Language and Linguistics (TROLLing),[23] an open repository for research data and statistical code in the field of language and linguistics. TROLLing now aims to become a CLARINO repository, and therefore aims to provide metadata also in CMDI-format.

The documentation of Best Practice guidelines for CLARINO partners who are to fill in metadata is in progress. The preliminary website is available from the left-hand menu on the COMEDI website.[24] The general CLARINO policy for metadata can be summed up as follows:

1. The set of profiles and components created by CLARINO is recommended.
2. Any CLARINO partner may, according to needs, create their own components and profiles.
3. All profiles used in CLARINO should contain the component resourceCommonInfo.[25]

The obligatory component *resourceCommonInfo* contains fields for general information which is relevant for all resource types, and is required in CLARINO to facilitate search across profiles by ensuring that

[16] Profile ID: clarin.eu:cr1:p_1407745711925

[17] Profile ID: clarin.eu:cr1:p_1428388179419

[18] Profile ID: clarin.eu:cr1:p_1422885449322

[19] Profile ID: clarin.eu:cr1:p_1422885449331

[20] Profile ID: clarin.eu:cr1:p_1407745712024

[21] http://elmcip.net/knowledgebase

[22] Profile ID: clarin.eu:cr1:p_1447674760331

[23] http://opendata.uit.no/dvn/dv/trolling

[24] http://clarino.uib.no/comedi/page?page-id=clarino-best-practice

[25] http://catalog.clarin.eu/ds/ComponentRegistry?registrySpace=private&itemId=clarin.eu:cr1:c_1396012485126

some basic information is always present, similarly to the minimal metadata schema in META-SHARE. This includes basic information such as resource type, resource name, identifiers (e.g. PIDs), licenses, origin, owner, contact information, metadata details, version info, validation info and relation to other resources. Moreover, using the *resourceCommonInfo* component ensures an optimal display of metadata in the national metadata registry at the National Library of Norway.[26]

When sufficient metadata are provided for a resource within the CLARINO Bergen Centre, a CLARIN compatible persistent identifier (*handle*) is created which redirects to a landing page displaying the metadata. These metadata are available in compact and full views. It may still be noted that in its effort to host as many treebanks as possible, INESS is currently also hosting treebanks for which documentation has not yet been fully supplied.

3.4 Access policies and licenses

The CLARINO Bergen Centre, including INESS, hosts a large number of resources which often integrate linguistic annotations with different provenance, which makes it necessary to accommodate licenses from different sources. INESS, like the entire CLARINO Bergen Centre, therefore writes license agreements for newly added treebanks using CLARIN depositor's license agreements as the default, and follows the CLARIN recommendation to make resources as freely and openly available as possible. However, INESS also accommodates treebanks with legacy licenses from different sources which may impose restrictions. In line with CLARIN recommendations, INESS streamlines the description of licenses using the CLARIN license categories, as mentioned above.[27]

Treebanks which are not publicly available require that users provide proper user credentials by logging in. Like LINDAT, the CLARINO Bergen Centre has implemented SAML 2.0-based single sign-on covering the CLARIN ID Provider and federations participating in eduGAIN[28] and the CLARIN Service Provider Federation (SPF).[29] Selection of one's ID provider is achieved through integration of the DiscoJuice[30] discovery service by Uninett and the CLARIN SPF.

Many treebanks have a complex provenance; futhermore, the license conditions may vary according to the type of access (more open license for access through the INESS search interface, more limited access for downloadability). Therefore, INESS is able to handle multiple licenses. Specifically, the *Distribution info* component in the CMDI metadata may contain more than one *License info* component. This is the case, for instance, in the *Distribution info* for BulTreeBank,[31] which has different licenses for users wishing to search in it (*Distribution access medium: accessibleThroughInterface*) and for users wishing to download it (*Distribution access medium: downloadable*).

Authorization for the use of treebanks handled locally in the infrastructure by asking the user to authenticate (by logging in with proper user credentials) and then click on a button to agree to the conditions for use specified in the resource's license. As illustrated in Figure 3, the INESS interface for accepting a license allows the user to click on the license to read the full text before accepting. Since the user's acceptance of the conditions is connected to their user identity, their license acceptance for the resource in question will be remembered for future sessions.

3.5 INESS as a K-centre

INESS and LINDAT were in 2015 approved as a joint virtual K-centre in the CLARIN Knowledge Sharing Infrastructure (KSI).[32] This implies that knowledge will be shared with users in a systematic way, so as to assist researchers in managing and using resources efficiently and in line with good practice. To that effect, the INESS website contains a menu link to an overview page for getting started, while an

[26] Currently available as a BETA version at http://www.nb.no/sprakbanken/repositorium#ticketsfrom?collection=clarino.

