# CLARIN 2015

## Wrocław

# Selected Papers from the CLARIN Annual Conference 2015

October 14–16, 2015
Wrocław, Poland


Edited by Koenraad De Smedt

Published by
Linköping University Electronic Press, Sweden
Linköping Electronic Conference Proceedings, No. 123

NEALT Proceedings Series, Volume 28

eISSN 1650-3740 • ISSN 1650-3686 • ISBN 978-91-7685-765-6

# Preface

These proceedings present the highlights of the CLARIN Annual Conference 2015 that took place in Wrocław, Poland. It is the second volume of what we hope is becoming a series of publications that constitutes a multiannual record of the experience and insights gained within CLARIN.

The selected papers present results which were obtained in the projects conducted within and between the national consortia and which contribute to the realization of what CLARIN aims to be: a European Research Infrastructure providing sustainable access to language resources in all forms, analysis services for the processing of language materials, and a platform for the sharing of knowledge that can stimulate the use, reuse and repurposing of the available data. As an international research infrastructure, CLARIN has the potential to lower the barriers for researchers to entry into digital scholarship and cutting edge research. With the growing number of participating countries and languages covered, CLARIN will help to establish a truly connected and multilingual European Research Area for digital research in the Humanities and Social Sciences, in which Europe's multilinguality reinforces the basis for the comparative investigation of a wide range of intellectual and societal phenomena.

This volume illustrates the various ways in which CLARIN supports researchers and it shows that there are more and more scholarly domains in which language resources are a crucial data type, whether approached as big data or nor so big data. Language resources can play a multitude of roles, such as carrier of information, record of the past, means of literary expression, social signal, or object of linguistic study. As diverse as the roles of language data are in the various research domains, so diverse is the community of scholars collaborating within CLARIN. The Annual Conference is one of the instruments for organizing the communication between those who build and maintain the infrastructure, those who provide data and tools, and those who use the CLARIN infrastructure in their scholarly projects. For all these groups, the sharing of insights in problems, solutions, failures and successes is a prerequisite for taking the next steps. Even more importantly, the Annual Conference offers a platform for the sharing of inspiring examples of investigations which have become feasible because of the infrastructure.

These proceedings once again demonstrate the potential impact of CLARIN on scholarly work. With the increasing volumes and diversity of services offered, the potential for impact on the research agendas addressing societal challenges is growing as well. Hopefully this volume will help to attract new categories of users, with ideas and requirement for use cases that can help us identify the directions and next steps to take in the further development of the CLARIN infrastructure.

April 4, 2016
Utrecht

Franciska de Jong
CLARIN ERIC Executive Director

# Introduction

This volume contains a selection of papers presented at the CLARIN Annual Conference 2015 which was held in Wrocław, Poland, from the 14th to the 16th of October 2015.

CLARIN has been organizing its Annual Conference since 2012. The aim of these conferences is to exchange ideas and experiences on the CLARIN infrastructure. Topics include the infrastructure's design, construction and operation, the data and services that it contains or should contain, its actual or potential use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Sharing Infrastructure.

Since 2014, CLARIN changed the format of these events by launching an open call for contributions, subjecting submissions to peer review, and publishing Selected Papers after the event. The program of the conference is the responsibility of the CLARIN National Coordinators' Forum (NCF). Each submission was reviewed by at least three members of the NCF or people delegated by NCF members.

In 2015 the program consisted of 22 presentations, accepted on the basis of extended abstracts which were published in a Book of Abstracts. Ten presentations were accepted to be presented orally and twelve as posters. Some presentations were supplemented by demonstrations. The program also included an invited talk *Interaction and dialogue with large-scale textual data: Parliamentary speeches about migration and speeches by migrants as a use case* by Andreas Blätte (University of Duisburg-Essen). This talk brought methodologies relevant to CLARIN into the context of current societal concerns.

After the conference, the authors of all presentations were asked to submit full papers. Ten full papers were received, each of which was evaluated by four reviewers. Nine papers were selected for publication, involving authors from Austria, the Czech Republic, Estonia, Finland, Germany, The Netherlands, Norway and Sweden.

A substantial range of topics is covered by the papers in this volume. These include the construction of the CLARIN infrastructure, the standards that underpin it (in particular as regards metadata and data concepts), and the enrichment of data and tools that populate it. There are also use cases of its scientific exploitment in the fields of Dutch linguistics and rhetorical history, with appropriate methodological considerations. Research data workflows, data curation, data management plans and the regulatory and contractual framework governing the use of data are also addressed.

I thank all reviewers for their evaluation of the submissions. I also thank Peter Berkesand at Linköping University Electronic Press for the efficient digital publication of this volume.

April 4, 2016
Bergen

Koenraad De Smedt
Program Committee Chair

# Program Committee

Lars Borin
António Branco
Koenraad De Smedt
Tomaz Erjavec
Eva Hajičová
Erhard Hinrichs
Bente Maegaard
Karlheinz Mörth
Jan Odijk
Rūta Petrauskaitė
Maciej Piasecki
Stelios Piperidis
Kiril Simov
Kadri Vider
Martin Wynne

## Additional reviewers

Daniël de Kok
Maria Gavrilidou
Paweł Kamocki

# Table of Contents

The CLARINO Bergen Centre: Development and Deployment . . . . . 1
*Koenraad De Smedt, Gunn Inger Lyse, Rune Kyrkjebø, Hemed Al Ruwehy, Øyvind Liland Gjesdal, Victoria Rosén and Paul Meurer*

The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure . . . . . 13
*Aleksei Kelli, Kadri Vider and Krister Lindén*

Variability of the Facet Values in the VLO – a Case for Metadata Curation . . . . . 26
*Margaret King, Davor Ostojic, Matej Ďurčo and Go Sugimoto*

A Use Case for Linguistic Research on Dutch with CLARIN . . . . . 45
*Jan Odijk*

CLARIN Concept Registry: The New Semantic Registry . . . . . 62
*Ineke Schuurman, Menzo Windhouwer, Oddrun Ohren and Daniel Zeman*

DMPTY – A Wizard For Generating Data Management Plans . . . . . 71
*Thorsten Trippel and Claus Zinn*

How Can Big Data Help Us Study Rhetorical History? . . . . . 79
*Jon Viklund and Lars Borin*

Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions . . . . . 94
*Tanja Wissik and Matej Ďurčo*

Enriching a Grammatical Database with Intelligent Links to Linguistic Resources . . . . . 108
*Ton van der Wouden, Gosse Bouma, Matje van de Kamp, Marjo van Koppen, Frank Landsbergen and Jan Odijk*

# 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

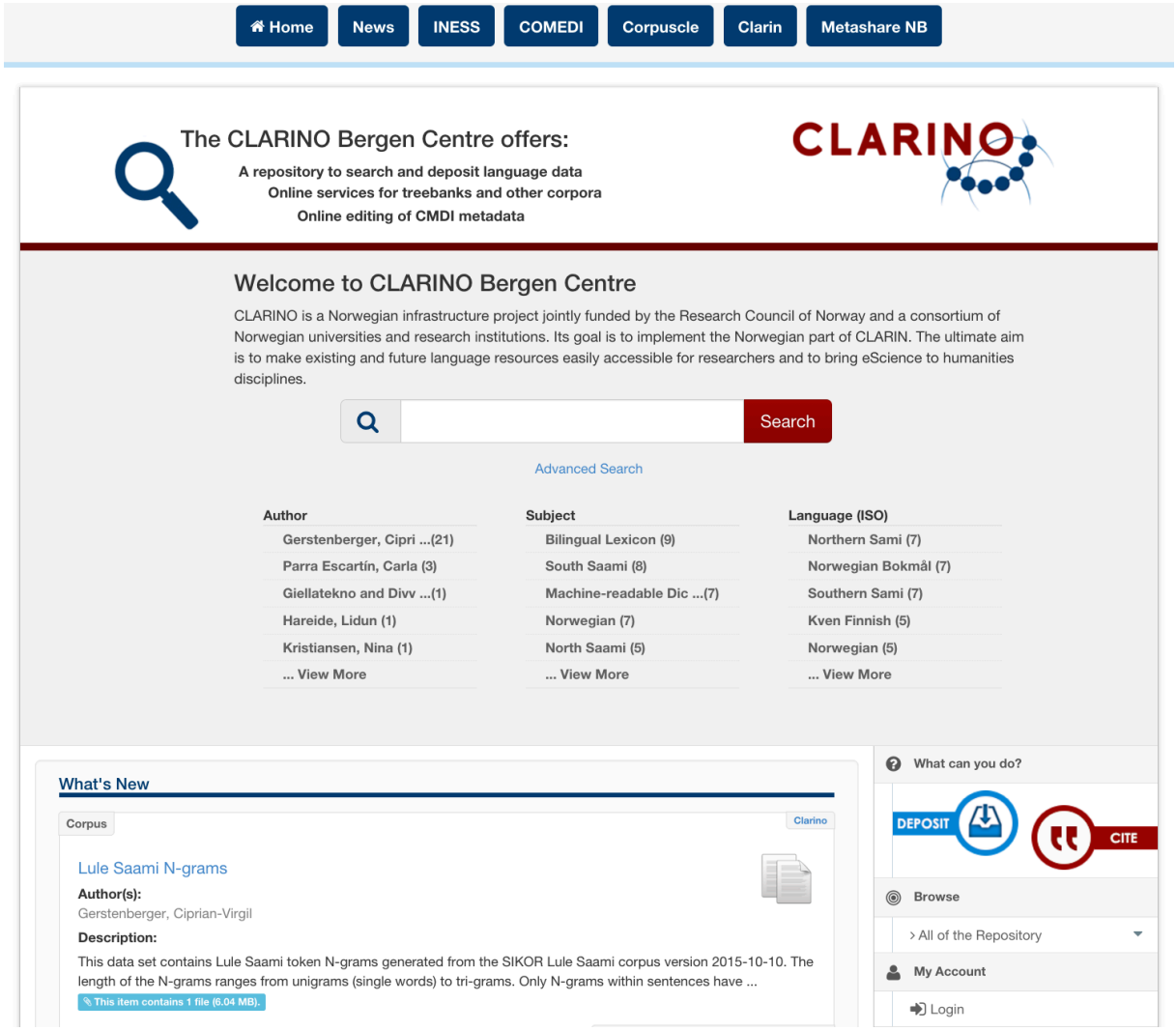



Figure 1: Main page of the CLARINO Bergen Centre website.

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 free-standing 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.

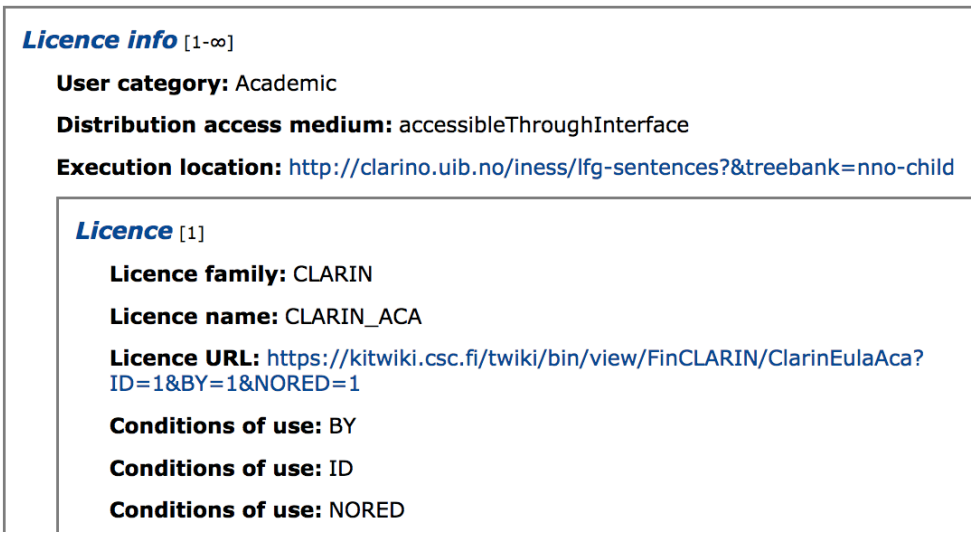



Figure 2: An example of a CLARINO license component describing a resource licensed under a CLARIN ACA license.

- 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`

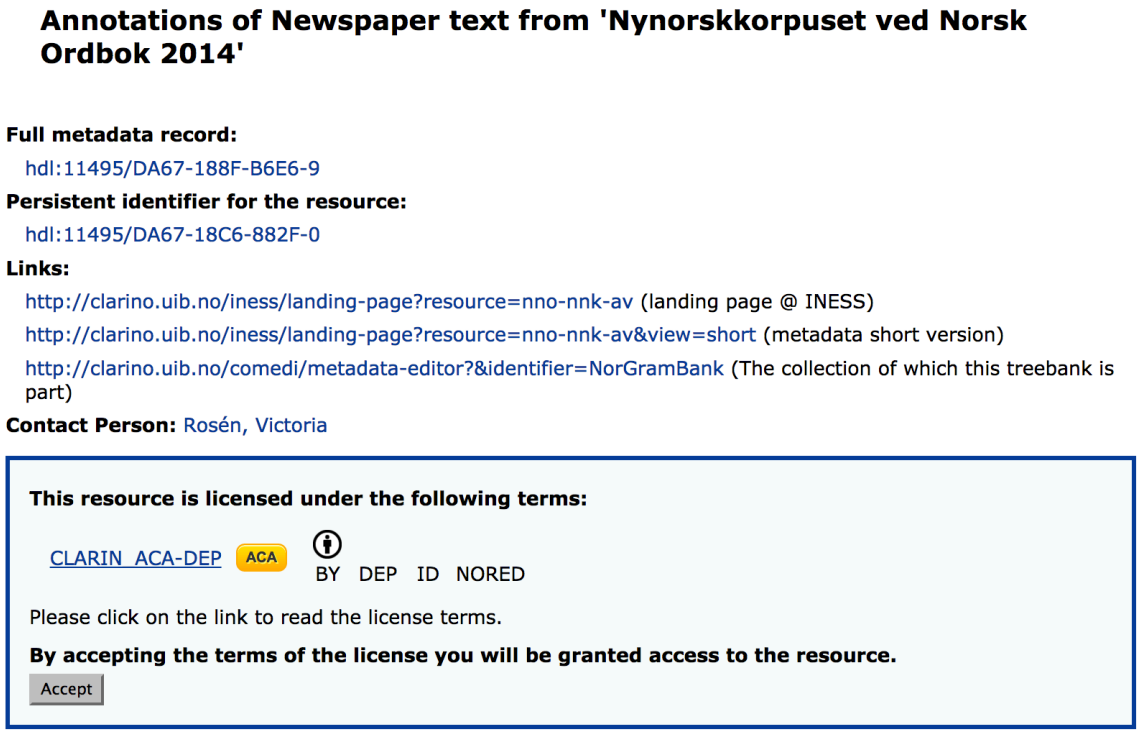



Figure 3: Interface in INESS for accepting a license.

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 nation-wide 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.

# The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure

**Aleksei Kelli**
Department of Private Law
University of Tartu, Estonia
aleksei.kelli@ut.ee

**Kadri Vider**
Centre of Estonian Language Resources
University of Tartu, Estonia
kadri.vider@ut.ee

**Krister Lindén**
Department of Modern Languages
University of Helsinki, Finland
krister.linden@helsinki.fi

**Abstract**

The article focuses on the regulatory and contractual frameworks in CLARIN. A process analysis approach has been adopted to allow an evaluation of the functionality and shortcomings of the entire legal framework applicable to language resources and technologies. The article discusses and provides background information to amendments of key provisions of CLARIN license templates. The authors also address issues relating to the research exception allowing for the development of language resources without the copyright holder's consent. The article introduces some practical information on a new version of the license category calculator. The article reflects the authors' personal understanding and insights gained by examining the legal aspects of language resources and technologies in Estonia and Finland.

## 1 Introduction

The nature of language resources (hereinafter resources) is defined[1] as software, applications and/or databases[2]. This definition can be analysed from a wide range of perspectives such as technological, linguistic, ethical and legal. We focus on the legal challenges relating to the development and distribution[3] of language resources and technologies. In view of this, the regulatory and contractual framework (hereinafter legal framework) constitutes one of the core infrastructures of CLARIN.

We base the discussion on our analysis of the process for developing and distributing language resources. In this paper we employ the analysis to facilitate our evaluation of the functionality and shortcomings of the entire legal framework concerning language resources. The process commences with the development and results in the distribution of language resources. However, we do not address different process phases separately since they are clearly intertwined. Instead, we identify and analyze legal issues across individual phases.

We resort to traditional methods in social sciences and draw on the previous legal research conducted by the authors. The analysis incorporates the Estonian and Finnish experience. Both countries are CLARIN ERIC members. The article reflects the personal understanding and insights of the authors gained while studying the legal aspects of language resources in Estonia and Finland. Subject to our insights, we have amended the CLARIN license agreement templates and terms of service. We provide background information to the amendments by offering suggestions how to improve the existing legal framework. Since the article outlines the practical legal issues pertaining to the management of resources, it can be of use to other CLARIN members as well.

We construe the Terms of Service (TOS) to mean the general conditions for using a CLARIN service, the End-User License Agreement (EULA) to mean the conditions designed for an end-user to use



[1] See Article 3 of a CLARIN Deposition License Agreement (e.g. CLARIN-DELA-PUB-v1.0).

[2] Note that from a legal perspective, in a research infrastructure like CLARIN, researchers primarily use works as databases from which they extract facts, in contrast to libraries from which they primarily borrow copies of works.

[3] Distribution means, inter alia, making available to the public, communication to the public and distribution to the public by wire or wireless means.

a language resource, and the Deposition License Agreement (DELA) to mean the conditions on which a CLARIN service can distribute a language resource to end-users. For an overview of the relationship between these concepts, see Figure 1. We designate main license category to mean one of the three broad categories of usage rights and restrictions conferred on the end-user by TOS, DELA and EULA, see Section 2. In addition, several subcategories entailing specific rights and restrictions have been defined, see Section 4. A laundry symbol[4] is an icon attributed to a license category.

The paper is organized into three main sections. The first section focuses on the establishment of institutional control over the existing language resources. The second addresses the issue of the development of language resources and deals with their distribution and potential subsequent utilization. We also study the case of providing public access to fragments of resources in a concordance service *versus* distributing resources in full for research purposes in light of a research exception in the copyright regulation and in the CLARIN contractual framework. The third section explains the concept of a license category calculator.

## 2 Establishment of the institutional control over language resources and technologies

The distribution and utilization of language resources and technologies depends on several conditions such as technological capabilities, the existence of resources, etc. The institutions and organizational units managing resources must also have the legal capacity to enter into valid transactions and obtain sufficient rights to distribute language resources. In order to avoid a purely abstract discussion of the legal framework characteristic of national CLARIN consortiums, we rely on Estonia as an example when addressing these issues.

Estonia set up the Center of Estonian Language Resources (CELR) as a consortium of 3 institutions at the national level on December 2, 2011. The consortium consists of the University of Tartu (UT) (as the leading partner in CELR), the Institute of Cybernetics at Tallinn University of Technology, and the Institute of the Estonian Language. The consortium constitutes an organizational framework for the coordination and implementation of the obligations of Estonia as a member in CLARIN ERIC.

The Estonian consortium agreement regulates issues relating to the background and foreground intellectual property[5] (IP) of the Estonian partners. However, it does not provide a clear framework for the resources developed and owned by persons outside the consortium. To acquire their language resources, the consortium partners have to conclude individual agreements with them. However, this poses a serious problem as the Estonian national consortium is not deemed a legal person in private or public law (also called legal entity). The consortium is an agreement between independent partners to cooperate on specific issues.[6] In theory, the right of representation could be derived from the consortium agreement and the partners could represent each other. However, this may create legal uncertainties as to the scope of the right of representation. Therefore, a possible way forward is to develop an understanding where each consortium partner concludes agreements governing certain types of language resources with external partners. This is compatible with the current situation where each partner is responsible for certain types of resources. Another option would be to establish a legal entity (e.g., a non-profit association, a private limited company, etc.).

The acquisition and distribution of resources within the CLARIN framework by its members has to be aligned and standardized as much as possible. Standard agreements constitute a key infrastructure of CLARIN, and CLARIN has developed standard agreement templates (Licenses, Agreements, Legal Terms) which can be used for this purpose.

CLARIN standard agreement templates are based on a conceptual division of all language resources into three main categories, i.e. resources which are publicly or openly available (PUB), those which are available for research or academic use (ACA) and those which are restricted to individual use, e.g.

[4] In the textile industry, a laundry symbol represents the maximum permitted treatment. The CLARIN symbols indicate in a similar vein a "maximum treatment" of language resources that can safely be permitted to an end-user by the CLARIN repository based on its DELAs, TOS and EULAs.

[5] Background IP are typically resources that existed before the consortium and foreground IP are resources created within the consortium.

[6] These include the items mandated for national consortia by the CLARIN ERIC Statutes (http://www.clarin.eu/sites/default/files/OJ-2012-136-EU-Decision.pdf)

due to personal data content, (RES). According to Oksanen & al. (2010) explaining the conceptual background and the evolution of this approach, this categorization is based on an extensive survey indicating that it is possible to group licenses in this manner. The PUB category allows wide distribution. The ACA category is designed to make resources available for research purposes[7] and RES permits limited use with additional requirements relating to data protection, e.g. research plan, etc. In addition to the three main categories, there are several subcategories for a more nuanced picture of the conditions of use as discussed in Section 4 on the License Category Calculator.

The authors support the ideology of CLARIN having a tripartite[8] main division of resources integrating this into a contractual framework for several reasons. Firstly, the current division contains more or less all language resources. Its appropriateness has been proven in practice[9] and no major problems have been identified. Secondly, if adopting a different main division, CLARIN would have to address the legal status of language resources and technologies having already entered CLARIN. Subsequently, a need would arise for the reclassification of already deposited and distributed resources. Thirdly, PUB, ACA and RES categories are already integrated into the CLARIN infrastructure. They are accepted by CLARIN stakeholders and thereby socially embedded. The potential adoption of a new main division would create unnecessary confusion. Nevertheless, the authors are not asserting that CLARIN should not develop the current categorization further. The key idea is that potential changes have to be made within a clear conceptual framework, be mature enough, absolutely necessary and outweigh potential negative aspects. An evolutionary approach, where the categorization is changed incrementally and gradually (i.e., the existing categories are specified, subcategories adopted, etc.), should be preferred over a radical approach (i.e., a totally new categorization). Stability as a value is a valid consideration.

A starting point for the analysis of the CLARIN license templates is the deposition license agreements (DELA) which resource right-holders can form when depositing resources in national CLARIN repositories. They are divided into three categories:

1) CLARIN-DELA-PUB-v1.0 (for publicly or openly available resources);
2) CLARIN-DELA-ACA-v1.0 (for resources for research or academic purposes);
3) CLARIN-DELA-RES-v1.0 (for resources restricted to individual use).

The acceptance of a general conceptual framework dividing language resources and agreements into PUB, ACA and RES does not preclude amendments to the existing templates. The process of translating the CLARIN agreement templates (DELAs and Terms of Service) into Estonian and bringing them into conformity with the Estonian legislation offered a good opportunity to scrutinize once again the existing agreement templates. The results and observations are discussed below.[10] An outline of the relationship between the CLARIN agreement templates is provided in Figure 1.

[7] In its fundamental form, ACA covers both commercial and non-commercial research.
[8] CLARIN is not the only community to have a tripartite division of resources, e.g. the ORCID community (http://orcid.org/) has a similar division with the ORCID Privacy Settings defining access for all, for a trusted community, or for the owner (http://support.orcid.org/knowledgebase/articles/124518-orcid-privacy-settings). While ORCID takes the perspective of the data producer, CLARIN has an end-user perspective on data access.
[9] The categories are currently in use as license metadata in the Virtual Language Observatory by CLARIN (https://www.clarin.eu/content/virtual-language-observatory)
[10] The new templates: http://www.helsinki.fi/finclarin/calculator/ClarinLicenseCategory.html (24.11.2015).

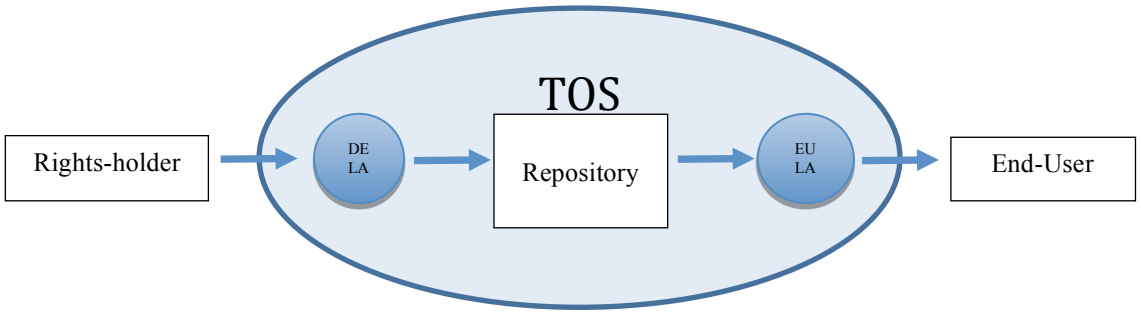



Figure 1. An outline of how the CLARIN agreement templates confer access rights

The first observation concerns the structure of DELAs. All DELAs have almost identical provisions. The main difference emanates from the provisions concerning intellectual property rights and access rights (Section 7 of DELA).[11] Therefore, it would be practical to have one standard deposition agreement and three different annexes regulating intellectual property (IP) matters. IP provisions mainly[12] determine the categorization of resources into PUB, ACA or RES.

According to the second observation, the provisions on the warranties and indemnity are deemed essential clauses[13] of DELAs and therefore require special scrutiny. In the previous versions of DELAs, Section 10 regulates liability and indemnity. The provision was revised to simplify the regulation. In the following table, the previous and amended Section 10 are presented in parallel:

| The previous provisions | The amended provisions |
|---|---|
| 10. Legal Obligations | 10. Warranties and indemnity |
| 10.1 The Copyright holder shall be responsible for holding copyright or a sufficient license and/or other rights based on intellectual property law to the Resource and that any use of the Resource for the purposes compliant with this Agreement does not in any form violate any third party copyright or any other rights based on intellectual property law or other incorporeal right. | 10.1 The Depositor warrants and represents that (i) it possesses all proprietary rights, title and interest in the Resource and has full authority to enter into this Agreement. The Depositor shall be responsible for holding copyright, related rights and other rights or a sufficient license and/or other rights to the Resource and that any use of the Resource for the purposes compliant with this Agreement does not in any form violate any third party copyright, related rights or any other rights. |
| 10.2 The Copyright holder is held liable for all damages and costs he causes CLARIN or the Trusted Centres in the CLARIN Service by breaching any of the obligations in 10.1. | 10.2 The Depositor undertakes to indemnify and hold harmless the Repository for any liability, directly or indirectly, resulting from the use and distribution of the Resources, including but not limited to claims from third parties. The Depositor is held liable for all damages and costs he causes CLARIN or the Trusted Centres in the CLARIN Service by breaching any of the obligations in 10.1. |
| 10.3 Should a third party present a justified claim that the Resource violates the obligations in 10.1., the Resource can be removed from the CLARIN Service. | 10.3 Should a third party present a claim that the Resource violates the obligations in 10.1., the Resource can be removed from the CLARIN Service. |

[11] There are also some differences in the annexes, which are easily brought into conformity.
[12] The use of personal data also has an impact on distribution of resources and their categorization.
[13] Another important part is the license granted to repositories by the owners of resources and technologies.

In the previous version, Section 10 is named "Legal Obligations". This is not the most appropriate wording since all obligations arising from a contract are legal. Therefore, Section 10 should be called "Warranties and indemnity". Subsection 10.1 and 10.2 are elaborated further to increase clarity. Subsection 10.3 was amended to provide sufficient grounds for the removal of resources if a third party files a claim due to the infringement of her rights, and the repository is under no obligation to prove that the claim was justified. In addition, Subsection 10.3 must be compatible with the CLARIN Notice and Take Down Policy.

The amended version also reflects a new terminological approach. The DELA terms identifying the parties to the agreement are replaced as follows: the Copyright curator (CLARIN Centre receiving LR and LT) is replaced with "repository" and the Copyright holder (person licensing LR and LT) with "depositor".

DELAs use the term "distribution" in a broad sense. This could cause misinterpretations since international conventions, the EU directives and national laws usually confine "distribution" to the context of tangible goods. For instance, Article 6 (1) of the WIPO Copyright Treaty (WCT) defines "distribution" as making works or their copies available to the public through the sale or other transfer of ownership. According to the agreed statements concerning Articles 6 and 7 added to WCT, the right of distribution under the said Articles refers exclusively to fixed copies that can be put into circulation as tangible objects (WCT 1996). Article 4 of the Directive on the harmonization of certain aspects of copyright and related rights in the information society (Information Society Directive 2001) has a similar approach. Language resources, however, are not distributed in a tangible form but are made available on-line. Acknowledging this discrepancy, there are two options: 1) replace the term "distribution" with "communication to the public" which is an umbrella term encompassing any communication to the public, by wire or wireless means, including making available to the public in such a way that members of the public may access them from a place and at a time individually chosen by them (Article 3 (1) of the Information Society Directive) or 2) define the term "distribution" widely so that it is compatible with the actual practice. Opting for the first could have created new issues. Some resources have already clearly been deposited using DELAs containing the term "distribution". Should we now decide to modify DELAs and replace "distribution" with "communication to the public" then this could give rise to a question whether resources deposited prior to the change only provide for the distribution of resources in a tangible form (e.g., on CD, print-outs, etc.). Any chances of a misunderstanding ought to be avoided at the outset. Therefore, the second option was preferred. The current version of DELA (Section 3) defines "distribution". According to the definition "*Distribution* means, inter alia, making available to the public, communication to the public and distribution to the public by wire or wireless means."

## 3 Development and distribution of language resources

Two tiers of rights are applicable to language resources: 1) the rights of the persons who put their intellectual effort into developing the resources (within employment transferred to the employer) and 2) the rights of the persons whose copyright-protected content (sometimes also content with related rights) was used for creating the resources. For an illustration of the tiers, see Figure 2.

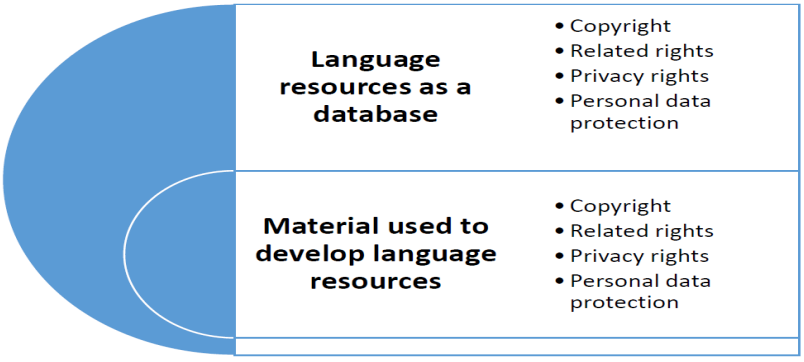



Figure 2. The two tiers of rights covering language resources.

From a legal perspective, language resources constitute copyright protected databases (Kelli et al. 2012; Tavast et al. 2013). The creation of language resources often requires the use of copyright protected works. The use of copyrighted material can be based on two models: 1) the contract model and 2) the exception model. Both models have their strengths and weaknesses.

### 3.1 The contract model

The contract model means that a person developing language resources acquires a permission (a license) to use copyrighted works (books, journal articles, etc.). The contract model allows negotiation of suitable terms for commercial use of copyrighted material to develop resources. This model contains two major problems. Firstly, the contract model involves high administrative costs relating to negotiation of contractual terms and management of contracts (especially in the absence of standard agreements). Secondly, incompatibilities between different contracts could restrict the development and distribution of resources. It should also be borne in mind that there are *de facto* orphan works (anonymous web posts, blogs etc.). Their authors are not reasonably identifiable and there is no one who can grant permissions for their use. Therefore, the contract model is regarded as expensive and non-functional.

Finland has opted for the contract model. FIN-CLARIN has refrained from paying for resources but has contributed a minimal sum towards the collective extended license[14] for the Finnish billion-word newspaper corpus which has been scanned and OCRed by the National Library of Finland comprising newspapers from 1792 to date. In December 2015, the FIN-CLARIN collective license has been extended to cover the copyright of all printed works in Finland subject to the consent of the editor and the publisher. However, most *de facto* orphan works are still not included but require separate agreements. FIN-CLARIN provides access to the full corpora for non-commercial research purposes and access to anyone for small excerpts based on search results. Similarly, the billion-word blog Suomi24 maintained and distributed by the commercial company AllerMedia is available through a separate agreement in full for non-commercial research purposes via FIN-CLARIN and as excerpts for anyone. The motivation for this by AllerMedia is that the company welcomes and encourages ideas developed by the research community by facilitating access to the data, and looks forward to providing access to the data for commercial companies against payment of a fee.

### 3.2 The exception model

The exception model is based on a copyright exception allowing the free use of works for research purposes. For instance, Section 19 of the Estonian Copyright Act provides a general research exception allowing, inter alia, the use of copyright protected works for the development of language resources. The development of language resources in Estonia takes place within the framework of the research exception. It should be noted, however, that there is no case law regarding the exact scope of the exception. For the sake of legal clarity the draft Copyright and Related Rights Act introduces a specific exception for data mining and text analysis worded as follows: "reproduction and processing of an object of rights for the purpose of text analysis and data mining, on the condition of attributing the name of the author of the used work, the name of the work and the source of publication, except if such attribution is impossible, and on the condition that such use is not carried out for commercial purposes" (for further discussion on copyright reform in Estonia, see Kelli 2015).

The exception model has several advantages. First of all, there is no need to ask for permissions from the right-holders to use copyrighted content. There is no administrative burden to negotiate licenses. It is equally possible to use works of identified authors and works of unidentifiable authors (*de facto* orphan works). The main disadvantage is that it is not possible to use the developed resources for

[14] The National Library is authorised by law to make the works available within its premises. Works whose copyright has expired can be displayed on the National Library website, but anything after 1911 is still regarded as potentially containing copyright and therefore needs a license by the collected extended license provided by the collecting society Kopiosto (http://www.kopiosto.fi/kopiosto/).

commercial purposes[15] or make the entire resource available in the PUB category. Distribution of the entire resource is possible only in the ACA and RES categories.

In this context, it is necessary to consider policy issues as well. Data originally created by researchers for research purposes should have public or open licenses. Data created for other purposes can enter the research domain by licensing or by a statutory research exception. The license template for ACA states that the teaching, education and research-purpose is the underlying criterion for ACA resources. The question is how willing the right-holders are to provide resources for research without any further control over the usage and distribution. This is also the fundamental concept behind a statutory research exception.

Typically, a statutory research exception makes data available only for non-commercial purposes. In addition, statutory research exceptions do not usually recognize intermediaries such as research infrastructures hosting the data on behalf of right-holders. Statutory exceptions often only allow data to be used by (but not distributed by) individuals, so a research infrastructure like CLARIN may need to license the data to distribute it even in the case of such a statutory exception for research in the EU copyright legislation.

### 3.3 Right-holders concerns and an interim solution

A relevant question is what additional conditions right-holders could impose to let CLARIN distribute their data for them? To answer this question we should examine the level of verification of the teaching, education and research-purpose the right-holders themselves apply to someone seeking their data. Is it sufficient for them that the user has a researcher status, an IP address at a research institution, a research plan or a self-declared research-purpose?

Currently an IP address at a research institution seems to be enough for many publishers to recognize a vague identity in order to give access to individual works. An identified user with researcher status would provide much stricter identification, and it would have the advantage of not tying the researcher to a location. However, if the researcher status or research purpose cannot be reliably verified, a proxy such as an IP address at a research institution also seems to suit the publishers.

If a mere IP address is acceptable, then maybe it is rather secondary that the user attributes (provided by an identity federation like EduGain) only roughly match a researcher identity. It probably suffices to distinguish between persons affiliated with non-commercial research institutions as opposed to guest users in order to convince the right-holders that a research infrastructure like CLARIN can safely distribute data for non-commercial research purposes. Distributing to a limited, albeit a substantial, number of users so as not to destroy the remaining market in the process is the main concern of right-holders regarding a broad research exception.

While waiting for the adoption of a broad statutory research exception in the EU-wide copyright legislation that will include infrastructure facilities like CLARIN ERIC, we propose the ACA license category to serve the same purpose. Until such a statutory research exception has been implemented, the resources in the ACA category require explicit agreements between CLARIN Centers and the right-holders.

### 3.4 Distribution of resources

Language resources are made available within CLARIN through a specific contractual framework. Firstly, a person interested in using the resources agrees to accept the Terms of Service (TOS) further specified in DELAs and EULAs. DELA is a resource specific set of usage permissions and restrictions while TOS is the general framework. When DELA shifts all liability regarding language resources to depositors, TOS disclaims and limits CLARIN's liability regarding resources to the maximum extent allowed by law. The users can access resources on as-is and as-available basis. CLARIN should not be held liable for potential errors and "bugs" in the resources. Resources are considered work in progress. Drawing on public licenses such as the European Union Public Licence (EUPL), the GNU General Public License (GPL) and Creative Commons (CC), we amend Section 5 of the TOS so that it is abso-

[15] The issue whether commercial research is allowed under the exception model remains somewhat controversial. Since the author does not get any remuneration when her works are used, then it is more likely that the research for the commercial purposes is not allowed.

lutely clear that the resources are provided on an as-is and as-available basis and no liability is assumed. In addition to TOS, the prospective user also has to consent to EULA annexed to the language resources and technologies.

As to the distribution of language resources, it is useful to remember the maxim of Roman law stating "*Nemo plus iuris ad alium transferre potest, quam ipse haberet*" (Dig. 50.17.54). This means you cannot grant others more rights to something than you have yourself (see Zimmermann, 1996). In other words, resources developed based on the research exception cannot be licensed in the PUB category.

In this context there are three issues which need to be addressed: 1) the acceptance of resources on as-is basis 2) the distribution of fragments of resources and 3) the distribution of virtual resources.

A question that CLARIN Centers could face is whether they should accept resources on an as-is basis. In case the depositor does not have rights to a resource as a database (the first tier of rights), it is ultimately out of the question. If the depositor has rights to a resource as a database but its development was based on the copyright exception and/or it includes personal data, the resource can probably be accepted. It can most likely be made available as an ACA or RES resource.

Another frequently raised relevant issue is the difference between licensing whole works as opposed to fragments of works. Copyrighted fragments are derived works of the original and therefore might hold the same license as the original unless otherwise agreed. In some cases, the original right-holders may be willing to permit distribution of derived fragments, e.g. sentences or paragraphs through a concordance tool such as https://korp.csc.fi, while being reluctant to provide a license to distribute the full work. The question is how to indicate this in the Deposition License Agreement (DELA) and in the metadata description of the resource without giving a misleading representation of the end-user's rights.

If a resource cannot be distributed under any circumstances, it has a proprietary license that does not even fall into the CLARIN RES category. In practice, the original may not be available to anyone besides a search tool provider. However, search results consisting of fragments such as sentences or paragraphs may still be available to anyone in the CLARIN PUB category. Describing the resource as being accessible in the CLARIN PUB category would be a misrepresentation, because it gives a false impression that the whole work is available for download.

One solution is to explicitly state in DELA that derived fragments can be distributed through a search interface and specify the license applicable to the search results. Even if the original resource cannot be distributed, the search results, i.e. the derived works, can be distributed in e.g. the CLARIN PUB category based on DELA.

The remaining problem is technical because such a virtual derived work cannot necessarily be given a persistent identifier (PID) in advance, even if the underlying resource has a PID. A new derived work will arise every time someone makes a search, i.e. a search produces a virtual corpus. One solution is to provide the search parameters with a PID on the fly. The search parameters are a combination of the query, the PID of the search engine and the PID of the underlying corpus. The search parameters are comparable with a small program which operates a search engine on a corpus, and can be regarded as a work on its own. The search program can be efficiently stored and persistently identified to be shared with others reproducing the virtual corpus as a search result. Similar solutions are needed for distributing virtual corpus collections in a federated search environment.

## 4 License category calculator

The license category calculator is a tool for assigning metadata to a license and now also helps the depositor determine the right DELA when depositing a resource in a CLARIN repository. The old license category calculator only provided the license metadata. The new calculator also proposes a DELA that can be signed by the depositor and the repository.

To fully grasp what an end-user can do with a resource, CLARIN provides a resource with license metadata, also known as "laundry tags" or license categories. The goal is to have icons for each laundry tag to make the permissions and restrictions visually recognizable for the end-user. The idea for license subcategories was adopted from the Creative Commons (CC) initiative. Based on a survey and a trial labeling of licenses of several hundred resources from various European countries (Oksanen et

al. 2010), subcategories were developed based on frequently occurring access conditions for language resources. The subcategories are detailed in the list of questions in Figure 3.

Result: **CLARIN PUB**

[TOS] [EULA] [DELA] [see also Open License Selector]

| | | |
|---|---|---|
| **Identification and Access conditions** | | |
| | Does the user need to be authenticated, i.e. identified? | ○ Yes ◉ No |
| | Does the user need to be affiliated with some specific community, e.g. through a university or research institution (EDU) or a community of language resource and technology researchers more generally (META)? | ○ EDU<br>○ META<br>◉ No |
| | Can the user only be given permission to use the resource on a case-by-case basis, e.g., based on a mandatory fee or a research plan? | ○ Yes ◉ No |
| **FF** | Is a fee required to get access to the resource? | ○ Yes ◉ No |
| **PLAN** | Does the right holder require a research plan for granting access? | ○ Yes ◉ No |
| **General use conditions** | | |
| **BY** | Is attribution, i.e. acknowledgement of authorship, required? | ○ Yes ○ No |
| **NC** | Is the content available only for non-commercial purposes? | ○ Yes ○ No |
| **INF** | Is informing the rights owner about the use of the resource required? | ○ Yes ○ No |
| **LOC** | Is the content available only at a single location, center, or site? | ○ Yes ○ No |
| **LRT** | Is the content available only for language research and technology development? | ○ Yes ○ No |
| **PRIV** | Are there personal data in the resource? | ○ Yes ○ No |
| **Distribution conditions** | | |
| **NORED** | Can the user distribute the original resource to third parties? | ○ Yes ○ No |
| **ND** | Can the user distribute derived works, i.e. works containing copyrighted parts of the original? | ○ Yes ○ No |
| **SA** | If the user can distribute derived works, should the same license be used, i.e. is the license reciprocal? | ○ Yes ○ No |
| **DEP** | If the user cannot distribute derived works, is the user still allowed to distribute modified versions via CLARIN? | ○ Yes ○ No |
| **Other conditions** | | |
| * | Are there other non-standard conditions in the license that the user should pay attention to? | ○ Yes ○ No |

Figure 3. License Category Calculator

To assist CLARIN Centers with the labeling and classification of licenses when the depositor offers a language resource for distribution via CLARIN, a license category calculator has been developed: https://www.clarin.eu/content/clarin-license-category-calculator. A new version of the license calculator also provides CLARIN Centers with a set of license templates outlined in the previous sections including the Terms of Service (TOS), the End-User License Agreement (EULA) and the Deposition License Agreement (DELA). The variable content of the templates corresponds to the conditions that can be identified in an existing resource license or in the wishes of the depositor of a new language resource. The conditions include commonly occurring restrictions and permissions with regard to user identification and access as well as resource usage and distribution. By answering the yes-no questions in the license category calculator, the laundry tags on the top line of the calculator are interactively updated. When the relevant questions have been answered, the depositor can click on the EULA and

DELA buttons to view the EULA and to print and sign the corresponding category-specific DELA before submitting the resource to a CLARIN repository[16]. For an illustration, see Figure 3.

Other license templates may be used by a CLARIN center if such templates effectively comply with e.g. national legislation. The templates provided by CLARIN ERIC are intended to serve as a first-aid kit for CLARIN Centers that have not yet developed an electronic resource submission workflow. In an electronic submission workflow, many of the boilerplate provisions in the deposition agreement template (DELA) can be incorporated into the Terms of Service (TOS) and the end-user conditions can be verified by an interface like the license category calculator so as to leave virtually only the end-user license conditions (EULA) for approval by a depositor.

## 5 Conclusion

The regulatory and contractual frameworks constitute an integral part of the CLARIN infrastructure. The regulatory framework is adopted by the CLARIN member states of the European Union. CLARIN can influence its development by issuing policy recommendations and using other means of lobbying. However, the contractual framework is adopted by CLARIN itself and enacted by CLARIN members, and therefore, CLARIN has direct control over its contractual framework.

The CLARIN contractual framework is based on a conceptual division of language resources together with their corresponding agreements into three main categories: PUB (publicly and openly available), ACA (research and academic use) and RES (restricted to individual use). The authors support this tripartite categorization since it comprises more or less all language resources, it has proven its practical value and is well-integrated and socially embedded in the CLARIN infrastructure.

A review of CLARIN agreement templates has to be conducted within the context of development and distribution of language resources and take into the account the legal nature of resources. Resources are usually governed by two tiers of rights. The first tier of rights covers the resources used as a database. The second tier encompasses the intellectual effort used for creating the database.

The creation of resources usually requires extensive use of copyrighted material. It can be grounded in two models: 1) the contractual model (a license for use of copyrighted material is acquired); and 2) the exception model (the use of copyrighted material is based on a statutory exception). The selected model affects how resources can be used further. Estonia has opted for a research exception covering the development of language resources. For the sake of clarity, the Estonian draft Copyright and Related Rights Act introduces a specific exception for data mining and text analysis. Finland relies on the contractual model. It is in the interest of CLARIN to lobby for an EU-wide mandatory statutory exception for research purposes.

Since the CLARIN license agreement templates have to be integrated and evaluated as a uniform functional system, we proceeded from the three deposition agreements. They are almost identical. The main difference resides in the intellectual property and access provisions. Therefore, it is more expedient to consolidate these provisions into three annexes. This way we have only one deposition agreement and the depositor can choose from among the three annexes, which determine whether the deposited resource is PUB, ACA or RES.

The deposition agreement purports to shift the liability for the resources to their depositors. The wording of the relevant provisions is amended accordingly so that it becomes very explicit that the depositor is liable for the resources and indemnifies CLARIN from any damage claims.

A key issue is how to provide public access to fragments of resources *versus* distributing resources in full for research purposes using the CLARIN contractual framework. A set of excerpts (i.e. search results) may be considered derived works subject to the same conditions as the original work unless otherwise agreed. We may, therefore, still need a deposition agreement to acquire the right to distrib-

[16] Note that also publicly available repositories like GitHub or Sourceforge have elaborate terms of service, e.g. https://help.github.com/articles/github-terms-of-service/ or https://slashdotmedia.com/terms-of-use/, that require depositors of resources to adhere to strict policies even if the deposition processes are streamlined to make the deposition as smooth as possible. GitHub in particular requires the resource to have an open license in order for the resource to be freely and openly distributed by GitHub. For a more limited distribution, GitHub charges a monthly fee. CLARIN distributes public and open resources with, e.g. CC licenses in the PUB category, and resources with more limited licenses within the confines of the ACA and RES categories.

ute the search results publicly. In most cases, right-holders are willing to make excerpts publicly available while the full corpus is only distributed for academic or restricted purposes. In case there is no research exception, this can still be agreed on in a deposition agreement with a relevant annex (PUB, ACA, RES). In both, the resource needs two metadata records: one applicable to the PUB excerpts and the other applicable to the original ACA/RES resource, i.e. we have data with two different uses provided by two licenses in one agreement.

The next central agreement in the CLARIN system is the Terms of Services (TOS). Before users can access resources, they have to agree to the Terms of Services. The objective of TOS is, inter alia, to limit CLARIN's liability towards users. Resources and technologies are offered on an as-is and as-available basis. The wording of the provisions regulating liability in TOS is amended to limit CLARIN's liability to the maximum extent permitted.

A license category calculator has been developed to assist CLARIN Centers with the labelling and classification of licenses when language resources are offered for distribution. The article provides information on a new version of the calculator automatically generating CLARIN End-User and Deposition License Agreements with reference to the Terms of Service.

## References


[CC] Creative Commons. Available at http://creativecommons.org/licenses/ (24.11.2015);

[DELA]. CLARIN PUB Deposition License Agreement Template 1.1. Available at https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinDelaV11?PUB=1 (24.11.2015);

[Estonian Copyright Act] Autoriõiguse seadus (valid since 12.12.1992). RT I 1992, 49, 615; RT I, 29.10.2014, 2 (in Estonian). Unofficial translation available at https://www.riigiteataja.ee/en/eli/531102014005/consolide (12.07.2015);

[EUPL] European Union Public License. V. 1.1. Available at https://joinup.ec.europa.eu/sites/default/files/eupl1.1.-licence-en_0.pdf (24.11.2015);

[GPL] GNU General Public License. V. 3. Available at http://www.gnu.org/licenses/gpl-3.0.en.html (24.11.2015);

[Information Society Directive 2001] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society. - OJ L 167, 22.6.2001, p. 10-19;

[Kelli 2015] Aleksei Kelli (2015). The Conceptual Bases for Codifying Estonia's IP Law and the Main Legislative Changes: From the Comparative Approach to Embedding Drafted Law into the Socio-Economic Context. – International Comparative Jurisprudence 1 (1), 44-54. Available at http://www.sciencedirect.com/science/article/pii/S2351667415000050 (24.11.2015);

[Kelli et al. 2012] Aleksei Kelli, Arvi Tavast, Heiki Pisuke (2012). Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach. – Juridica International (19), 40-48;

[Dig. 50.17.54]. Available at http://www.thelatinlibrary.com/justinian/digest50.shtml (13.7.2015);

[Licenses, Agreements, Legal Terms]. Available at http://clarin.eu/content/licenses-agreements-legal-terms (13.7.2015);

[Oksanen et al. 2010] Ville Oksanen, Krister Lindén, 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. Available at https://helda.helsinki.fi/handle/10138/29359 (18.11.2015);

[Tavast et al. 2013] Arvi Tavast, Heiki Pisuke, Aleksei Kelli (2013). Õiguslikud väljakutsed ja võimalikud lahendused keeleressursside arendamisel (Legal challenges and possible solutions in developing language resources). – Eesti Rakenduslingvistika Ühingu Aastaraamat (9), 317-332;

[The draft Copyright and Related Rights Act] Autoriõiguse ja autoriõigusega kaasnevate õiguste seaduse eelnõu. Versioon: 21.7.2014 [The Estonian draft Copyright and Related Rights Act. Version: 19.7.2014]. (in Estonian), https://ajaveeb.just.ee/intellektuaalneomand/wp-content/uploads/2014/08/AutÕS-EN-19-7-2014.pdf, (accessed on 5 May 2015);

[Zimmermann, 1996] Reinhard Zimmermann. The Law of Obligations Roman Foundations of the Civilian Tradition. – Oxford University Press, 1996.

[WCT 1996] WIPO Copyright Treaty. Adopted in Geneva on December 20, 1996. Available at http://www.wipo.int/wipolex/en/treaties/text.jsp?file_id=295166 (20.11.2015).

# Variability of the Facet Values in the VLO – a Case for Metadata Curation

**Margaret King**
ACDH-OEAW
Vienna, Austria
marga-ret.king@oeaw.ac.at

**Davor Ostojic**
ACDH-OEAW
Vienna, Austria
davor.ostojic@oeaw.ac.at

**Matej Ďurčo**
ACDH-OEAW
Vienna, Austria
matej.durco@oeaw.ac.at

**Go Sugimoto**
ACDH-OEAW
Vienna, Austria
go.sugimoto@oeaw.ac.at

## Abstract

In this paper we propose a strategy for metadata curation especially with respect to the variability of the values encountered in the metadata records and hence in the facets of the main CLARIN metadata catalogue, the VLO. The approach concentrates on measures on the side of the infrastructure and on the interaction between human curators and the automatic processes.

## 1 Introduction

CLARIN runs a mature well-established metadata infrastructure, harvesting metadata from more than sixty providers on a weekly basis using the standardised OAI-PMH[1] protocol. Up to a million records[2] are being collected and provided via the main metadata catalogue, the Virtual Language Observatory or VLO (Van Uytvanck et al, 2010). It aims to provide access to a broad range of linguistic resources from many disciplines and countries based on the flexible metadata framework CMDI (Broeder et al., 2010; Broeder et al., 2012). After a few years of intensive use by the community and continuous growth of the body of data made available via this service, a number of issues have been identified (Broeder et al., 2014) concerning the functionality of the catalogue, but mainly the quality of the metadata provided by the data providers such as the variation in metadata values. These irregularities seriously hamper the discoverability of resources.

After reviewing the work done on this issue in other institutional contexts and within the CLARIN community until now, we concentrate on the issues underlying the problem of variant values within the facets in the VLO, exemplified primarily by the Resource Type facet, and propose a strategy for the implementation of a metadata curation workflow that could rectify (some of) the described problems.

## 2 State of research

The general problem of the curation, harmonisation and normalisation of metadata has been central to libraries, academic and cultural institutions for many years. Thus, before looking at CLARIN's approach to the curation of metadata and normalising facet values within the VLO, we reflect on other institutions' treatment of this issue.

### 2.1 Cases of other communities

Within the library community several approaches have been implemented in dealing with similar issues. Calarco et al. (2014) elaborate at a theoretical level on the role of metadata normalisation in resource discovery. They outline three principles as a basis of a good metadata curation strategy: 1. rigid standards, 2. cooperation with data providers, and 3. technological enhancements. Rigid standards reduce ambiguity and variation, facilitating user discovery. Communicating and including the data providers in



[1] https://www.openarchives.org/pmh/

[2] With considerable fluctuations

the process is the most effective way to avoid future problems. Planning for ongoing technological enrichment of the metadata allows for future development and technical sustainability.

Huffman (2015) shares a case study on normalising variant subject and name values in EAD files. Using OpenRefine[3] and some XSLT processing, He quickly analyses the variability of the values (applying OpenRefine's "cluster and edit" feature) and combines manual inspection and automatic application to reduce the number of unique values by 7%. This procedure resembles in some ways the approach to be proposed later in the paper for the VLO (manual inspection and normalisation of distinct values, followed by automatic application of a normalisation map, see Section 5.4). While OpenRefine in general can be geared toward large-scale "messy" data, being a standalone application which incorporates database-like spreadsheets and on-the-fly facet construction, it relies too much on its own structures, which would make it difficult to integrate it into the VLO workflow which requires an on-going, collaborative process (issues like how to sustain the process over multiple iterations, how to reach agreement on mapping in a large group, etc. would be difficult to resolve).

On the other side of the spectrum, Europeana deals with massive amounts of data in a broad range of data types and formats from many cultural sectors, libraries, museums, archives etc., from many European countries, requiring a robust processing infrastructure. In line with the principles by Calarco et al., Europeana produced comprehensive documentation of the native metadata schema (originally ESE Europeana Semantic Elements (Europeana, 2009), and its successor EDM, the Europeana Data Model (Europeana, 2014)), including definitions of all classes and properties, as well as a number of case studies for mapping from existing formats to EDM[4]. While we can only focus on some aspects relevant to the topic of our paper, we would like to emphasise that the well-defined data model as well as the extensive user and data provider oriented documentation represent best practices that the CLARIN community should use as inspiration for the work on CMDI and metadata curation, especially in the light of the planned harvesting of parts of Europeana data into the VLO.

Europeana's normalisation workflow as it is represented in the guidelines provides several clear applications to the problem of normalising the VLO's variant values. First they designate certain elements as mandatory (and another few as recommended). Starting with required elements is indispensable to ensure a minimal common denominator for describing the resources. The short list of required elements as well as the allowed alternatives is a savvy and pragmatic strategy to balance the data provider-specific situations and the need for a minimal, consistent, descriptive information set.

As for restricting the allowed values of the metadata elements, EDM – in accordance with the Semantic Web principles – instructs data providers to use URL references wherever possible, with some elements allowing both literals and references and others allowing only references. It utilises restrictions of the usable references where possible (e.g. the important required field *edm:rights* that has to take a reference to one of the rights statements endorsed by Europeana[5]). Thus Europeana provides controlled vocabularies in a way that is compatible with the Semantic Web. In one case there is an explicit list of allowed values given. The only element with an explicit list of allowed values is the *edm:type* with the vocabulary: `TEXT, VIDEO, SOUND, IMAGE, 3D`.

Since *edm:type* is semantically closely related to the VLO facet *ResourceType* which serves as the primary example in this paper and is being dealt with intensively by the Curation taskforce, it will be dealt with in more detail here. The type element is innately prone to varying interpretations. Europeana pre-empted this problem by only allowing the five values mentioned above. Initially (in the prototyping phase) Europeana had their providers submit a spreadsheet with mapping for each object to one of the (then) four values (3D was only added in EDM) and applied these centrally. This process evolved toward the content providers supplying the correct term and contacting Europeana only in difficult cases. Meanwhile most of the content of Europeana is provided via the intermediate country- or domain-specific aggregators (The European Library, OpenUp, CARARE, etc.[6]), who take over certain curation tasks. An example of the professionalisation of the aggregation and curation process is also the Europeana's MINT (Metadata Interoperability) platform[7] that aims to "facilitate aggregation initiatives for cultural heritage content and metadata in Europe". Note that next to *edm:type* EDM features also the widely

[3] http://openrefine.org/
[4] http://pro.europeana.eu/share-your-data/data-guidelines/edm-case-studies
[5] http://pro.europeana.eu/web/available-rights-statements
[6] http://www.europeana.eu/portal/browse/sources
[7] http://labs.europeana.eu/apps/mint

used *dc:type*. While *edm:type* occurs exactly once with one of the predefined values, *dc:type* is optional and repeatable and can be used much more broadly, drawing values from any custom vocabulary.

Interestingly the *europeana:unstored* element defined in ESE and recommended to be used for any information that that does not fit into any of the predefined elements, the *edm:unstored* element has been removed as of version 2.2 (Europeana, 2014, p. 46), without any indication of a substitute.

Next to *edm:type* and *dc:type*, EDM also adopts SKOS[8] for classifying resources and to represent controlled vocabularies. This allows for standard-conforming normalisation and enrichment on the part of the content provider allowing them to clean their values and align their metadata to any given thesaurus. This is demonstrated for example in the curation of the new Europeana Sounds collection, or enriching Europeana data with AAT (Art and Architecture Thesaurus)[9].

Some insights can be drawn from this survey of other institutions' approaches to value normalisation with regard to CLARIN's quest to improve the VLO's metadata and especially its facet values. From a smaller case study to value normalisation we gleaned guidelines from another commonly used technology for curation of messy data, OpenRefine, and from a theoretical library study on clean metadata values, we echo the three key ingredients namely, strict adherence to standards, cooperation with data providers, and technological enhancements, finally by an in depth investigation of Europeana's approach, we saw a way to implement these principles on a large scale.

### 2.2 Case of CLARIN

The CLARIN community is acutely aware of the problems concerning metadata quality, especially the variation of metadata values. It has discussed the question of how to curate metadata and especially normalise the VLO's facet values on multiple occasions. A Metadata Curation Taskforce was established in 2013 by the Centre's Committee (SCCTC) with delegates from member countries, however this taskforce until now could only collect ideas, describe the situation and try to remedy some of the encountered problems. It has not yet been able to sustain a sufficient level of concerted activity to systematically approach this problem.

CLARIN-D established a separate VLO Taskforce in October 2013 (Haaf et al., 2014) which worked out recommendations for the VLO facets in an attempt to provide more guidance and clarity regarding the usage and meaning of the facets to the data providers. The VLO Taskforce meetings throughout 2014 and 2015 have brought about small steps towards a solution. However the Taskforce has concentrated on recommendations and sound definitions, the actual implementation is not seen as one of its tasks.[10] A sound definition of the facets and recommended values for the facets is certainly a necessary condition and a good starting point towards answering the problem under consideration. However such definitions are only of use when it is integrated into the infrastructure and taken up by data providers.

In 2014, Odijk conducted an in depth survey of the VLO from the point of view of discoverability of linguistic resources (Odijk, 2014). The comprehensive report identifies a number of concrete issues and some proposed solutions. These identified problems pertain both to the schema level (e.g. crucial elements not obligatory), to the instance level of the data (fields not filled, variation of the values), and also to the functionality provided by the VLO (missing facets, multi-selection). He also underscores the aspect of granularity, a related point currently much discussed throughout CLARIN but one which falls outside the scope of this paper.

In an unpublished, follow-up, internal CLARIN report in 2015, Odijk lays out a strategy for metadata curation, concentrating on the main goal of achieving clean facets. Based on the assumption that "the providers in general case cannot improve their metadata" (Odijk, 2015) the main actor in the curation process is the curation task force operating on the harvested metadata. The main reason why the metadata in CLARIN domain cannot be improved on the side of the data providers seems to be the lack of resources available for improving legacy data. CMDI in its complexity may also pose a steep challenge to data providers with limited resources. It is perhaps an unreasonable expectation for data providers to select the right CMD profile without guidance. Finally, in the provider's own realm the metadata may be perfectly consistent and homogeneous, it is just through aggregation that inconsistencies arise.

[8] http://www.w3.org/2004/02/skos/

[9] http://pro.europeana.eu/share-your-data/data-guidelines/edm-case-studies/europeana-aat

10 as indicated in informal talks with members of the taskforce

## 3 VLO metadata: a closer look

Thus the mission of the CLARIN metadata curation task force in (in normalising the variant facet values) is twofold. In the first place it must analyse the different problems of variation and its effect on discoverability. The second practical aim is to create and implement a strategy for curation within the framework of CLARIN's social structures.

### 3.1 Variation of values

We can identify different types of variation. These vary from trivial ones like case or whitespaces ("WrittenCorpus" vs. "Written Corpus"), to a combination of multiple values in one field with arbitrary (or even no) delimiters (e.g. "AddressesAnthologiesLinguistic corporaCorpus"), similar concepts ("text" vs "written") and, most problematically, complex (confusing) values that carry information that should be assigned to another facet (e,g. "bioscoop" (cinema in Dutch) and "bible" as *ResourceType*).

Odijk points to the data provider's isolation as a main cause for the variation of values (Odijk, 2014). Indeed, it is clear that different people describe things in different ways. Some providers assign the value "text" to "Tacitus' Annals" while others choose to create a new value called "Annals". This assumption is also supported by the fact that once the data is restricted to a single collection or organisation the values in facets mostly "clear up" and appear as a consistent set.

The obvious solution to the problem from the infrastructure point of view is to reach better coordination between the data providers, basically applying shared controlled vocabularies (Durco and Moerth, 2014). Presently the only guidance regarding recommended vocabularies for individual facets is provided in the Recommendations by the VLO Taskforce. Even these vocabularies are rarely used. In the *ResourceType* facet only 15,000 records use one of the 25 recommended values. All in all round 250 different values are used in the *ResourceType* facet. The most common reason for variation is the inclusion of extra information (unrelated to *ResourceType* but to some other facet). For example Shakespeare's King Lear is described by the *ResourceType* "poem", a value which would belong in the Genre facet while the most suitable value for the *ResourceType* is "text". A controlled vocabulary could help data providers to assign the details to the correct facet.

### 3.2 Missing values

Even worse than the variation of the values is the fact that many records do not provide any value for certain facets. Odijk attributes this mainly to the lack of obligatory metadata elements in CMDI and the fact that the metadata authors are often 'blind' to the 'obvious' aspects of their resources, like language or type. For the special case of the *ResourceType* one reason for omitting it may be that it is implicitly provided in the name of the underlying CMD profile (e.g. *TextCorpusProfile*, *LexicalResourceProfile*).

Whatever the reasons, the extent of the problem is alarming. Some facets cover only about one third of the records, so that of 631000 records found in the VLO at the time of writing, typically around five hundred thousand are not visible and findable in each facet (except for the automatic/obligatory ones: *Collection*, *Data Provider*, as well as well-recorded *Format* and *National Project*). Figure 1 lists the number of null values for each facet. Given these alarming figures, it is clear that facet browsing, one of the most significant functionalities of the VLO, is not effective for resource discovery, calling for urgent action.

A minimal remedy (or rather a "patch") to the problem of facets without specified values is to make this information explicit to the end-users (e.g. with a default value "[missing value]"). A more advanced solution is to borrow values for certain facets from other facets or metadata fields, for example filling Continent facet based on values from Country facet. We aim for complete coverage, i.e. every record should be represented at least once.

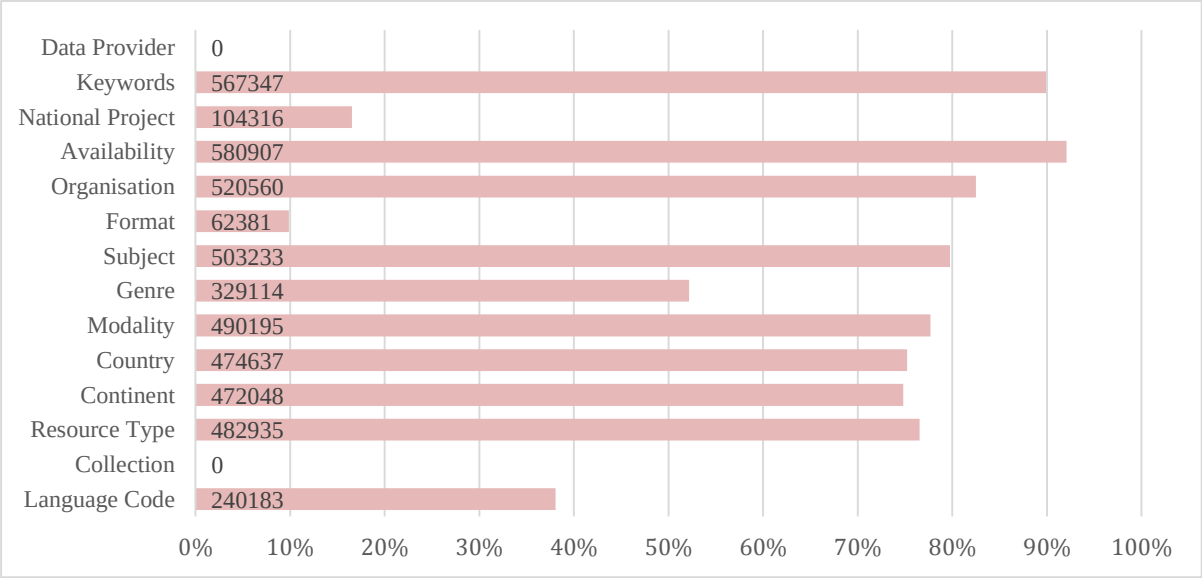



Figure 1 Number of records not covered within given facet in the VLO (on a sample of 631 000 records)

### 3.3 Missing facets

One source of the problem of confusing values may be the lack of appropriate facets. When normalising the values of the *ResourceType* facet it is sometimes unclear, in dealing with an overloaded value, exactly where the extra information should go. For example, the medium of information such as radio, internet, mobile phone as well as more technical entries do not have a clear value among the recommendations for this facet. This lack of facets is also identified by Odijk (2014), who suggests adding a dedicated facet *Linguistic Annotation*, as well as by the VLO task force, proposing new facets *Lifecycle Status, Rights Holder and License*. However adding more facets also raises the complexity of the user interface and mapping, so that the impact of such additions would need to be carefully examined.

### 3.4 Missing guidance

A recurring theme in the CMDI/VLO analysis is the lack of coherent guidance for the data providers. Some of the problems with VLO facets may be caused by the data providers not being aware enough of complex mappings taking place in the VLO ingestion process (see Section 4), especially the concepts-to-facets mapping. Their main concern (and a requirement to submit data to CLARIN) is to comply with CMDI and not with VLO facets, CMDI is not specifically designed for the VLO and its scope is broader, but it is indeed a precondition of the latter. Therefore, it is absolutely essential to provide clear guidance for the two aspects. In this respect the facet map checking tool provided by Windhouwer[11] is very useful, but it is not widely circulated nor is it integrated into a coherent set of guidelines for the data providers. The issue of guidance and proposed solutions are further discussed in Section 6.1.

### 3.5 Need for an efficient curation workflow

As mentioned above much effort has been done to establish the types of problems that exist in the area of facet value normalisation, most notably in Odijk (2014). While some of the trivial problems in value variation can be solved programmatically (case folding, whitespace normalisation), all of the more complex issues like synonyms and complex values require human input – a mapping of variant values to recommended ones. A few tentative mappings have been created as a result of the analysis done by Odijk or the team of authors. Besides the question of the reliability of and broader agreement about such mappings, the next challenge is how to integrate such a mapping into the established harvesting and ingestion workflow, especially how to ensure a sustainable and consistent process over time.

[11] https://lux17.mpi.nl/isocat/clarin/vlo/mapping/index.html

Some automatic curation steps have been applied during the ingestion of the metadata into the indexer for some time (the so-called "post-processing"). Initially, this was limited to simple programmatic corrections of values. Gradually mappings between actual and normalised values were applied to individual facets (*Organisation, Availability, Language, nationalProject*). What is especially missing is a procedure to ensure that the mappings are kept up to date (new previously unseen values are added and mapped) and that the curation process has access to the most current version of the mappings. Meanwhile a more elaborate process is being implemented which is described in, Section 5.4.

## 4 The mapping and normalisation mechanism

The previous chapter introduced and analysed the issues of metadata quality based on the observation and statistics of the records. This chapter focusses on the underlying mapping and normalisation mechanisms and their impact on the present problems of metadata variability. This section presents the current setup followed by three approaches which will be evaluated by the ability to achieve an improvement in data integrity, the discoverability of resources, and the usability of the VLO.

Before discussing the details of the mapping and normalisation, it is important to touch upon the underlying mechanisms delivered by the CMD framework. The principal interoperability mechanism devised by the CMD framework is the linking of individual CMD elements defined in the schema (or CMD profile) to well-defined concepts (Broeder et al. 2010). This delivers sound semantic grounding of the defined elements independent of the structural aspects of the schemas. Moreover by reusing the same concepts in multiple schemas they can serve as crosswalks. Moving from traditional pair-wise crosswalks (between each pair of schemas) to a conceptual pivotal layer to map individual schemas is a far-reaching paradigm shift.

Even though the CLARIN community and this paper concentrate on the problems of the CMDI framework, some 200 defined CMD profiles, a number of concepts linked from dozens or even hundreds of profiles, and the VLO (and other exploitation applications) relying on this semantic interoperability layer, are a solid demonstration that this approach indeed works. Thus when exploring the alternative mappings in the following sections, we need to bear in mind that CMD, through the use of concept links, already delivers a first (substantial) reduction of variability (on the schema level) using a many-to-one mapping between metadata elements and concepts (see part 1 in Figure 2).

When presenting the mapping scenarios we distinguish three factors or degrees of freedom which have an effect on the display of CMD records in the VLO: 1. *schema mapping* (elements in different schemas referring the same concept as described above), 2. *facet mapping* (multiple concepts mapped to one facet in the VLO), 3. *value mapping* (or "normalisation", values encountered in the metadata replaced with values from a controlled vocabulary according to a normalisation map).

### 4.1 Scenario 1 (current configuration)

The current procedure of concept mapping and value normalisation is illustrated with an example. A given CMD record contains *ms:OrganisationName* "Summer Institute of Lingustics" and *olac:Type* "Diaries". Like all other CMD records, the format/structure of this record is defined in a XMLSchema derived from a specific CMD profile. A record from another data set defined by another CMD profile includes ex:AgencyName "RADIO ORANJE" and ex:ItemType "plainText". The abovementioned mechanism of concept links (schema mapping) ensures that these elements are also semantically grounded. For instance, *ms:OrganisationName* is linked to the CCR concept *ccr:C-2459*, whereas *olac:Type* element corresponds to a concept *dc:Type*[12]. Similarly, *ex:AgencyName* is mapped to *ccr:C-2979*, whereas *ex:ItemType* is defined as an equivalent to *ccr:C-5424*.

In the ingestion process the framework uses a facet mapping file which defines the mapping between concepts and VLO facets (facet mapping). In our case, *ccr:C-2459* as well as *ccr:C-2979* are mapped to the VLO Organisation facet. Likewise, *ccr:C-5424* and *dc:type* are mapped to VLO *ResourceType* facet[13]. In parallel, normalisation of the values takes place for selected facets (value mapping). Values

[12] CCR is (by design) not the only conceptual reference used for VLO. Especially DCMI was since the beginning handled as equally valid source of concepts.

[13] The labels Resource Type and Resource Class have been used inconsistently and seemingly synonymously for the facet and the corresponding concept. We strongly encourage the standardisation of the labelling convention. We maintain "ResourceType" throughout this paper.

in the metadata elements are mapped into values from controlled vocabularies according to a manually maintained normalisation file. For instance, "Summer Institute of Linguistics" and others including "SIL of University of Oklahoma" in the Organisation facet will be normalised as "SIL". The values such as "Fictions" and "Diaries" are mapped to "plainText" in the Resource Type facet. We will not discuss whether those mappings are correct or not, but we merely present here how data is curated in the VLO ingestion pipeline.

At first glance it is not problematic, however considering that the semantics of the VLO facets are not yet clearly defined and agreed upon, we face some semantic problems for resource discovery. Cases may arise where the Organisation facet includes not only organisations providing language resource, but also some auxiliary organisations involved in text processing, or as sponsor, depositor or license holder. The same holds true for other facets, like *Country* and *Date*.

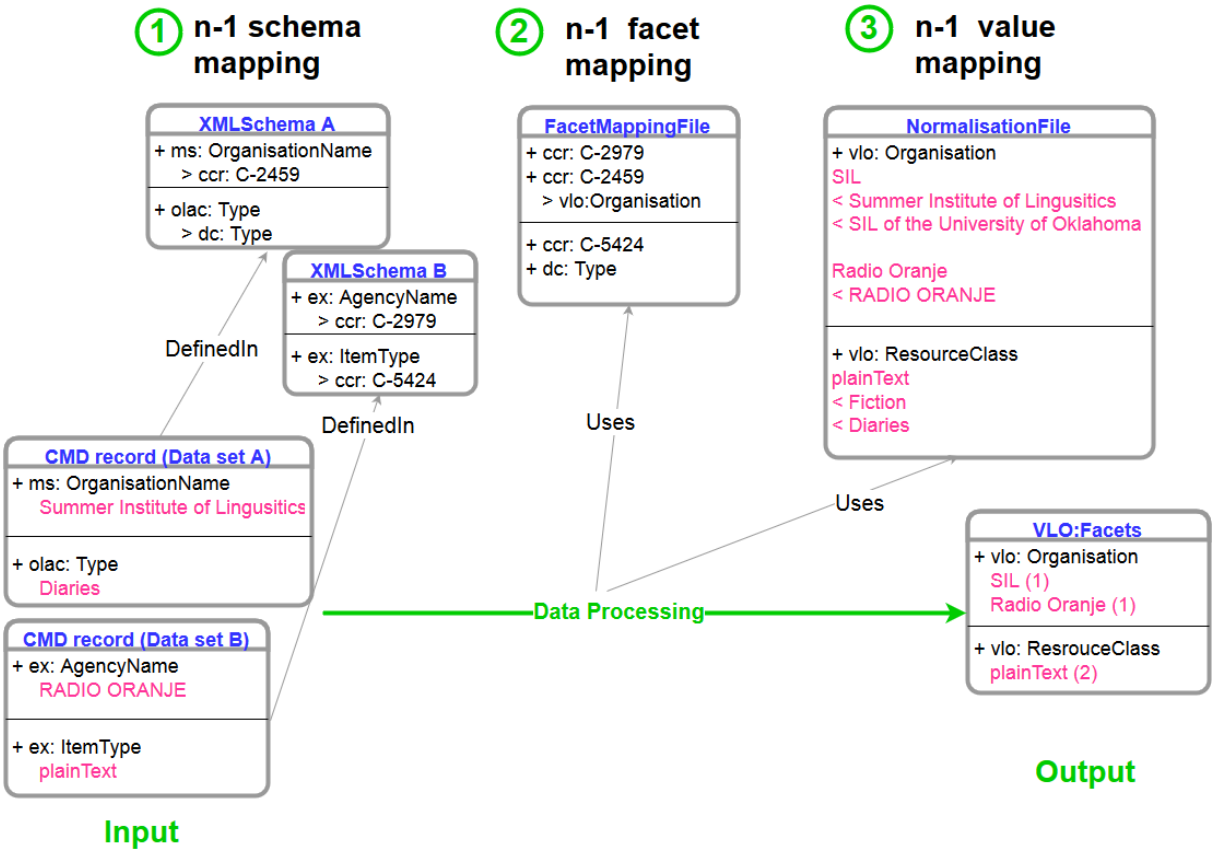



Figure 2. Current scenario with two example records.

Principally, it is acceptable to have a very broad definition of the facets, as long as this definition is made clear to the providers and users. The CCR working group is currently formulating such solid definitions based on their previous work. However, there is a strong sentiment in the CCR and metadata task forces that the facets' definitions should be narrower and stricter, in order to provide clarity to the users and offer them the (supposedly) most popular semantics of the facets. There are good arguments for this direction, but it further undermines the issue of (already poor) facet coverage. In addition, as long as we apply a many-to-one concepts-to-facet mapping in the VLO configuration, the definition of the latter is necessarily broader than the former. By enlarging the scope of the definitions, we deliver a wide spectrum of concepts defined by the data providers and at least sustain the coverage of records for the VLO facets. In any case it is necessary to assess the impact of restricting the semantic scope of the VLO facets on the facet coverage.

### 4.2 Scenario 2 (one-to-one facet mapping)

In scenario two, the mapping between the concepts and VLO facets will be one-to-one (Figure 3). In the *FacetMappingFile*, one concept corresponds to exactly one counterpart facet. (In all probability in this scenario the *FacetMappingFile* will be obsolete and the CCR concepts will be used directly to define

the indices/facets in the VLO.) For instance, the first CMD record of the previous scenario has the same XML schema, but this time *FacetMappingFile* defines that *ccr:C-2459* will be mapped to *vlo:Organisation* and *ccr:C-5424* will be mapped to *vlo:ResourceType*. The slightly different second CMD record includes "University X" in *ex:PublicationInstitution* field and "plainText" in *ex:Category* field. It will use its own XMLSchema to define a mapping to *ccr:C-xxxx* and *ccr:C-yyyy* concepts. These concepts will further reach *vlo:PubOrganisation* and *vlo:ResourceCategory* facets.

The one-to-one mapping relieves the semantic mismatch problem – every facet carries the semantics of the underlying concept. The problem is that hundreds of concepts are linked from the many defined CMD profiles/schemas, many of them semantically similar (which is exactly the reason why a facet mapping is being applied in the first place), so this option will pass the ambiguity and semantic proximity problems to the user. Furthermore, a user interface with dozens of facets would not be user-friendly. Usually, faceted search interfaces do not employ more than 10 facets (e.g. Europeana six, Gallica nine), and 100% or high data coverage is expected.

With this in mind, a joint effort of the curation and CMDI task forces agreed in October 2015 to explore alternative display methods and user interface layouts, in order to accommodate a substantially higher number of facets/indices, especially concentrating on dynamic or conditional and hierarchical facets. Conditional facets are displayed only under certain circumstances, e.g. bound to a certain resource type, or collection, a given coverage ratio or explicit user selection. This would allow (advanced) users to customize which facets should be displayed.

Hierarchical facets have a potential to resolve some of the limitations of the semantic mapping. For example the user will be able to search in a broad facet *vlo:Organisation*, but will have the option to narrow down to *vlo:PubOrganisation*, as illustrated in Figure 3. Although the implementation would be very challenging and would require substantial development resources, the feature has a potential to become an innovative practice. The development will definitely need to be accompanied by an extensive analysis of the usability of the hierarchical facets, because it is not a very widespread functionality, thus it may cause a reverse effect, confusing the end users.

Also, the facet hierarchy (the hierarchical relations between the concepts) as well as the other dependencies (conditions) between the facets still needs to be defined somewhere, basically moving the facet mapping challenge to another level. Nevertheless this approach may be useful as it makes the concept-to-facet mapping more explicit and transparent to the users. Indeed it would allow to move the mapping from indexing time to query time, yielding a much more flexible exploration interface. The faceted browser developed by the Meertens Institute[14] already adopts this strategy to a certain extent.

We also need to take into account, that the one-to-one facet mapping would have a strong impact on the value mapping, as the normalisation maps are currently defined per facet. When the number of facets would grow considerably, so would potentially also the number of needed normalisation mappings.

### 4.3 Scenario 3 (dumb-down schema mapping)

The third option features the same one-to-one facet mapping, but it differs from the previous one in that the number of relevant concepts used are reduced to the minimum by ensuring that the relevant concepts used in the schemas exactly match the VLO facets (see part 1 of Figure 4) The *FacetMappingFile* will become obsolete, given the one-to-one facet mapping. The disadvantage is obvious. During the mapping from original metadata elements to the CCR concepts, some semantics are lost. For example, *ms:Organisation* and *ex:PublicationOrganisation* are dumbed down to the general *vlo:Organisation* facet. With regard to value mapping, this scenario would result in even higher variability and ambiguity of the values in each facet. We consider this scenario a purely theoretical option, as it would require the change of the definitions of most of the existing CMD profiles and would introduce a significant loss of semantic precision in the schema definitions contrary to the basic principles of CMDI.

### 4.4 Scenario 4 (many-to-many value mapping – multi-facet decomposition)

This scenario further develops the idea of value mapping introduced above, introducing the "multi-facet decomposition", i.e. one value can be mapped to multiple values in different facets.

[14] http://www.meertens.knaw.nl/cmdi/search/

For example, a metadata element with values "diaries" or "Bibles" in the *olac:Type* field that is mapped assigned to the *vlo:ResourceType* facet (using the facet mapping). Normally, the *vlo:Resource-Type* facet should contain similar values as DCMI type vocabulary[15]. This suggests a problem with the semantics. A proposed remedy is to normalise the values as well as to re-map them to other facets such as subject and genre. The result for the "diaries" would be, for instance, "plainText" in *vlo:Resource-Type* facet, plus "diary" in *vlo:Genre* facet, and, for the "Bibles", "plainText" in *vlo:ResourceType*, while "Bible" in *vlo:Subject* (or *vlo:ResourceTitle*) and "religious text" in *vlo:Genre*. Regarding schema and facet mapping, this scenario is the same as the current configuration (Figure 2).

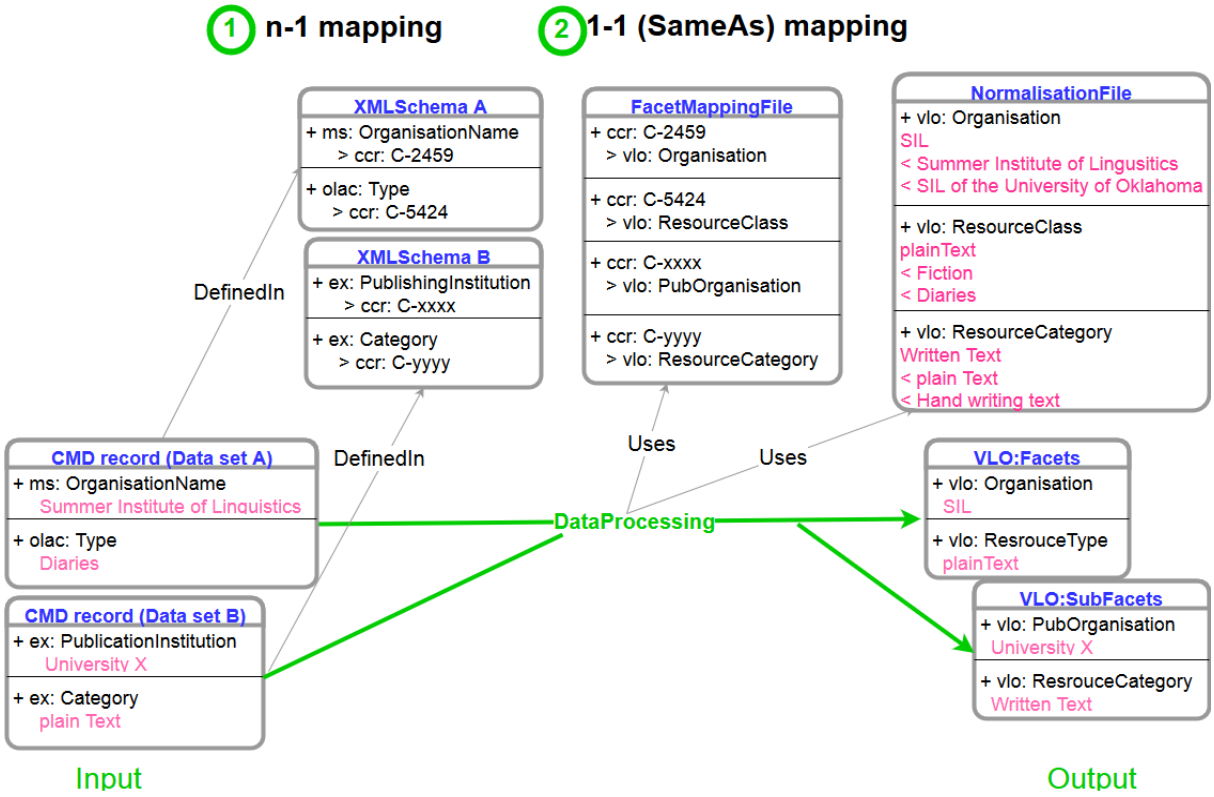



Figure 3. Scenario 2. One-to-one facet mapping.