[27] http://www.clarin.eu/content/license-categories

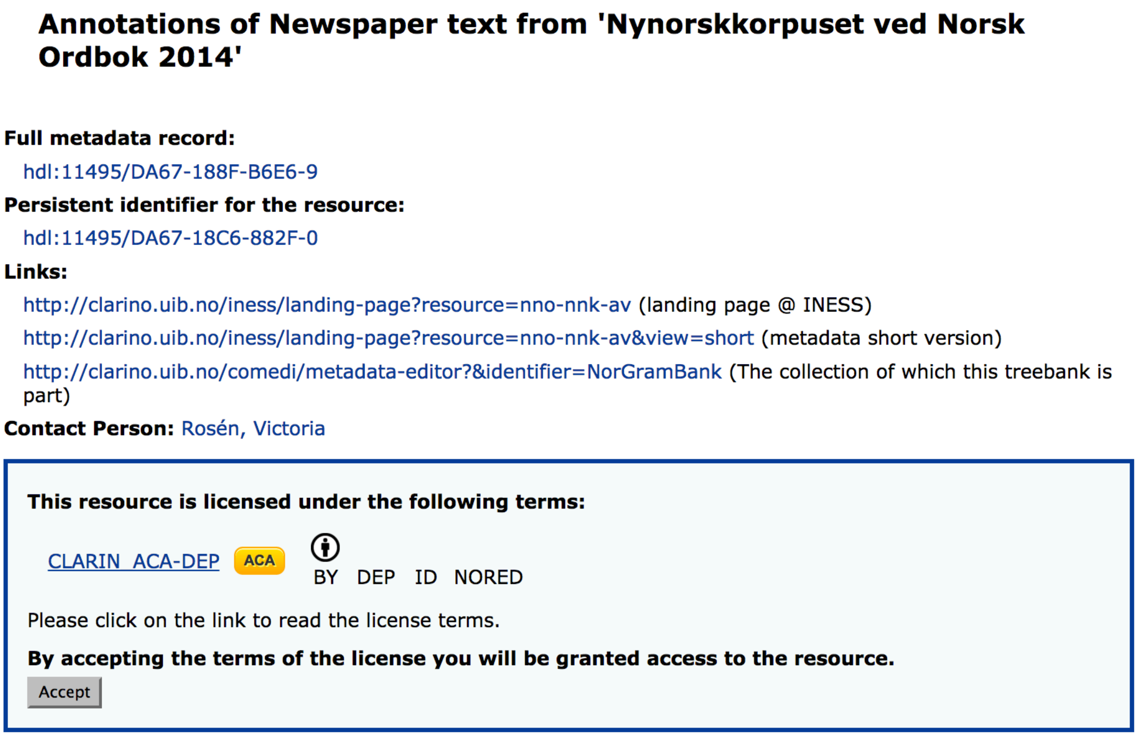
[28] http://services.geant.net/edugain

[29] https://www.clarin.eu/content/service-provider-federation

[30] http://discojuice.org

[31] http://hdl.handle.net/11495/D918-1214-6A7E-1

[32] http://www.clarin.eu/content/knowledge-centres



Annotations of Newspaper text from 'Nynorskkorpuset ved Norsk Ordbok 2014'

Full metadata record:
hdl:11495/DA67-188F-B6E6-9
Persistent identifier for the resource:
hdl:11495/DA67-18C6-882F-0
Links:
http://clarino.uib.no/iness/landing-page?resource=nno-nnk-av (landing page @ INESS)
http://clarino.uib.no/iness/landing-page?resource=nno-nnk-av&view=short (metadata short version)
http://clarino.uib.no/comedi/metadata-editor?&identifier=NorGramBank (The collection of which this treebank is part)
Contact Person: Rosén, Victoria

This resource is licensed under the following terms:

CLARIN ACA-DEP ACA BY DEP ID NORED

Please click on the link to read the license terms.

By accepting the terms of the license you will be granted access to the resource.

Figure 3: Interface in INESS for accepting a license.

FAQ intends to answer common questions for troubleshooting. There is also a link to extensive documentation about grammars, the query language, the web interface, annotation guidelines, and sharing of treebanks. Furthermore, there are links to publications and to internal and external resources. Users can interact through a user forum, while a contact email address is also provided. The K-centre also organizes treebanking tutorials. The first such event was organized by INESS in Warsaw on February 2 and 6, 2015.[33]

3.6 Users and use cases

INESS fills the gap between two groups targeted by CLARIN: those who have resources but need a place to make them available, and those who wish to use resources and who need advanced online tools to explore them. Several projects and organizations, including also philology and historical linguistics initiatives such as Menotec,[34] ISWOC[35] and PROIEL,[36] have in INESS found both an archive to deposit their resources and a virtual laboratory for exploring resources.