[15] http://dublincore.org/documents/2003/11/19/dcmi-type-vocabulary/

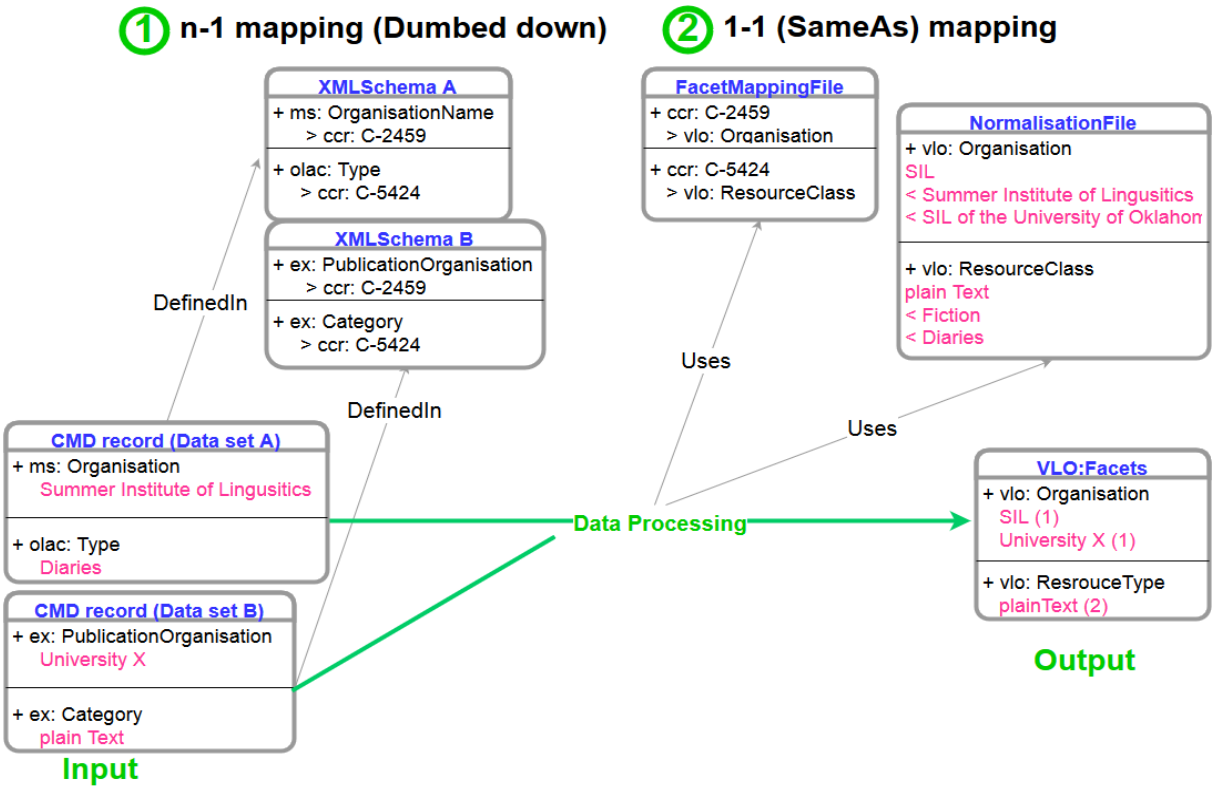



Figure 4. Scenario 3 Dumbed-down schema mapping

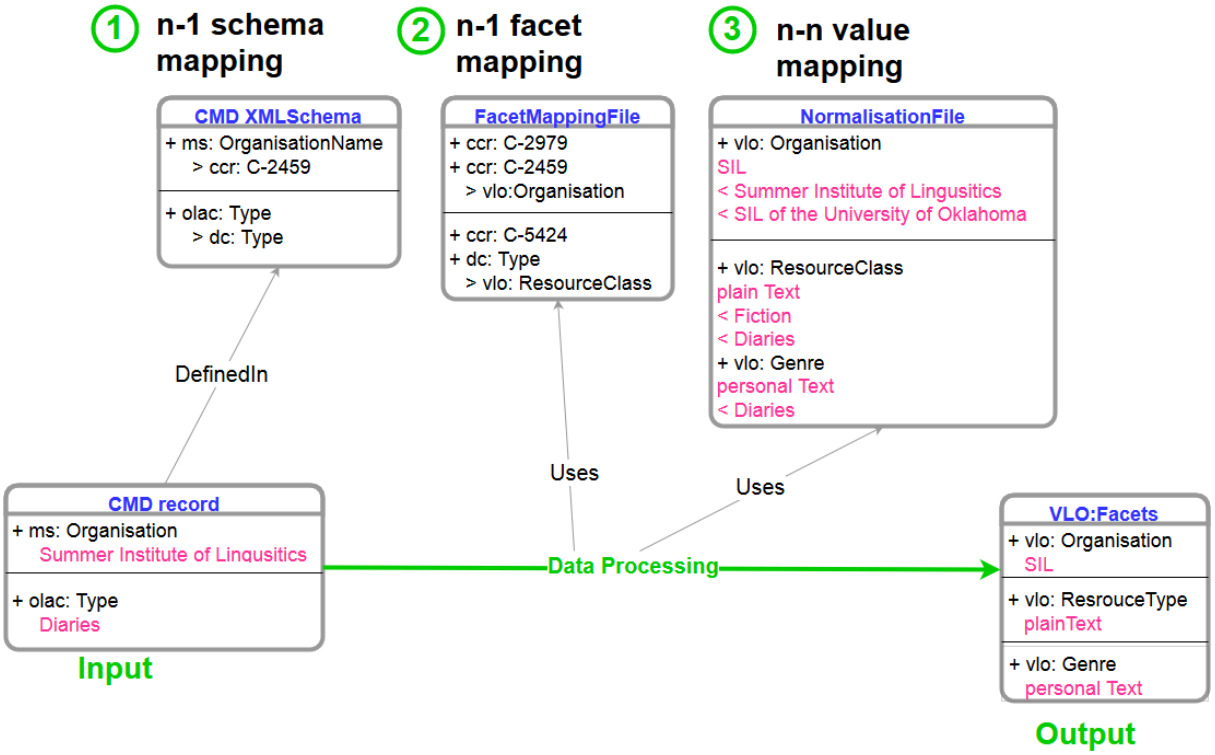



Figure 5. Scenario 4. Many-to-many value mapping

This strategy is motivated by the view that is unlikely that the normalisation would happen on the side of the data providers and that the central curation team has to take charge. On the surface (in the VLO), the approach is effective, because the facets appear clean and more understandable for the end-users. However, this approach implies a significant interpretative addition on the side of the curation. The modification may be far from the data provider's intention and the end user's expectation and could

negatively affect the resource discoverability. This can serve as a short-term solution for maintaining data consistency in VLO but extensive steps must be put into place for the curators to ensure that resource discoverability is not compromised. See Section 5.5 for some details on the case study mapping values for the *vlo:ResourceType* facet.

### 4.5 Summary of the mapping scenarios

In this section, we have elaborated on different scenarios for concept mapping and value normalisation. The focus of the discussion is the second part of the data processing (mapping between concepts and VLO facets). It is currently a many-to-one model. If we do not change this relationship (scenario 1), broader definitions for the VLO facets are recommended in order to accommodate various types of similar concepts in one facet. If we move to a one-to-one relationship, there are theoretically two paths to be followed. Scenario 2 maintains (and maybe also fine-tunes) all the existing concepts. As it is not possible to guarantee the same number of facets as concepts, an extensive investigation is necessary on how to accommodate this complexity in the user interface without hampering the usability. Dynamic and hierarchical faceting is a promising strategy here. Scenario 3 also aims at one-to-one facet mapping, but shifts the semantic reduction towards the schema mapping, which is both not feasible (all the profiles must be changed) and unacceptable (contrary to core CMDI principles). The last scenario introduces the idea of multi-facet decomposition on the value mapping level, inspired by the "messy" (overloaded) values of the metadata in some facets. However, this mechanism may lead to an unintended manipulation and interpretation of data, which would distort resource discovery. Thus, it can only be applied with great caution and in a conservative manner. The above analysis makes explicit the complexity of the process of ingesting and mapping CMD records into the VLO. Even if we disregard the problems coming from the data providers – there are three levels of semantic engineering in this process, which makes it a very demanding task to trace back to the source the different problems with quality of metadata and in the VLO.

## 5 The data processing workflow/pipeline

After investigating qualitative and quantitative aspects of the question at hand as well as mapping and normalisation mechanisms in detail, in this section we focus on the actual implementation of the ingestion and curation workflow and propose optimisations aiming mainly at a more integrated, more ergonomic setup and better communication with the data providers.

### 5.1 Current setup

Figure 6 illustrates a simplified view of the current workflow. It is a well-established chain of actions starting from a data submission through data processing to indexing and publishing on the VLO website. The starting point is when a data provider accesses a metadata authoring tool often hosted at a CLARIN national centre to design and create their records. Typical examples are ARBIL[16] in the Netherlands, COMEDI[17] developed in Norway, and the custom submission form of DSpace as implemented in Czech Republic[18] or in Poland[19]. Most of the tools are tightly integrated with the underlying data repository, where the metadata is stored together with the digital resources. Some allow for the use of any (or multiple) CMD profiles, some are tailored towards one specific profile. The metadata is exposed via an OAI-PMH endpoint from where it is fetched by VLO harvester on a regular basis. OLAC[20] and CMDI are the two major metadata formats that can be imported into the VLO environment, and the former is converted to CMDI by a predefined mapping. When CMDI is ready, it is being ingested into the Solr/Lucene index, governed by a set of configuration files: a facet mapping file and value mapping and normalisation files described in Section 4. The processed data is indexed and published on the VLO website, where the end users can browse and search the data.

While the authoring tools try to provide a local control over the quality of the metadata, offering a custom auto-complete functionality based on local controlled vocabularies and various consistency

[16] https://tla.mpi.nl/tools/tla-tools/arbil/
[17] http://clarino.uib.no/comedi/page
[18] https://github.com/ufal/lindat-dspace
[19] https://clarin-pl.eu/dspace/
[20] http://www.language-archives.org/OLAC/metadata.html

checks, and there are also already individual infrastructural services available for data providers to check their metadata (OAI-endpoint validator, schema validator), a coherent, formal and rigorous mechanism for VLO data ingestion is lacking.

The CLAVAS service especially dedicated to shared management of controlled vocabularies, although advocated on several occasions, is not yet fully functional as an authoritative source of controlled vocabularies in the CMD infrastructure, mainly due to lack of well-defined organisational procedures. An integrated user interface for editing the facet mapping file and the value normalisation maps is missing (see Section 5.4 for details of current usage) There is also no automatic (and very little manual) feedback from the VLO team after data ingestion, thus the data providers are required to exert a significant amount of effort to improve the metadata quality by individual consultation. Both the VLO curators and the data providers can examine the quality/integrity of the metadata only on the public website.

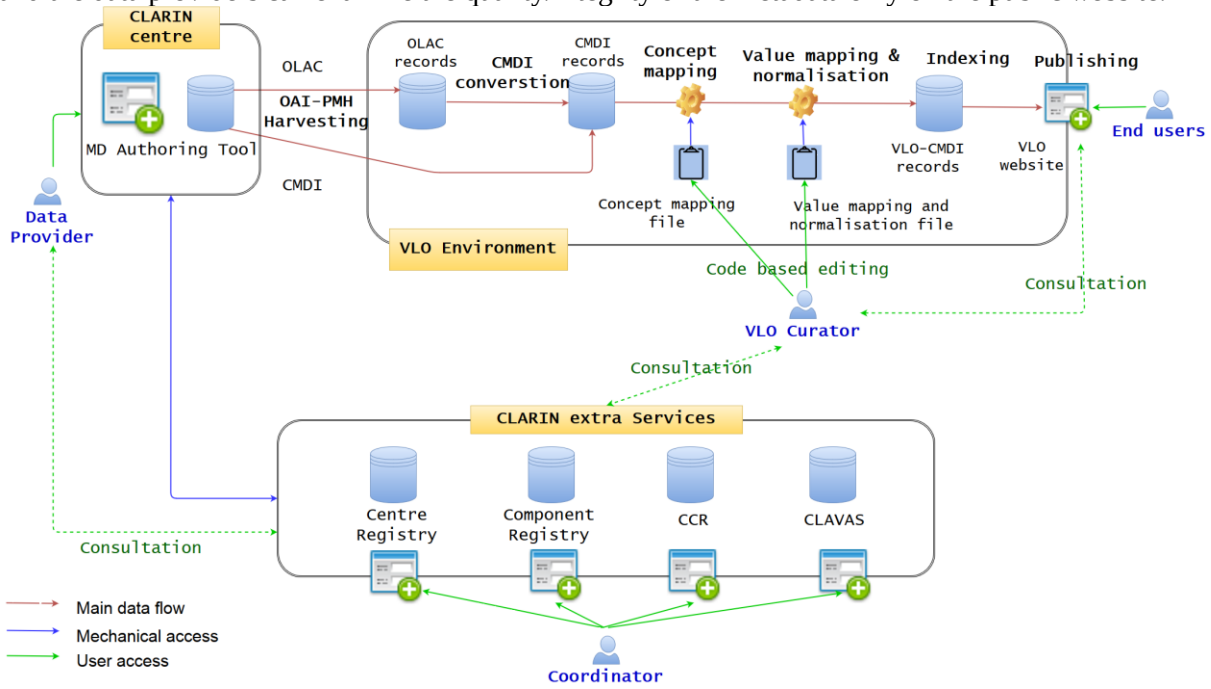



Figure 6. Current workflow.

## 5.2 Dashboard – an integrated data management system

In order to better manage the full VLO workflow, we propose a dashboard component (Figure 7) as a central interface or single entry point that will integrate all the procedural information from the individual steps of the data processing pipeline into one user-friendly GUI web interface with which the VLO curators, administrators and data providers can work on data management much more efficiently and coherently in a uniform manner, fostering the metadata quality and the VLO user experience. It should offer an intuitive monitoring view, illustrating each stage of the entire process, starting from harvesting, converting, and validating, to indexing and distributing. Note, that the dashboard would not perform any of the tasks in the process itself, but rather interact with the individual components of the VLO framework – harvester, converter, validator, mapper/normaliser, and indexer/publisher. The functionalities of the dashboard should include (but are not limited to):

F1. List of the datasets (OAI-PMH sets), optionally grouped per data provider and per CLARIN centres/countries (MUST)[21]

F2. Status and statistics of the datasets within the ingestion pipeline (errors, progress indicator) (MUST) (export as PDF, XML, CSV etc. (SHOULD))

F3. Simple visualisation of the statistics in F2, including pie charts, bar charts etc. (COULD)

[21] Suggestions of priorities are made using MoSCoW method.

F4. Browse the data quality reports per set (MUST) (export as PDF, XML, CSV etc (SHOULD)) including a link checker which lists broken links (COULD)
F5. Deliver the data quality report to the data provider/CLARIN centre (via automatic email, and/or via a web interface) (SHOULD)
F6. Edit the concept to facet mapping (MUST)
F7. Edit the value mapping and normalisation (MUST)
F8. Manual data management (deactivate indexing of the sets, delete the data sets, invoke harvesting of data sets etc) (MUST)
F9. Browse the log files of the VLO systems (COULD)
F10. Browse the Piwik web traffic monitoring (COULD) (do it per data set (COULD))

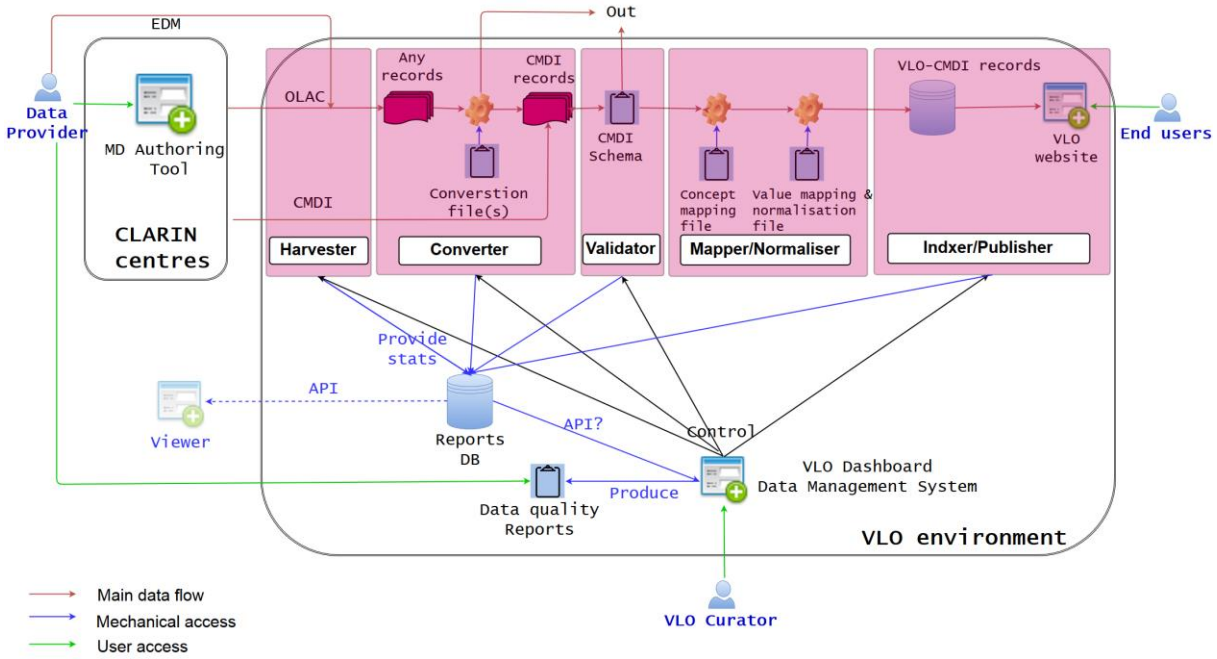



Figure 7. Proposed workflow – integrated dashboard.

With regard to the metadata problems, the dashboard is not supposed to resolve them per se. Rather it helps to identify them and keep track of them by means of a user-friendly interface (F1, 2, 3). In this sense the metadata quality report (F4) is probably the most prominent function. It compiles information from the harvesting, validation and curation step and makes it available to both VLO curators and data providers (F5). It is crucial that the users are able to edit/configure the complex mapping and normalisation within the integrated environment in sync with the core infrastructural registries Component Registry, CCR and CLAVAS. Additionally, the dashboard environment could allow for alternative instantiations of the facet mapping, so that the curator can experiment with different versions, approaches etc. before publishing the metadata records. The dashboard would also allow us to invoke actions on datasets directly in the interface (F8):

- (Re-)harvesting of the data set
- Disable indexing
- Delete the data set
- Show the data quality report (see Section 5.3) (download them as XML, PDF, etc.)
- Show the error messages (download them as PDF etc.)
- Show the metadata records
- Show the schema/profile (with the link to Component Registry, CLAVAS, and CCR)
- Send an email to the data provider (e.g. data quality report)

Figure 8 visualises the idea of the dashboard user interface in which the VLO curators can monitor the whole data processing workflow. In the mock-up, OAI-PMH data sets are listed as rows, and can easily be sorted per country and data provider. Following the ID and title, there is a date of latest update (e.g. harvesting date or latest actions). The status of data processing is clearly visible with green and red circles (harvested, converted to CMD, validated against CMD). When the dataset is indexed and published, the number of records is shown. Should the data be offered for further distribution (e.g. OAI-PMH, Linked Open Data etc.) in the future, the status will also be indicated here. Finally, an indication of the data quality could be provided with stars (or the metric delivered by quality assessment). Different actions will be selectable per set, according to the status of the data (F8).

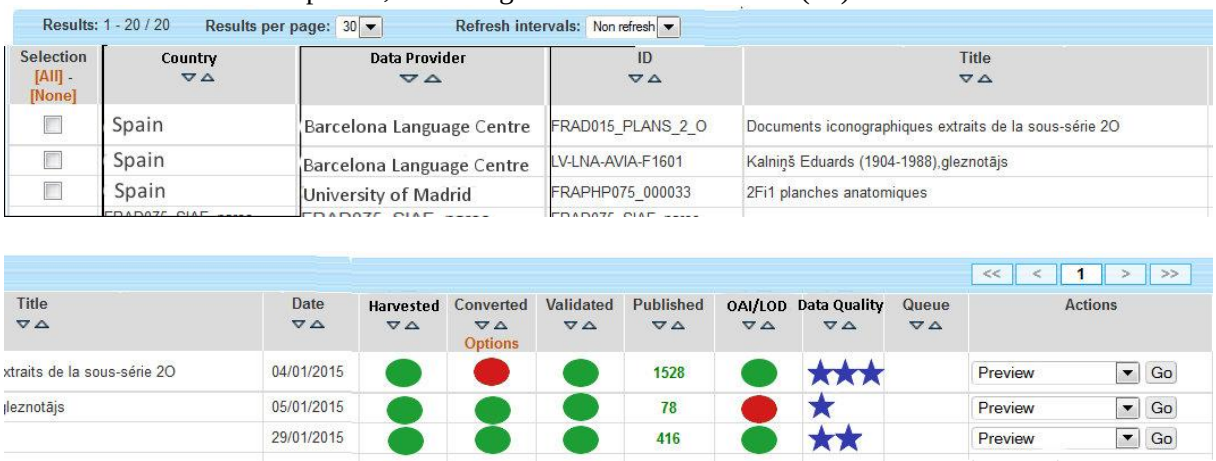



Figure 8. Dashboard mock-up (table cut into two parts for readability).

All actions can be applied both to a single data set as well as to multiple data sets in batch mode. The user can search, sort, and filter the information in this table view. In addition, s/he can select multiple data sets by clicking the checkboxes on the left. With this table view, the user can see the overall statistics (and/or selected data sets), including the number of datasets, countries, data providers, the status, the number of records indexed and distributed. These figures are important performance indicators (for CLARIN board, funders, but also the data providers themselves). The user should also be able to export the statistics as PDF and CSV or directly into an online spreadsheet. Equally important is the availability of historic information, i.e. statistics from previous harvests, allowing the curator to spot immediately sudden quality drops, or dramatic changes in the amount of records provided.

In summary, the dashboard will provide manual data processing functions, complementing and governing the automatic data processing. It will allow the VLO curators to monitor the data and manually interact with it without any knowledge of behind-the-scene codes and scripts.

### 5.3 Curation module

We are aware that the dashboard cannot be built overnight. Thus, it has to be developed block by block. The most imminent block of development will be curation module. Currently, the curation steps are implemented within the VLO ingestion application. The CMDI curation and VLO teams agreed to extract this functionality into a separate module that can be reinserted into the VLO ingestion pipeline, but can be used in other contexts as well. The specification for the Curation Module is based on previous work by Kemps-Snijders (2014), Trippel et al. (2014) and other (technical) documentation created by CLARIN metadata team in the last few years. The curation module will validate and normalize single metadata records, as well as whole collections, assess their quality and produce reports with different information for different actors in the VLO workflow. The module will integrate with the harvester that is being re-implemented and especially will also be tightly integrated with Dashboard application. The following four main use cases were already identified:

a) Metadata creator checks the validity of newly created records

b) Metadata modeller checks the quality of profiles

c) Repository administrator checks quality of metadata in his repository

d) Continuously check all metadata harvested and indexed by VLO

The module is being developed by the ACDH-OEAW implementing the task 2.2.1 of the CLARIN-PLUS project. First version is scheduled for February 2016. Some of the planned features are as follows:

- Schema validation,
- URL inspection,
- Value validation and normalisation against controlled vocabularies and normalisation.
- Assess facet coverage (of the profile and of the record)
- Feedback about errors, per record or per collection.
- Quality metrics
- Provision of instructions on improvement of the metadata optionally accompanied by already amended (normalised) CMD records
- Comparison of the curation results over time

### 5.4 Management of vocabularies and mapping

We need to take yet a closer look on the handling of the vocabularies in relation to the value mapping. A relatively simple (and partly implemented) approach to the management of the mappings is to maintain the vocabularies in the vocabulary repository CLAVAS, where, based on the SKOS data model, every entity or concept is registered as a separate item (*skos:Concept*), with a *skos:prefLabel* as the normalised label for a given concept and all variants encountered in the actual metadata stored as *skos:altLabel* (or *skos:hiddenLabel*). This information can be easily retrieved from CLAVAS via its REST-API and injected in the harvesting/curation workflow of the VLO. Until now, this has been done for Organisation names. The change introduced in CMDI 1.2 (Goosen et al., 2014) allows the indication of a controlled vocabulary for a given element in the CMD profile which will enable a more consistent handling of vocabularies in relation to the metadata elements.

What is still missing is an automatic procedure to add new previously unseen values to CLAVAS. The application underlying CLAVAS, OpenSKOS exposes a rich RESTful API that allows not only to query but also to manipulate the data. So technically it is possible for the curation module to add new candidate concepts. Human interaction is crucial here. These candidate concepts need to be clearly marked and kept in "quarantine" until they are checked and approved by a group of curators.

However, even if this whole process is set up, it does not offer a solution to more complex normalisation scenarios, like the multi-facet decomposition introduced in Section 4.4. Even though this specific scenario, is problematic in a certain respect (potentially severe interpretative intervention), it is clear that more advanced mapping mechanisms will be needed that cannot be served by the simple approach based on *skos:Concept* as proposed above. The current temporary solution for maintaining multi-facet mappings with which we have experimented is to use a simple spreadsheet with the encountered values in first column, and a separate column for the other facets, allowing the curators to assign values in multiple facets for any given value. These files are stored as *text/csv* file and maintained under version control in the CLARIN's code repository[22], so they can be edited by a team of curators, who can see who has done what when, but also retrieved and processed by any application, most notably the curation module. However this is still a very cumbersome and not well integrated process. Ideally, the functionality for value normalisation has to be well integrated into the dashboard (see F7 in Section 5.2). Thus, in addition to the data management view, the Dashboard has to offer a user interface to create and edit the concept mapping and the value mapping and normalisation (F6, 7). This functionality should completely hide the internal mapping mechanism, freeing the VLO curators from the manual editing and tedious syncing of CSV or XML files stored in different places, as is the case currently. However developing an interactive web-based table or spreadsheet application that features at least a minimal set of functionalities is a resource-intensive task itself. The value normalisation interface has to integrate with

[22] https://github.com/clarin-eric/VLO/tree/vlo-3.3-oeaw/vlo-vocabularies/maps/csv

CLAVAS and ideally also other sources of controlled vocabularies, offering the curator normalised values via autocomplete or similar functionality.

### 5.5 Normalisation example: resource type

Let us take a closer look at the example facet *ResourceType*. Currently, the facet encompasses around 300 different values. There were multiple attempts to define a controlled vocabulary for this facet, among others by the CLARIN-D VLO taskforce and Odijk (2015). These proposals stand next to a number of existing controlled vocabularies from other domains like Europeana (*edm:Type*, see Section 2.1), or DCMI Type. All the controlled vocabularies have some overlapping terms, some omissions and some slight differences in the semantics of the terms.

The curation task force is working from a vocabulary of some ten to twelve terms that tries to accommodate all of the above. The governing aspects are: high-level distinction to keep the number of values low (no more than 15, ideally under 10), and decomposition, i.e. each term signifies a certain "atomic" aspect of the resource and it is allowed to use a combination of values to describe one resource. Example:

```
AnnotatedTextCorpus = collection, text, annotation
Audio recording with transcription = audio, annotation
```

The currently proposed draft vocabulary

- annotation
- audioRecording
- collection
- structuredData
- grammar
- image
- lexicalResource
- physicalObject
- text
- videoRecording
- software
- service/interactiveResource

A crucial aspect of any controlled vocabulary is a sound definition of individual terms. A trial set of definitions is currently being worked out making use of the existing definitions from existing vocabularies. Once a coherent set of definitions is available this vocabulary will be circulated among the relevant CLARIN bodies and colleagues and especially will be discussed with the CLARIN Concept Registry group, in order to achieve a broad agreement for the vocabulary. Next, after adapting the normalisation maps against this authoritative list, a thorough examination of the soundness of the mappings will be undertaken and finally the mappings will be applied in the VLO and the vocabulary will be exposed for public use. We especially plan to use this vocabulary to make *vlo:ResourceType* a primary, prominent facet in the VLO, adorned by appropriate icons, potentially influencing also the customisation of display (different resource type ask for different facets, as proposed in Scenario 2 in Section 4.2).

## 6 Upcoming work

In this section we place the proposed partial solutions introduced in the previous sections, into a bigger picture. The metadata quality issues cannot be addressed by a single measure, but by a comprehensive set of measures. We point especially to the importance of the social dimension of the solutions. For example, the maintenance of vocabularies and mappings can only be effective if a broad agreement can

be reached in a collaborative manner, if they are to be integrated into the automatic curation process, and widely adopted by the data providers.

The technical solutions comprise the following elements: adaptation of the facet mapping, concepts directly available in the VLO accompanied by advanced user interface features like dynamic or hierarchical facets (scenario 2 Section 4.2); (conservative) advanced value normalisation based on shared normalisation maps (scenario 4 Section 4.4); a dashboard as an integrated interface for managing the ingestion and curation workflow (Section 5.2); a curation module assessing various aspects of metadata quality (Section 5.3). It is crucial to ensure that all changes applied during the processing (i.e. the mapping of the records to facets and the value normalisation) are transparent to the data provider and to the user of the VLO. Another requirement is to make the workflow more modular, especially allowing for the curation module to be encapsulated enough to be reusable in other contexts. A final technical issue is the testing phase. In order to ensure that the metadata quality and VLO discoverability are improved by the curation module, test cases have to be designed by experts. Each class of identified problems should be covered and generated reports should be used by metadata curators and software developers for further improvements.

It is impressive that CLARIN has developed a unique approach and system for their data aggregation. It has developed CMDI to facilitate the heterogeneity of the metadata for a linguistic domain, accompanied by the impressive automation of the data processing from the harvesting to indexing. It is to some extent effective. However, precisely due to this combination of data ingestion mechanisms, it leaves space for problems. It is evident that satisfactory results cannot be achieved with purely automatic measures. There has to be always a human curation, whether it is by a data provider or central curation team. Considering that the facets are the main selling point of VLO, it is imperative to find a complete solution to the extremely low coverage of records which has not been recognised until recently, and to tackle the issue from a structural point of view.

A crucial ingredient to the proposed strategy is the question of governance, i.e. who is going to steer the process and persistently remind data providers of the problems encountered and propose solutions. CLARIN has well-defined organisational structures and a number of bodies with delegates from all member countries where decisions can be agreed upon at different levels. In the described case, the primary operative unit is definitely the metadata curation task force with representatives from national consortia, in a tight collaboration with the CMDI task force, both reporting to the SCCTC, which in turn reports to the Board of Directors. Thus both the horizontal coverage over the member countries is ensured, so that national metadata task forces can report shortcomings they have identified, as well as the vertical integration of the decision-making bodies, allowing the application of small, practical, technical solutions as well as to propose substantial structural changes, if needed.

### 6.1 Prevention – fighting the problem at the source

While we pessimistically stated before that we cannot expect the providers to change their metadata, we cannot give up on them, as it is clearly better to combat the problem at the source. There are indeed a number of measures that can (and need to) be undertaken on the side of the data provider:

a) best practices guides and recommendations (like the CLARIN-D VLO Taskforce recommendations on the VLO facets), especially a list of recommended profiles (one or two per resource type) need to be provided, with profiles that have good coverage of the facets and use controlled vocabularies wherever possible

b) provision of metadata quality reports to the providers

c) provision of curated/amended metadata records directly back to the data providers

d) availability of controlled vocabularies via a simple API (as is provided by the OpenSKOS-API) to be combined with metadata authoring tools. This functionality has been in planning to be introduced through at least two metadata editors used by the CLARIN community: Arbil (Withers, 2012) and COMEDI (Lyse et al., 2014)

### 6.2 Semantics and relations of the concepts

We recognise that CCR is currently undergoing a restructuring, recently replacing the ISOcat in February 2015. However we must not forget that ISOcat had been in intensive use as a semantic layer of the

VLO for several years to build and develop the data aggregation systems. It is now high time to seriously discuss all the concepts in CCR and finalise them in order to ensure the semantic data integrity of concepts used within CLARIN. In particular, this paper has shown the emerging issues of mapping problems between CCR concepts and VLO facets. Broeder et al. (2010, 2014) reiterated that ISOcat forms the basis for semantic interoperability and will enable semantic search over the dataset. However this requires answers to following questions:

- What are the relationships between CCR concepts?
- How do we relate CCR concepts and external concepts such as DCMI?
- What are the needs and requirements for VLO and CCR to implement Semantic Web?

The first and second questions were indeed picked up right from the beginning (Broeder et al, 2010), referring to the need for a Relation Registry (Windhouwer, 2012), accompanying the concept registry. Adopting SKOS, a decision come to on account of the migration to CCR, as data model for the concepts allows us to define relationships such as the hierarchical structure of the concepts (skos:broader, skos:narrower). This at least provides good technical means to help clarify the semantics and relations of CCR concepts and VLO facets, especially given the support for defining the SKOS relations in OpenSKOS[23] the software underlying CLAVAS and CCR.

On the other hand, it remains unclear how these relations relate to the semantics of the defined CMD components and how they can be exploited for resource discovery. It is at least clear that it will add an extra dimension (maybe confusion) to the already complex concepts-facets mapping. It would be worthwhile to explore whether SKOS relations could replace, or serve as basis for the VLO facet mappings. Moreover, SKOS can only define (hierarchical) relations between concepts. For a comprehensive description of properties and relations we need to look to RDFS. This task is already being taken up within the CLARIAH-NL project, building on previous work by Durco and Windhouwer (2014) on expressing the whole of CMD in RDF[24].

## 7 Conclusion

In this paper we evaluated the VLO metadata quality issues at different levels. First of all, we outlined our observations and statistical analysis of the value variation. Secondly, we pointed out the volatile situation of the current mapping and normalisation mechanisms. Thirdly, the fragmented ingestion workflow was revealed. Different elements of the VLO environment all contribute to the problems. Therefore, our proposal is a comprehensive set of technical and social solutions. We proposed a concrete strategy for curation and normalisation of values in the facets of the VLO. We elaborated on the ways to establish and sustain a data ingestion mechanism and workflow that combines systematic, automatic, transparent curation of the metadata with continuous input from human curators providing the mappings from actual values encountered in the metadata to recommended normalised values. An integral part of the process must be a suite of test cases that ensures the quality of the mappings and the whole curation process. Finally, all output of the curation (corrections and amended metadata records) must be recycled to the data providers in the hope of preventing problems in the future and the entire work cycle must repeat as new resources are added. Thus the need for metadata curation is perpetual.

VLO is a living infrastructural service designed to provide a single access point to the European language resources. Thus, it will adopt new ideas and technologies and continue to evolve for the needs of the end-users. In this sense, we always take a heuristic approach. This paper has aimed to deliver a detailed analysis of the current metadata issues in the context of data ingestion mechanism and workflow. We can continue to improve the service accordingly and continuously. In a close collaboration with CLARIN partners in other European countries, we strive to make a regular contribution to the development of high-quality stable and sustainable VLO services.

[23] http://openskos.org
[24] https://github.com/TheLanguageArchive/CMD2RDF

## References


[Broeder et al.2010] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. 2010. A data category registry-and component-based metadata framework. In *Procedings of the Seventh Conference on International Language Resources and Evaluation* [LREC2010]. Pp. 43-47.

[Broeder et al.2012] D. Broeder, M. Windhouwer, D. Van Uytvanck, T. Goosen, and T. Trippel. 2012. CMDI: a component metadata infrastructure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* [LREC2012]. Pp. 1387-1390.

[Broeder et al.2014] D. Broeder, I. Schuurman, and M. Windhouwer. 2014. Experiences with the ISOcat Data Category Registry. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* [LREC 2014]. Pp. 4565-4568.

[Calarco et al.2014] P. Calarco, L. Conrad, R. Kessler, and M. Vandenburg. 2014. Metadata Challenges in Library Discovery Systems. In *Proceedings of the Charleston Library Conference*. Purdue University e-Pubs. Pp. 533-540.

[Durco and Moerth2014] M. Ďurčo, and K. Mörth. 2014. Towards a DH Knowledge Hub - Step 1: Vocabularies. Presented at *Clarin 2014 Conference* [CAC2014].

[Durco and Windhouwer2014] M. Ďurčo and M. Windhouwer. 2014. From CLARIN Component Metadata to Linked Open Data. In *Proceedings of the Third Workshop on Linked Data in Linguistics* [LDL 2014]. Pp. 13-17.

[Europeana2009] Europeana. 2009. Metadata Mapping & Normalisation Guidelines for the Europeana Prototype: Europeana Version 1.2. Europeana: Think Culture, Den Haag, Netherlands.

[Europeana2014] Europeana. 2014. EDM Mapping Guidelines: Europeana Version 2.2. Europeana: Think Culture, Den Haag, Netherlands.

[Goosen et al.2014] T. Goosen, M. Windhouwer, O. Ohren, A. Herold, T. Eckart, M. Ďurčo and O. Schonefeld. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure [CAC2014]. In *Selected Papers from the CLARIN 2014 Conference* [CAC2014]. Pp. 36-53.

[Haaf et al.2014] S. Haaf, P. Fankhauser, T. Trippel, K. Eckart, T. Eckart, H. Hedeland and D. Van Uytvanck. 2014. CLARIN's Virtual Language Observatory (VLO) under scrutiny-The VLO taskforce of the CLARIN-D centres. Presented at *Clarin 2014 Conference* [CAC2014].

[Huffman2015] N. Huffman. 2015. Adventures in metadata hygiene: using Open Refine, XSLT, and Excel to dedup and reconcile name and subject headings in EAD. In *Bitstreams: Notes from the digital projects team.* Duke University Libraries, N.C.

[Kemps-Snijders2014] M. Kemps-Snijders. 2014. Metadata quality assurance for CLARIN. Technical report.

[Lyse et al.2014] G. Lyse, P. Meurer, and K. De Smedt. 2014. COMEDI: A New Component Metadata Editor. In *Papers from the CLARIN 2014 Conference* [CAC2014]. Pp. 82-88.

[Odijk2014] J. Odijk. 2014. Discovering Resources in CLARIN: Problems and Suggestions for Solutions. Utrecht University Repository, Netherlands.

[Odijk2015] J. Odijk. 2015. Metadata curation strategy. Internal document, unpublished.

[Palmer2014] W. Palmer, 2014. Fits metadata normalisation API? Github Repository.

[Sofou and Tzouvaras2015] N. Sofou, and V. Tzouvaras. 2015. MS28: Sounds thesaurus and metadata cleaning and normalization module complete. Europeana Sounds 620591, Den Haag, Netherlands.

[Trippel et al.2014] T. Trippel, D. Broeder, M. Ďurčo, and O. Ohren. 2014. Towards automatic quality assessment of component metadata. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* [LREC 2014]. Pp. 3851-3856.

[Van Uytvanck2010] D. Van Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, and M. Gardelleni. 2010. Virtual Language Observatory: The portal to the language resources and technology universe. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* [LREC 2010]. Pp. 900-903.

[Van Uytvanck2012] D. Van Uytvanck, H. Stehouwer, and L. Lampen. 2012. Semantic metadata mapping in practice: The Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* [LREC2012]. Pp. 1029-1034.

[Windhouwer2012] M. Windhouwer. 2012. RELcat: a Relation Registry for ISOcat data categories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* [LREC2012]. Pp. 3661-3664.

[Withers2012] P. Withers. 2012. Metadata management with Arbil. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* [LREC2012]. Pp. 72–75.

# A Use Case for Linguistic Research on Dutch with CLARIN

**Jan Odijk**
Utrecht University, the Netherlands
j.odijk@uu.nl

## Abstract

In this paper I describe a particular Dutch linguistic problem and I show that it can be addressed in a better, more efficient, and more user-friendly manner than ever before, thanks to CLARIN. Most of the data that are used in the investigation could only be used by technical experts a few years ago but are now available to all linguists through a variety of easily accessible web applications developed in CLARIN with interfaces dedicated to their intended users. However, it also shows that still a lot of further extensions and improvements can and must be made. Fortunately, most of these are being implemented in currently running projects.

## 1 Introduction

In this paper I describe a particular Dutch linguistic problem and I show that it can be addressed in a better, more efficient, and more user-friendly manner than ever before, thanks to CLARIN. Most of the data that are used in the investigation could only be used by technical experts a few years ago but are now available to all linguists through a variety of easily accessible web applications developed in CLARIN with interfaces dedicated to their intended users, linguists.[1]

The relevant problem was first defined in unpublished work (Odijk, 2011). This report also specified what kinds of search actions would be needed to address this problem. At the time, almost none of these search actions were possible, or only with great difficulty, and they required expert knowledge on the relevant resources and programming or scripting skills. In 2014, (Odijk, 2014a) showed in a lecture that many of the desired search actions had become possible, in a simple manner, and through applications with interfaces dedicated to the targeted users, linguists. At the same time, it was observed that not everything was possible yet in an easy way, and new requests arose by using the relevant applications. Since neither (Odijk, 2011) nor (Odijk, 2014a) was published, I report on their findings in this paper, and I will show new functionality created to accommodate the newly arisen needs. This paper thus serves as an example of a report on a *research pilot*: a project to use functionality offered by the infrastructure with the twin goals of furthering the research but also of identifying novel functionality that the infrastructure should offer to be able to further the research.

I introduce the basic facts to be investigated in section 2, make an assessment of these facts in section 3, and list a few of the many research questions that these facts raise in section 4. I then show that a variety of web applications developed in CLARIN for searching in linguistic resources (lexicons and corpora), for enriching corpora and for analysing search results make research into this problem possible that is based on more data, which are found faster and easier than was possible ever before. The web applications considered are OpenSONAR (section 5.1), the LASSY Word Relations Search engine (section 5.2), GRETEL (section 5.3), CORNETTO (section 5.4), COAVA (section 5.5), and PaQU (section 5.6). All applications mentioned are available in the CLARIN infrastructure and can be accessed via the CLARIN-NL portal[2]. I summarize the conclusions in section 6.



[1] CLARIN as a whole of course targets all humanities researchers, but the applications discussed in this paper target linguists.

[2] http://portal.clarin.nl/

This paper shows that great progress has been made since 2011 in the number of applications offered in the CLARIN infrastructure, and it shows a significant increase in the functionality that they offer, but it also identifies functionality that was desired from the start as well as novel desired functionality that have not been implemented yet. Section 7 describes future work that can and must be done to address the research questions.

## 2 Basic facts

In this section I introduce the basic facts related to the problem that I want to investigate. It is a specific case of the problem of the acquisition of lexical properties by first language (L1) learners.

The three Dutch words *heel*, *erg* and *zeer* are near-synonyms meaning 'very', i.e. (stated informally) they modify a word that expresses a gradable property or state and specify that its modifiee has the property or state it expresses to a high degree. Of these, *heel* can modify adjectival (A) predicates only, while *erg* and *zeer* can modify not only adjectival, but also verbal (V) and adpositional (P) predicates. This is illustrated in example (1)

(1) a. Hij is daar heel / erg / zeer blij over
he is there very / very / very glad about
'He is very happy about that'

b. Hij is daar *heel / erg / zeer in zijn sas mee
he is there very / very / very in his lock with
'He is very happy about that'

c. Dat verbaast mij *heel / erg / zeer
That surprises me very / very / very
'That surprises me very much'

In (1a) the adjectival predicate *blij* 'glad' can be modified by each of the three words. In (1b) the (idiomatic) prepositional predicate *in zijn sas* can be modified by *zeer* and *erg* but not by *heel*. The same holds in (1c) for the verbal predicate *verbaast*.[3] In English, something similar holds for the word *very*: it can only modify adjectival predicates. For verbal and prepositional predicates one cannot use *very* but one can use the expression *very much* instead:

(2) a. He is very happy about it
b. He is very *(much) in love with her
c. It surprised me very *(much)

There is a lot more to say about these data, and there are more relevant data to consider and some qualifications to be made. Some of these will be discussed below. I refer to (Odijk, 2011), (Odijk, 2014a) and (Odijk, 2015a) for further details.

## 3 Assessment of the facts

The distinctions I illustrated in the preceding section are purely syntactic in nature. The words *heel*, *zeer* and *erg* are synonyms or near-synonyms, and the expressions *blij* and *in zijn sas* are near-synonyms as well, which makes it unlikely that the difference can be derived from semantic properties.[4] It is also not in any way obvious how the differences could follow from universal principles of language or language acquisition.

There are other differences among the words *heel*, *erg* and *zeer*. (Odijk, 2015a, section 4) describes these differences. If any of these differences is somehow related to the difference under investigation then it must be a difference in which *heel* opposes the other two words *erg* and *zeer*. However, (Odijk, 2015a, section 4) shows that this is not the case for any of these differences.

[3] or maybe the whole VP *verbaast mij*.
[4] See (Odijk, 2011) for more examples supporting this conclusion.

I conclude that the differences in modification potential of the words *heel*, *erg* and *zeer* cannot be derived from other facts and must be acquired by learners of Dutch.

## 4 Research questions

The simple facts described in the preceding sections are interesting for a number of reasons. First, they constitute a kind of minimal pair in first language acquisition: though *heel* on the one hand and *zeer*, *erg* on the other are (near-)synonyms, their syntactic modification potential differs. They also illustrate acquisition of a negative property: L1 learners must learn that *heel* canNOT modify verbal or prepositional predicates. These facts therefore raise many research question related to language acquisition. Examples of these research questions are:

(3) a. How can children acquire the fact that *erg* and *zeer* can modify A, V and P predicates (in L1 acquisition)?
b. How can children acquire the fact that *heel* can modify A but **canNOT** modify V and P predicates (in L1 acquisition)?
c. What kind of evidence do children have access to for acquiring such properties?
d. Is there a relation with the time of acquisition?
e. Is there a role for indirect negative evidence (i.e., absence of evidence interpreted as evidence for absence)?

Obviously, this paper cannot address all these questions. The main purpose of this paper is to show that, by using CLARIN, research questions such as the ones in (3) can be addressed in a better and more efficient manner than without CLARIN. In this paper, I will focus on research question (3c)

In order to address these research questions, data are needed that can provide evidence on these questions. Fortunately, many such data exist. We will mention several relevant sets in the coming sections. However, though most of these data existed before CLARIN, they were hardly usable for supporting linguistic research at the time.

## 5 Search and Analysis with CLARIN web applications

I described the problem of section 2 in (Odijk, 2011), as an example user scenario for search applications to be developed in CLARIN. At the time, many of the search actions I would like to be able to carry out were not possible yet, or could only be carried out with great difficulty and only with expert knowledge of the relevant data sets and query options. Some queries suggested there involve search in metadata, an area where much progress has been made since then, though some of the specific queries suggested are still not possible (and there are many other problems with searching for data via metadata through the Virtual Language Observatory[5], as described in (Odijk, 2014b)). We will not discuss this here anymore. Other suggested queries involve search in the data themselves. I list most of them here, together with an indication where they will be dealt with in this paper:

- search for synonyms, hyponyms, and co-hyponyms for the words *heel*, *erg* and *zeer* (discussed in section 5.4)
- search for bi-grams in corpora with linguistic annotations on tokens (discussed in section 5.1)
- search in the Dutch CHILDES corpora, in the children's speech, and the speech by adults addressing children (discussed in section 5.5)
- search in treebanks (discussed in sections 5.2 and 5.3)
- search in CHILDES corpora enriched with syntactic structures / PoS-tags (discussed in section 5.6)

[5] `https://vlo.clarin.eu/?0`

In March 2014, I investigated what was possible at the time, and reported on that in a lecture (Odijk, 2014a). Some crucial functionality which was still lacking was identified, which led to plans for the creation of two new applications, *PaQu* (see section 5.6) on which (Odijk, 2015a) reported extensively, and *AutoSearch* (briefly discussed in section 5.6). The easiest way to get a first overview of what kind of applications developed in CLARIN-NL can be used for humanities research is via the CLARIN-NL portal[6], which allows faceted search by *research domain*, *tool task*, *language* and other facets. For a more detailed assessment of the suitability of a certain application for a specific research question, the application has to be studied in more detail through its documentation or via a tutorial (see the CLARIN-NL portal's *Educational Packages Section* for available educational material.)

Several suggested queries can be now carried out, but many are not yet possible. We will take up this issue in section 7.

### 5.1 OpenSONAR

OpenSONAR[7] is a web application that enables search in and analysis of the large scale Dutch reference corpus SONAR[8] and SONAR New Media[9] (Oostdijk et al., 2013). In part because of the size of the corpus (500 million tokens[10]), accessing the information contained in the data set has proven to be difficult. OpenSONaR facilitates the use of the SoNaR corpus by providing a user-friendly online web interface tuned to the intended users, linguists. No software or data need to be downloaded, no programs installed, and no programming knowledge is required.

SONAR is a reference corpus of contemporary written Dutch for use in different types of linguistic (incl. lexicographic) and HLT research and the development of applications. It was created in the STEVIN (Spyns and Odijk, 2013) funded SONAR project (2008-2011) that built on the results obtained in the earlier STEVIN projects D-Coi and Corea.

SONAR contains over 500 million tokens of full texts from a wide variety of text types from conventional media. SONAR New Media contains texts from the social media (Twitter, Chat, SMS) with about 35 million tokens. These corpora have been tokenized, tagged for part of speech and lemmatized, and Named Entities have been labelled. All annotations were produced automatically, no manual verification took place.

OpenSONAR is an online application for exploration of and searching in the SoNaR corpus. In the *Exploration* interface one can look into the corpus distributions, request statistics from sub-corpora, retrieve n-grams from sub-corpora and search for specific documents using the SoNaR document ID. In the *Search* interface one can use any of four different search strategies: simple, extended, advanced or expert .

OpenSONAR makes it easy to search for two adjacent tokens (bigrams) via their properties *lemma*, *word*, and *part-of-speech* (*pos*). For example, one can search for a token with lemma="heel" immediately followed by a token with pos="preposition", or the same with lemma="zeer" instead of "heel", or for a token with lemma="heel" immediately followed by a token with pos not equal to adjective.

Adjacency of tokens does of course not imply a grammatical relation of modification. Therefore the search results will contain many false hits. Nevertheless the search results are useful, in particular because the search results can be sorted and grouped in various ways, which reduces the effort of separating correct from false hits.

Analysis of the search results yield several new results. Firstly, it turns out that *heel* **can** modify certain PPs, in particular certain adverbial PPs, such as

(4) a. heel in het begin
very in the beginning

[6] http://portal.clarin.nl/

[7] http://portal.clarin.nl/node/4195

[8] https://vlo.clarin.eu/record?q=SONAR&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-1498_64_format_61_cmdi

[9] https://vlo.clarin.eu/record?q=SONAR&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-1502_64_format_61_cmdi

[10] I use the term *token* in this paper as a term for *occurrence of an inflected word form.*

in the very beginning

b. heel af en toe
very off and to
very infrequently

c. heel in het bijzonder
very in the particular
very specially

d. heel op het laatst
very on the last
at the very last moment

e. heel in de verte
very in the distance
very far away

f. heel uit de verte
very from the distance
from very far away

g. heel in het algemeen
very in the general
very generally

These examples do not undermine our earlier claims on the data, but do add a new set of data that clearly should be incorporated in the analysis.

Secondly, *heel* does indeed also occur with predicative PPs in SONAR as in (5):

(5) a. heel buiten zijn verwachting
very outside his expectation
completely unexpectedly

b. heel in de mode
very in the fashion
completely fashionable

c. heel in de vakantiestemming
very in the holiday-mood
completely in the mood for a holiday

d. heel in het zwart
very in the black
completely without paying taxes

e. heel in orde
very in order
completely OK

I find all examples of (5) ill-formed. The mere occurrence of such examples in a corpus need not be significant, since corpora contain examples of actual language use, which may contain errors. However, their number is sufficiently large to suspect that we are dealing here with a genuine instance of variation in the Dutch language. Though I glossed the word *heel* here as *very*, I think that people who use such expressions intend *heel* here in the sense of *geheel* or *helemaal* ('completely'), and the translations I provided in (5) reflect this. Obviously, one would like to investigate further properties of these utterances (e.g., genres that they occur in, origin of the utterer (Netherlands or Flanders), his/her gender and age etc.), but that is not so easy with the current version of OpenSONAR: The search output contains many false hits. Though one can cross-classify *all* search results with metadata information, one cannot mark a subset of search results for such a cross classification with metadata. An extension of OpenSONAR is required for this (see section 7).

### 5.2 LASSY Word Relations Search Engine

As mentioned above, adjacency of two words does not imply that these two words entertain a grammatical dependency. What one would actually want is a database in which grammatical dependencies between words are represented and are searchable. This information is available in treebanks, but the databases that contain this information are much smaller than SONAR. The LASSY Word relations Search Engine (LWRS)[11] (Tjong Kim Sang et al., 2010) enables one to search for such grammatical dependencies in certain treebanks. Actually, LWRS already existed when I described the linguistic problem for the first time. It was originally not developed in the CLARIN-NL project, but clearly inspired by the desire in CLARIN to provide web applications for search in corpora with interfaces that are tuned to linguists as users.

LWRS has a dedicated interface that enables a user to specify a query that searches for utterances containing two words entertaining a grammatical dependency by providing the properties of these two words (lemma, word form, part of speech) and the label of their grammatical dependency (e.g. subject, object, etc.).[12] This makes it easy to search for utterances that e.g. contain a word with lemma *heel* that is a modifier of a word with pos *verb*, and many similar examples.

Such queries carried out on the 1 million token manually verified written language treebank LASSY-Small Corpus[13] (van Noord et al., 2013) yield the following results:

- LASSY-SMALL contains examples where *heel* appears to modify a verb, but in all cases these are adjectives that happen to be identical in form to the participles of verbs. In such cases, LASSY, by convention, always analyzes these as verbs.

- LASSY-SMALL incorrectly analyzes *heel* in *heel open staan for* lit. very open stand for, 'be very receptive for' as modifying the verb *staan*, while in fact it modifies the adjective *open*.

- LASSY-SMALL contains examples where *erg* or *zeer* modifies verbs. In most cases, this also involves adjectives that happen to be identical to participles of verbs, but there are also several cases of modification of a real verb.

In short, these findings confirm our initial assumptions of the facts, which are now backed by a large amount of empirical material.

### 5.3 GRETEL

GrETEL is web application that enables a user to provide an example sentence of a construction that he/she is interested in and to specify which aspects of this example sentence are crucial for identifying the construction. The system then automatically generates a query and applies it to a treebank (LASSY-SMALL or the Spoken Dutch Corpus treebank, each manually verified and containing 1 million tokens).[14]. The query is generated by parsing the example sentence with the same parser that was used in the creation of the treebank (Alpino[15] (van der Beek et al., 2002)), thus increasing the chances of providing a query that finds instances of the construction searched for. The GrETEL application has been described in detail elsewhere (Augustinus et al., 2012; Vandeghinste and Augustinus, 2014).

Applying it to the Spoken Dutch Corpus (Oostdijk et al., 2002) yields the following results:[16]

- The word *heel* occurs as a modifier of a verb in 61 cases. However,

[11] `http://portal.clarin.nl/node/1966`

[12] Not specifying any properties matches with every word or relation, so this functions basically as a variable in the query.

[13] `https://vlo.clarin.eu/record?q=LASSY-SMall&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-1493_64_format_61_cmdi`

[14] And since recently, also the automatically parsed SONAR-500 corpus.

[15] `http://www.let.rug.nl/vannoord/alp/Alpino/`

[16] The full GrETEL functionality is not necessary for the problem at hand, though it can be used for it. In fact, the analysis described here has been carried out with PaQu (see section 5.6), since its options for analyzing the search results are more extensive than GrETEL's. Extension of GrETEL's analysis options is planned for the future. See section 7.

  - in 53 of these, the word is actually an adjective that happens to be identical to the participle of a verb (as above in LASSY-Small);
  - in 3 cases *heel* actually modifies a substantivised infinitive (and, as as modifier of a noun, has the meaning 'whole');
  - in 2 cases I find the sentence ill-formed. Maybe *heel* is intended here as 'completely'. Both utterances are of Flemish origin;
  - in 3 cases the analysis in the treebank is incorrect;

- The word *heel* occurs as a modifier of a preposition in 6 cases:
  - in 4 cases these are adverbial PPs that we also encountered with OpenSONAR (see section 5.1, the examples in (5);
  - in one case I find the sentence ill-formed. Maybe *heel* is intended here as 'completely'. The utterance is again of Flemish origin;
  - in one case *heel* modifies the expression *voor de hand liggen* lit. in-front-of the hand lie, 'be obvious'. I find the example marginal, except when the verb in the expression is a present participle. In that case, however, we are arguably dealing with an adjectival expression.[17]

- The word *heel* occurs as a modifier of an *MWU* (multi-word unit). These MWUs have no other part of speech code, but further analysis shows that they involve
  - adjectives in 3 cases[18];
  - nouns in 4 cases (e.g. *heel Den Haag*, lit. whole the Hague) and *heel* means 'whole' in these cases;
  - adverbial prepositional phrases in 2 cases (*heel af en toe*, lit. very off and to, 'very infrequently')
  - incorrect analyses in 3 cases

  In summary, these facts are consistent with our findings on the basis of OpenSONAR and with our initial assumptions on the data, and they suggest that the use of *heel* as a modifier of predicative PPs might be possible for certain Flemish speakers.

### 5.4 CORNETTO

(Odijk, 2011) suggested that analysing the modification potential of (near-)synonyms, co-hyponyms, and hyponyms of the words *heel*, *erg* and *zeer* may contribute to an understanding of the problem at hand. At the time, searching for synonyms or near-synonyms, let alone for words with other semantic relations for a given word, was very difficult. Obviously, one would want to use the Cornetto database for this purpose.

The Cornetto database is a lexical resource for the Dutch language which combines two resources with different semantic organisations: the Dutch Wordnet with its synset organisation and the Dutch Reference Lexicon which includes definitions, usage constraints, selectional restrictions, syntactic behaviours, illustrative contexts, etc. The Cornetto database contains over 92K lemmas and almost 120K word meanings.

At the time, an interface to Cornetto existed, but it often did not work, required an old version of the Firefox browser[19], and the interface itself was not well-designed. Searching for semantically related

[17] For example, it can be used predicatively and be modified by *te* 'too'

(1) Dat is te voor de hand liggend
That is too in-front-of the hand lying
That is too obvious

which is not possible for verbal present participles.

[18] In *heel ver weg*, lit. very far away, *ver weg* is analyzed as a MWU, though clearly here *heel* modifies the adjective *ver*, and together they modify the word *weg*.

[19] Arguably, this is a defect of Firefox. Producing upgrades that are not backwards compatible should be banned!

words has become easy with CLARIN, since a web application with a dedicated interface to the Cornetto database has been created.

The Cornetto web application[20] offers 3 different interfaces: Simple search for lexical entries[21], Advanced search for lexical entries[22], and Search for synsets[23].

Searching for (near-)synonyms of *zeer* in the relevant sense (Cornetto sense identifier *zeer-adv-3*) yields the following set of sense identifiers from Cornetto:[24]

(6) allemachtig-adv-2, beestachtig-adv-2, bijzonder-a-4, bliksems-adv-2, bloedig-adv-2, bovenmate-adv-1, buitengewoon-adv-2, buitenmate-adv-1, buitensporig-adv-2, crimineel-a-4, deerlijk-adv-2, deksels-adv-2, donders-adv-2, drommels-adv-2, eindeloos-a-3, enorm-adv-2, erbarmelijk-adv-2, fantastisch-adv-6, formidabel-adv-2, geweldig-adv-4, goddeloos-adv-2, , godsjammerlijk-adv-2, grenzeloos-adv-2, grotelijks-adv-1, heel-adv-5, ijselijk-adv-2, ijzig-a-4, , intens-adv-2, krankzinnig-adv-3, machtig-adv-4, mirakels-adv-1, monsterachtig-adv-2, moorddadig-adv-4, oneindig-adv-2, onnoemelijk-adv-2, ontiegelijk-adv-2, ontstellend-adv-2, ontzaglijk-adv-2, ontzettend-adv-3, onuitsprekelijk-adv-2, onvoorstelbaar-adv-2, onwezenlijk-adv-2, , onwijs-adv-4, overweldigend-adv-2, peilloos-adv-2, reusachtig-adv-3, reuze-adv-2, schrikkelijk-adv-2, sterk-adv-7, uiterst-adv-4, verdomd-adv-2, verdraaid-a-4, verduiveld-adv-2, , verduveld-adv-2, verrekt-adv-3, verrot-adv-3, verschrikkelijk-adv-3, vervloekt-adv-2, vreselijk-adv-5, waanzinnig-adv-2, zeer-adv-3, zeldzaam-adv-2, zwaar-adv-10

The word *heel*, in one of its senses (with Cornetto sense identifier *heel-adv-5*) is included here.

Similarly, the near-synonyms of *erg*, in the relevant sense (with Cornetto sense identifier *erg-a-2*) are listed in (7):

(7) erg-a-2, ernstig-a-2, fel-a-1, hard-a-4, heftig-a-1, hevig-a-1, krachtig-a-3, sterk-a-4, stevig-a-2, straf-a-2, vet-a-5, vurig-a-1, zwaar-a-3

And the hyponyms of these senses can be retrieved easily as well. Cornetto thus offers, in a very simple way, a list of word senses (and therefore words) that are semantically related to the word sense queried.

Now one would like to use the corpus search interfaces described above to investigate the modification potential of the words associated with these meanings. This is possible, but currently requires making a separate query for each of the words associated with the meanings listed above (some 70 words). One can also write a single query with each of the relevant words as an alternative, but the analysis options of the current corpus search and analysis applications do not enable e.g. a grouping by the modifier lemma and the modifiee part of speech. For example OpenSONAR's analysis options enable one to group the results by the part of speech of the immediately adjacent word but do not allow sorting the results by the lemmas searched for at the same time. An alternative approach, suggested by (Odijk, 2011), is parameterized search, but this has not yet been implemented in any of the search applications (see section 7).

The analysis of the modification potential of these words is therefore work for the future. It is already clear that many of the synonyms are untypical for children and are probably acquired rather late. It is therefore interesting to investigate whether there is a relation between the timing of acquisition of these words and their modification potential.

### 5.5 COAVA

Since the problem we are interested in concerns (first) language acquisition, it is obvious that data that directly concern language acquisition must be taken into account. The most important data set for language acquisition is the CHILDES data set.

[20] http://portal.clarin.nl/node/1944

[21] http://cornetto.clarin.inl.nl/simple_search.xql

[22] http://cornetto.clarin.inl.nl/advanced_search.xql

[23] http://cornetto.clarin.inl.nl/wordnet.xql

[24] It is pretty difficult and often quite arbitrary to add translations to these words, and they are not needed for understanding the current paper, so I left them out.

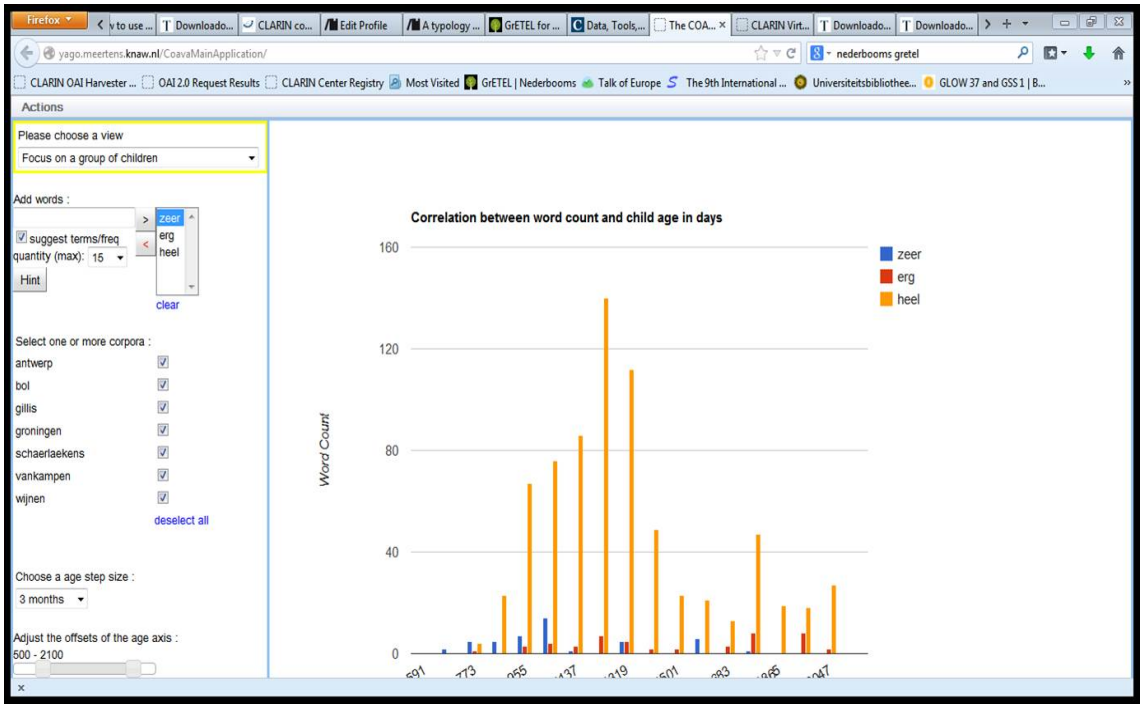



Figure 1: Frequency of the word *heel*, *erg* and *zeer* in the children's speech in the Dutch CHILDES corpora. The X-axis specifies time intervals of three months, the Y-axis indicates the frequency of the word. Each word has a separate color (blue=*zeer*, orange=*heel*, red=*erg*)

The Dutch CHILDES corpora[25] are accessible via the CLARIN Virtual Language Observatory (VLO)[26] or directly via Talkbank[27] and contain relevant material to investigate the research questions formulated in section 4. They contain transcriptions of dialogues between children acquiring Dutch on the one hand, and adults (mostly parents) and in some cases other children on the other hand, and a lot of additional information about the context, setting, age of the child, etc.

I investigated the occurrence of the words *heel*, *erg* and *zeer* in the CHILDES data through the COAVA web application[28]. The COAVA[29] web application provides combined access to two sets of databases: one with historical dialect data (the databases WBD[30] and WLD[31] with lexical data of the Brabantish and Limburgian dialect between 1880-1980) and one with first language acquisition data.

Though COAVA offers many facilities for research into the relation between language acquisition and lexical variation, my main interest is in the occurrence, and especially the first occurrence of the words *heel*, *erg* and *zeer* in the children's utterances. Figure 1 shows this.

From this figure, we can conclude that the word *zeer* occurs first, followed by *heel*, and *erg*. However, each of the words *heel*, *erg* and *zeer* is ambiguous. COAVA does not take this into account, so we do not know whether the first occurrences observed concern the relevant sense ('very') of these words. In (Odijk, 2014a) I therefore made a manual analysis, which yields different results, as shown in Table 1.[32]

From this table, one can conclude that the first occurrence of *heel* in the sense 'very' is used very early by children (before their second birthday); the first occurrence of *erg* appears only about a year later,

[25] I considered the subcorpora DeHouwer, Gillis, Groningen, Schaerlaekens, VanKampen, Wijnen and Zink, but not CLPF.

[26] `http://catalog.clarin.eu/vlo/search?fq=languageCode:code:nld&fq=collection:TalkBank`

[27] `http://childes.talkbank.org/data/Germanic/Dutch/`

[28] `http://portal.clarin.nl/node/1928`

[29] Acronym for *Cognition, Acquisition and Variation Tool*

[30] `https://vlo.clarin.eu/search?3&fq=collection:Dictionary+of+the+Brabantic+dialects`

[31] `https://vlo.clarin.eu/search?2&fq=collection:Dictionary+of+the+Limburgian+dialects`

[32] The table specifies the age of the child in days, followed by the CHILDES notation for children's ages in the format *(year;month)*.

| First Occurrence | heel | erg | zeer |
|---|---|---|---|
| **Day (Year;Month)** | 705 (1;11) | 1048 (2;10) | 1711 (4;8) |

Table 1: First Occurrence of *heel*, *erg* and *zeer* in the relevant sense ('very') in the Dutch CHILDES Children's speech

and *zeer* occurs only very late (far in the fourth year). The latter may be related to the fact that *zeer* is considered rather formal by many people, and also occurs rather infrequently in adult child interactions in CHILDES. Note that the very early occurrence of *zeer* in Figure 1 involves a different sense of this word, viz. as *pain* or *painful*.

Clearly, it is desirable to have the manual analysis carried out here supported or even completely replaced by an automatic procedure. The next section describes a first step towards this goal.

### 5.6 PaQU

As we saw in the preceding section, a serious problem for the investigation is that the words being investigated are, as any decent word in natural language, highly ambiguous. Table 2 describes the ambiguity. For example, the word *heel* is 6-fold ambiguous. This ambiguity is partly solved by taking into account morpho-syntactic and syntactic factors. For *heel* as a finite verb (Vf) the ambiguity reduces to 2, which cannot be further resolved by morpho-syntax or syntax: 'heal' and 'receive' (of stolen goods). As an adjective (A) *heel* is 4-fold ambiguous. The ambiguity is partially resolved by taking into account its syntactic properties with regard to modification: if it modifies an adjective (mod A), the ambiguity is resolved to the single meaning 'very'; if it modifies a noun (mod N), the ambiguity is reduced to 3: 'whole', 'in one piece' or 'large'. If it is used as a predicative complement, it can only mean 'in one piece'.[33]

| Word | Morphosyntax | Syntax | Meaning |
|---|---|---|---|
| *heel* | A | Mod N | 1. 'whole'<br>2. 'in one piece'<br>3. 'large' |
| | | predc | 'in one piece' |
| | | Mod A | 'very' |
| | Vf | | 1. 'heal'<br>2. 'receive' |
| *erg* | N | uter | 'erg' |
| | | neuter | 'evil' |
| | A | Mod N, predc | 'bad', 'awful' |
| | | Mod A V P | 'very' |
| *zeer* | N | | 'pain' |
| | A | Mod N, predc | 'painful' |
| | | Mod A V P | 'very' |

Table 2: Ambiguity of the words *heel*, *erg* and *zeer*

[33] I use the following notation in the table: *Mod X* means that the word can modify a word of category X; *Mod X Y Z* means that a word can modify words of any of the categories X, Y, or Z; *predc* stands for *can occur as predicative complement*; Dutch distinguishes two values for gender: *uter* (i.e., common gender) and *neuter*. *Vf* stands for *finite verb form*.

The Dutch CHILDES corpora do not contain any information about the meanings of its word occurrences. Fortunately, as is clear from Table 2, most of the ambiguities can be resolved by taking into account morpho-syntactic and syntactic properties of the word occurrences. However, as observed above, the Dutch CHILDES corpora do NOT have (reliable) morpho-syntactic information (part of speech tags) or syntactic information for the utterances either.

One would want to be able to automatically parse the CHILDES corpora, and to upload the resulting treebank in a search and analysis application. PaQu was developed for this purpose.[34]

The web application PaQu[35] was developed by the University of Groningen. It enables one to upload a Dutch text corpus. This text corpus is either already parsed by Alpino, or if not, PaQu can have it automatically parsed by Alpino. After this, it is available in the word relations search interface of PaQu (an extension of the LASSY Word Relations Search application[36] originally developed by (Tjong Kim Sang et al., 2010) and discussed in section 5.2), as well as via PaQu's XPATH interface.

For the specific problem dealt with here, we need, for each of the words *heel*, *zeer* en *erg*, a characterisation of the part of speech of the head word it is a dependent of and the label of the dependency relation (grammatical relation) holding between them. PaQu offers a dedicated interface precisely for this (see Figure 2). The relevant queries are not easily expressed in XPATH[37], which makes GrETEL (after it has been extended with corpus upload facilities) less suited for this particular problem (but it might be more suited for other problems).

The output of PaQu is a list of utterances that match the query, and (partially user-definable) statistics on properties of matched words and matched triples of the form (property of dependent word, grammatical relation, property of head word).[38] See Figure 3. Each of the matches and each of the statistical aggregates contains links with automatically generated queries for exploring specific subcases in more detail.

PaQu accepts as input plain text (in multiple varieties) or a text corpus parsed by Alpino in the LASSY XML[39] format. It currently does not allow a CHILDES corpus (in CHAT format (MacWhinney, 2015)) directly as input. This clearly requires an extension of PaQu (see section 7). For the experiments described below I wrote an ad-hoc script to select and clean utterances from CHILDES corpora (see (Odijk, 2015a) for details).

PaQu offers full parses of sentences in a corpus, but these parses have been generated in a fully automatic manner, so they will contain errors. It is therefore required to evaluate the quality of the automatically generated parses. (Odijk, 2015a) describes the results of such an evaluation for the words *heel*, *erg* and *zeer* dealt with here in the CHILDES Van Kampen subcorpus.[40] The results are summarized in Table 3, both for the adult speech (column *Adults*) and for the children's speech (column *Children*)

The results for the adults' speech and the children's speech shows a similar distribution, though the results for the children's speech are lower. For the adult speech, the results for *heel* and *erg* are very good with over 90% accuracy compared to the gold standard. The results of *zeer* appear to be very bad. Further analysis reveals that most errors are made for the construction *zeer doen*, lit. *pain do*, 'to hurt', which Alpino really does not know how to analyze. The word *zeer* in this expression is correctly analyzed by Alpino as a noun, an adjective, or an adverb[41], but the grammatical functions assigned vary widely and

[34] Analogously, the *AutoSearch* application was developed to support search in corpora with annotations on tokens. AutoSearch is a web application developed by INL. Here FoLiA or TEI formatted Dutch text corpora containing (extended) PoS codes (e.g. as created by the Frog (van den Bosch et al., 2007) part of speech tagger in TTNWW) can be uploaded and searched via a Corpus of Contemporary Dutch -like search interface. This application will not be discussed in this paper any further.

[35] `http://portal.clarin.nl/node/4182`

[36] `http://www.let.rug.nl/~alfa/lassy/bin/lassy`

[37] Such a query has to take into account not only headed structures but also coordinated structures and co-indexed nodes in the syntactic structure. In addition, the dependent word can be contained in a phrase that is a dependent of the head word.

[38] Where *properties* include *word form*, *lemma*, and *part of speech*.

[39] `http://www.let.rug.nl/vannoord/Lassy/alpino_ds.dtd`

[40] If one logs in into the PaQu application, one actually finds the parsed corpora with the cleaned Van Kampen adult sentences, since I shared the corpora with everyone. They are called *VanKampenHeel*, *KampenErg*, and *VanKampenZeer*, resp. The children's utterances in Van Kampen are in the corpus *VanKampen-child-heelergzeer*.

[41] Alpino distinguishes adverbs from adjectives in some cases by means of the syntactic category. The gold standard does not distinguish adverbs from adjectives by syntactic category.

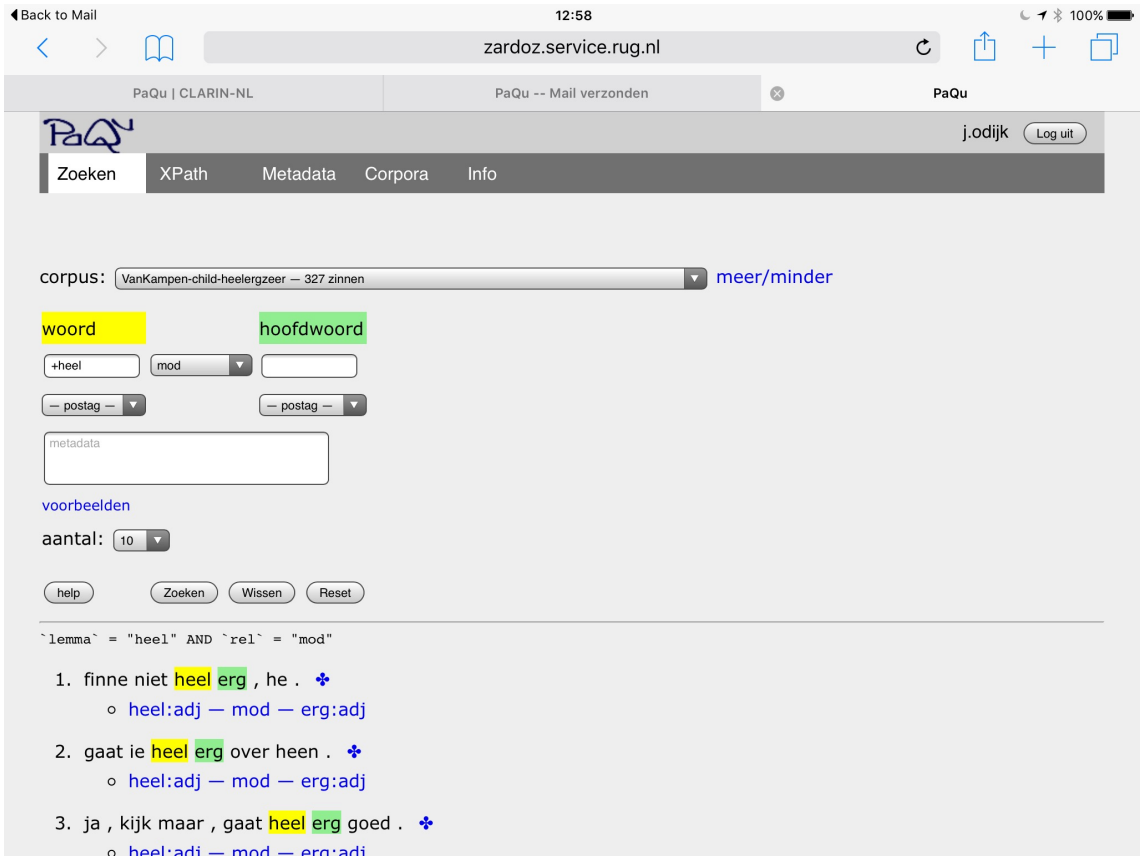



Figure 2: PaQu web interface with a query for occurrences of the lemma *heel* as modifier

| word | Adults | Children |
|---|---|---|
| *heel* | 0.95 | 0.90 |
| *erg* | 0.91 | 0.73 |
| *zeer* | 0.21 | 0.17 |

Table 3: Accuracy of Alpino parses for the words *heel*, *erg* and *zeer* in the CHILDES Van Kampen subcorpus

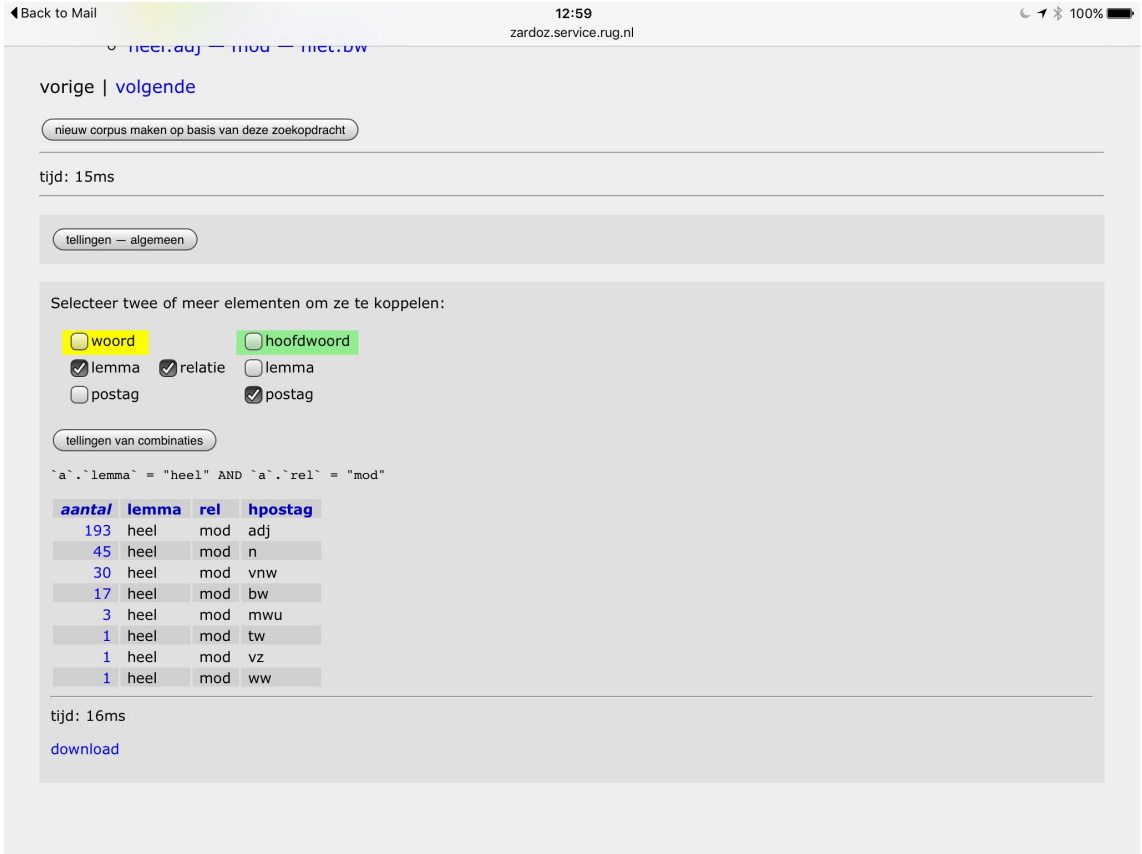



Figure 3: PaQu analysis: count of occurrences of the lemma *heel* as modifier by part of speech of the modifiee.

are mostly incorrect: *direct object*, *predicative complement*, *modifier*, and even *subject*. For a linguist, the analysis is also not obvious, but I have analyzed *zeer* in this construction in all cases as a predicative complement to the verb *doen*. Whether *zeer* is a noun or an adjective is often indeterminable, and this distinction has not been taken into account in making the comparison.

Since the bad results for *zeer* are mainly caused by one type of construction, which can be easily identified in PaQu[42], the results of PaQu are still very useful.

Though (Odijk, 2015a) correctly warns against generalizing these results to other cases, they are nevertheless promising: high accuracy in some cases, and the low accuracy examples are easily identifiable.

The results of an analysis of the words *heel*, *erg* and *zeer*, based on an automatic parse of all adult utterances in the Dutch CHILDES corpora are given in Table 4.[43] It specifies, for each of the three words, the counts of their occurrences in specific grammatical roles that concern us here, the counts of their occurrences in other grammatical roles (*other*), and of cases where the grammatical role could not be determined (*unclear*).[44]

| **Results** | **mod A** | **mod N** | **Mod V** | **mod P** | **predc** | **other** | **unclear** | **Total** |
|---|---|---|---|---|---|---|---|---|
| *heel* | 881 | 51 | 2 | 2 | 14 | 0 | 2 | **952** |
| *erg* | 347 | 27 | 109 | 0 | 187 | 5 | 0 | **675** |
| *zeer* | 7 | 1 | 83 | 0 | 19 | 21 | 7 | **138** |

Table 4: Analysis of *heel*, *erg* and *zeer* in adult utterances in Dutch CHILDES

(Odijk, 2015a) analyzes these findings in some detail, and the results can be summarized as follows:

- *Heel* is most frequent (almost 54%)
- *Heel* as mod A is overwhelming: ($>$ 93%)
- *Heel* as mod V, mod P are analyzed incorrectly
- For *erg*, the distribution between Mod A and mod V is more balanced than for *heel*
- Evidence for *zeer* is mostly lacking. The examples of *zeer* as Mod V are mostly wrong analyses
- Evidence for Mod P is mostly lacking, though there is some evidence for *erg* en *zeer* (4 occurrences)

This example clearly shows the advantages of using PaQu for manual verification of hypotheses, and shows that, if some care is exercised, it can also be used for automatic verification of hypotheses. However, PaQu, in its current state, is not yet able to derive Table 1 or a variant of Figure 1 for the words *heel*, *erg* and *zeer* in the relevant sense. That requires an analysis of the search results in terms of a mix of linguistic annotations and metadata pertaining to the whole utterance or the whole session. See section 7.

## 6 Conclusions

We can draw two types of conclusions from the work presented in this paper: conclusions with regard to the linguistic problem, and conclusions with regard to CLARIN as a research infrastructure.

Starting with the linguistics, any conclusions here must be very preliminary, given the small scale of the research done here. Nevertheless, the observations made in the preceding section are suggestive of further research. For example, they suggest that the overwhelming amount of occurrences of *heel* as a modifier of an adjective in comparison to its occurrence as a modifier of a verb (881 v. 2), perhaps in combination with its early occurrence (see section 5.5), might play a role in fixing the modification

[42]Through the query `http://zardoz.service.rug.nl:8067/?db=childesadultsheelerga&word=zeer&rel=&hword=%2Bdoen&postag=&hpostag=`; login is required to access the corpus.

[43]The results reported here deviate slightly from what (Odijk, 2015b) reported. In the current table the wrong mapping of the pronoun *wat* has been corrected, and changed from *mod A* to *mod N*. This concerns 5 examples, all modified by *heel*. This small correction does not affect the overall results.

[44]For example, in incomplete or ungrammatical utterances.

potential of this word to adjectives. In contrast, the occurrences of the word *erg* as a modifier of adjectives and verbs are more balanced: 347 v. 109.

The fact that there are hardly any examples for *zeer* make it difficult to draw any conclusions. In any case, the current CHILDES data give no clue how the use of *zeer* as a modifier of A, V, P is acquired, simply because there are hardly any data. This most probably means that the current Dutch CHILDES databases are insufficiently large as a sample of first language acquisition.[45]

Concerning CLARIN, (Odijk, 2011) defined a linguistic problem and specified what kinds of search actions would be needed to address this problem. At the time, almost none of these search actions were possible, or only with great difficulty, and they required expert knowledge on the relevant databases and programming skills. In 2014, (Odijk, 2014a) showed that many of the desired search actions had become possible, in a simple manner, and through applications with interfaces dedicated to the targeted users, linguists. At the same time, it was observed that not everything was possible yet in an easy way, and new requests arose by using the relevant applications. Since neither (Odijk, 2011) nor (Odijk, 2014a) was published, I report on their findings in this paper, and I showed new functionality created to accommodate the newly arisen need. This paper thus serves as an example of a report on a *research pilot*: a project to use functionality offered by the infrastructure with the twin goals of furthering the research but also of identifying novel functionality that the infrastructure should offer to be able to further the research.

This paper shows great progress in the number of applications offered in the CLARIN infrastructure, and a significant increase in the functionality that they offer, but I have also identified functionality that was desired from the start as well as novel desired functionality that have not been implemented yet.

## 7 Future Work

There is a lot of work that can (and should) be done in the near future. Firstly, the same words could be investigated in other corpora that are relevant for language acquisition, in particular the Basilex corpus[46]. Secondly, similar experiments can be carried out for other tuples of (near-)synonymous words with different syntactic selection or modification properties. One example is *te* v. *overmatig*, which both mean 'too' but differ in modification potential (*te* only A, *overmatig* at least A and V). Another example concerns the copular verbs *worden* 'become' v. *raken* 'get', in which *worden* can only take NP, AP and a very limited number of PP predicates, while *raken* can take only AP and PP predicates, very similar to their English translations *become* and *get*. Of course, as usual in natural language, most of these words are ambiguous.[47] Most of these ambiguities can be resolved by the syntactic contexts, so treebanks can (and must) be used to find the relevant examples and their statistics.

It surely also makes sense to manually verify and where needed correct (parts of) parses for CHILDES corpora, improving the reliability of the annotations on these data.

I have identified many instances of desired functionality that is not available yet. (Odijk, 2011) suggested parameterized search, but this has not yet been implemented. The functionality of uploading one's own corpus should also be added to other treebank search applications, in particular the GReTEL[48] application (Augustinus et al., 2012). All search engines that allow uploading one's own corpus must be extended to support input in all formats commonly used in linguistics. For example, PaQu only allows plain text as input, but it should actually support, e.g, the CHILDES CHAT format, the FoLiA[49] format (van Gompel and Reynaert, 2013) and TEI[50]. In addition, it should take in not only the actual data, but also the metadata of the corpus, its subcorpora or textual units such as utterances, paragraphs etc.

Search applications should offer extensive options for analyzing the search results. Such analysis options are available in PaQU and OpenSONAR, but hardly in GrETEL, and the PaQU and OpenSONAR

[45] A rough count shows that the Dutch CHILDES corpora dealt with here contain 534 k utterances and approx. 2.9 million inflected word form occurrences ('tokens').

[46] `http://tst-centrale.org/nl/producten/corpora/basilex-corpus/6-158`

[47] For example, *te* is an adjective, a preposition, and an infinitive marker; *raken* is not only a copula but also a transitive verb (with two meanings); *worden* is not only a copula but also a passive auxiliary.

[48] `http://nederbooms.ccl.kuleuven.be/eng/gretel`

[49] `http://proycon.github.io/folia/`

[50] `http://www.tei-c.org/index.xml`

analysis options must be extended as well. In particular, the search applications should enable users to carry out analyses not only on the data but on arbitrary combinations of search result data and and their metadata.

It is also essential that the search results can be further annotated by users, or at least categorized. This is important since most search actions in practice do not yield exactly the set the researcher is interested in (there are problems of recall and of precision). With a categorisation option, one can use a broader query and then categorize the results (e.g. to exclude some).[51] And these newly added categories should participate as first class citizens in the analysis options offered.