INESS has also proved useful, for instance, in the context of the Parseme COST action, which aims, among other things, at investigating how multiword expressions are annotated in treebanks.[37] Since searches for specific dependency relations can be performed in several treebanks simultanously, INESS-Search is a practical tool for making comparisons and checking consistency (De Smedt et al., 2015). The treebank building facilities in INESS have also been used by researchers at IPI-PAN (Warsaw) who have set up their own instance of the software and have developed a Polish LFG treebank (POLFIE) (Patejuk and Przepiórkowski, 2015).

[33] http://pargram.b.uib.no/meetings/spring-2015-meeting-in-warsaw/
[34] http://link.uib.no/4s0i6
[35] http://www.hf.uio.no/ilos/english/research/projects/iswoc
[36] http://www.hf.uio.no/ifikk/english/research/projects/proiel/
[37] http://parseme.eu

4 Concluding remarks and outlook

The CLARINO Bergen Centre is fully operative[38] and is open to other partners in CLARINO and to the whole CLARIN community. We believe that the decision to locate the CLARINO Bergen Repository at the UBL is a step towards sustainability, since this library is committed to support data publication and has permanent staff at its section for digital resources. In terms of human resources our solution requires UBL to have at least one programmer, who is backed up with planning and support from management and colleagues, and from the institution's IT department. UBL currently has two programmers in its digital systems section. The time and effort spent on the installation is estimated to five person-months of programmer work, and two person-months on the graphical design, branding and content adaptations of the site.

The CLARINO Bergen Repository is being populated by resources produced not only in Bergen but also by several other CLARINO consortium partners. Furthermore, INESS, Corpuscle and COMEDI are populated and used by an international audience of CLARIN members. All services provide OAI-PMH endpoints for metadata harvesting.[39] Metadata at the OAI-PMH endpoints are now periodically harvested by the CLARIN VLO as well as by the National Library of Norway. The latter is constructing a nationwide catalogue. Furthermore, one can also view metadata for the individual resources by specifying the metadata format and the handle of the associated resource.[40]

The construction of the CLARINO Bergen Repository has been greatly facilitated by the willingness of the LINDAT partner to share their systems with other consortia. Using CLARIN-compliant software as a basis has undoubtably been cost and time saving. The mobility actions proved to be a useful supporting measure.

The integration of INESS in the centre followed a different route. INESS was initiated before CLARINO, but has gradually been incorporated in CLARINO and has to a large extent become compliant with good practice in CLARIN. Corpuscle and COMEDI, in contrast, were constructed in the CLARINO project. The main software for INESS, Corpuscle and COMEDI has been written in Common Lisp for expressiveness, extensibility and rapid development and updating, and is available to others. INESS has so far been installed at two sites: the Bergen CLARINO Centre and the IPI-PAN centre in Warsaw, which has become an associated partner in INESS.

Among possible future extensions to INESS, we are considering user-serviced uploading of treebanks, of corpora to be parsed, and of grammars. However, in our experience, there are often unpredictable divergences from standard formats which need to be manually solved.

Access to a language resource based on authorization granted by an external rightsholder remains an interesting challenge in the wider context of CLARIN. This is illustrated by the following cases in INESS. The use of LASSY-Klein, a treebank distributed by TST-Centrale,[41] is conditional upon the user signing a license agreement exclusively with TST-Centrale. Thus, end user licensing for this treebank cannot be handled by INESS. Since licensed users can download and search this treebank as they wish, they can request access to this resource through INESS, but only if they present a piece of paper signed by TST-Centrale — a procedure which is not practical. There is no automated way of verifying if potential users have obtained a license for this resource from the rightsholders. Similarly, our research group has a license for TüBa-D/Z,[42] for which a paper-based signed agreement is required as well, but we are not allowed to give users access to this treebank unless they are explicitly authorized by the University of Tübingen. Again, there is no easy way of verifying if potential users have signed a license for this resource.

The adoption of a common resource entitlement management system such as REMS[43] would make authorization a more streamlined process, not only for treebanks, but for any restricted resources which may be accessible through services at more than one CLARIN centre. In such a scheme, any authorization

[38] http://clarino.uib.no

[39] For the repository: https://repo.clarino.uib.no/oai/request

[40] For example, https://repo.clarino.uib.no/oai/cite?metadataPrefix=cmdi&handle=11509/3

[41] http://tst-centrale.org/nl/producten/corpora/lassy-klein-corpus/6-66?cf_product_name=Lassy+Klein-corpus

[42] http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html

[43] https://confluence.csc.fi/display/REMS/Home

given by a rightsholder (e.g. TST-Centrale) to a user would be recorded in a secure database, which in turn could be consulted by service providers (such as INESS). The use of such a shared AAI architecture will be an important step for CLARIN in reaching a truly European dimension. It will, however, only be effective if it is centrally promoted by the CLARIN ERIC and and widely adopted by CLARIN resource rightsholders and service providers alike.