Fortunately, most of the possible future work mentioned here is actually planned in the CLARIAH-CORE[52] project or in the Utrecht University project *AnnCor*, and part of it is already being carried out.[53] With these projects, we hope to be able to run queries such as the ones already suggested in (Odijk, 2011) but currently not possible yet (with *heel*, *erg* and *zeer* only in the relevant sense 'very'):

- For each child, give list of pairs (session, age) of the child
- For each child, give me #sessions by period, where period is e.g. every month, week, half year, year
- For each child give me the list of new words uttered by period
- For child and each session, give #occurrences of *zeer*, *heel*, *erg*;
- Idem, by period
- Give me utterances containing occurrences of *zeer*, *erg*, *heel* uttered by the child before any adult used any of these words
- Give me #occurrences of *heel* uttered by the parent before the child utters it (idem for *zeer*, *erg*, etc.)

and many others that might be needed to address the research questions of section 4.

## Acknowledgements

This work crucially uses data and/or tools made available through the CLARIN infrastructure. The work was financed by CLARIN-NL[54], an NWO project in the Netherlands.

## References

[Augustinus et al.2012] Liesbeth Augustinus, Vincent Vandeghinste, and Frank Van Eynde. 2012. Example-based treebank querying. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

[MacWhinney2015] Brian MacWhinney. 2015. Tools for analyzing talk, electronic edition, part 1: The CHAT transcription format. Technical report, Carnegie Mellon University, Pittsburg, PA, April27. `http://childes.psy.cmu.edu/manuals/CHAT.pdf`.

[Odijk2011] Jan Odijk. 2011. User scenario search. internal CLARIN-NL document, `http://www.clarin.nl/node/166`, April 13.

[Odijk2014a] Jan Odijk. 2014a. CLARIN: What's in it for linguists?, March 27. Uilendag Lecture, Utrecht, `http://dspace.library.uu.nl/handle/1874/295277`.

[51] The Lancaster web access to the British National Corpus offers such categorisation options.

[52] `http://www.clariah.nl`

[53] For example, in AnnCor manual verification and correction of Alpino parses for CHILDES utterances is worked on, and since a few months, PaQu enables analysis of search results in combination with metadata, at least for the Spoken Dutch Corpus. And it already supports more input formats than just plain text, among them FOLIA and TEI.

[54] `http://www.clarin.nl`

[Odijk2014b] Jan Odijk. 2014b. Discovering resources in CLARIN: Problems and suggestions for solutions. unpublished article, Utrecht University, http://dspace.library.uu.nl/handle/1874/303788, August.

[Odijk2015a] Jan Odijk. 2015a. Linguistic research with PaQu. *Computational Linguistics in the Netherlands Journal*, 5:3–14, December.

[Odijk2015b] Jan Odijk. 2015b. Linguistic research with PaQu. Lecture held at CLIN 2015, Antwerp, http://www.clarin.nl/sites/default/files/Poster%20Odijk%20CLIN%202015%202015-02-02.pdf, February 6.

[Oostdijk et al.2002] N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen. 2002. Experiences from the Spoken Dutch Corpus project. In M. González Rodriguez and C. Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 340–347. ELRA, Las Palmas.

[Oostdijk et al.2013] N. Oostdijk, M. Reynaert, V. Hoste, and I. Schuurman. 2013. The construction of a 500 million word reference corpus of contemporary written Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, pages 219–247. Springer, Berlin. http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1.

[Spyns and Odijk2013] P. Spyns and Jan Odijk. 2013. *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*. Springer. http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1.

[Tjong Kim Sang et al.2010] Erik Tjong Kim Sang, Gosse Bouma, and Gertjan van Noord. 2010. LASSY for beginners. Presentation at CLIN 2010, Utrecht, February 5.

[van den Bosch et al.2007] A. van den Bosch, G.J. Busser, W. Daelemans, and S. Canisius. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. Van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste, editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114. Leuven, Belgium.

[van der Beek et al.2002] Leonoor van der Beek, Gosse Bouma, and Gertjan van Noord. 2002. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7:353–374.

[van Gompel and Reynaert2013] Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, 12/2013.

[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. http://dx.doi.org/10.1007/978-3-642-30910-6_9.

[Vandeghinste and Augustinus2014] Vincent Vandeghinste and Liesbeth Augustinus. 2014. Making large treebanks searchable. The SoNaR case. In Marc Kupietz, Hanno Biber, Harald Lüngen, Piotr Bański, Evelyn Breiteneder, Karlheinz Mörth, Andreas Witt, and Jani Takhsha, editors, *Proceedings of the LREC2014 2nd workshop on Challenges in the management of large corpora (CMLC-2)*, pages 15–20. ELRA, Reykjavik. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-CMLC2%20Proceedings-rev2.pdf.

# CLARIN Concept Registry: The New Semantic Registry

**Ineke Schuurman**
KU Leuven, Belgium
and Utrecht University, The Netherlands
ineke@ccl.kuleuven.be

**Menzo Windhouwer**
Meertens Institute
Amsterdam, The Netherlands
menzo.windhouwer@meertens.knaw.nl

**Oddrun Ohren**
National Library of Norway
oddrun.ohren@nb.no

**Daniel Zeman**
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic
zeman@ufal.mff.cuni.cz

## Abstract

The CLARIN Concept Registry (clarin.eu/conceptregistry) is the place in the CLARIN Infrastructure where common and shared semantics of, but not limited to, linguistic concepts are defined. This is important to achieve semantic interoperability, and to overcome to a degree the diversity in data structures, either in metadata or linguistic resources, encountered within the infrastructure. Whereas in the past, CLARIN has been using the ISOcat registry for these purposes, nowadays this new registry is being used, as ISOcat turned out to have some serious drawbacks as far as its use in the CLARIN community is concerned. The main difference between the two semantic registries is that the CCR is a concept registry whereas ISOcat is a data category registry. In this paper we describe why the decision to switch to a concept registry has been made. We also describe the most important other characteristics of the new (Open)SKOS-based registry, as well as the management procedures used to prevent a recurrent proliferation of entries, as was the case with ISOcat.

## 1 Introduction

One of the foundations of the CLARIN Component Metadata Infrastructure (CMDI; Broeder et al. 2012; clarin.eu/cmdi) is a semantic layer (Durco and Windhouwer, 2013) formed by references from CMDI components or elements to entries in various semantic registries. Popular have been references to the metadata terms provided by the Dublin Core Metadata Initiative (DCMI; dublincore.org) and the data categories provided by ISO Technical Committee 37's Data Category Registry (DCR; ISO 12620, 2009) ISOcat (isocat.org). For describing more content-related data, like describing the components of a morpho-syntactic tagset, ISOcat was also used. Using entries either based on (ISO) standards, *de facto* standards (for example generally used in the Netherlands or, broader, in the Dutch-speaking regions) or even made by users themselves (esp. in legacy data).

Although using ISOcat has been encouraged by CLARIN, it has certain drawbacks. As pointed out by Broeder et al. (2014) and Wright et al. (2014), ISOcat, with its rich data model combined with a very open update strategy, has proved too demanding, at least for use in the CLARIN context. Among other things, confusion on how to judge whether a candidate ISOcat entry adequately represents the semantics of some CMDI component or element, has led to proliferation far beyond the real need. This resulted in a semantic layer of questionable quality. Therefore, when ISOcat, due to strategic choices made by its Registration Authority, had to be migrated and became, for the time being, static, CLARIN decided to

look for other solutions to satisfy the needs of the infrastructure. As a result the Meertens Institute is now hosting and maintaining the CLARIN Concept Registry (CCR; `clarin.eu/conceptregistry`).

This paper motivates and describes this new semantic registry, its model, content and access regime, indicating the differences from ISOcat where appropriate. Proposed management procedures for the CCR are also outlined, although not in detail.

## 2 Issues in ISOcat

One of the principles behind ISOcat was that it was an open registry, so it was very easy to get a login and to get the rights to enter new data categories. Such entries could remain private, only you yourself could read and edit them, plus the persons you had given permission to do so. Or they could be made public, i.e. everybody can read them (but not edit). Still the content of the registry was out of control. People were, for example, urged not to provide entries that were more or less copies of already existing ones, but a) there was no way to prohibit it, and b) people sometimes copied an entry, just in order to make sure that the original owner would not change the entry without them knowing it. So the first issue to be addressed was:

- **Proliferation**, due to
  - access: too many people having write access;
  - quality: ignorance or negligence of ISOcat/CLARIN requirements (esp. with regard to definitions);
  - reliability: (public) entries being changed in a semantically meaningful way, i.e. changing their meaning.

The proliferation issue can, to a large extent, be solved by giving far less people permission to contribute new entries and to change existing ones in the new environment, i.e. the CCR should not be an open registry.

Another issue, already referred to in Section 1, concerns the complexity of a Data Category Registry like ISOcat. For example, in ISOcat different entries were needed for the data category /*genre*/, depending on how the needed value domain is defined, even when the definition as such would be the same (cf. Section 3). In other cases one had to specify in which format the 'data type' could be described, with over 40 options to select from. In most cases the default was chosen, as most users did not understand all options. Strictly speaking several parts of the DCR data model are not really needed for CLARIN-purposes, like the 'data type' mentioned. So the second issue to be addressed was:

- **Complexity**, due to
  - forced duplicates (data category type);
  - options only useful for experts in a specific field (data type);
  - obligatory parts of the data model not really necessary for CLARIN purposes.

These issues have been taken care of in the CCR (cf. the next sections). We will first discuss the issue last mentioned (Complexity), as the choice made strongly influences the design of the new registry: dealing with concepts instead of data categories. Avoidance of the causes of proliferation described under the first bullet is mainly related to procedural measures (cf. Section 7).

## 3 From data categories to concepts

As indicated above, the shift from ISOcat to a new format also involved changing the main entity of the registry. Instead of focusing on data categories as before, the new registry contains concepts. The transition is illustrated in Figure 1 depicting core features of the data categories and their abstraction into concepts.

A data category as modeled in ISOcat is an elementary descriptor in a linguistic structure or an annotation scheme (ISO 12620, 2009), implying it is a descriptor of something. In the CMDI universe this

something is most often a language resource or related objects. Moreover, there are several types of data categories, notably

- Data categories representing attributes or properties of something, in the sense that they are to be assigned values. In ISOcat these were called complex data categories. Their value sets may either be closed (controlled value vocabularies), open (to be specified freely by the user) or constrained (must follow specific rules). Their domains (the set of objects to which they may be applied) are not formally specified, but are often implied by the definition.

- Data categories representing atomic elements to be included in the value set of some complex data category. In ISOcat these were called simple data categories.

This means that a data category is defined not only by its meaning, but also by how it may be used. For example, consider a data category */genre/*, defined as a complex category (may be assigned a value) with value to be chosen from a finite set of genres. Any annotation/metadata scheme needing a */genre/* data category must agree to the same set of values. In cases where additional values or a completely different value set is called for, or where the user should be able to specify genre freely, we are in effect talking about different data categories, - the existing */genre/* cannot be reused. The same is true if genre is needed as a value of another complex data category. Consider for example the data category */subject type/* defined with value set including topic, genre, temporal coverage a.o. In this case yet another */genre/* data category will have to be defined, this time as a simple data category.

While such a model is very rich in expressive power, it is notoriously hard to maintain consistency and requires a high degree of understanding and alertness from users to be successful.

In the CCR we want to take advantage of the fact that groups of data categories, although applicable in different contexts have the same meaning in terms of a more or less similar definition. Following the genre example, the core idea of the concept *genre* is the same, whether it is to be used as an attribute with values (and irrespective of value set) or itself as a value of some other attribute, e.g. subject type. By disregarding information on application domain and value range and focusing only on definition and conceptual relations, the registry should be leaner and easier to maintain. On the other hand, the resulting semantic layer spanning the collective set of CMDI records inevitably will be coarser and thereby less informative.

## 4 An OpenSKOS registry

In CLARIN-NL the Meertens Institute had already developed (and continues hosting) the CLAVAS vocabulary service (`openskos.meertens.knaw.nl`) based on the open source OpenSKOS software package (Brugman and Lindeman, 2012; `openskos.org`), which was originally created in the Dutch CATCHPlus project. The OpenSKOS software provides an API to access, create and share thesauri and/or vocabularies, and also provides a web-based editor for most of these tasks. The software is used by various Dutch cultural heritage institutes. The Meertens Institute joined them to collectively maintain and further develop the software.

Based on the experiences with ISOcat OpenSKOS was evaluated to see if it would meet the needs of the CLARIN community and infrastructure. The major aim was to improve the quality of the concepts by having a) a much simpler data model and b) a less open, but also less complicated, procedure for adding new concepts or changing existing ones and recommending them to the community. In addition, certain technological requirements of the CLARIN infrastructure had to be met. Based on this evaluation the Meertens Institute extended the OpenSKOS software in various ways:

- Concepts in the CCR get a handle as their Persistent IDentifier (PID);

- The CCR can easily be accessed by the CLARIN community via a faceted browser (cf. Figures 2 and 3; `clarin.eu/conceptregistry`);

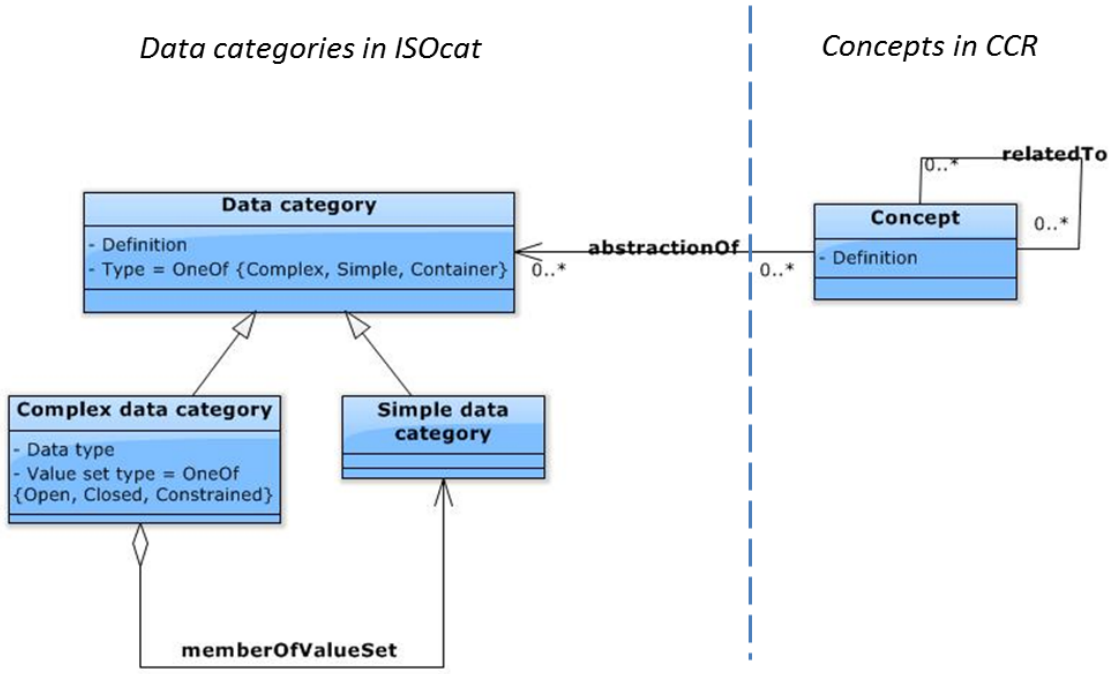



Figure 1: Core features of data categories and their abstraction into concepts

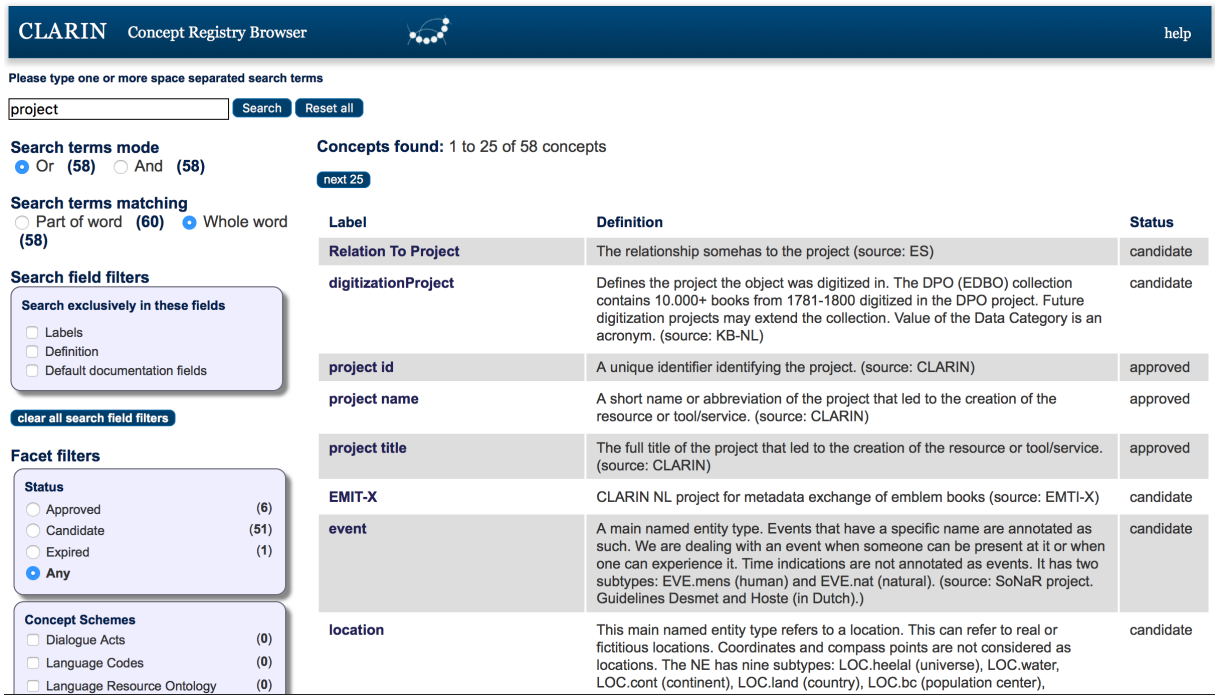



Figure 2: The CCR faceted browser

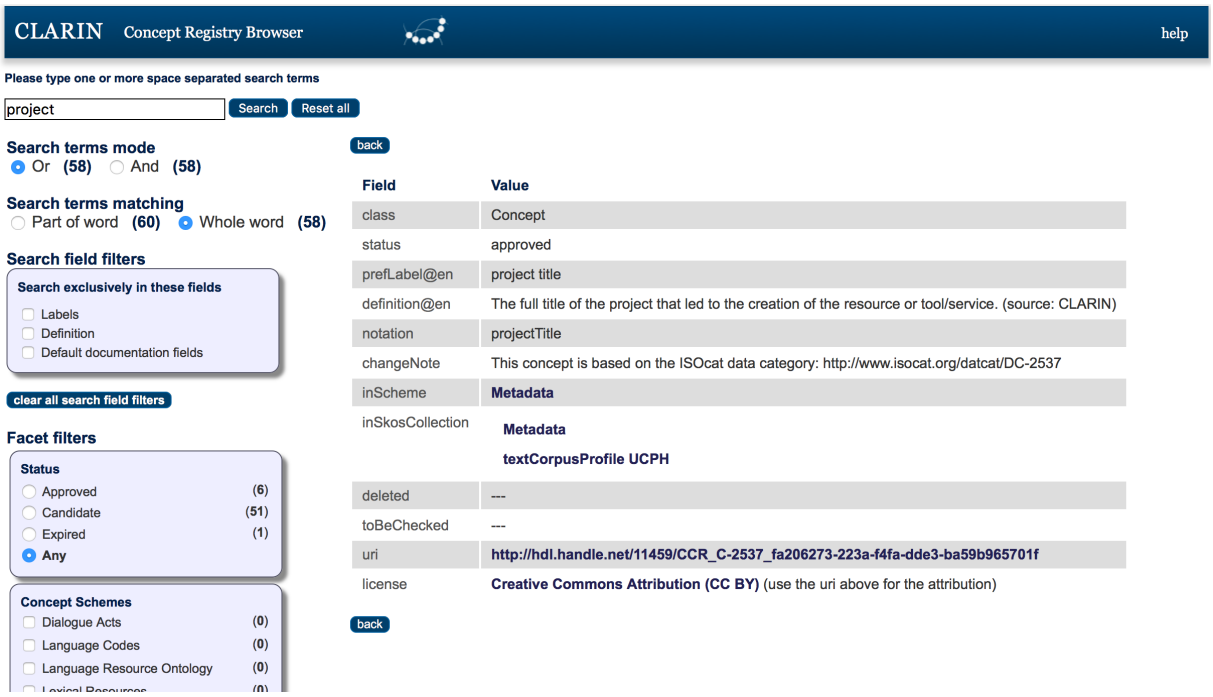



Figure 3: The CCR faceted browser - concept view

- Support for SKOS collections;
- Shibboleth-based access to the CCR.

Currently these extensions reside in a private Meertens Institute source code repository, but as part of the CLARIN-PLUS project (`clarin.eu/content/factsheet-clarin-plus`) these extensions (and more) will be integrated with the next version of OpenSKOS now under development.

## 5 Representing CCR concepts in the SKOS model

The data model supported by OpenSKOS is a substantial part of the Simple Knowledge Organization Scheme (SKOS) recommendation by W3C (`w3.org/skos`). SKOS is typically used to represent thesauri, taxonomies and other knowledge organization systems. At the Meertens Institute support for collections was added and currently Picturae, a Dutch service provider within the cultural heritage domain and the original developer of OpenSKOS, works on supporting the extended labels of SKOS-XL.

The work done by the CLARIN community in ISOcat was made available in the CCR by importing selected sets of data categories as new concepts (cf. Section 6). This made it possible to start a round of clean-up and creating a coherent set of recommended concepts (cf. Section 7). This import is not lossless as data category specific properties like the data category type and data type are lost. However, these properties have turned out to be one of the main causes of confusion and proliferation in the use of ISOcat (Broeder et al., 2014; Wright et al., 2014). In general SKOS appears to be a suitable model for the CCR. Each CCR concept may be assigned preferred labels (at most one per language,[1] alternative labels, definitions, examples and various kinds of notes. Moreover, the ISOcat thematic domains and data category selections could be maintained by importing them to SKOS concept schemes and collections, respectively. Only one import decision turned out to be problematic: converting the data category

[1]For the moment all entries are in English. Only when the entries have been approved by the national CCR coordinators other languages may be added. This is a lesson learned from ISOcat where translations often were not in sync.

identifier into a concept notation. SKOS notations are required to be unique within their concept scheme, whereas this constraint did not apply to data category identifiers in the DCR data model. A first clean-up round to remedy this has been finished successfully.

The SKOS model provides the possibility to express semantic relationships between concepts, e.g. broader than, narrower than and related to. In contrast, the DCR data model only contained relationships based on the data category types, e.g. a simple data category belonged to the value domain of one or more complex data categories. These domain-range relationships do not correspond well to any of the SKOS relationship types. Careful manual inspection would be needed to determine if any mapping can be made. Hence, for now these relationships have not been imported into the CCR. At a later date these facilities of the SKOS model and OpenSKOS can be exploited and could eventually take over the role originally envisioned for RELcat (Windhouwer, 2012). For now the initial focus is on the concepts themselves.

Neither SKOS itself nor OpenSKOS yet provides an extensive versioning model, i.e. concepts can be expired but there is no explicit link to a superseding concept. This is now on the wishlist for the next version of OpenSKOS as developed in CLARIN-PLUS.

Being RDF-based SKOS also brings the potential to more easily join forces with the linked data and semantic web communities. However, our current focus is on cleaning up the registry, thus gradually obtaining a coherent hub to be offered to the linked data cloud.

## 6 The CCR content

In the past few years, many national CLARIN teams made an effort to enter their data in ISOcat. This work has not been useless as all entries of relevance to a specific CLARIN group have been imported into the CCR. Leaving out redundant entries already means a considerable reduction in number of entries (from over 5000 in ISOcat (Broeder et al., 2014) to 3139 in CCR (June 2015). Although the imported concepts received new handles, care was taken to retain a link with their ISOcat origin. Automated mapping is thus possible and can be used to convert references to ISOcat data categories into references to CCR concepts. A mapping tool[2] for this has been developed and especially used for the existing CMDI components and profiles. But the tool is generic and can be used for other types of resources.

## 7 Maintaining the CCR: procedures and actors

Just like ISOcat the CCR can be browsed and searched by anyone, member of the CLARIN community or not, and anyone can refer to the concepts. However, contrary to ISOcat, only specifically appointed users, namely the national CCR content coordinators[3] are given rights to update the CCR (cf. Figure 4). These coordinators were appointed by their respective CLARIN national consortia when the problems with the usage of ISOcat became apparent. Their mission is to improve the quality of the data categories (now concepts) used within CLARIN. With the CCR in place the national CCR content coordinators have teamed up more actively and established procedures around the CCR to fulfill this mission.

To deal with the ISOcat legacy the coordinators are doing a round of clean-up with the aim to deprecate[4] low quality concepts and recommend high quality concepts. Notice that, just like in ISOcat, deprecated concepts remain accessible, i.e. their semantic descriptions are not lost, but their active usage is discouraged. The main focus is on providing good definitions. A good definition should be ”as general as possible, as specific as necessary” and should therefore be:

1. Unique, i.e. not a duplicate of another concept definition in the CCR;
2. Meaningful;
3. Reusable, i.e. refrain from mentioning specific languages, theories, annotation schemes or projects;
4. Concise, i.e. one or two lines should do;

[2] `github.com/TheLanguageArchive/ISOcat2CCR`

[3] Cf. `clarin.eu/content/concept-registry-coordinators`

[4] In the OpenSKOS status model deprecation means a concept gets the status expired.

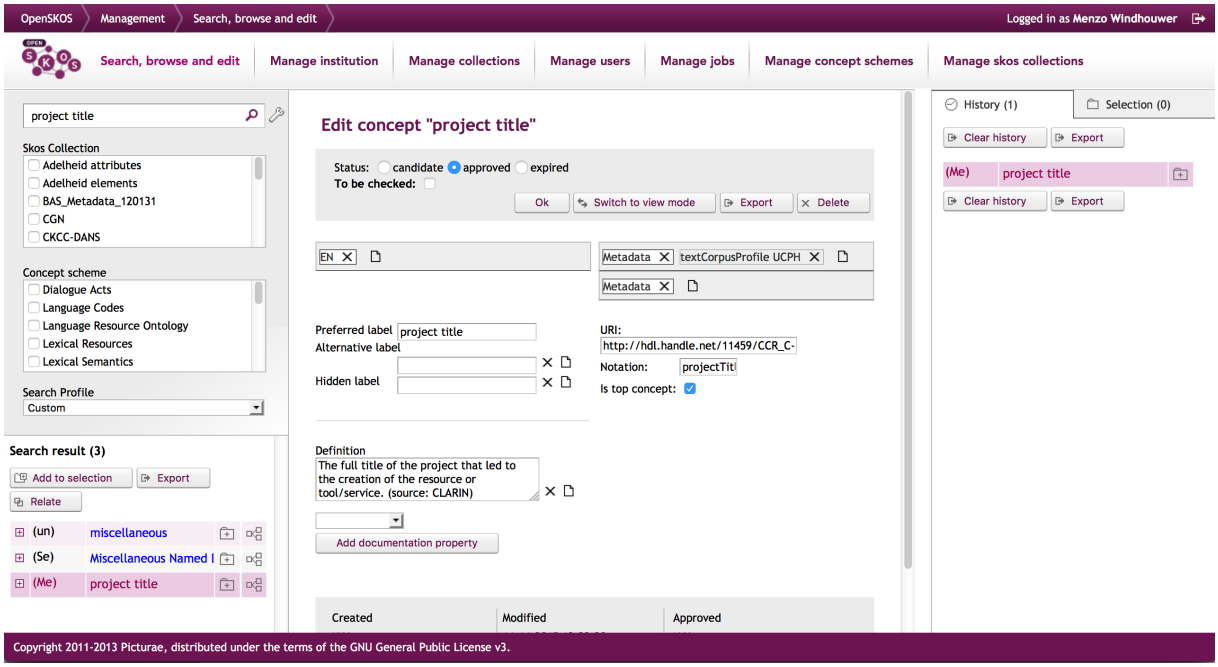



Figure 4: The CCR editor

5. Unambiguous.

As far as point 5 is concerned, a concept used in the entry of another concept should be referred to by its handle. Detailed guidelines are under development by the coordinators and will become generally available in due course. Apart from defining best practice for the coordinator group, such guidelines will benefit users directly, enabling them to issue informed requests to the CCR coordinators (see below).

The changes the coordinators can do to existing concepts are limited, i.e. they should not change the meaning. Only typos, unclear formulations, etc. can be remedied. Otherwise a new concept has to be created, and the original one may be deprecated.

All the coordinators or their deputies are involved in these changes. In cases where they do not agree a vote might take place and the change will be performed if 70% or more of the coordinators agree. A book keeping of the results of votes is maintained at the CCR section of the CLARIN intranet. The time frame within which the discussions and possibly a vote have to reach a decision is 2 weeks. In the holiday seasons and during the initial start-up phase a longer time period can be agreed upon by the coordinators.

Members of the CLARIN community wanting new concepts or changes to existing ones need to contact their national CCR content coordinator (`clarin.eu/content/concept-registry-coordinators`). Users from countries with no content coordinator should use the general CCR email address (ccr@clarin.eu) to file their requests. These requests will then be discussed within the national CCR content coordinators forum as described above. Note that in OpenSKOS any changes made to concepts are directly public. Therefore new entries or changes will only be entered after their content has been approved by the content coordinator forum. This procedure will take some time, but should result in a registry containing concepts with a better quality and with less proliferation. And therefore the CCR content should eventually deserve a higher level of trust as was the case for the ISOcat content. One can also expect that the need for new concepts will diminish over time due to the CCR covering more and more of the domains relevant to CLARIN.

## 8 Future work

In the Netherlands two other linguistic projects are also using concept registries. We want to investigate what the possibilities are with respect to interoperability, e.g. migrate concepts from these registries into the CCR. But keeping in mind that the maintenance of the collections of concepts should be kept strictly separate in order not to run into troubles à la ISOcat, i.e. it should remain clear which concepts are recommended by CLARIN.

Furthermore, as mentioned before the CLARIN-PLUS project, which started in the second half of 2015, aims to strengthen the CLARIN infrastructure on various fronts, including the CCR. Although the focus is mainly technical the improvements will also help the CCR coordinators with their task and improve the expressiveness of the CLARIN semantic interoperability layer. The CLARIN-PLUS CCR work is done by the Meertens Institute and is split into 4 phases (CE-2015-0688):

**Phase 1** Monitor and test the stability of the new OpenSKOS version currently under construction by the OpenSKOS user community, especially Sound & Vision and Picturae;

**Phase 2** Merge the currently existing various OpenSKOS forks into one trunk, so everyone in the user community can benefit from new features and stability improvements;

**Phase 3** Implement features that fully support the concept life cycle, e.g. referring to a succeeding concept from a deprecated concept;

**Phase 4** Implement features to support internal and external relationships of any type, i.e. not only SKOS relations and not only between concepts, and attribution thereof.

Currently, early 2016, phases 1 and 2 are ongoing and focus on strengthening the technical basis of OpenSKOS, and thus the CCR. The upcoming phases 3 and 4 will add new functionality needed by the CLARIN, and OpenSKOS, community.

## 9 Conclusions

Although CLARIN just started working on the new OpenSKOS-based CLARIN Concept Registry and there is still a lot of ISOcat legacy to deal with, the new registry looks promising. Our feeling is that it will be able to provide a more sustainable and higher quality semantic layer for CMDI. An important lesson from the ISOcat experience is that technology is not always the main problem, although a complicated data model or interface never helps. What we do believe in, is establishing robust, yet simple management procedures, as outlined in Section 7. These rely on teamwork in the national CCR content coordinators forum, together with active involvement of the user community.

## Acknowledgments

The authors like to thank the national CCR content coordinators forum for the fruitful discussions on the procedures around the CCR. They also like to thank the Max Planck Institute for Psycholinguistics, CLARIN-NL and the Meertens Institute for their support to realize a smooth transition from ISOcat to the CCR.

## References

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a Component Metadata Infrastructure. Proceedings of LREC Workshop *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*. Istanbul, Turkey.

Daan Broeder, Ineke Schuurman, and Menzo Windhouwer. 2014. Experiences with the ISOcat Data Category Registry. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Hennie Brugman and Mark Lindeman. 2012. Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service. *Proceedings of the Describing Language Resources with Metadata workshop (LREC 2012)*, Istanbul, Turkey.

CE-2015-0688. *CLARIN-PLUS CCR analysis*. CLARIN ERIC, Utrecht.

Matej Durco and Menzo Windhouwer. 2013. Semantic Mapping in CLARIN Component Metadata. In E. Garoufallou and J. Greenberg (eds.), *Metadata and Semantics Research (MTSR 2013)*, CCIS Vol. 390, Springer.

ISO 12620:2009. *Specification of data categories and management of a Data Category Registry for language resources*. International Organization for Standardization, Geneve.

Menzo Windhouwer. 2012. RELcat: a Relation Registry for ISOcat data categories. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman and Daan Broeder. 2014. Segueing from a Data Category Registry to a Data Concept Registry. *Proceedings of the 11th international conference on Terminology and Knowledge Engineering (TKE 2014)*, Berlin, Germany.

# DMPTY – A Wizard for Generating Data Management Plans

**Thorsten Trippel** and **Claus Zinn**
Seminar für Sprachwissenschaft
Universität Tübingen, Germany
thorsten.trippel@uni-tuebingen.de
claus.zinn@uni-tuebingen.de

## Abstract

To optimize the sharing and reuse of existing data, many funding organizations now require researchers to specify a management plan for research data. In such a plan, researchers are supposed to describe the entire life cycle of the research data they are going to produce, from data creation to formatting, interpretation, documentation, short-term storage, long-term archiving and data re-use. To support researchers with this task, we built DMPTY, a wizard that guides researchers through the essential aspects of managing data, elicits information from them, and finally, generates a document that can be further edited and linked to the original research proposal.

## 1 Introduction

All research depends on data. To address a research question, scientists may need to collect, interpret and analyse data. Often the first phase of scientific activity, data collection, is the most decisive, and also a time-consuming and human-resource-intensive task. It must be planned well enough so that a significant number of data points are available for subsequent inspection so that underlying research questions can be analysed thoroughly. When the analysis of data is yielding results of significant interest, the study is described in scientific parlance, and then submitted to a scientific conference or journal. Once reviewed and accepted for publication, the resulting article constitutes the formal act of sharing research results with the scientific community, and most articles in reputable publication outlets are archived for posteriority. While the results are now public, the underlying research data often remains private, and usually stays with the individual researcher or the research organization. This makes it hard for other researchers to find and to get access to the data, and limits the opportunity for them to reproduce the results, or to base secondary studies on the same data. In brief, the sharing and long-term archiving of research data has these four main benefits:

**Reproducibility:** One of the main principles of the scientific method is reproducibility: it shall be possible to replicate experimental results, in preference by redoing the analysis on the existing data rather than on newly collected data. This discourages fraud and tempering with research data.

**Facilitation of secondary studies:** With researchers having access to existing data sets, there is no need for a costly collection of new data, and therefore, it becomes easier for researchers to explore similar research questions, contrastive studies, or meta studies.

**Attribution:** It should be good scientific practise to give an explicit acknowledgement of ownership or authorship to the one who has collected the data. Scientific reputation shall not only be merited by findings, but also by having acquired underlying data.

**Economy:** Funding money and researchers' time shall not be wasted for collecting data sets if comparable data already exist. Open access to existing data also allows researchers to add to existing data sets, and hence might contribute towards a "Wikipedia effect", yielding increasingly rich resources.

To reap these benefits, research data should be accessible in public repositories, properly documented, with generous access rights, and possibly, in an easy-to-read, non-proprietary data format.

Funding agencies increasingly require grant applicants to complement their research plan with a plan for managing and sharing the research data that is going to be created during their research. In the United Kingdom, the Digital Curation Centre maintains a list of the major British funding bodies and their data policies [1]. In Germany, the national research foundation (DFG) expects data management plans, at least for larger collaborative research projects, and often funds personnel and necessary hard- and software to ensure that data is sustainably archived. The situation is similar in Switzerland where the Swiss National Science Foundation requires researchers that "the data collected with the aid of an SNSF grant must be made available also to other researchers for secondary research and integrated in recognised scientific data pools" [2]. Some universities see data management issues as an integral part to good scientific practise. In the *Guidelines for Research Integrity of the ETH Zurich*, Article 11 is about the collection, documentation and storage of primary data, primarily, with the purpose that all "research results derived from the primary data can be reproduced completely" [3]. The *Research Data Management Policy* of the University of Edinburgh states that "All new research proposals [...] must include research data management plans or protocols that explicitly address data capture, management, integrity, confidentiality, retention, sharing and publication" [4].

Successful research proposals are centred around one or more research questions, the applicants apply sound research methodologies, and in empirically-driven research, this usually includes a convincing approach to gather and analyse research data to address the research questions. Individual research projects, by their very nature, take a short-term view: at the end of the project's lifetime, research questions have been addressed in the best possible way, and results properly written up and published. Usually, there is no long-term view, in particular, with respect to the research data gathered. What is the research data life cycle, how will data be stored (*e.g.*, using which format) and adequately documented? How can the data be cited and made accessible for the scientific community for the long term (archival)? With regard to accessibility, how is personal or confidential data be taken care of? Which licence should be chosen for the data to ensure that other researchers can access the data in the future? Does the data collection ensure that research results can be reproduced, or that follow-up studies can use the data with ease?

We must not expect researchers to handle such questions, which are secondary to the research question, entirely on their own. Taking the long-term perspective requires a different set of skills, and it calls for a cooperation between researchers and a permanent research infrastructure. It is advisable that such cooperation is initiated at the early stage of a research project so that all aspects of the data life cycle are properly taken care of. Infrastructures such as CLARIN can assist researchers in managing their data.

The remainder of this paper is structured as follows. In Sect. 2 we describe the common elements of data management plans. Sect. 3 sets data management in the CLARIN context and defines the division of labour and shared responsibilities between data producer and data archive. In Sect. 4, we present the DMPTY wizard for data management planning. In Sect. 5, we report on a preliminary evaluation of DMPTY, and in Sect. 6, we discuss related work and conclude.

## 2 Common elements of data management plans

In Britain, the Digital Curation Centre (DCC) "provides expert advice and practical help to anyone in UK higher education and research wanting to store, manage, protect and share digital research data" (see `http://www.dcc.ac.uk/about-us`). The DCC, for instance, has a good summary page that overviews and links to data management plan requirements of a number of British funding agencies [5]. The DCC has also published a checklist for devising a data plan [6]. The checklist seems to be an amalgamation of the various plans, and with its broad base takes into account requirements from different scientific fields such as the Natural Sciences or the Humanities. The checklist is divided into eight different parts, and covers all of the essential aspects for managing research data:

1. **Administrative Data:** nature of research project, research questions, purpose of data collection, existing data policies of funder or research institution;

2. **Data Collection:** type, format and volume of data; impact on data sharing and long-term access; existing data for re-use; standards and methodologies, quality assurance; data versioning;
3. **Documentation and Metadata:** information needed for the data to be read and interpreted in the future; details on documenting data acquisition; use of metadata standards;
4. **Ethics and Legal Compliance:** consent for data preservation and sharing; protection of personal data; handling of sensitive data; data ownership; data license;
5. **Storage and Backup:** redundancy of storage and backup; responsibilities; use of third party facilities; access control; safe data transfer;
6. **Selection and Preservation:** criteria for data selection; time and effort for data preparation; foreseeable research uses for the data, preservation timeframe; repository location and costs;
7. **Data Sharing:** identification of potential users; timeframe for making data accessible; use of persistent identifiers, data sharing via repositories and other mechanisms;
8. **Responsibilities and Resources:** for DMP implementation, review, and revision at plan and item level, potentially shared across research partners; use of external expertise; costs.

The Directorate-General for Research & Innovation of the European Commission has also published data management guidelines for Horizon 2020 projects [7]. Following the guidelines, the DMP must describe how research data is handled during *and* after the project; how the data will be collected, processed or generated; and what methodology and standards will be followed. It must say whether and how data will be shared (preferably open access), and how data will be curated and preserved. While the use of a DMP is "required for projects participating in the Open Research Data Pilot", other projects are "invited to submit a DMP if it is relevant to their planned research" [7].

An interesting aspect of the EC guidelines is the emphasis on the dynamics of research data management: "the DMP is not a fixed document, but evolves during the lifespan of the project" [7, page 5]. As a consequence, projects that submit a DMP do so multiple times: they must provide a first version of the DMP within the first six months of the projects' start, and later on, periodically update their DMPs during the projects' lifetime. Moreover, in Horizon 2020, all costs related to research data management can be accounted for.

## 3 Data Management in the CLARIN Infrastructure

The CLARIN shared distributed infrastructure aims at making language resources, technology and expertise available to the Humanities and Social Sciences research communities. To streamline the inclusion of new data and tools, and to help researchers with managing their data, CLARIN-D now offers advice on data management plans and supports their execution. The CLARIN-D plan template mirrors the structure of the DCC checklist, but has a number of adaptations to best profit from the CLARIN infrastructure. As a first step, researchers are invited to select the CLARIN-D centre whose expertise matches best the type of resource being created during the project. This aims at ensuring that researchers get the best possible advice from a CLARIN-D centre of their choice.[1] Following the plan template, researchers are asked to contact their CLARIN centre of choice when starting to devise their research data plan.

With regards to the DCC plan, our plan adjusts to the CLARIN infrastructure as follows:

**Data Collection:** a policy on preferred non-proprietary data formats for all types of language-related resources (in line with the CLARIN-D User Guide, see [8]).

**Documentation and Metadata:** the selection of existing CMDI-based metadata schemes, or if necessary, the adaptation of existing ones to best describe the research data.

**Ethics and Legal Compliance:** encouraging the use of the CLARIN License Category Calculator[2].

[1] For identifying the most suitable CLARIN-D centre, researchers can consult the link `http://www.clarin-d.net/de/aufbereiten/clarin-zentrum-finden`.

[2] Available online at `https://www.clarin.eu/content/clarin-license-category-calculator`

Note that, while CLARIN encourages open access to research data, it is not imposed by the DMP.

**Responsibilities and Resources:** a budget estimate that accounts for all the personnel and financial resources, and which are shared between data producer and CLARIN-D archive.

Moreover, there is a ninth plan component that describes a time schedule that data producer and data archivist agree on: when is the research data (i) described with all metadata, (ii) ingested in a data repository, and (iii) made accessible to interested parties or the general public? Moreover, it defines how long the research data should be held, and when, if applicable, it should be deleted. The data management plan shall also be complemented with a *precontractual agreement* between the data producer and the CLARIN centre for archiving, and which captures the rights and obligations of each partner.[3]

## 4 The DMPTY Wizard for the Generation of CLARIN-supported DMPs

DMPTY is a browser-based wizard available at the German CLARIN-D portal and encodes the CLARIN-D plan template.[4] The wizard makes use of the Javascript framework *AngularJS*, see `angularjs.org`, where each of the nine plan steps is presented to the user as an HTML form, and where navigation between the nine forms is easily possible so that information can be provided in flexible order (see Fig. 1). Associated with the forms is an HTML document that represents the contents of the data management plan template. Whenever the user enters information into one of the web form elements, the underlying plan is instantiated appropriately. At any given time (but within a browser session), it is possible to generate the plan as a text document in one of the formats Word, rtf and LaTeX, using the *pandoc* tool, see `pandoc.org`.

Researchers can then edit the document to enter additional information to address, for instance, institution-specific ethical or legal questions, or to state a cooperation with third parties, or to respond to specific requirements of funding agencies that we have not anticipated. Researchers may also want to add cross-links to relevant parts of the corresponding research proposal, change the formatting *etc*.