References

[Broeder et al.2010] Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A data category registry- and component-based metadata framework. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[De Kok et al.2014] Daniël De Kok, Dörte De Kok, and Marie Hinrichs. 2014. Build your own treebank. In *CLARIN Annual Conference 2014 (abstracts)*.

[De Smedt et al.2015] Koenraad De Smedt, Victoria Rosén, and Paul Meurer. 2015. Studying consistency in UD treebanks with INESS-Search. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 258–267, Warsaw, Poland. Institute of Computer Science, Polish Academy of Sciences.

[Losnegaard et al.2013] Gyri Smørdal Losnegaard, Gunn Inger Lyse, Anje Müller Gjesdal, Koenraad De Smedt, Paul Meurer, and Victoria Rosén. 2013. Linking Northern European infrastructures for improving the accessibility and documentation of complex resources. In Koenraad De Smedt, Lars Borin, Krister Lindén, Bente Maegaard, Eiríkur Rögnvaldsson, and Kadri Vider, editors, *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013, May 22–24, 2013, Oslo, Norway. NEALT Proceedings Series 20*, number 89 in Linköping Electronic Conference Proceedings, pages 44–59. Linköping University Electronic Press.

[Lyse et al.2015] Gunn Inger Lyse, Paul Meurer, and Koenraad De Smedt. 2015. COMEDI: A component metadata editor. In Jan Odijk, editor, *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*, number 116 in Linköping Electronic Conference Proceedings, pages 82–98, Linköping, Sweden. Linköping University Electronic Press.

[Martens2013] Scott Martens. 2013. TüNDRA: A web application for treebank search and visualization. In Sandra Kübler, Petya Osenova, and Martin Volk, editors, *Proceedings of the Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 133–144. Bulgarian Academy of Sciences.

[Meurer et al.2013] Paul Meurer, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Gunn Inger Lyse, Gyri Smørdal Losnegaard, and Martha Thunes. 2013. The INESS treebanking infrastructure. In Stephan Oepen, Kristin Hagen, and Janne Bondi Johannessen, editors, *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16*, number 85 in Linköping Electronic Conference Proceedings, pages 453–458. Linköping University Electronic Press.

[Meurer2012a] Paul Meurer. 2012a. Corpuscle – a new corpus management platform for annotated corpora. In Gisle Andersen, editor, *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian*, number 49 in Studies in Corpus Linguistics. John Benjamins Publishing Company.

[Meurer2012b] Paul Meurer. 2012b. INESS-Search: A search system for LFG (and other) treebanks. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG '12 Conference*, LFG Online Proceedings, pages 404–421, Stanford, CA. CSLI Publications.

[Mišutka et al.2015] Jozef Mišutka, Amir Kamran, Ondřej Košarko, Michal Josífko, Loganathan Ramasamy, Pavel Straňák, and Jan Hajič. 2015. Linguistic digital repository based on DSpace 5.2. http://hdl.handle.net/11234/1-1481. LINDAT/CLARIN Digital Library at Institute of Formal and Applied Linguistics, Charles University in Prague.

[Oksanen et al.2010] Ville Oksanen, Krister Lindén, and Hanna Westerlund. 2010. Laundry symbols and license management – practical considerations for the distribution of LRs based on experiences from CLARIN. In *Proceedings of LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*.

[Patejuk and Przepiórkowski2015] Agnieszka Patejuk and Adam Przepiórkowski. 2015. POLFIE: an LFG grammar of Polish accompanied by a structure bank. In *CLARIN Annual Conference 2015 (abstracts)*.

[Rosén et al.2012] Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

[Telljohann et al.2012] Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2012. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Department of General and Computational Linguistics, University of Tübingen, Germany.

[van Noord et al.2013] Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer, Berlin/Heidelberg.

[Vandeghinste and Augustinus2014] Vincent Vandeghinste and Liesbeth Augustinus. 2014. Making a large treebank searchable online. The SoNaR case. In Marc Kupietz, Hanno Biber, Harald Lüngen, Piotr Bański, Evelyn Breiteneder, Karlheinz Mörth, Andreas Witt, and Jani Taksha, editors, *Challenges in the Management of Large Corpora (CMLC-2)*, Reykjavik, Iceland.