At the time of writing, a beta version of DMPTY is publicly available; the wizard generates plans in German only, and it only lists CLARIN-D centres as cooperation partners. Upon a successful evaluation, DMPTY will also be capable of generating plans in English, and listing all CLARIN centres as archiving partner. So far, the scope of DMPTY is restricted to its application within the CLARIN world.

We are currently preparing an evaluation of DMPTY and seek interested researchers to participate in our evaluation study. We also intend do make DMPTY available on an HTTPS-enabled server to protect the privacy and integrity of all data exchanged between the web browser and the server.

## 5 First Evaluation Results and Discussion

DMPTY has been demonstrated in depth to interested researchers at a tutorial on data management plans at the November 2015 meeting of the German Research Data Alliance in Potsdam. During the 90 minutes session the following questions were brought forward from the audience (answers by the second author):

***Question:*** Is the data provided by DMPTY users transmitted to the server, and how long is it kept there, and is the data safe?

***Answer:*** For the generation of the plan document in Word, LaTeX, or RTF format, user data is transmitted to the server, and temporarily stored on the server's `tmp` directory. For the time being, an unsecured HTTP connection is used.

[3] The precontractual agreement precedes the *deposition agreement* that is signed just before research data is deposited at the CLARIN centre. The content of the two agreements may substantially differ. The precontractual agreement is seen similar to a letter of intent; and intentions may shift given the outcome of the research grant application, the reviewers' comments on the proposal, and the many unforeseen events that may happen during a research project. As a result, assumptions about the research data (*e.g.*, formats, methodologies, access policies) that were held at the time of the precontractual agreement may need to be revised at the depositing stage, and hence require a rewording in the deposition agreement.

[4] See `http://www.clarin-d.net/de/aufbereiten/datenmanagementplan-entwickeln`.


Figure 1: Screenshot of the DMPTY entry page. In the upper part, all nine parts of the plan template can be accessed; the lower part shows elements of the selected first part of the plan template.

The confidentiality issue is a serious one, as some researchers may fear to expose their grant application to an untrusted third party. Users asked if it were possible to develop and make available a stand-alone version of DMPTY that users can install on their local machine.

One attendee suggested to define a sophisticated Microsoft Word or LaTeX template that users could fill out using their favoured word processing software, without fearing Browser-related session time-outs or data breaches (data loss or theft).

Another attendee suggested more advanced but desktop-based software, where data management plans are devised by starting in terms of the data that is being used or created during a research project. Such software should be less text-driven; rather, a visual programming language should be defined where users create data management plans by manipulating plan elements graphically, following the visual programming paradigm.

***Question:*** How about the completeness and adequateness of the plan template across different disciplines (in light of our previous discussion of different requirements for DMPs across European

funding agencies and research institutions)?

***Answer:*** DMPTY's plan management template largely draws upon the DCC checklist, which we described in Sect. 2, and which seems to be an amalgamation of the various data management plan templates provided by different UK funding agencies. As such, it takes into account requirements from different scientific fields. Our adaptation of the template for the CLARIN infrastructure is seen both as advantage and drawback, depending on whether potential users are part of the CLARIN infrastructure, or not.

We have seen that British funding organisations offer detailed checklists for writing data management plans. The situation in Germany is rather less developed; the German Research Foundation and the Federal Ministry of Education and Research offer little guidance in this respect. Here, the use of DMPTY gives researchers a hands-on approach to become aware of all the different aspects of data management in a research project, but completeness and adequateness issues must be addressed between scientists, their research institutions and their funding organisations, and by taking into account the nature of the research proposals in question.

***(Rhetorical) Question:*** What is the purpose of data management plans, other than securing funding?

***Answer:*** While it is hard to plan a three to five-year research project, it is also hard to carry it out as planned. The same holds for data management plans. For the research aspect, many funding agencies require researchers to submit regular progress reports (in Germany, often mid-term and final progress reports). Should the reporting be extended to include the data management aspects of research projects?

Clearly, a data management plan should be more than just a "letter of intent" that is not acted upon once funding has been secured. The plan should be regularly consulted, its steps executed, and the execution monitored. In DMPTY, the plan includes steps where interactions with third parties (the CLARIN centres) are necessary at given points in time. This puts some pressure on researchers with regards to plan execution and monitoring. Time will tell whether funding agencies will ask grant holders to complement their research reports with the addressing of data management issues.

In this respect, DMPTY (and other data management plan wizards) might develop into more sophisticated planning and plan execution systems, supporting researchers in devising a plan, but also executing it, and when necessary, in adapting the plan given changing requirements. DMPTY may hence develop into some sort of project management software that helps managing the entire life cycle of a project's research data.

It must be noted that none of the twenty-plus participants of the tutorial session had any experience with writing data management plans. Some reported having experimented with a number of DMP tools (such as DMPTY), in part, to prepare for the tutorial. Overall, participants seemed to appreciate the importance of addressing data management issues in a systematic manner. Also, most participants seemed to wish receiving more guidance from their funding organsations, but this may be due to the tutorial's German audience, and thus specific to the German research sector.

## 6 Related Work and Conclusion

With data management plans becoming increasingly necessary to attract funding, there are now a number of tools available that help researchers to address all relevant data management issues. The DCC provides a web-based wizard (DMPOnline) to help researchers devising data management plans, see [9]. Once researchers have selected a funding organisation for their research proposal, and optionally, their home research organisation, a corresponding data management plan template is created, which the wizard then follows step by step. DMPOnline users can thus devise a DMP that adheres to both funder and institutional requirements. The DMPOnline software is written in Ruby on Rails (see `rubyonrails.org`), is open source, and is available on Github for download, see [10]. The software's design facilitates the adaption of DMPOnline to cater for institution-specific customization (*e.g.*, customized logo, guidance,

and boilerplate text); in fact, over a dozen of UK institutions have already customized DMPOnline to better fit their needs.

The second DINI/nestor workshop was devoted to data management plans [11]. The workshop's programme featured, among others, a talk from the German Research Foundation on policy issues, several presentations on the nature and benefits of data management plans, and also talks on examples of successful data management, *e.g.*, for biological data and earth system science data. During the workshop, there were also a number of tools presented: the DMP Webtool from the University of Bielefeld, see [13], and the TUB-DMP tool of the Technical University of Berlin, see [14]. Both tools offer a plan template that largely draws from the WissGrid checklist [12]; also, both tools support the Horizon 2020 checklist [7]. Both tools are in house developments that aim at connecting data management planning with the existing research infrastructures of the respective institutions. The DMP Webtool is based upon the content management system Drupal (see `drupal.org`); the TUB-DMP tool of the TU Berlin is developed in php (see `php.net`).

With a considerable number of different initiatives in the area of data management plans, researchers may be confronted with an increasing number of requirements for a "proper" management of research data. There is a danger that requirements defined by the researcher's home institution may conflict with those of the funding organisation, and that both require the use of respective DMP wizards. When research data is deposited at a third institution, say a CLARIN centre, a potential third wizard (DMPTY) enters the scene. Here, however, we anticipate that requirements for data management plans will stabilize across research institutions, funding organisations, other parties involved, and also across disciplines. Also, a single DMP tool will need to be able to process multiple templates and guidelines from different research institutions and funding organisations. The DMP Online Tool of the DCC already follows this direction, and we anticipate a market consolidation where few, feature-rich tools will survive. DMPTY's design makes it easy to adapt a given plan template, or to include secondary ones. Also, DMPTY's design anticipates a follow-up editing phase, where researchers use their favoured text processing software to adapt the resulting DMP for their needs.

**Conclusion.** On a grand scale, managing research data poses significant challenges for research infrastructures, see [15]. On the individual level, researchers face additional work. With data management plans becoming an accepted part of good scientific practice, researchers must take into account all questions concerning their research data at an early stage of their research projects. Clearly, specifying and executing data management plans consumes resources, but the investment will pay off. DMPTY lowers the burden for researchers to develop their own plan, it guides them through all relevant aspects of such plans, and helps streamlining the cooperation with the CLARIN infrastructure. With CLARIN involved, researchers get support for the management of their data during the data's entire life cycle; touching base at regular intervals with CLARIN guarantees that the plans are up to their needs and properly executed. It also ensures that the appropriate resources (personnel, equipment) are accounted for. As a result, the number of high-quality accessible research data is bound to increase, which makes it easier for researchers to reap the benefits of sustainable data archiving and data re-use.

**Acknowledgements.** In November 2015, the authors co-organised a tutorial on data management plans, which was held during the German RDA meeting in Potsdam. We would like to thank all participants for their valuable feedback on DMPTY, and their comments on research data management in general. Also, we would like to thank the anonymous referees for their comments, which helped improve this paper considerably.

## References


[1] Digital Curation Centre (DCC). Funders' data plan requirements. See http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements.

[2] Guidelines at the Swiss National Science Foundation. See http://www.snf.ch/sitecollectiondocuments/allg_reglement_e.pdf, Article 44(b).

[3] Guidelines for Research Integrity at the ETH Zurich. See https://www.ethz.ch/content/dam/ethz/main/research/pdf/forschungsethik/Broschure.pdf.

[4] Research Data Management Policy at the University of Edinburgh. See http://www.ed.ac.uk/schools-departments/information-services/research-support/data-management/data-management-planning.

[5] Funder Requirements at the Digital Curation Centre (DCC). See http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements.

[6] Checklist for a Data Management Plan. v.4.0, Edinburgh: Digital Curation Centre, 2013. Available online: http://www.dcc.ac.uk/resources/data-management-plans.

[7] Guidelines on Data Management in Horizon 2020 (Version 2.1). European Commission, Directorate-General for Research & Innovation. See http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, February 2016.

[8] CLARIN-D User Guide, v.1.01. See http://www.clarin-d.net/de/sprachressourcen-und-dienste/benutzerhandbuch

[9] Digital Curation Centre. DMPOnline, a tool to devise data management plans. See https://dmponline.dcc.ac.uk.

[10] DMPonline_v4. Software maintained at https://github.com/DigitalCurationCentre/DMPonline_v4.

[11] Second DINI/nestor Workshop, Berlin, 2015. See http://www.forschungsdaten.org/index.php/DINI-nestor-WS2.

[12] J. Ludwig and H. Enke (Eds.) Leitfaden zum Forschungsdaten-Management. Verlag Werner Hülsbusch, Glückstadt, 2013. See http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf.

[13] The DMP Webtool, University of Bielefeld, see https://data.uni-bielefeld.de/de/data-management-plan.

[14] The TUB-DMP Tool, Technical University of Berlin, see https://www.szf.tu-berlin.de/menue/dienste_tools/datenmanagementplan_tub_dmp

[15] B. Almas *et al.*, Data Management Trends, Principles and Components - What Needs to be Done Next? V6.1. EUDAT, 2015. See http://hdl.handle.net/11304/f638f422-f619-11e4-ac7e-860aa0063d1f.


*All links were accessed on March 01, 2016.*

# How Can Big Data Help Us Study Rhetorical History?

**Jon Viklund**
Department of Literature
Uppsala University, Sweden
jon.viklund@littvet.uu.se

**Lars Borin**
Språkbanken/Dept. of Swedish
University of Gothenburg, Sweden
lars.borin@svenska.gu.se

## Abstract

Rhetorical history is traditionally studied through rhetorical treatises or selected rhetorical practices, for example the speeches of major orators. Although valuable sources, these do not give us the answers to all our questions. Indeed, focus on a few canonical works or the major historical key figures might even lead us to reproduce cultural self-identifications and false generalizations. However, thanks to increasing availability of relevant digitized texts, we are now at a point where it is possible to see how new research questions can be formulated – and how old research questions can be addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small samples, but where a methodology has not yet established itself. The aim of this paper is twofold: (1) We wish to demonstrate the usefulness of large-scale corpus studies ("text mining") in the field of rhetorical history, and hopefully point to some interesting research problems and how they can be analyzed using "big-data" methods. (2) In doing this, we also aim to make a contribution to method development in e-science for the humanities and social sciences, and in particular in the framework of CLARIN.

## 1 Introduction

### 1.1 Background

Like many other scientific disciplines, the humanities and social sciences (HSS) are now entering the age of big data. The modern digital world and the mass digitization of historical documents together provide unprecedented opportunities to all research fields relying on text as primary research data. This includes almost all HSS disciplines. In fact, we are now at a point where it is possible to see how new HSS research questions can be formulated – and how old research questions can be addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small samples, but where a methodology has not yet established itself. As larger amounts of texts are digitized and made searchable, we are able to see and investigate abstract patterns in large text masses that produce, partly, a new kind of knowledge.

In 2010, a team of American researchers published a study based on the enormous amount of texts made public by Google, a corpus of over 5 million digitized books. They named the new field of study *Culturomics* (Michel et al., 2011; Aiden and Michel, 2013), since it purported to uncover cultural and linguistic developments over time by large-scale computational processing of words, the basic building blocks of texts, in a way analogous to the methodology developed in the biomedical-informatics subfields of *genomics* and *proteomics*.

However, this development comes with considerable methodological challenges. The studies published so far are both tantalizing and disappointing. One can hardly deny their potential, but most often they are arguably more in the way of proof-of-concept showcases than seriously intended HSS research efforts, and have in fact been conducted primarily by non-HSS researchers (e.g., physicists and computer scientists). The field has been characterized by a conspicuous lack of "deep" research questions. The results of these studies affirm what we already knew. They argue for the advantages of different kind of

"text mining" in large collections of data, but so far they have not produced significant results that answer important research questions in the different fields of the humanities.

Culturomics and similar high-profile big-data initiatives drawing on vast amounts of digitized text have been criticized – justifiedly so, in our opinion – both for the low level of linguistic sophistication of the basic text processing underlying them (e.g., Zimmer 2013; see also Tahmasebi et al. 2015), and for the questionable suitability, in terms of representativity, of the investigated data sets for the particular research questions asked and claims made in this body of work (e.g., Pechenick et al. 2015). Thus, critics have pointed out that the bulk of this work reveals a quite simplistic "folk-linguistic" notion of language on the part of its authors, quite far removed from how the various branches of modern linguistics construe their object of study. The practitioners of culturomics have also so far shown little awareness of the issue of validity of their big data, one which humanists and social scientists have been contending with for a long time (Franzosi, 1987).

Having said this, however, we hasten to add that the basic premise of culturomics and similar big-data HSS initiatives appears to us both sound and exciting, and capable of bringing a new dimension to HSS research. The way to overcome the methodological obstacles mentioned above may simply be a matter of arranging for – in the words of Zimmer (2013) – "better communication between disciplines that previously had little to do with each other". This is exactly where CLARIN fits into the picture, as a provider of an e-science infrastructure allowing for linguistically sophisticated large-scale text processing, explicitly directed towards addressing HSS research questions, an infrastructure which crucially includes "experts", language-technology researchers who in dialogue with HSS scholars support the latter in their use of the infrastructure..

### 1.2 Towards HSS e-science – a concrete case study utilizing big textual data

The aim of the work presented here is twofold.

**Firstly**, in this paper we will try to demonstrate the usefulness of large-scale computational methods in the field of rhetorical history, and hopefully point to some interesting research problems and how they can be analyzed. In rhetorical studies, and more so in the field of rhetorical history, these quantitative, statistical methods have barely been tested. Our main historical interest lies in Swedish 19th century rhetorical culture, and we will demonstrate how large-scale quantitative analysis of a suitable text material can be used to learn more about ideas of and attitudes towards eloquence in this period.

These issues are not easily investigated through traditional research perspectives used by historians of rhetoric. Usually the materials under investigation in such research are either rhetorical treatises, or selected rhetorical practices, for example the speeches of major orators. These are valuable sources, but they represent only a minor part of all relevant historical documents, and, consequently, they give us but a limited view of the historical period at hand. In addition since these texts conform to the traditional material for rhetorical historiography, we also tend to pose traditional research questions, which often render predictable answers. In the worst case focus on a few canonical works or the major historical key figures might even lead us to reproduce cultural self-identifications and false generalizations (Malm, 2014).

Take the idea of "the death of rhetoric", repeated in handbooks for many years. This narrative is easily confirmed following the canonical texts of, e.g., enlightenment philosophers and authors from the romantic period. It has led many rhetorical scholars to conclude that the discipline of rhetoric in some sense really was dead in the 18th and 19th century, which of course is untrue (Fischer, 2013). With a large-scale textual base, we can see beyond this old master narrative, and find historical trends pointing to other narrative paths in history.

**Secondly**, we aim to make a methodological contribution to HSS e-science, and concretely to the research infrastructure for HSS being developed within CLARIN ERIC.

Following Tangherlini (2013, 8), the four logical stages of HSS e-research can be characterized as: (1) collection and archiving; (2) indexing and classification; (3) visualization and navigation; and (4) analysis. In our ongoing work, we rely on existing digitized text collections for the first stage, and stage 4 is the realm of the HSS scholar, of course. Stage 2 is largely the domain of text-mining and

computational-linguistic analysis and annotation – a central concern for CLARIN, and fortunately also a very lively research area in its own right.

In the present phase of our work, the focus is on the third stage which is often underdeveloped in digital humanities projects (Warwick et al., 2008, 99f), despite the fact that humanities researchers "need to be able to orient themselves digitally as well as they can in a physical library (including being able to estimate the total size of a digital resource)" (Burrows, 2013, 578). The development of methodologies and accompanying interactive tools for visualization and navigation which expand the "exploratory space" between stage 2 and 3 is where we aim to make a concrete methodological contribution. There are ample indications that data visualization and visual analytics have an extremely important role to play here (e.g., Havre et al. 2000; Smith 2002; Allen et al. 2007; Lee 2007; Schilit and Kolak 2008; Keim et al. 2010; Chuang et al. 2012b; Chuang et al. 2012a; Hirschmann et al. 2012; Broadwell and Tangherlini 2012; Chen et al. 2012; Krstajić et al. 2012; Oelke et al. 2012; Oelke et al. 2013), but also that the involvement of the end-users – HSS researchers – in the design process is crucial for acceptance of the final solution (e.g., Ramage et al. 2009; Chuang et al. 2012b; Rohrdantz et al. 2012). The present study is part of such an initiative, where a rhetorical scholar (Viklund) is working together with a researcher in natural language processing (Borin) and an e-science infrastructure unit (SWE-CLARIN/Språkbanken) on designing, developing and evaluating language-technology based e-science tools for HSS.

The traditional methodology in HSS is qualitative, corresponding to what literary scholars sometimes refer to as "close reading", and most HSS communities are still uncomfortable with large-scale quantitative approaches imported from language technology (LT) and text mining (Gooding, 2013). In this connection it is highly relevant that initiatives like culturomics and, in literary studies, "distant reading"/"macroanalysis" (Moretti, 2005; Moretti, 2013; Jockers, 2013) have so far generally not heeded Shneiderman's (1998, 523) well-known "visual information-seeking mantra": "Overview first, zoom and filter, then details on demand". These initiatives have tended to emphasize the bird's-eye aspect and have generally not provided the means of referring back to and studying individual text passages at close range. We believe that real progress in big-data HSS will only be forthcoming using methodology which combines distant and close reading, quantitative and qualitative research, and allows the researcher to move effortlessly between the two modes of enquiry (e.g., Schöch 2013).

Importantly, the actual mechanisms realizing this methodological point must work both ways. The observed broader statistical regularities must be translated into concrete and detailed research questions, but results of investigating the latter must also be made to inform the basis for future quantitative analysis. We consider it a major advancement in HSS research methodology, if qualitative close-reading methods can be applied to texts and text passages selected not on the basis of convenience, chance or tradition, but by mining very large text collections using explicit and reproducible operationalizations of principled criteria, utilizing information contributed by statistical and linguistic analysis tools and text mining tools, and where the tools are successively attuned to and informed by the research results.

The present work represents a first step in this direction, whereby a corpus infrastructure designed according to the principles described above, but aimed specifically at linguistic research, is pressed into service as a tool for introducing big-data methodology to the study of the history of rhetoric, in the hope that we will learn something new about the history of rhetoric, as well as learn more about which kind of digital tools could be useful for studying it effectively in very large volumes of text.

## 2 The research question: *doxa* in everyday discourse

Obviously, some questions are better suited than others for text mining. For example, we believe that it might be productive to raise issues that concern expressions of *doxa* (roughly: belief and opinion) in everyday discourse: What did people talk about, and believe was true, good, or beautiful? For instance, what were people's attitudes towards *eloquence* as a cultural phenomenon? How did people talk about eloquence, what ideas where brought forwards in relation to rhetoric, what kind of stereotypes were used, and how were they transformed?

Previous work has showed the cultural importance of of eloquence in Sweden up until the 18th century, as well as the centrality of rhetorical theory in schooling, aesthetics, sermons, political discourse etc.

(e.g., Johannesson 2005). There is a general conception that the importance of rhetoric as a discipline diminishes during the 19th century, only to increase again in our days. However, we have no significant studies that confirm the image of a general demise of rhetoric or eloquence in terms of people's attitudes towards these issues, and probably we need to revise our view also in relation to more specific materials, such as rhetorical manuals (Viklund, 2013). As demonstrated by Fischer (2013) in his study of the status of rhetoric and eloquence in the 18th century debate, the talk of rhetoric's "demise" or "death" in the period is based on an anachronistic view; one didn't conceive of rhetoric in those terms. The status of rhetoric can therefore best be evaluated through studies of the manner in which one talked about eloquence and rhetorical issues in general.

These issues are better investigated taking as the point of departure the large mass of everyday public discourse than through these singular and exceptional books and speeches. The literary scholar Franco Moretti (2013) calls these computational analyses of large masses of texts "distant reading", as opposed to "close reading", generally used in the humanities (see also Moretti 2005; Jockers 2013). Through abstraction and reduction, these quantitative analyses reveal patterns that only emerge from a distance. Of course, the results are not meaningful in themselves. The data are not explanations; these you need to supply yourself. However, what the big-data infrastructure brings to the research by these statistical readings of predominantly yet unread texts, is a tool for understanding historical transformations in a new way. What we don't know about the transformation of rhetoric is precisely the *gradual changes* of opinions in everyday discourse, the mindset of people toward an issue and how it changed.

From about 1840 we see the start of the democratization of politics and public life in Sweden, a process that culminates 1921 in the first election with universal suffrage. During these decades we have the enactment of a number of parliamentary reforms, social movements are changing the public agenda, and new groups of people are entering the public stage, moving the constraints of public debate. The long-term goal of the work in which the present study forms a part is to study the rhetorical formation of public debate on politics and social issues in this time of change. How were political issues talked about, and how did this discourse mutate over time? How did the key concepts of the debates change, and how did the rhetorical framing of these concepts change?

In the history of rhetoric studies – whether they concern rhetorical practice or theory – arguments are generally based on specific examples, which are set in relation to general notions of the "rhetorical tradition". With the large amounts of texts now at our disposal, we can now begin at the other end: to induce patterns of discourses from a vast material that can serve as starting points for much more systematic descriptions of rhetorical practices, as well as analyses of attitudes displayed in the rhetoric of the debate.

From a historical point of view, in order to understand rhetorical practices of a specific period, we need to know about doxa, about opinions and values, social cognitions formulated in language and practiced by groups of individuals in society (Amossy, 2002; Rosengren, 2002). From this angle, argumentation in public debate can be studied in terms of *topoi*, commonplaces or argumentative themes that reflect a system of public knowledge and thereby support the argument (Anscombre, 1995; Angenot, 1982). These topoi, based on values and general opinions, change over time and are difficult to describe systematically. One of our aims is to develop methods for such a systematic diachronic study.

A revision of 19th century rhetorical historiography is long overdue. Scholars have studied rhetorical performance of the period in relation to major authors (e.g., Johannesson et al. 1987; Viklund 2004) or to the specific rhetorical practices of the social movements of late 19th and early 20th century (e.g., Josephson 1991; Mral 1993).

As for the 18th century, scholars have studied public debate and political rhetoric (e.g., Skuncke 1999; Skuncke 2004; Öhrberg 2001; Öhrberg 2010). In addition, a newly finished research project on the attitudes to rhetoric in 18th century newspapers and periodicals (Fischer, 2013; Öhrberg, 2014), and rhetorical practices in the so called cultures of politeness (Öhrberg, 2011), has demonstrated the fruitfulness of meta-rhetorical studies. The status of rhetoric between late 18th and mid 20th century has been considered low, but few empirical studies have been made to support this claim. Lately Viklund (2013)

has begun to revise this simplified historiography, in a study that proposes that there is a renaissance of elocutionary rhetoric parallel to the process of democratization in Sweden.

## 3 Methodology: Towards a big-data infrastructure for HSS

The digital infrastructure used for this investigation is an advanced corpus search tool called *Korp* (Borin et al., 2012).[1] It is developed and maintained by Språkbanken (the Swedish Language Bank) at the University of Gothenburg. Språkbanken is a national language technology research and infrastructure development center, and the coordinating node of SWE-CLARIN, the national Swedish CLARIN ERIC organization,[2] and Korp is a central component of the Swedish CLARIN infrastructure. Its development started in 2010, drawing on several decades of experience in collecting and processing Swedish text corpora, and making them available for researchers and the public. Korp is a mature corpus infrastructure with modular design and an attractive and flexible web user interface, which is also used by other national CLARIN consortia.[3] Notably, a guiding principle for its design has been exactly the requirement that the user be able to move at will between high-level and abstract overview visualizations and individual data points.

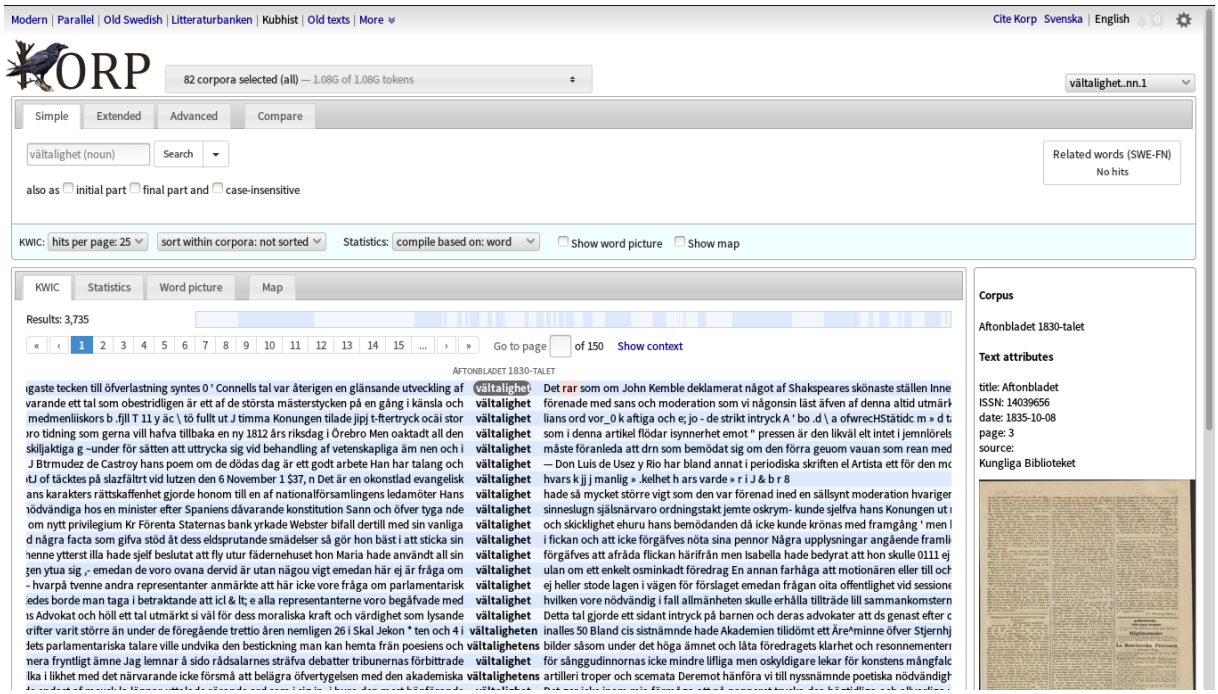



Figure 1: Korp: KWIC view for the lemma *vältalighet* 'eloquence' in Språkbanken's 19th century newspaper corpus (∼1 billion words)

Språkbanken offers access to a large amount of annotated corpora[4] and Korp offers the opportunity to make simple word searches as well as more complicated combined searches utilizing the automatic linguistic annotations present in the corpora.[5]

The results are presented in three different result views: as a list of hits with context (***k**eyword **i**n **c**ontext*: KWIC; see Figure 1); as statistical data with relative and absolute occurrence frequencies in

[1] See <http://spraakbanken.gu.se/korp/#?lang=en>.

[2] See <http://sweclarin.se>.

[3] Korp is used at least in Finland <https://korp.csc.fi>, in Estonia <https://korp.keeleressursid.ee>, and in Norway <http://gtweb.uit.no/korp/> (for the Sami languages).

[4] At the time of writing, the corpora searchable through Korp amount to over 10 billion words, out of which about 1 billion words are historical texts.

[5] Most of the corpora have annotations for part of speech, lemma, and dependency syntax. There is actively ongoing research at Språkbanken on extending and improving the annotations. For the experiments presented here, the lemma annotations have been used, and the dependency-syntax annotations are the basis for Korp's "word picture".

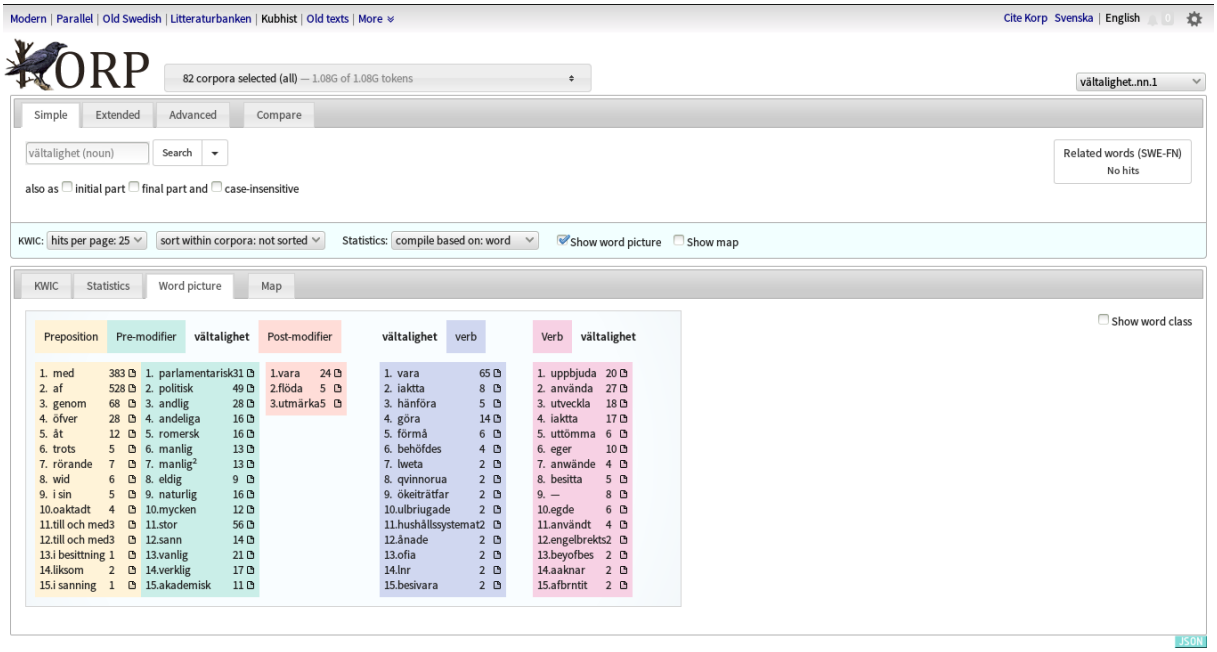



Figure 2: Korp: Word picture for the lemma *vältalighet* ‘eloquence’ in Språkbanken’s 19th century newspaper corpus (∼1 billion words)

subcorpora, which for example give you the opportunity to create a trend graph – relative frequency plotted over time – for one or several words or lemmas (see Figures 3–5 below); and, thirdly, as a “word picture”, which shows the most typical fillers of selected syntactic dependency relations of a word (most typical subjects and objects of a verb, most typical adjectival modifiers of a noun, etc.; see Figure 2). In line with the general interface design principles referred to earlier, the Korp user can move freely between the more comprehensive views and the KWIC view. Thus, clicking a data point on the trend graph, or the document symbol in one of the word picture items, will open a new KWIC view showing the corresponding search hits and their contexts. In the case of non-copyrighted material – e.g., the historical press texts used here – the KWIC view context can be expanded to a longer text passage. For this corpus and some other historical corpora, there is also a link to the digitized page image (see Figure 1).

Although originally devised for the purposes of linguistic analysis of texts, the word picture can be used as a kind of abstract topical maps that guide you to closer readings of the corpus. The corpus used in this study is a collection of historical newspapers from the late 18th to early 20th century, digitized by the Swedish Royal Library. The total corpus contains about one billion words, or almost 70 million sentences. On the one hand this is small in comparison with the Google Books dataset, but, as already mentioned, our corpus is annotated with linguistic information, including lemmatization made using high-quality Swedish lexical resources (modern and historical), which goes a long way towards compensating for the smaller size of the corpus by providing much greater accuracy (Borin and Johansson, 2014; Tahmasebi et al., 2015). Notably, however, and importantly, it is still far larger – by orders of magnitude – than any material previously used for studying the development of Swedish public discourse during this period.[6]

[6]For example, studying the historical linguistic development of the texts in the early Swedish Social Democratic press, Ledin (1995) works with a sample comprising less than 0.5% of the issues published during the 21-year period he examines, and Byrman (1998; 2001) samples short news items from less than 0.03% (issues) of the investigated newspapers for her study of the diachronic development of the language of short news items and public notices. Similarly, the corpus used by Lagerholm (1999) for his investigation of “orality in writing”, is made up of 16,000-word samples collected at 50-year intervals over a period of 200 years, 64,000 words in total.

## 4 Preliminary findings

So how can this search tool help us to learn more about the history of rhetoric? As previously noted, big-data methods facilitate studies of historical transformations. When did certain words come into use, and when did they disappear? How does the interest in certain topics change over time? More particularly, the Korp infrastructure helps us to visualize certain trends in what people talked about in daily newspapers, which is of great value when studying the topics of rhetoric, eloquence and debate.

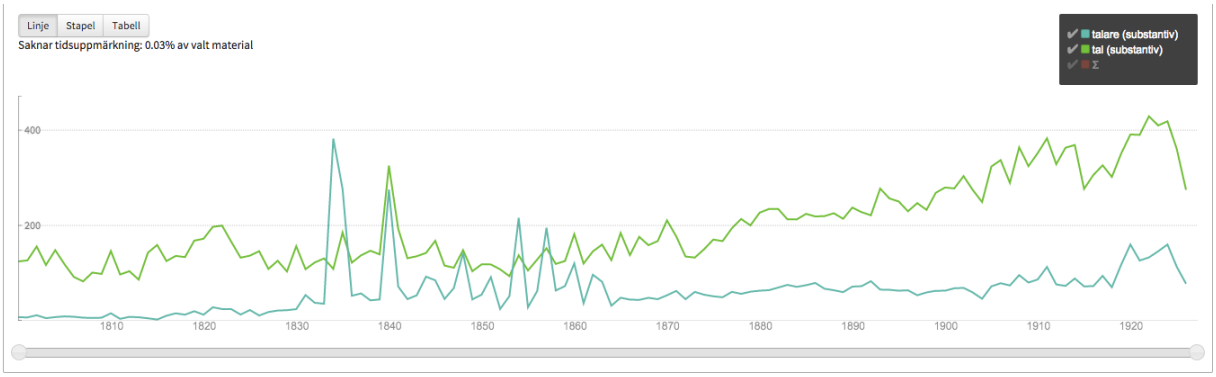



Figure 3: *Talare* 'speaker' and *tal* 'speech', 1800–1920s

The trend graphs in Figure 3 show the use of the words *talare* 'speaker' and *tal* 'speech' between 1800 and the 1920s in the Swedish press corpus. The nine peaks between the 1830s and 1860s coincide with the parliament sessions held every third year. In 1866 the old parliament where only the upper strata of society were represented, was dissolved, in favor of a new, more egalitarian parliamentary system. Apparently the newspapers were very keen on reporting the discussions in the old parliament, but less so thereafter, when the parliament met annually. The graph prompts a number of interesting questions: Why the sudden rise of interest in speakers around 1835, and why did the interest in the debates diminish? And still one notices a slow but steady increase of interest in the 'speaker' and in 'speech' in newspapers from the 1860s to the 1920s. Whereas the first peaks point to institutional changes and how they were reflected in press reports, it seems probable that the subsequent rise might suggest a more general interest in public debate and public opinion.

The testing of another hypothesis might support this assumption, one that seems plausible in light of what we know of the emergence of social movements in this period. We ought to be able to see a correlation between on the one hand democratization and the rising interest in politics, and, on the other, an increasing interest in rhetorical practices: oral performances and debate. As a simple way of testing this one might see to what degree people talked about 'politics' and 'democracy'. That is, through these search queries one can get an idea of the topical changes in the newspapers. The graphs in Figure 4 seem to confirm that there is an increasing interest in politics towards the end of the investigated period.

The next step would be to investigate some terms that we associate with rhetorical performance: *föreläsning*, *föredrag* 'lecture'; *tal* 'speech'; *declamation* 'declamation'; *debatt* 'debate'; *framförande* 'delivery, performance'; *agitation* 'agitation'; *anförande* 'speech'. The distribution in our corpus material of some of these is shown in Figure 5.

## 5 'Eloquence' in 19th century Swedish public discourse

All these results point to an increase in talk about rhetorical practices. As opposed to the words mentioned earlier that become markedly more frequent during the 19th century, for example *talare* 'speaker', *debatt* 'debate', and *deklamation* 'declamation', the word *vältalighet*, 'eloquence', has a relatively stable trend curve. One can simply surmise that it is a constant cultural concern all through the period. Navigating through a large part of the examples generated by the word picture function of Korp (see Figure 2) it was obvious that although the context in which the word was used did change to some extent, it was not worth investigating at this point. For example, the use of the word in a political context increased, but that

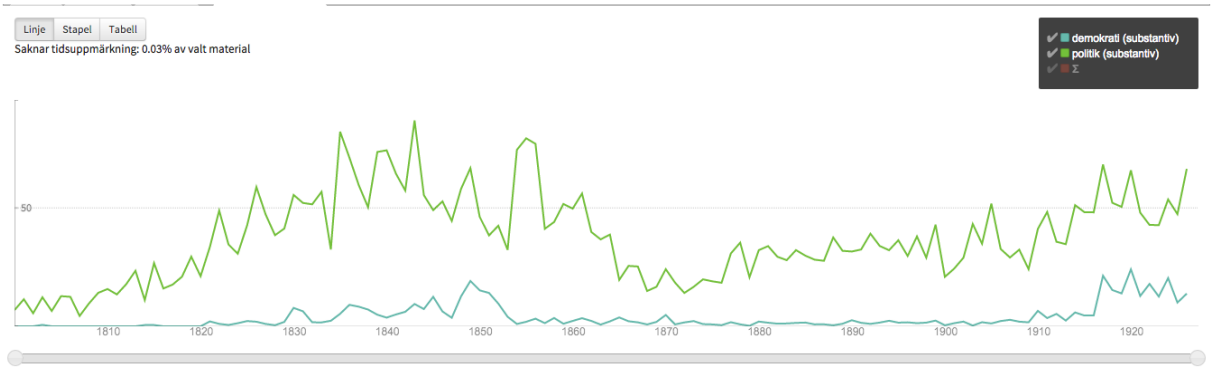



Figure 4: *Demokrati* ‘democracy’ and *politik* ‘politics’, 1800–1920s

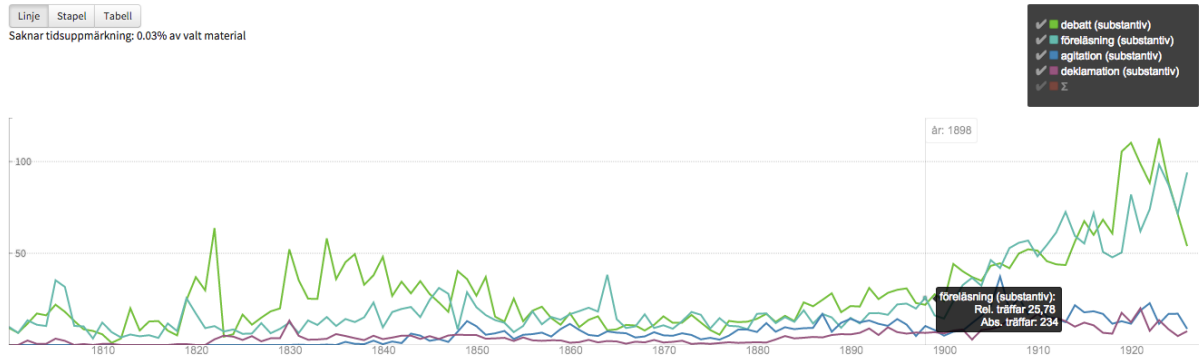



Figure 5: Terms associated with rhetorical performance, 1800–1920s

was only predicable, due to the process of democratization mentioned above. So initially, it seemed most productive to see what kind of generalizations one might make about the use of the word over the whole period, generalizations that could be the starting point for a study of the transformation of the concept of rhetoric in the 20th century. What kinds of attitudes toward *vältalighet* ‘eloquence’ are revealed in the newspaper texts?

The term “attitude” is used in a broad sense, as a manner of thinking about an object or phnomenon, often in terms of a positive or negative evaluation. In modern rhetorical criticism attitudes are often analyzed in relation to style and word choices (e.g., Burke 1969; Billig 1996): the way we use language reveals our attitudes toward a thing or person. Attitudes in this sense are not necessarily connected to a person’s stated intentions. In the discussion below, looking for attitudes is, on the one hand, a perspective that makes it easier to sort out central meanings in the use of the word ‘eloquence’, and, on the other, one that highlights a certain framing of the word.

So what kinds of attitudes toward eloquence and rhetoric are displayed in the material? The methodological gain with this broad starting point is that it does not exclude any new information the searches in Korp might produce. As indicated earlier the word picture function in Korp can be used as a kind of topical search tool: what words go together with this noun, *vältalighet* ‘eloquence’, and what do they reveal about the uses of the word? The searches produce for example the most common, or the most representative, modifiers, generally adjectives, and when studied in greater detail, one could observe a pattern of notions that describe certain qualities of ‘eloquence’. They can be divided into three main groups, plus an “other” category:

- **Genre:** ‘parliamentary’, ‘spiritual’, ‘political’, ‘Roman’, ‘academic’, ‘marital’, etc.
- **Truthfulness and naturalness:** ‘real’, ‘true’, ‘right’, etc.; ’manly’, ’natural’, ‘artless’, ‘unpretentious’, ‘simple’, etc.

- **Conceptual metaphors:** 'fiery', 'glowing', 'burning' (eloquence is fire); 'flowing', 'fluent', 'pouring' (eloquence is water)
- **Other:** 'mute', 'normal', 'mighty', 'great', 'irresistible', 'bold'

The categories retrieved from the word picture help us to better understand the linguistic contexts in which one talked about eloquence, as well as the attitude towards rhetoric evidenced in the corpus. The words prompt a number of interesting explorations. One might for example investigate positive vs. negative connotations of the concept of eloquence, or one might look into gender differences in relation to the various examples and the categories produced by the computational reading.

We found the conceptual metaphors interesting. It was surprising to see that the majority of metaphors connected to the notion of 'eloquence' so clearly could be divided in two main categories: either *eloquence is fire* or *eloquence is water*. In these kind of metaphors, described first by cognitive linguists, one conceptual domain – the target domain, here *eloquence* – is mapped onto another – the source domain, here *fire* and *water*. These concepts are often analyzed in terms of image schemas, which, in simplified terms can be said to be mental patterns that structure the metaphorical expression. Talking about eloquence as warmth, and as sparks, for example, demonstrates how the metaphor can be based on motor or sensory experience. And water, to add another example, is in these expressions often *flowing from one container to another*; so the expressions are orientational, and often they are emphasizing the quality of *fluidity* or *fluency*. (Lakoff and Johnson, 1980; Geeraerts, 2006) These image schemas can help us to understand what the metaphors are saying about the attitudes expressed in the examples.

Methodologically we here used all the context material retrieved in connection to the word *vältalighet* 'eloquence' and searched for the most common words in the conceptual metaphor clusters, and in that way ended up with many examples of these metaphorical expressions that were not necessarily associated with the specific word 'eloquence', but with the more general concept of 'rhetoric'.

So what can be learnt about how 'eloquence' was perceived from these clusters of words? After studying the many instances of these metaphors we found that they generated a deeper understanding of the attitudes toward 'eloquence' during the period under study, and we have summarized these insights into four points.

**1. Positive values:** A general knowledge retrieved from these expressions concerns which positive values are emphasized. The fire metaphors generally express images of force: pathos, burning hearts, passion, and energy.

> *Hans herravälde öfver sinnena måste man, såvida man icke är alltför intagen af fördomar, uteslutande tillskrifva hans lågande vältalighet, hans alltid slagfärdiga dialektik och särskildt det glansfulla sätt på hvilket han vet att försvara de demokratiska grundsatserna.*
>
> 'His mastery over the senses must, unless one is too captivated by prejudice, be ascribed exclusively to his fiery eloquence, his invariably witty dialectics, and especially the glittering manner in which he is able to defend the democratic tenets.'

The metaphors build on eloquence as a force of nature. Fire, heat, glow and lightning are not only natural phenomenon, they are also overpowering, i.e. they have the power to overwhelm the senses of the listener. Another positive quality is the association to genius:

> *[. . . ] parlamentarisk vältalighet. Såsom lyrisk skald lyser Béranger i synnerhet genom ingifvelsens eld och äkta originalitet [. . . ]*
>
> '[. . . ] parliamentary eloquence. As lyric poet Béranger shines especially through the fire of inspiration and genuine originality [. . . ]'

One should note that the conceptual metaphor *eloquence is fire* affirms the distinction between the two main modes of persuasion: with reason, *logos*, or with emotion, *pathos*, since it always expresses the force of the latter. One could say that these expressions reflect one mode of rhetorical proof while deflecting another. That dichotomy is, of course, generally only implicitly present in the examples, but

once in a while you can find it thematized in the texts, as in the example where someone is contrasting eloquence in Norway and Sweden: the heat of passion is contrasted with the calmer and colder nature of reason:

> *Norska tribunens vältalighet är af en lugnare natur i det man mer lägger an på att verka genom skäl än granna ord mera söker verka på förståndet än känslan hvars villfarelse man fruktar; denna tribun är icke derföre mindre mäktig och verksam. Om den ock ej disponerar tordönet och ljungelden så har den likväl grunder kallblodighet ståndaktighet och mod. Debatterna gå sällan utom den lugna diskussionens gränser men föredragen som nästan alltid äro muntliga och improviserade ersätta som oftast genom grundlighet och öfvertygande kraft hvad dem brister i liflighet och värma.*
>
> 'The eloquence of the Norwegian tribune is of a quieter nature in that one makes a point of rather acting by reason than gaudy words, more seeking to act on the mind than the feeling, the delusion of which is feared; this tribune is not therefore less powerful and effective. Even though it may not possess thunder and lightning, it is nonetheless grounded in coolness, steadfastness, and courage. The debates rarely exceed the limits of calm discussion, but the talks that almost always are oral and improvised, usually substitute what they lack in vividness and heat with thoroughness and persuasive force.'

**2. Attitudes toward gender:** The concept *eloquence is fire* also highlights attitudes toward gender. The fire metaphors are clearly coded as a male feature – there are no women described or speaking in this category. This is not surprising; force and genius are qualities that traditionally have been seen as male. But an awareness of the consistency is important; it would be interesting to see at what time in history this trend is broken. In the other metaphorical concept – *eloquence is water* – we have examples that refer to both men and women.

**3. Attitudes toward eloquence as an art:** The *eloquence is water* concept frames the attitudes to rhetoric in a way that it can be used either positively or negatively (ironically), as opposed to the *eloquence is fire* concept that almost always is used in an unambiguously positive sense. When eloquence is described as flowing, streaming etc. the semantic orientation has more to do with rhetorical ability than degree of pathos.

> *Hans klara och lätta diction flöt rikt ex tempore och blef jemväl full af eld när det behöfdes*
>
> 'His clear and easy diction flowed abundantly ex tempore and yet became full of fire when needed'
>
> *Alla dessa enskildheter flödade från timmermannens läppar i en ström af enkel vältalighet, hvilken inga lektioner i "uttalslära" hade kunnat föröka med ett grand af ytterligare effekt*
>
> 'All these particularities flowed from the carpenter's lips in a stream of simple eloquence, which no lessons in "pronunciation" could multiply with a mote of additional power'

Here the image schema describing the technical ability of delivery – diction – has to do with fluency. A person characterized with flowing eloquence has the rhetorical skill of speaking naturally without displaying too much art. Both thoughts and feelings come "streaming from the speaker", and this fluency can be a sign of natural ability – opposed to the art of rhetoric as in the latter example – or just a sign that one masters the art. For this reason, the metaphorical concept is often used negatively when there is a conflict between art and genuine thoughts and feelings:

> *Med glödande öfvertygelse, med prålande ord och flytande vältalighet framhålla de, nationernas representanter, var och en på sitt håll, sanningen sådan den bäst passer sig för dem*
>
> 'With glowing conviction, with pompous words and fluent eloquence, those representatives of the nation emphasize, each in their own way, the truth as they see fit'

*Han ägde en obeskriflig vältalighet, men jag tyckte nästan han talade med alltför stor lätthet – det flödade öfver som en flod*

'He possessed a tremendous eloquence, but I almost thought he was speaking with too much ease – it flowed over like a river'

*Men inte ens den mest ifriga anhängare af »saken», kunde i hettan uthärda 2 timmars flödande vältalighet från 20 talarstolar på en gång*

'But not even the most ardent supporter of the "cause", could in the heat endure two hours of flowing eloquence from 20 podiums at once'

**4. Attitudes toward eloquence: from one heart to another:** The most characteristic image schema of the two metaphorical concepts concerns the pattern *from one container to another*. A fire is burning *from* the soul, and a flame, a bolt of lightning or an electric spark is coming *from* the speaker to the audience and sets the listener's mind on fire. Likewise, the stream of eloquence is sometimes described as coming, with a cliché often used, from the heart of a speaker.

*Han älskade alltid att dröja vid denna stora tanke; och äfven nu sökte han med all sin brinnande vältalighet att inskrifva den outplånligt i sina åhörares hjertan.*

'He always loved to dwell on this great idea; and also this time he sought with all his fiery eloquence to write it indelibly into the hearts of his listeners.'

*Engelbrekts flammande vältalighet hade bland det mäktiga Söderköpings talrika borgerskap upptändt en eld som ännu glödde under askan [. . . ]*

'Engelbrecht's flaming eloquence had among the mighty and numerous burghers of Söderköping kindled a fire which still glowed under the ashes . . . '

*Vi voro dock nog djerfve att trotsa denna ordförandens uppmaning och stannade alltså , afvaktande om något af den visdom , som flödade öfver hans läppar, möjligen kunde tränga in i våra dumma hufvuden och vi derigenom sättas i stånd att begripa hvad en utgift på 2,000 kr [. . . ]*

'We were, however, bold enough to defy the chairperson's request and therefore stayed, awaiting to see if some of the wisdom that flowed from his lips, could possibly penetrate our stupid heads so that we could be able to comprehend what an expenditure of 2,000 kr . . . '

*Då Champagnen började flöda , då sprungo också alla snillets källor upp och den mest eldiga vältalighet slog i hvart ögonblick sin elektriska gnista i förvånade åhörares sinnen*

'When the champagne began to flow, then all the sources of genius also erupted and the most fiery eloquence struck, in every passing moment, its electric spark in the minds of the astonished listeners'

The last, more ironical instances of the metaphor seem to suggest that the writers have no problems turning the concept of the ideal orator upside down. But despite of this one can assume that the very nature of these orientational metaphors suggests a cognitive frame indicating that eloquence has a special status as a communicative tool. But for how long are these expressions used? Today this idea of communication still exists, but one would not primarily find these metaphors in connection with the words 'eloquence' or 'rhetoric'. More likely one would find them in sentimental discourses about love or friendship.

Once again, it would be interesting to investigate how these verbal conventions transform over time, because of course they do. Even though we do not yet have enough comparable data for most of the 20th century to show the details of the changes that have taken place, we have large amounts of evidence from the latest form of written public discourse, i.e., social media such as blogs and online discussion forums, represented by close to 8 billion words accessible through Korp in Språkbanken. Thus, we can compare today's attitudes towards rhetoric to those uncovered in the 19th century material.

Today, for one, *vältalighet* 'eloquence' is rarely used – the word *retorik* 'rhetoric' has taken its place – and both these words show quite different distributions in modern corpora, such as newspapers and

blogs from the last four decades, as compared to the 19th century material. Looking at the word pictures generated on basis of the modern material, we find almost exclusively two kinds of modifiers: (1) *type of rhetoric* (for example: 'political', 'feminist', 'religious', 'social democratic'); or (2) *negative qualities* (for example: 'empty', 'made-up', 'aggressive'); and of course the two categories combined (for example: 'racist', 'populist', 'scholastic').

## 6 Conclusions and future work

Above we have described some initial experiments where a very large historical corpus of Swedish newspaper text and the state-of-the-art corpus infrastructure of SWE-CLARIN have been brought to bear on research questions in the field of rhetorical history.

A central and important purpose of this work was to investigate the method itself: Does it produce new knowledge? Yes and no. For instance, the fact that two conceptual metaphors – *eloquence is fire* and *eloquence is water* – were so dominant in the material definitely adds to our knowledge of *doxa* during the 19th century. Most other results were expected, or at least not surprising. But even if the results only confirm old knowledge it is worthwhile: if the method reveals results that confirm old knowledge then it might be able to see also new things which until now have not been acknowledged.

The method is promising, and we see several natural directions in which this work can be continued. The material studied here covers the 19th and the beginning of the 20th century. As described above, we also did a preliminary study using 21st century social-media and news text. The two studies together indicate that there have been considerable changes in the attitudes toward eloquence and rhetoric as expressed in Swedish public discourse over the last two centuries, but the details of these changes (including their timing) remain beyond our ken for the time being. With a corpus that covered also the 20th century it would be possible to advance our knowledge on this score.

Finally, the present corpus infrastructure – intended mainly for linguistically oriented research – proved useful for the research questions that we have described above. However, the old adage about what happens when you have a hammer is certainly valid here: We did adapt our research questions so that they could be addressed using the linguistic annotations, search functions and visualizations that Korp makes available. These should be complemented by text processing and interfaces more geared towards supporting more general digital humanistic inquiry. In particular, to the form-oriented search useful to linguists we would like to add *content-oriented* search modes – based on information-retrieval or information-extraction techniques or content-classification technologies such as topic modelling, vector-space models, or word embeddings – accessible through interfaces that would still allow the user to move easily and effortlessly between various forms of macro view visualization ("distant reading") and individual instances ("close reading"). To repeat: This we believe to be a crucial – even necessary – feature of any such tool.

## Acknowledgements

The research described here has been supported in part by a framework grant from the Swedish Research Council (*Towards a knowledge-based culturomics* 2012–2016; project no 2012-5738). The basic research infrastructure development involved has been made possible by the Swedish Research Council's funding of SWE-CLARIN (2014–2018), the Swedish node of the CLARIN ERIC.

## References

Erez Aiden and Jean-Baptiste Michel. 2013. *Uncharted: Big data as a lens on human culture*. Riverhead Books, New York.

Robert B. Allen, Andrea Japzon, Palakorn Achananuparp, and Ki Jung Lee. 2007. A framework for text processing and supporting access to collections of digitized historical newspapers. In M. J. Smith and G. Salvendy, editors, *Human interface, Part II, HCII 2007*, number 4555 in LNCS, pages 235–244. Springer, Berlin.

Ruth Amossy. 2002. How to do things with doxa: Toward an analysis of argumentation in discourse. *Poetics Today*, 23(3):465–487.

Marc Angenot. 1982. *La parole pamphlétaire: Typologie des discours modernes*. Payot, Paris.

Jean-Claude Anscombre. 1995. *Théorie des topoï*. Kimé, Paris.

Michael Billig. 1996. *Arguing and thinking: A rhetorical approach to social psychology*. Cambridge University Press, Cambridge.

Lars Borin and Richard Johansson. 2014. Kulturomik: Att spana efter språkliga och kulturella förändringar i digitala textarkiv. In Jessica Parland-von Essen and Kenneth Nyberg, editors, *Historia i en digital värld*.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Peter M. Broadwell and Timothy R. Tangherlini. 2012. TrollFinder: Geo-semantic exploration of a very large corpus of Danish folklore. In *The Third Workshop on Computational Models of Narrative*, pages 50–57, Istanbul. ELRA.

Kenneth Burke. 1969. *A rhetoric of motives*. University of California Press, Berkeley.

Toby Burrows. 2013. A data-centred 'virtual laboratory' for the humanities: Designing the Australian Humanities Networked Infrastructure (HuNI) service. *Literary and Linguistic Computing*, 28(4):576–581.

Gunilla Byrman. 1998. Tidningsnotisen i förändring 1746–1997. Institutionen för nordiska språk, Lunds universitet. Svensk sakprosa, rapport nr 15.

Gunilla Byrman. 2001. Municipalstämma hölls igår i Tomelilla .... Svenskt notisspråk 1746–1997. In Björn Melander and Björn Olsson, editors, *Verklighetens texter. Sjutton fallstudier*, pages 443–483. Studentlitteratur, Lund.

Annie T. Chen, Ayoung Yoon, and Ryan Shaw. 2012. People, places and emotions: Visually representing historical context in oral testimonies. In *The Third Workshop on Computational Models of Narrative*, pages 45–49, Istanbul. ELRA.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012a. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*.

Jason Chuang, Daniel Ramage, Christopher D. Manning, and Jeffrey Heer. 2012b. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*.

Otto Fischer. 2013. *Mynt i Ciceros sopor. Retorikens och vältalighetens status i 1700-talets svenska diskussion*, volume 1 of *Södertörn Retoriska Studier*. Södertörns högskola, Huddinge.

Roberto Franzosi. 1987. The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 20(1):5–16.

Dirk Geeraerts, editor. 2006. *Cognitive linguistics: Basic readings*. De Gruyter, Berlin.

Paul Gooding. 2013. Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing*, 28(3):425–431.

S. Havre, B. Hetzler, and L. Nowell. 2000. ThemeRiver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*, pages 115–123, Salt Lake City.

Hagen Hirschmann, Anke Lüdeling, and Amir Zeldes. 2012. Measuring and coding language change: An evolving study in a multilayer corpus architecture. *ACM Journal on Computing and Cultural Heritage*, 5(1):article 4.

Matthew L. Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, Urbana/Chicago/Springfield.

Kurt Johannesson, Eric Johannesson, Björn Meidal, and Jan Stenkvist. 1987. *Heroer på offentlighetens scen. Politiker och publicister i Sverige 1809–1914*. Tidens förlag, Stockholm.

Kurt Johannesson. 2005. *Svensk retorik. Från medeltiden till våra dagar*. Norstedts, Stockholm.

Olle Josephson. 1991. *Diskussionsskolan 1886: Språkmiljö, argumentation och stil i tidig arbetarrörelse*. Number 1 in Arbetarrörelsen och språket. Avdelningen för retorik, Uppsala universitet, Uppsala.

Daniel A. Keim, Leishi Zhang, Miloš Krstajić, and Svenja Simon. 2010. Solving problems with visual analytics: Challenges and applications. *ACM Transactions on Embedded Computing Systems*, 4(4):article 39.

Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann, and Daniel A. Keim. 2012. Incremental visual text analytics of news story development. In *Proceedings of Conference on Visualization and Data Analysis (VDA '12)*.

Per Lagerholm. 1999. *Talspråk i skrift. Om muntlighetens utveckling i svensk sakprosa 1800–1997*. Number A 54 in Lundastudier i nordisk språkvetenskap. Lunds universitet, Institutionen för nordiska språk, Lund.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

Per Ledin. 1995. *Arbetarnes är denna tidning. Textförändringar i den tidiga socialdemokratiska pressen*. Number 20 in Acta Universitatis Stockholmiensis: Stockholm Studies in Scandinavian Philology, New Series. Almqvist & Wiksell International, Stockholm.

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 472–479, Prague. ACL.

Mats Malm. 2014. Digitala textarkiv och forskningsfrågor. In Jessica Parland-von Essen and Kenneth Nyberg, editors, *Historia i en digital värld*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, (331).

Franco Moretti. 2005. *Graphs, maps, trees: Abstract models for a literary history*. Verso, London/New York.

Franco Moretti. 2013. *Distant reading*. Verso, London/New York.

Brigitte Mral. 1993. *Kommunikation och handlande i Malmö kvinnliga diskusionsklubb 1900–1904*. Number 6 in Arbetarrörelsen och språket. Avdelningen för retorik, Uppsala universitet, Uppsala.

Daniela Oelke, Dimitrios Kokkinakis, and Mats Malm. 2012. Advanced visual analytics methods for literature analysis. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 35–44, Avignon. ACL.

Daniela Oelke, Dimitrios Kokkinakis, and Daniel A. Keim. 2013. Fingerprint matrices: Uncovering the dynamics of social networks in prose literature. *Computer Graphics Forum*, 32(3):371–380.

Ann Öhrberg. 2001. *Vittra fruntimmer. Författarroll och retorik hos frihetstidens kvinnliga författare*. Gidlunds, Hedemora.

Ann Öhrberg. 2010. "Fasa för all flärd, konstlan och förställning". Den ideala retorn inom 1700-talets nya offentlighet. *Samlaren*, 131.

Ann Öhrberg. 2011. Between the civic and the polite. Classical rhetoric, eloquence and gender in late eighteenth century Sweden. In Otto Fischer and Ann Öhrberg, editors, *Metamorphoses of Rhetoric. Classical Rhetoric in the Eighteenth Century*, number 3 in Studia Rhetorica Upsaliensia. Uppsala University, Uppsala.

Ann Öhrberg. 2014. *Samtalets retorik. Belevade kulturer, offentlig kommunikation och kön i svenskt 1700-tal*. Symposions förlag, Höör.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, 10(10):e0137041, 10.

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D. Manning, and Daniel A. McFarland. 2009. Topic modeling for the social sciences. In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada.

Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli, and Daniel A. Keim. 2012. The world's languages explorer: Visual analysis of language features in genealogical and areal contexts. *Computer Graphic Forum*, 31(3):935–944.

Mats Rosengren. 2002. *Doxologi. En essä om kunskap*. Rhetor förlag, Åstorp.

Bill N. Schilit and Okan Kolak. 2008. Exploring a digital library through key ideas. In *Proceedings of JCDL'08*, pages 177–186, Pittsburgh. ACM.

Christof Schöch. 2013. Big? Smart? Clean? Messy? Data in the humanities. *Journal of Digital Humanities*, 2(3). <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/>.

Ben Shneiderman. 1998. *Designing the user interface*. Addison-Wesley, Reading, Mass., 3rd ed edition.

Marie-Christine Skuncke. 1999. Den svenska demokratidebatten 1766–1772. In Rut Boström Andersson, editor, *Ordets makt och tankens frihet. Om språket som maktfaktor*. Uppsala universitet, Uppsala.

Marie-Christine Skuncke. 2004. Press and political culture in Sweden at the end of the Age of liberty. Enlightenment, revolution and the periodical press. In Hans-Jürgen Lüsebrink and Jeremy D. Popkin, editors, *SVEC 2004:06*. Voltaire Foundation, Oxford.

David A. Smith. 2002. Detecting and browsing events in unstructured text. In *Proceedings of SIGIR'02*, Tampere. ACM.

Nina Tahmasebi, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik Johansson, Richard Johansson, Mikael Kågebäck, Olof Mogren, Pierre Nugues, and Thomas Risse. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2–4):169–187.

Timothy R. Tangherlini. 2013. The folklore macroscope. Challenges for a computational folkloristics. *Western Folklore*, 72(1):7–27.

Jon Viklund. 2004. *Ett vidunder i sitt sekel. Retoriska studier i C.J.L. Almqvists kritiska prosa*. Gidlund, Hedemora.

Jon Viklund. 2013. Performance in an age of democratization: The rhetorical citizen and the transformation of elocutionary manuals in Sweden ca. 1840–1920. Paper presented at ISHR [International Society for the History of Rhetoric] biannual conference in Chicago.

Claire Warwick, Melissa Terras, Paul Huntington, and Nikoleta Pappa. 2008. If you build it will they come? The LAIRAH study: Quantifying the use of online resources in the arts and humanities statistical analysis of user log data. *Literary and Linguistic Computing*, 23(1):85–102.

Ben Zimmer. 2013. When physicists do linguistics. Is English 'cooling'? A scientific paper gets the cold shoulder. *Boston Globe*, February 10.

# Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions

**Tanja Wissik**
ACDH-OEAW
Vienna, Austria
tanja.wissik@oeaw.ac.at

**Matej Ďurčo**
ACDH-OEAW
Vienna, Austria
matej.durco@oeaw.ac.at

**Abstract**

In this paper we will present an institutional research data workflow model covering the whole lifecycle of the data and showcase the implementation of the model in a specific institutional context. We will present a case study from the Austrian Centre for Digital Humanities, a newly founded research institute for digital humanities of the Austrian Academy of Sciences, which also supports researchers in the humanities as service unit. The main challenge addressed is how to harmonize existing processes and systems in order to reach a clear division of roles and achieve a workable, sustainable workflow in dealing with research data.

## 1 Introduction[1]

Institutions like universities and academies have an increasing obligation to manage and share research data. For the majority of scholars these endeavours, especially in the humanities, are relatively new and not deeply integrated into their existing working practices: for example, only recently have funding bodies started to request a data management plan which follows open access policies for publications and research data as part of a project proposal[2]. Whereas the traditional non-digital research process consisted only of project planning, data acquisition and data analysis and the publication, in e-research, data sharing, data preservation and data reuse are added to the lifecycle (Briney, 2015).

However, recent studies (e.g. Bauer et al., 2015; Akers and Doty, 2013; Corti et al., 2014) found out, that sharing and reuse of research data is not yet always an integral part of good research practice and that researchers are not familiar with data management plans etc.

A survey carried out in Austria in 2015 (3016 questionnaires) showed significant variations in researchers' data management practice and needs: "Access to self-generated research data by third parties is usually allowed to a limited degree by researchers. While slightly more than half of the respondents stated they allowed access only on request, only one in ten provides their research data as open data for the public; the same number of researchers deny access altogether." (Bauer et al., 2015). The study also reported that 49% of the respondents would need help with project-specific research data management, e.g. creation of data management plan. In a survey study at Emory University in the USA, Akers and Doty (2013) found that "most (~82%) faculty researchers are only somewhat or not at all familiar with requirements for data management or data sharing plans" and "arts and humanities researchers are most likely to be completely unfamiliar with these funding agency requirements for data management plans." A study in the UK in 2008 showed a similar picture: "Only 37% of studied researchers shared their data with collaborators in their own circles and only 20% shared more widely outside of their own network." (Corti et al., 2014: 9). Most concerns about sharing data arise from a lack of knowledge on how to make digital research data sharable for the longer term and a lack of



[2] E.g. Austrian Research Fund (FWF), https://www.fwf.ac.at/en/research-funding/open-access-policy/ (accessed 28.12.2015).

suitable infrastructure in their own institutions. However, knowledge about data management, training and working infrastructure and services are part of successful research data management that has to be adopted by individual researchers as well as institutions like universities and academies. The survey in Austria (Bauer et al., 2015: 11) also found out that "[t]he majority of researchers desire technical infrastructure and project-specific support for research data management. In addition, more than one-third show interest in legal advice, a general help desk, as well as training programs."

This analysis is based on already existing lifecycle or workflow models, taking into account the existing working practice and institutional requirements. Therefore research workflows and the related data management processes vary not only from discipline to discipline but also from one institutional context to another. Once a workflow model is in place, it can serve also as a quality assurance mechanism.

In this paper we will present a case study from the Austrian Centre for Digital Humanities, a newly founded research institute for digital humanities of the Austrian Academy of Sciences, which also supports researchers in the humanities as service unit, where an institutional research data workflow model is being implemented based on already existing lifecycle models.

The context-specific challenge for this undertaking was to bring all the stakeholders together in order to create a model which can simultaneously meet the unique needs of the various sub-disciplines, departments and researchers as well as those of the institute as a whole. Another challenge was being general enough to be applicable to different scenarios including national and international contexts. At the international level the institute is heavily involved in the infrastructure consortium CLARIN-ERIC. One can see the necessity of sound digital management practice at this level, most notably in the institute's role as national coordinator and as service provider of the CLARIN B Centre. This involvement implies a domain specific repository providing depositing services for language resources, the CLARIN Centre Vienna[3]. Furthermore, creating a workflow and data management model that can be applied to the wide variety of different types of data and sources in the arts, humanities and social sciences to be dealt with is a major challenge.

## 2 Research data lifecycle models

The data lifecycle has becoming an ever more important factor in the researcher's scientific work. This is even more the case given the increasing emphasis on data sharing in research. (Corti et al., 2014). "Life cycle models are shaping the way we study digital information processes. These models represent the life course of a larger system, such as the research process, through a series of sequentially related stages or phases in which information is produced or manipulated." (Humphrey, 2006). They "help to define and illustrate these complex processes visually, making it easier to identify the component parts or distinct stages of the research data" (Carlson, 2014) and the responsible persons or entities. There is a wide range of data lifecycle models, each with a different focus or perspective. The research data lifecycle models can be classified according to the form (linear, circular, non-linear or other models) or (Carlson, 2014) according to the context of the model (individual-based, organisation-based and community-based models) as described by Carlson (2014). In this section, we will present and discuss existing data lifecycle models.

### 2.1 Models classified according to visualisation form

An example of the linear type is the USGS Science Data Lifecycle Model (see Figure 1). This model describes the data lifecycle from the perspective of research projects and the activities that are performed in phases, e.g. planning, collection, processing, analysis, preservation, publication and sharing of the data for others to reuse. In addition to these activities, there are others that must be performed continually across all phases of the lifecycle, such as the documentation of the workflow process, and the provision of metadata, as well as the backup of data in order to prevent the possibility of physical loss (Faundeen et al., 2013).

[3] http://clarin.oeaw.ac.at

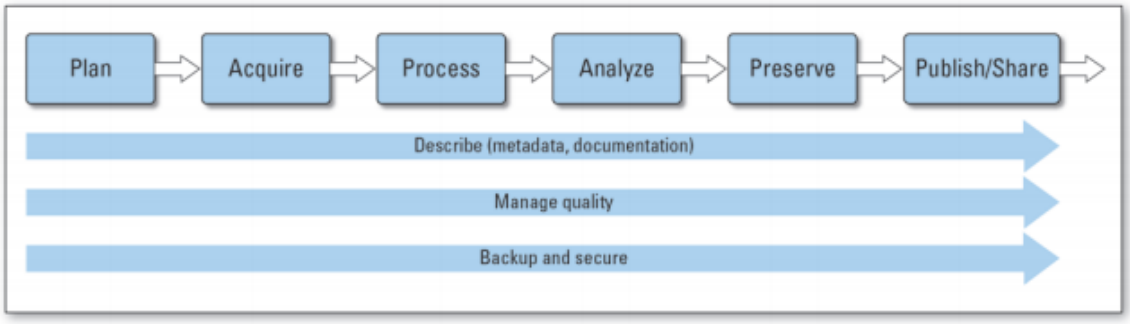



Figure 1: USGS Science Data Lifecycle Model (Faundeen et al., 2013).

There are also circular models which try to reflect the iterative nature of the research process where each step builds on existing material. Circular models seem better suited to describe current research practices increasingly relying on sharing and reuse of data (beyond one researcher or group), an example of which below (Figure 2) shows the e-research and data and information lifecycle (Allan, 2009) with a focus on sharing of data and information.

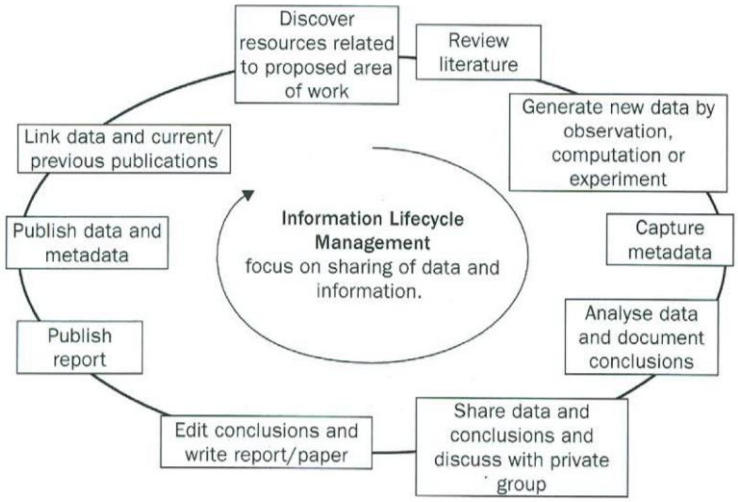



Figure 2: e-research and data and information lifecycle (Allan, 2009).

There are also other types of lifecycle models or workflow models, for example, the non-linear GLBPM model (Barkow et al., 2013) or the OAIS, the Open Archival Information System Reference Model (Lavoie, 2004) which is a concept model for digital repository and archival system. Given that this system does not intend to represent the whole research workflow, it does not fit in the classification above.

### 2.2 Models classified according to the creator or user of the model

Carlson (2014) describes three different types of life cycle models: individual-based life cycle models, organisation-based models and community-based models. Individual-based models are project-specific (Carlson, 2014) and are often not at an abstract level but contain project related detailed information. Such individual-based models can be helpful in elaborating a data management plan for a specific project. Organisation-based life cycle models "are produced by organi[s]ations offering services or assistance to researchers" (Carlson, 2014). These organisations include universities, libraries, data repositories, publishers etc. An example of an organisation-based model is the University of Oxford Research Data Management Chart (see Figure 3). Compared to individual-based life cycle models, organisation-based life cycle models generalise the different phases of the data lifecycle more since they are not focused on a specific project. From Figure 3, it becomes apparent

that this chart, in contrast to Figure 2, is not organised alongside the research process, but alongside the services that the organisation can offer the researchers in the different stages of the data lifecycle. The University of Oxford has a "Policy on the Management of Research Data and Records". In this policy it is stated, that the university "is responsible for: Providing access to services and facilities for the storage, backup, deposit and retention of research data and records that allow researchers to meet their requirements under this policy and those of the funders of their research; Providing researchers with access to training, support and advice in research data and records management; Providing the necessary resources to those operational units charged with the provision of these services, facilities and training." (University of Oxford, 2014). The model shows that, in compliance with the above mentioned policy, the support offered provides a data management planning checklist as well as services for data backup and data archiving.


Figure 3: University of Oxford Research Data Management Chart (CEOS, 2011).

Models from the third type are called the community-based life cycle models. They have been developed to support the needs of a particular research community and convey recommended best practices in a way that leads to a shared understanding and adoption of these practices in the interested community (Carlson, 2014). An example of a community-based lifecycle model is the DCC Curation Lifecycle Model (Higgins, 2008) (see Figure 4), which describes the different stages of data curation in detail but does not locate the curation process within a research project lifecycle. The model "offers a graphical high-level overview of the lifecycle stages required for successful curation. Generic in nature, the model is indicative rather than exhaustive. When used as an organisational planning tool, it is adaptable to different domains, and extensible to allow curation and preservation activities to be planned at different levels of granularity. It can be used to: define roles and responsibilities; build frameworks of standards and technologies; and ensure that processes and policies are adequately documented. The model identifies: curation actions which are applicable across the whole digital lifecycle; those which need to be undertaken sequentially if curation is to be successful; and those which are undertaken occasionally, as circumstances dictate" (Higgins, 2008). The DDC model (Higgins, 2008) is structured around data (digital objects or databases) and actions. It divides the actions in full lifecycle actions (description and representation information, preservation planning, community watch and participation, curate and preserve), sequential actions (conceptualise, create and receive, appraise and select, ingest, preservation actions, store, access, use and reuse, transform) and occasional actions (dispose, reappraise, migrate).

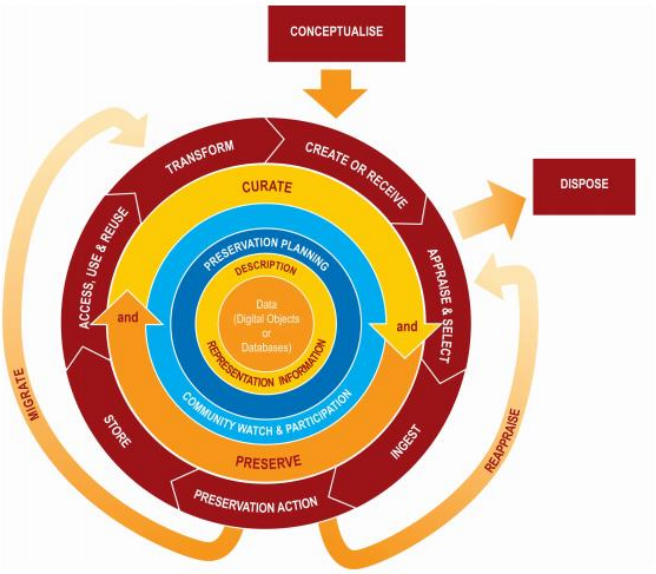



Figure 4: DCC Curation Lifecycle Model (Higgins, 2008).

## 3 Research Data Management

In this section, we will first define some key terms followed by a description of the institutional case study, the stakeholders and the workflow model. Additionally, we will explain the relation to Clarin and delineate the current status of the implementation.

For this paper, we define research data management as "all data practices, manipulations, enhancements and processes that ensure that research data are of a high quality, are well organized, documented, preserved, sustainable, accessible and reusable" (Corti et al., 2014). Even though the definition of data and research data, especially in the humanities, is subject to intensive discussion (e.g. Sahle, 2015; Kennan and Markauskaite, 2014), it will not be further discussed here. For this paper, we use the definition given by the Consortia Advancing Standards in Research Administration Information (CASRAI)[4].

As mentioned before, data lifecycles are a high level presentation of processes. On the other hand, data management workflows should be specific and detailed enough to serve as blueprint. In order to design the workflow, the stakeholder, the different steps, and their dependences have to be identified for every task/scenario. As Carlson (2014) stated: "Applying life cycle models to support services for managing research data has several benefits". Because "[f]rom its inception to its use and completion, research data will likely undergo multiple transformations in its format, application, use and perhaps even its purpose. Through identifying and naming the transformations that data will undergo as stages in a larger life cycle, organi[s]ations can better target their services […]." (Carlson, 2014).

While the abstract lifecycle models can serve as guidance, they have their limitations. In practice the workflows will usually be more complex with possible variations due to context-specific constraints and because lifecycle models tend to present an idealized version of the processes (Carlson, 2014).

[4] Data: Facts, measurements, recordings, records, or observations about the world collected by scientists and others, with a minimum of contextual interpretation. Data may be in any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records. (http://dictionary.casrai.org/Data) [accessed 30.12.2015]

## 4 Case study: Institutional Data Management Service at the Austrian Academy of Sciences

There are a lot of publications dealing with institutional case studies, describing the setting up of data management services, as well as the different approaches and challenges encountered (e.g. Choudhury, 2014; Brown and White, 2014; Beitz et al., 2014; Akers et al., 2014; Henry, 2014). When applying data lifecycle models to data management services, different factors have to be taken into account, e.g. who is the target group, what are the best practices and standards in the relevant field or community and what are the real needs of the intended target group (Carlson, 2014). In the following, we will describe the case study of the development and implementation of an institutional data management service at the Austrian Academy of Sciences.

In 2014, the Austrian Academy of Sciences launched the "go!digital"[5] funding initiative supported by the Federal Ministry for Science, Research and Economy (BMWFW) as well as the funding initiative "Digital Humanities: long-term projects on cultural heritage"[6] in order to foster/boost the digital humanities research at the academy and in Austria in general, with special focus on scientific digitisation of cultural heritage as the indispensable base for DH research.

As a natural consequence of this initiative, there has been, as expected, a substantial rise in the amount of new data, leading to a corresponding need to manage the research data and support these projects. In 2015 different institutional stakeholders formed a working group tasked with coordination of the development and implementation of research data management services at the institutional level; these services will be accompanied by technical support, training and workshops on best practice.

### 4.1 Stakeholders and target group

The following stakeholders are part of the working group *data services*: the Austrian Centre for Digital Humanities (ACDH-OEAW)[7]; the institutional publishing house Academy Press; the institutional computing centre of the Academy (ARZ) and the institutional library (BAS:IS). There are also other stakeholders that are at the moment not part of the working group but nevertheless, play a key role in the data management workflow: third-party service providers for digitisation. The intended target group for the data management services are the researchers[8], both within and outside the academy, especially in the arts, humanities and social sciences. Researchers in life sciences, physics, mathematics etc. use already well established infrastructure for archiving in their relevant fields and are not the main target group.

The ACDH-OEAW runs a domain specific repository for the arts and humanities, with a particular emphasis on language resources, for which we operate the Language Resources Portal which is part of the CLARIN Centre Vienna (*CCV/LRP*)[9]. The ACDH-OEAW also offers a range of applications and services for processing, analysing, visualising and querying different kinds of data.

The Press has been operating the institutional repository of the Academy, *epub.oeaw*[10] that is designated to hold primarily scientific publications, but increasingly also research data. The repository serves a double role: publication and archiving, data in the repository being replicated to the Austrian National Library (Stöger et al., 2012). So, while there is some overlap in the task description of *epub.oeaw* and *CCV/LRP*, there are distinct features, that justify the co-existence of the two repositories.

Currently, the stakeholders are elaborating a common strategy to act as a coordinated network of providers for data-related services, with clear division of roles. In this plan, ACDH-OEAW will concentrate more on the interaction with the researchers (consulting, data modelling), development

[5] http://www.oeaw.ac.at/en/fellowship-funding/promotional-programmes/godigital/

[6] http://www.oeaw.ac.at/en/fellowship-funding/promotional-programmes/digital-humanities-long-term-projects-on-cultural-heritage/

[7] http://www.oeaw.ac.at/acdh

[8] In this paper, as researchers we mean research staff of the Austrian Academy of Sciences as well as non-members of the Austrian Academy of Sciences who are conducting research in collaboration with the Academy or are making use of the offered services and are willing to deposit data in one of the described repositories.

[9] https://clarin.oeaw.ac.at/

[10] http://epub.oeaw.ac.at/

and provision of tools for processing, analysing, visualising the data. The Press will keep running the repository for archiving and publishing of publications and certain types of research data. However, not all kinds of resources are equally well suited for the digital asset management system underlying *epub.oeaw*, particular examples of which are: relational databases, corpora and graph-based data. Thus, the working group still needs to work a strategy for archiving for this kind of data. Furthermore, there are plans to establish in-house capacities for digitisation at the institutional library that also serves as an important content provider.
One of the challenges was to bring all the stakeholders together and to develop a common strategy how to deliver a data management service together, since these stakeholders haven't worked together until recently. One of the peculiarities of the present case study is that in contrast to the usual setup, where the institutional libraries are the driving forces in the process and deliver most of the services related to data management (Choudhury, 2014; Brown and White, 2014; Beitz et al., 2014; Akers et al., 2014; Henry, 2014), in our case the coordinating unit, the Austrian Centre for Digital Humanities, is a research institute that also functions as service unit, and therefore, the institute is involved also as research partner.
In the following section we will explain the workflow model.

### 4.2 Workflow Model

In Figure 5 the proposed research data management workflow is illustrated from the perspective of the institute, the ACDH-OEAW with a focus on projects from the arts, humanities and social sciences. The key roles in this model are taken by the researcher, the institute, the publishing house, the library and third party service provider. The institutional computing centre of the Academy (ARZ) is not present in the model, however it is still an indispensable partner as it runs the basic technical infrastructure (servers, storage, networks, etc.). If in Figure 5 for one task only one form is visible, then only one stakeholder is responsible for this task, if there are more overlapping forms in different grey tonalities, black or white then the responsibilities are shared.

In this model below, six different phases are shown which are as follows: the pre-processing (divided into proposal stage and granted stage), the processing, the storage, the publishing and the reuse phase as well as quality assurance. As shown in the model (Figure 5), not all the phases are clear-cut and they can overlap. The quality assurance process is special, as it accompanies and underlies the whole workflow. There are mainly two different scenarios to which the institutional research data management model has to be applied. The first scenario is when a new project proposal is written, here we call this scenario *new project* (Figure 5) the second is when the project is already over, here (Figure 5) we call this scenario *legacy data*. In the following we will describe the two scenarios in detail.

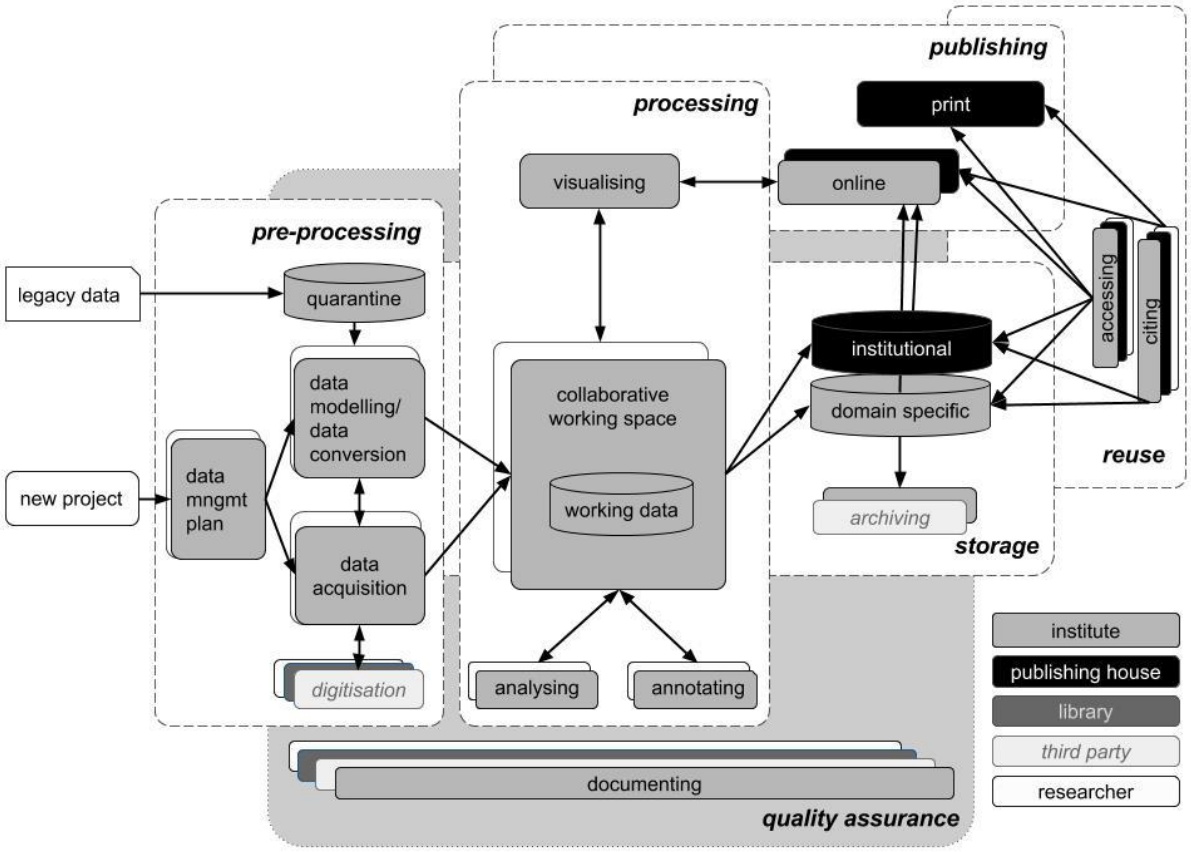



Figure 5: Proposed institutional research data management workflow.

### 4.3 First scenario: new project

In this first scenario, the researcher approaches the institute (as part of the *data services* team) for advice with a project idea in the proposal phase. There, the new project enters the pre-processing phase which itself has two different stages, the *proposal* and *granted* stage. The first step in the proposal stage is the elaboration of the data management plan that most of the funding agencies nowadays require for a grant application. The institute advises the researcher on data management issues, especially on the resources (people, equipment, infrastructure and tools) that have to be taken into account in the project budget. In the ideal case, the institute and the *data services* group is included into the project proposal. If the project gets funded, the project enters the *granted* stage at which time, the data collection starts. If the new project involves digitisation, this is also part of the pre-processing phase and is either done by the researchers themselves or by a third party service provider.[11] In parallel to collecting or acquiring the data, the institute elaborates the data model together with the researcher. As data model we understand "[a] model that specifies the structure or schema of a dataset. The model provides a documented description of the data and thus is an instance of metadata"[12] as defined by the Data Foundation and Terminology Working group of the Research Data Alliance (RDA). Based on the data model and the requirements of the project, formats for data and metadata are discussed and chosen in accordance with best practices and standards in order to avoid data loss and conversion problems in the future. If we compare our model with the previous discussed lifecycle models, the pre-processing phase in our model corresponds to the activities *plan* and *acquire* in the USGS Science Data Lifecycle Model (Figure 1) or to *generate new data* and *acquire metadata* in the model of Allen (2009) in Figure 2. In the model in Figure 3 it would

[11] Figure 5 illustrates that the Academy library also offers digitisation services. This service is not yet in place but it is expected to be enacted sometime this year.

[12] RDA Term Tool, entry "data model" available at http://smw-rda.esc.rzg.mpg.de/index.php/Data_Model ([accessed: 30.12.2015]

correspond to *data management planning* and in the DDC Model (Figure 4) it would correspond to *conceptualise* and *create.*

After the pre-processing phase the data enters the processing phase during which all research activities related to previously acquired data take place. In referring to *processing*, we specifically mean: "performing a series of actions in something (an input) in order to achieve a particular result (output)."[13] Some of the actions are mentioned in the model (analysing, annotating, visualising), but they are not exhaustive. If we take the NeDiMAH Methods Ontology as a reference point, annotating would be a subtype of analysing, but we decided to depict them at the same level, given the importance of the annotation step in the research process. Ideally, the researchers work in an integrated collaborative working space, where they get offered a series of tools for annotating, analysing, visualizing etc., run as a service by the data services working group. Data visualisation is helpful in detecting patterns and performing analysis, and therefore it is used in the collaborative working space during the processing phase and it is used in the publication phase for online publication of the data. In the model the visualising activity is in the overlapping of the processing and the publishing phase in order to reflect these two purposes. Currently the above mentioned portfolio of tools is being built up combining existing open source applications as well as specific solutions to a task. Thanks to the strong international involvement of ACDH-OEAW, the tool development is deeply embedded in the activities of the research infrastructures CLARIN & DARIAH as well as RI projects, most prominently the new H2020 project PARTHENOS[14]. The processing phase corresponds to the activities *process* and *analyse* in the USGS Science Data Lifecycle Model. The collaborative working space reflects the activities *analyse data and document conclusions* and *share data and conclusions and discuss with private group* in the data lifecycle by Allan (2009). In both lifecycle models, publishing activities are foreseen as well as in our proposed workflow.

An important activity, especially in relation to future reuse (Corti et al., 2014) of data, is documenting. Documenting is understood as "providing information regarding each and every step of the activities that took place in a project, in order to describe how everything was done and enable someone that was not initially involved to understand."[15] Data documentation includes information on data creation, content, structure, coding, anonymization etc. There are two types of documentation, the high level description, also known as study-level documentation and the data level documentation (Ibid). If we have a closer look at the model, the documenting can be found as part of the quality assurance, that runs alongside all the processes Already in the data acquisition and digitisation, documenting plays an important role in order to achieve reusable data at the end of the workflow.

It is important to underline that all the phases as well as the whole workflow cannot be seen as a simple step or linear sequence of steps, but rather a complex, non-linear, iterative process, both within one project as well as beyond the project boundaries

In the storage phase, underlying the whole workflow, the data and metadata are stored and archived. We need to distinguish different kinds of storage. In the pre-processing phase during the data collection, large amounts of data is produced that is the starting point/serves as base for the whole further process and needs to be secured and made accessible within the workspace. In the processing phase, a lot of additional data is produced, oftentimes of transitional nature. We call this "working data". Stable data – raw captured data as well as secondary data / enrichments contributed in the processing phase – aimed at long-term availability and/or publication is moved to the institutional or domain specific repository, which in the long run represents the main source for the datasets. Before the data will be ingested in one of the repositories, licence issues have to be discussed and agreements have to be signed. At the archiving stage, it is necessary to ensure long-term availability of the data even beyond a disaster scenario e.g. main repository is damaged through fire or similar. This involves geographically distributed replication/mirroring of the data to reliable providers of storage services, like scientific data centres. The data from the repository *epub.oeaw* is already being replicated to the Austrian National Library. Additionally, we build up alliances with national providers as well as

[13] Definition taken from the NeDiMAH Methods Ontology (NeMO) available at http://nemo.dcu.gr/index.php?p=hom [accessed: 30.12.2015].

[14] http://www.parthenos-project.eu/

[15] Definition taken from the NeDiMAH Methods Ontology (NeMO) available at http://nemo.dcu.gr/index.php?p=hom [accessed: 30.12.2015].

international players mainly in the context of the EUDAT initiative. Archiving and preservation activities are also mentioned in the USGS Model, in the Oxford Research Data Management Chart and in the DCC Model.
The publishing phase refers primarily to presentation, online and/or print, of the results of the project but also – in line with the open access policy and subject to copyright and ethical restriction – the provision of the underlying research data. Enabling discoverability and citability of the research data is a precondition for effective reuse. The institute and publishing house are providing infrastructure and user interfaces for researchers to search for data and publications and to access them e.g. via the interface of *epub.oeaw* (Stöger et al., 2012). Next to direct access to the data, it is crucial to ensure wide-spread dissemination of the data, again ensured by the combined competencies of Press, library and ACDH-OEAW. While Press ensures indexing of the resources by services like Google Scholar and OpenAIRE, ACDH-OEAW pushes into the more domain-specific channels in the context of CLARIN and DARIAH. One important issue in the reuse phase is proper citation. Proper citation of publications, in the humanities especially of print publications, is an integral part of good research practices. But not all the researchers in the humanities are yet familiar with citations of primary or secondary data sources or data sets or the citation of digital editions. One increasingly popular possibility to help researchers is to integrate citation recommendation within the online presentation of the resources[16]. For data sets the attribution of a unique persistent identifier is essential. While there are several standard persistent identifier (PID) systems (see Corti et al., 2014; Briney, 2015) so far the most relevant to the Academy are Digital Object Identifiers (DOI).The institutional repository *epub.oeaw* is assigning DOIs to each uploaded research result (Stöger et al., 2012). In LRP every resource is assigned a handle-based[17] PIDs in accordance to CLARIN requirements. However, it is essential to use the persistent identifier in the citation, because it helps tracking data citations (Briney, 2015) and use recommended formats of data citations, e.g. Starr and Gastl (2011) resembling traditional print publication citations.

### 4.4 Second scenario: legacy data

The second scenario, covered by the workflow, is the so called *legacy data* scenario. As legacy data we understand data that fall into the category of dark data or at-risk data. More often, we deal with at-risk data, that is data that are at risk of being lost due to the fact that the project is already over, and the stored data is not well or not at all documented (including missing metadata or the data has been detached from supporting data or metadata) and therefore not useable or reusable or it is stored on a medium that is obsolete or at risk of deterioration.[18]
When confronted with legacy data, in a first step, all the relevant data is stored, as shown in Figure 5, in a kind of "quarantine" repository to be further processed. Then the data and the data model/structure are examined, especially with respect to the suitability of the format, existence of metadata and documentation and internal structure of the data. Based on the analysis, it is decided if the data has to be converted and the data model needs to be adapted, transformed together with the estimation of the required resources of such transformation. Then the data is stored (see storage phase above) in the repositories and archived without going through the processing phase. Usually, there are only limited resources to deal with legacy data, the primary goal is to ensure a reliable deposition of the data and the accessibility for other researchers. Thus as long as no new user/project interested in this data arises, no interaction with the data is expected in the working space, nor is an online publication.

[16] E.g. in the ABaC:us – Austrian Baroque Corpus digital edition a citation suggestion is generated with each query: Abraham â Sancta Clara: Todten-Capelle. Würzburg, 1710. (Digitale Ausgabe) Vorrede [S. 14]. In: ABaC:us – Austrian Baroque Corpus. Hrsg. von Claudia Resch und Ulrike Czeitschner. <https://acdh.oeaw.ac.at/abacus/get/abacus.3_48> abgerufen am 3. 1. 2016
[17] http://www.handle.net/
[18] Modified definition taken from CASRAI Dictionary: legacy data available at http://dictionary.casrai.org/Legacy_data [accessed 07.03.2015]; dark data available at http://dictionary.casrai.org/Dark_data [accessed 07.03.2015]; at-risk data available at http://dictionary.casrai.org/At-risk_data [accessed 07.03.2015]

### 4.5 Relation to CLARIN

As mentioned before, the development or adaptation of an institutional-based model should take into account the relevant best practices and standards in the community of the intended target group. Given that ACDH-OEAW runs a CLARIN Centre[19] and is a national coordinator of CLARIN activities, many aspects of the workflow are strongly guided by the requirements expected by CLARIN-ERIC[20] – assignment of persistent identifiers, metadata in CMDI (Component Metadata Infrastructure) format (Broeder et al. 2010), OAI-PMH[21] (Open Archives Initiative Protocol for Metadata Harvesting) endpoint as a dissemination channel for the metadata harvested by the CLARIN harvester. One of the aims of the presented strategy is to make new resources automatically available via the CLARIN infrastructure.

Currently, for the resources we use 4 different CMDI profiles, and make the resources available in different forms, partly as raw data, partly within complex web applications that allow search and browsing through the data via different dimensions (linguistic, semantic). These steps are related to the reuse phase in the research data management model in Figure 5. The access to resources and services is managed through Federated Identity.

In 2016, CCV/LRP is scheduled for reassessment. In preparation for this, the repository solution will be overhauled, taking into account lessons learned in the last two years, aiming for tighter integration with the institutional repository *epub.oeaw* (eliminating redundancies). As part of this process, language resources already existing in the *epub.oeaw* repository shall be made accessible within CLARIN (primarily by providing appropriate CMDI records). One central challenge in this task will be to reflect the broader role that the ACDH-OEAW has lately assumed covering not just language resources but expanding to a broad spectrum of disciplines in the context of digital humanities (archaeology, history, art history, etc.). Here we aim – in accordance with the principles of the research infrastructures – for a setup with common/harmonized technical infrastructure in combination with domain- or project-specific solutions/views building on top of it.

With respect to the tools offered for use, there is a reciprocal relation to CLARIN, where tools from the CLARIN community are part of the portfolio, like *WebLicht* (Hinrichs et al. 2010) as well the solutions developed at the institute are made available to the whole CLARIN community, like the SMC Browser (Durco, 2013), Vienna Lexicographic Editor (Budin et al., 2013), or the corpus shell[22] framework

With respect to long-term archiving we plan to take advantage of the relation of CLARIN-ERIC to the EUDAT initiative.

### 4.6 Current status

Currently, the model is being implemented. Many parts/components of the model are already available (like the repositories, individual processing and visualisation tools, the publishing workflow), the main task in 2016 will be to provide the glue between these components by establishing the procedures inside the *data services* working group and make the services accessible/usable by the target audience – the researchers of the academy and of the broader Austrian DH community.

The usefulness and appropriateness is currently being tested on a number of research projects, especially from the calls *go!digital* and *Digital Humanities: long-term projects on cultural heritage*, all of which started last year. A few examples of types of projects and data we are dealing with include APIS project[23], which aims to enrich and convert a large biographical lexicon with the help of NLP tools into richly structured Linked Open Data. These data will then be made available for exploration through appropriate interactive visualisation means; similarly in *exploreAT!*[24] huge amounts of heterogeneous data gathered over more than a hundred years and available in different digitisation stages and formats will be harmonized (adhering to the LOD paradigm) and made available online,

[19] http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-105
[20] hdl:1839/00-DOCS.CLARIN.EU-77
[21] http://www.openarchives.org/pmh/
[22] https://clarin.oeaw.ac.at/corpus_shell
[23] Austrian Prosopographic Information System - http://www.oeaw.ac.at/acdh/apis
[24] http://www.oeaw.ac.at/acdh/exploreat

both as raw structured data and as rich explorative applications; in DEFC [25] a database of archaeological sites and finds is being developed;[26] aims at full linguistic and semantic enrichment of the historic texts of Baedeker using TEI/XML as the native format. These are just four out of a number of projects to sketch the variety of data and requirements the *data services* team is confronted with. From historic manuscripts to archaeological sites and finds - with each project we learn something new and update our data management workflow and with each new project we learn more about how best to manage, store and present online data from the humanities, arts and social sciences. Alongside the testing of the model, we also give training workshops in order to raise awareness and understanding and to improve research data management skills.

## 5 Conclusion and outlook

In this paper we presented an institutional workflow model for research data as it is currently being implemented at the Austrian Academy of Sciences, coordinated by the ACDH-OEAW, a newly founded research institute of the Academy that acts also as a service unit for researchers in the art and humanities in the institutional and national context. Starting from abstract (research) data lifecycle models, we discussed the stakeholders and scenarios for the specific institutional settings and elaborated a workflow model that caters to the specific situation of the Academy.
Just like Higgins (2008) stated that the DCC Model "is not definitive and will undoubtedly evolve", also the ACDH-OEAW model will evolve. Even once a service is fully functional, the evolution of data-dependent research practices and the changing research technologies have to be monitored in order to adapt the service to changing demands.
The paper shows that the elaboration of an institutional research data workflow model is important since there is no "one-size-fits-all-solution", e.g. Higgins (2008) mentioned "domain-specific variations" of the DCC model, but high level data lifecycle models are a good basis to start with and to adapt to the specific institutional context. The elaboration or adaptation of already existing models depends on different aspects like target group, relevant best practices and standards and real world needs of the intended target group. Once the workflow model is implemented, it can not only be used as quality assurance measure but it can also guide the researchers in the project planning phase, when and whom to approach for advice, assistance and support.

### Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions on this paper.

## References

[Allan, 2009] Robert Allan. 2009. Virtual Research Environments. From portals to science gateways. Chandos Publishing, Oxford, UK.

[Akers et al. 2014] Katherine G. Akers, F.C. Sferdean, Natsuko H. Nicholls, Jennifer A. Green. 2014. Building Support for Research Data Management: Biographies of Eight Research Universities. *International Journal of Digital Curation*. 9(2):171-191.

[Akers and Doty, 2013] Katherine G. Akers and Jennifer Doty. 2013. Disciplinary Differences in Faculty Research Data Management Practices and Perspectives. *The International Journal of Digital Curation*, 8(2):5-26.

[Barkow et al., 2013] Ingo Barkow, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, Wolfgang Zenk-Möltgen. 2013. *Generic longitudinal business process model. DDI Working Paper Series – Longitudinal Best Practice, No. 5.* DOI: http://dx.doi.org/10.3886/DDILongitudinal2-01

[Beitz et al., 2014] Antony Beitz, David Groenewegen, Cathrine Harboe-Ree, Wilna Macmillan and Sam Searle. 2014. Monash University, a strategic approach. In: Graham Pryor, Sarah Johnes and Angus Whyte. 2014.

[25] Digitizing Early Farming Cultures - http://www.oeaw.ac.at/acdh/defc
[26] https://acdh.oeaw.ac.at/acdh/traveldigital

*Delivering Research Data Management Services. Fundamentals of good practices*. Facet Publishing, London, UK. 163-189.

[Briney, 2015] Kristin Briney. 2015. Data Management for Researchers. Organize, maintain and share your data for research success. Pelagic Publishing, Exeter, UK.

[Bauer et al., 2015] Bruno Bauer, Andreas Ferus, Juan Gorraiz, Veronika Gründhammer, Christian Gumpenberger, Nikolaus Maly, Johannes Michael Mühlegger, José Luis Preza;, Barbara Sánchez Solís, Nora Schmidt andChristian Steineder (2015): Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung. Report 2015. Version 1.2. DOI: 10.5281/zenodo.32043. Online available at http://phaidra.univie.ac.at/o:407513

[Broeder et al. 2010] Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Philip Withers, Peter Wittenburgand Claus Zinn.2010. A data category registry-and component-based metadata framework. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]*. European Language Resources Association (ELRA). 43-47.

[Brown and White, 2014] Mark L. Brown and Wendy White. 2014. University of Southampton – a partnership approach to research data management. In: Graham Pryor, Sarah Johnes and Angus Whyte. 2014. Delivering Research Data Management Services. Fundamentals of good practices. Facet Publishing, London, UK. 135-161.

[Budin et al, 2013] Gerhard Budin, Karlheinz Moerth and Matej Durco. 2013. European Lexicography Infrastructure Components. In Iztok Kosem, JelenaKallas, Polona Gantar, Simon Krek, Margit Langemets and Maria Tuulik (eds.), Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013 (pp. 76–92). Tallin, Estonia: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut. http://eki.ee/elex2013/conf-proceedings/

[Carlson, 2014] Jake Carlson. 2014. The Use of Life Cycle Models in Developing and Supporting Data Services. In: Joyce M Ray (ed). 2014. *Research Data Management. Practical Strategies for Information Professionals*. Purdue University Press, West Lafayette, IN. 63-86.

[CEOS, 2011] CEOS. 2011. *CEOS Working Group on Information Systems and Services. Data Life Cycle Models and Concepts. CEOS Version 1.0*. Available at http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS_DSIG-Data-Lifecycle-Models-and-Concepts-v8_Sep2011.docx

[Choudhury, 2014] G. Sayeed Choudhury. 2014. John Hopkins University Data Management Services. In: Graham Pryor, Sarah Jones and Angus Whyte. 2014. *Delivering Research Data Management Services. Fundamentals of good practices.* Facet Publishing, London, UK: 115-133.

[Corti et al., 2014] Louise Corti, Veerle Van den Eynden, Libby Bishop and Mattew Woolard. 2014. *Managing and sharing research data. A guide to good practice*. SAGE, London, UK.

[Durco, 2013] Matej Durco. 2013. *SMC4LRT - Semantic Mapping Component for Language Resources and Technology*. Technical University, Vienna, Austria. http://permalink.obvsg.at/AC11178534

[Faundeen et al., 2013] John L. Faundeen, Thomas E. Burley, Jennifer A. Carlino, David L. Govoni, Heather S. Henkel, Sally L. Holl, Vivian B. Hutchison, Elizabeth Martín, Ellyn T. Montgomery, Cassandra C. Ladino, Steven Tessler, and Lisa S. Zolly 2013. *The United States Geological Survey Science Data Lifecycle Model. U.S. Geological Survey Open-File Report 2013–1265*, 4 p, http://dx.doi.org/10.3133/ofr20131265.

[Henry, 2014] Geneva Henry. 2014. Data Curation for the Humanities. Perspectives From Rice University. In: Ray, Joyce M. (ed). 2014. *Research Data Management. Practical Strategies for Information Professionals*. Purdue University Press, West Lafayette, IN. 347-374.

[Higgins, 2008] Sarah, Higgins. 2008. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation.* 3(1):134-140.

[Hinrichs et al., 2010] Marie Hinrichs, Thomas Zastrow and Erhard Hinrichs. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. Paper presented at LREC 2010, Valetta, MT.

[Humphrey, 2006] Charles Humphrey.2006. E-science and the life cycle of research. Available at http://www.usit.uio.no/om/organisasjon/uav/itf/saker/forskningsdata/bakgrunn/life-cycle.pdf

[Kennan and Markauskaite, 2015] Kennan, Mary Anne and Lina Markauskaite. 2015. Research Data Management Practices: A Snapshot in Time. *International Journal of Data Curation*. 10(2):69-95.

[Lavoie, 2004] Brian F. Lavoie. 2004. *The Open Archival Information System Reference Model: Introductory Guide*. OCLC Online Computer Library Center.

[ICPSR, 2012] Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to Social Science Data Preparation and Archiving: Best Practice throughout the Data Life Cycle* (5th ed.). Ann Arbor, MI. Available at http://www.icpsr.umich.edu/files/deposit/dataprep.pdf

[Sahle, 2015] Patrick Sahle. 2015. Forschungsdaten in den Geisteswissenschaften. *SAGW Bulletin*, 2015(4)4(2015):43-45.

[Starr and Gastl, 2011] Starr, Joan and Angela Gastl. 2011. A Metadata Scheme for DataCite. *D-Lib Magazin*, 17(1/2). doi:10.1045/january2011-starr

[Stöger et al., 2012] Herwig Stöger, Vittorio Muth - Georg Lasinger. 2012. *epub.oeaw Benutzerhandbuch. Das Publikationsportal der Österreichischen Akademie der Wissenschaften*. Österreichische Akademie der Wissenschaften. Vienna, AT. Available at http://epub.oeaw.ac.at/dokumentation14/0000_Epub.UserGuide_1.4_printable.pdf

[University of Oxford, 2014] University of Oxford. 2014. *Policy on the Management of Research Data and Records*. Oxford, UK. Available at http://researchdata.ox.ac.uk/files/2014/01/Policy_on_the_Management_of_Research_Data_and_Records.pdf

# Enriching a Grammatical Database with Intelligent Links to Linguistic Resources

**Ton van der Wouden**
Meertens Institute
Amsterdam, The Netherlands
Ton.van.der.wouden@meertens.knaw.nl

**Gosse Bouma**
Groningen University
The Netherlands
g.bouma@rug.nl

**Matje van de Camp**
De Taalmonsters
Tilburg, The Netherlands
matje@taalmonsters.nl

**Marjo van Koppen**
Utrecht University
The Netherlands
J.M.vanKoppen@uu.nl

**Frank Landsbergen**
Institute for Dutch Lexicography
Leiden, The Netherlands
Frank.Landsbergen@inl.nl

**Jan Odijk**
Utrecht University
The Netherlands
j.odijk@uu.nl

**Abstract**

We describe goals and methods of CLARIN-TPC, a project to enrich the on-line Taalportaal (Language Portal) grammatical database with intelligent links in the form of annotated queries to a variety of interfaces to on-line corpora and an on-line linguistic morphophonological database.

## 1 Introduction

We describe the on-line grammatical database Taalportaal (Language Portal) an particularly how it is being enriched with intelligent links in the form of annotated queries to a variety of interfaces to on-line corpora and an on-line linguistic morphophonological database. This database contributes to the use of the CLARIN research infrastructure in the following ways:

- It provides users with actual corpus examples for linguistic phenomena described in Taalportaal.
- It points out the existence and usefulness of search interfaces developed in the CLARIN infrastructure such as PaQu, GrETEL and OpenSONAR to linguists.
- By redirecting the user to the front-ends, it stimulates the further use of these applications in the CLARIN infrastructure for modifying queries or submitting new queries. Together with the multiple interfaces of most of these applications, this may also have a significant educational role.

## 2 Background

Linguistic data is everywhere. The working linguist is confronted with data any moment he/she reads a newspaper, talks to their neighbour, watches television, switches on the computer. To overcome the volatility of many of these data, digitized corpora have been compiled for languages all around the globe since the nineteen sixties. These days, there is therefore no lack of natural language resources. Large corpora and databases of linguistic data are amply available, both in raw form and enriched with various types of annotation, and often free of charge or for a very modest fee.

There is no lack of linguistic descriptions either: linguistics is a very lively science area, producing tens of dissertations and thousands of scholarly articles in a small country as the Netherlands only. An

enormous amount of this linguistic knowledge, however, is stored in paper form: in grammars, dissertations and other publications, both aimed at scholarly and lay audiences. The digitization of linguistic knowledge is only beginning, online grammatical knowledge is relatively scarce in comparison with what is hidden in the bookshelves of libraries and studies.

Of course, there are notable exceptions. One such exception is the Taalportaal (Language Portal) project, that is currently developing an online portal, containing a comprehensive and fully searchable digitized reference grammar, i.e. an electronic reference of Dutch and Frisian phonology, morphology and syntax. With English as its meta-language, the Taalportaal aims at serving the international scientific community by organizing, integrating and completing the grammatical knowledge of both languages. In contrast, the standard reference grammar for Dutch, the *Algemene Nederlandse Spraakkunst* (Haeseryn et al. 1997), is aimed at a broader (and other) audience than the international scientific community only, and is written in Dutch. The digital version[1] is essentially an XML-version of the paper edition.

To enhance the Taalportaal's value, the CLARIN project described here (NL-15-001: TPC) sought to enrich the grammatical information within the Taalportaal with links to linguistic resources. The idea was that the user, while reading a grammatical description or studying a linguistic example, was to be offered the possibility to find both potential examples and counterexamples of the pertinent constructions in a range of annotated corpora, as well as in a lexical database containing a wealth of morphophonological data on Dutch. Although links to raw text (including internet search) are offered as well, we here focus on resources with rich linguistic annotations explicitly, since we want to do more than just string searches: searching for construction types and linguistic annotations themselves is one way to reduce the problem of the massive ambiguity of natural language words and sentences.

In light of the restricted resources in terms both of time and money, this CLARIN project was not aiming at exhaustivity, that is, not all grammatical descriptions and not all examples are adorned with query links. TPC is explicitly to be seen as a pilot project, aiming for a proof of concept by showing the feasibility of efficient coupling of grammatical information with queries in a number of linguistic resources.

## 3 The Taalportaal

The Taalportaal project is a collaboration of the Meertens Institute, the Fryske Akademy, the Institute of Dutch Lexicology and Leiden University, funded, to a large extent, by the Netherlands Organisation for Scientific Research (NWO). The project is aimed at the development of a comprehensive and authoritative scientific grammar for Dutch and Frisian in the form of a virtual language institute (cf. Landsbergen et al. 2014). The Taalportaal is built around an interactive knowledge base of the current grammatical knowledge of Dutch and Frisian. Its prime intended audience is the international scientific community, which is why English is chosen as the language used to describe the language facts. The Taalportaal provides an exhaustive collection of the currently known data relevant for grammatical research, as well as an overview of the currently established insights about these data. This is an important step forward compared to presenting the same material in the traditional form of printed handbooks. For example, the three sub-disciplines syntax, morphology and phonology are often studied in isolation, but by presenting the results of these sub-disciplines on a single digital platform and internally linking these results, the Taalportaal contributes to the integration of the results reached within these disciplines.

As of January 2016, the first release of the Taalportaal is online[2]. Figure 1 shows the portal's opening screen.

[1] http://ans.ruhosting.nl/e-ans/index.html.
[2] http://www.taalportaal.org.

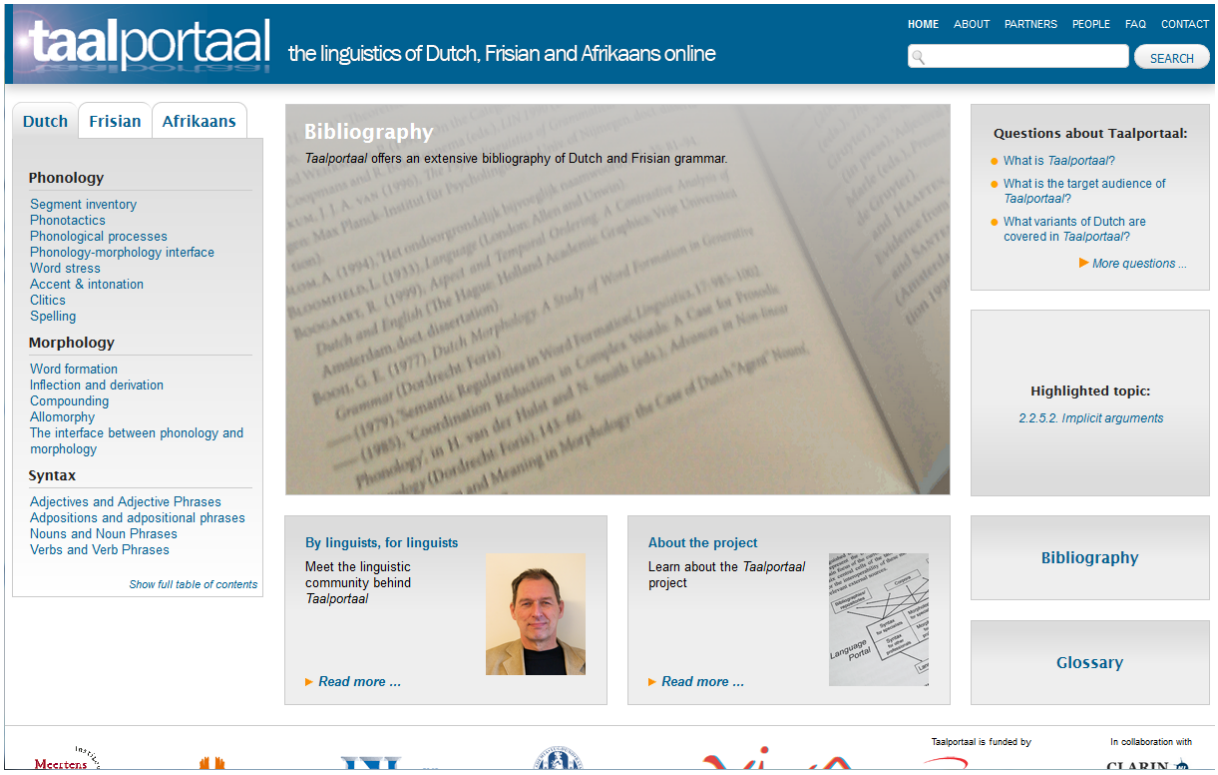



Figure 1: Opening screen of the Taalportaal.

Technically, the Taalportaal is built as an XML-database, organized as DITA-topics.[3] The data is freely accessible via the Internet using any standard internet browser. Organization and structure of much of the linguistic information is reminiscent of, and is to a certain extend inspired by, Wikipedia and comparable online information sources. An important difference, however, is that Wikipedia's democratic (anarchistic) model is avoided by restricting the right to edit the Taalportaal information to authorized experts. Figure 2 shows a small, introductory fragment concerning Dutch phonology.

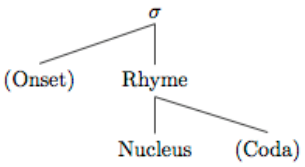
[3] https://en.wikipedia.org/wiki/Darwin_Information_Typing_Architecture.

## Phonotactics

PHONOTACTICS is the branch of PHONOLOGY dealing with the distribution of SEGMENTS within phonological and morpho-syntactic domains. It studies the restrictions on combinations of CONSONANTS, VOWELS and consonant-vowel-sequences depending on their phonological positions, both in a particular language and cross-linguistically.

The description of the phonotactics of Dutch will rely heavily on the concept of the SYLLABLE (σ). The SYLLABLE is assumed to consist of the following hierarchically ordered CONSTITUENTS:

*Figure 1*




[click image to enlarge]

The occurrence of vowels, consonants and CONSONANT CLUSTERS in Dutch is dependent on a variety of factors: many configurations only appear in specific contexts while they are prohibited in others. For example, the consonant cluster /kt/ is allowed in syllable codas, as in the word *pakt* /pɑkt/ 'pact' but it is prohibited in syllable onsets: accordingly, the hypothetical sequence /*ktɑm/ is not a possible Dutch word. The majority of the relevant generalizations can be expressed by making reference to the syllable and its constituents; there are, however, other factors that influence phonotactics, such as prosodic factors.

Figure 2: Introductory fragment on Dutch phonology.

The Netherlands are not the only country thinking of a virtual language institute like the Taalportaal. Recently, South Africa has started building a virtual language institute called **Viva!** [4] that aims at developing a digital infrastructure for Afrikaans. Among its goals are study and description of the Afrikaans language, and development of comprehensive tools and resources for written and spoken Afrikaans, including digital dictionaries and corpora; language advice is also supplied. Part of the Viva! portal is a comprehensive grammar of Afrikaans, which is based on the Taalportaal architecture, and will be part of the Taalportaal infrastructure.

Besides the grammar modules, the Taalportaal contains an ontology of linguistic terms (recently recast in the CLARIN Concept Registry, cf. Schuurman 2015) and an extensive bibliography. Note that the text in the pictures is full of words and phrases that are marked (bluish): these can be clicked, which results in definitions popping up and/or related topics being opened. Moreover, the texts are often amply illustrated, not only in the familiar ways with linguistic examples and tree-like drawings as in the fragment, but also with sound fragments – which is of course unfeasible in old-school paper books of reference.

## 4 Enriching the Taalportaal with links to linguistic resources

Another manner of enriching the Taalportaal's grammatical information that is also unfeasible in traditional printed grammatical works of reference is to enrich it with links to digital linguistic resources within the CLARIN infrastructure. In a collaborative effort of the Meertens Institute, the Institute of Dutch Lexicology, the Universities of Groningen and Utrecht, and Taalmonsters, a motivated selection of Taalportaal texts has been enriched with links that encompass queries in corpus search interfaces (project CLARIN-NL15-001).

[4] http://viva-afrikaans.org/.

The Taalportaal contains, among other things, an online edition of the *Syntax of Dutch* (SoD) (Broekhuis et al., 2012-2016), a descriptive grammar that goes well beyond the level of detail provided by other sources, including reference grammars. Although descriptive, the emphasis in the selection and presentation of the phenomena discussed is clearly guided by discussions in the theoretical, more specifically the generative, literature (cf. Bouma et al. 2015).

In his largely positive review of the first SoD volumes on NP syntax, Hoeksema (2013) points out that "There is a growing body of work in empirical studies of judgment variation [...] that future extensions of this grammar could benefit from, especially when coupled to studies of actual usage patterns in corpus material" and that "This particular reader would also have welcomed to see some more lists in the book". By enriching the on-line version of SoD with queries over syntactically annotated corpora, the current project tries to accommodate the needs of researchers like Hoeksema.

Queries are linked to the following:

- Linguistic examples
- Linguistic terms
- Names or descriptions of constructions.

The queries are embedded in the Taalportaal texts as standard hyperlinks to other resources within the CLARIN network, where CLARIN supplies guidelines for things like a common vocabulary, common annotation standards, common interfaces, and single log-in. Clicking these links brings the user to a corpus query interface where the specified query is executed — or, if it can be foreseen that the execution of a query takes a lot of time — the link may also connect to an internet page containing the stored result of the query. In general, some kind of caching appears to be an option worth investigating.

Two tools are available for queries that are primarily syntactic in nature:

- The PaQU web application[5] (cf. Odijk 2015)
- The GrETEL web application[6] (cf. Augustinus et al. 2013).

Both tools can be used to search largely the same syntactically annotated corpora, viz. the Dutch spoken corpus CGN (van der Wouden et al. 2003) and the LASSY corpus of written text (van Noord et al. 2006), but they offer a slightly different functionality. Both applications offer dedicated user-friendly query interfaces (word pair relation search in PaQu and an example-based querying interface in GrETEL) as well as XPATH as a query language,[7] so that switching between these tools is trivial. Moreover, it is to be foreseen that future corpora of Dutch (and hopefully for Frisian as well) will be embedded in the very same CLARIN infrastructure, using the same architecture (cf. Landsbergen et al. 2014), the same type of interface and the same kind of linguistic annotation; the latter is the annotation schema for the Dutch Spoken Corpus CGN (cf. Schuurman et al. 2003, van der Wouden et al. 2003) which has become a de facto standard for Dutch corpus annotation, thus allowing for the re-use of the queries on these new data.

Translation of a linguistic example, a linguistic term, or a name or description of a construction is not a deterministic task that can be implemented in an algorithm. Rather, the queries are formulated by student assistants. After proper training, they get selections of the Taalportaal texts to read, interpret and enrich with queries where appropriate. The queries are amply annotated with explanations concerning the choices made in translating the grammatical term or description or linguistic example into the corpus query. When necessary, warnings about possible false hits, etc. can be added. The student assistant's work is supervised by senior linguists.

Next to the annotated corpora mentioned above, access to two more linguistic resources have been investigated in TPC. On the one hand, there is the huge SONAR corpus (cf. Oostdijk et al. 2013). The size of this corpus (> 500 M tokens) makes it potentially useful to search for language phenomena that are relatively rare. In this corpus, however, (morpho-)syntactic annotations (pos-tags, inflectional

[5] http://portal.clarin.nl/node/4182.
[6] http://portal.clarin.nl/node/1967.
[7] https://en.wikipedia.org/wiki/XPath.

properties, lemma) are restricted to tokens (i.e., occurrences of inflected word forms). It comes with its own interface,[8] which allows queries in (a subset of) the Corpus Query Processing Language[9] and via a range of interfaces of increasing complexity. The original interface was not directly suited for linking queries as proposed here. For that reason, an update of this interface has been made to make the relevant queries possible.[10]

As the corpora dealt with so far offer little or no morphological or phonological annotation, they cannot be used for the formulation of queries to accompany the Taalportaal texts on morphology and phonology. There is, however, a linguistic resource that is in principle extremely useful for precisely these types of queries, namely the CELEX lexical database (cf. Baayen et al. 1995) that offers morphological and phonological analyses for more than 100.000 Dutch lexical items. This database is currently being transferred from the Nijmegen Max Planck Institute for Psycholinguistics (MPI) to the Leiden Institute for Dutch Lexicology (INL). It has its own query language, which implies that Taalportaal queries that address CELEX will have to have yet another format, but again, the Taalportaal user will not be bothered with the gory details.

As was mentioned above, the Frisian language – the other official language of the Netherlands, next to Dutch – is described in the Taalportaal as well, parallel to Dutch. Although there is no lack of digital linguistic resources for Frisian, internet accessibility is lagging behind. This makes it difficult at this point to enrich the Frisian parts of the Taalportaal with queries. It is hoped that this CLARIN project will stimulate further efforts to integrate Frisian language data in the research infrastructure.

Since the links with the queries always go via the corpus search applications' *front-ends*, the Taalportaal user will, when a link has been clicked, be redirected not only to actual search results but also to a corpus search interface. The user can, if desired, adapt the query to better suit his/her needs, change the corpus being searched, search for constructions or sentences that diverge in one or more aspects (features) from the original query, or enter a completely new one. Most applications used (viz. PaQu, GrETEL, and OpenSONAR) have multiple interfaces differing in pre-supposed background knowledge of the user, and we believe that such options will actually be used. In this way, the enrichment of the Taalportaal as described here not only provides linguist users with actual corpus examples of linguistic phenomena, but may also have an educational effect of making the user acquainted with the existing corpus search interfaces.

## 5 An example

To get the gist of our approach, we will discuss a little example here. The beginning of the Taalportaal's chapter on nominal complements of adpositions[11] discusses the fact that both full noun phrases and bare nouns are possible as complements of prepositions. This is explained in terms of referentiality: in the variant with the full noun phrase *Jan werkt op het kantoor* (Jan works at the office) 'Jan is employed at the office' the noun phrase *het kantoor* 'the office' just refers to a building, and it is claimed that Jan is working there, whereas in the variant with a bare noun *Jan werkt op kantoor* (John works at office) 'Jan is an office employee', the prepositional phrase *op kantoor* 'at office' does not refer to a specific location. Figure 3 shows the relevant fragment.

[8] OpenSONAR via http://portal.clarin.nl/node/4195.
[9] Cf. http://cwb.sourceforge.net/files/CQP_Tutorial.
[10] https://portal.clarin.inl.nl/opensonar_whitelab/search/.
[11] http://taalportaal.org/taalportaal/topic/link/syntax__Dutch__adp__adp2__p2_compl.2.1.xml.

## 2.1. Nominal complements

Complements of adpositions are normally noun phrases. A distinction must be made between noun phrases with a determiner and (singular) bare noun phrases, that is, noun phrases without a determiner. As is to be expected, the first are normally referential in nature; the noun phrase *het kantoor* 'the office' in (1a) just refers to a building, and it is claimed that Jan is working there. The bare noun phrase *kantoor* in (1a'), on the other hand, does not refer to a specific building, and the PP does not refer to a specific location; instead, it is claimed that Jan has an occupation that in some way is related to the noun: he may be an office or administrative worker. Similarly, (1b) expresses that Jan is located at the office, while (1b') simply expresses that Jan is at work.

*Example 1*

a. Jan werkt op het kantoor.
Jan works at the office
'Jan is employed at the office.'

a'. Jan werkt op kantoor.
Jan works at office
'Jan is an office employee.'

b. Jan zit op dit moment op het kantoor.
Jan sits at this moment at the office
'Jan is at the office at this moment.'

b'. Jan zit op dit moment op kantoor.
Jan sits at this moment at office
'Jan is at work at this moment.'

Figure 3: Nominal complements of adpositions.

The b-examples are enriched with corpus queries, as indicated by the little icons. Clicking the first icon opens a pop-up window, illustrated in Figure 4:

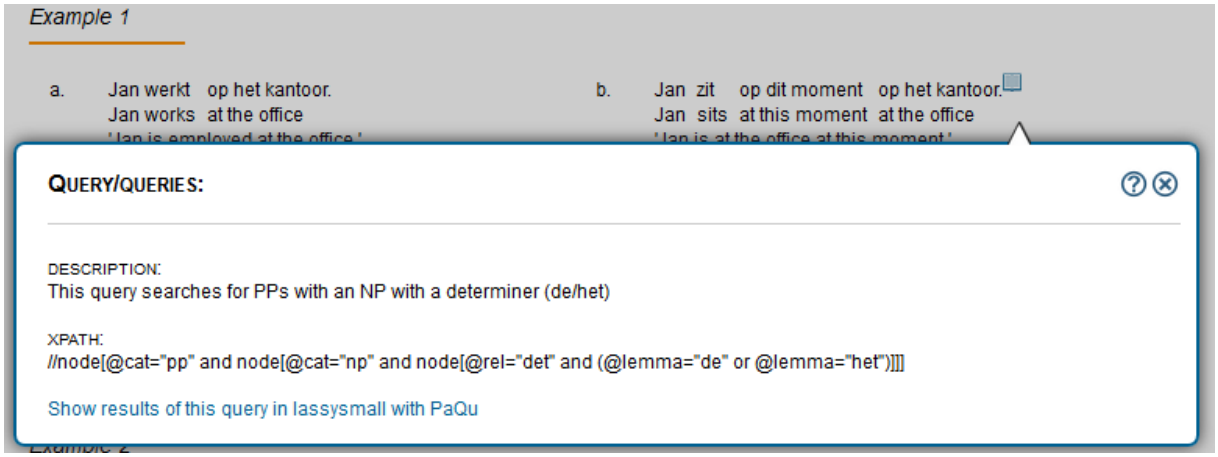



Figure 4: The query pop-up window.

The pop-window offers a brief description, showing the annotator's interpretation of the text fragment, and the exact query in XPATH. The final line of the window is a direct link to the exact query. Clicking this link opens a new window in the user's internet browser that shows the result of the query, as illustrated in Figure 5.

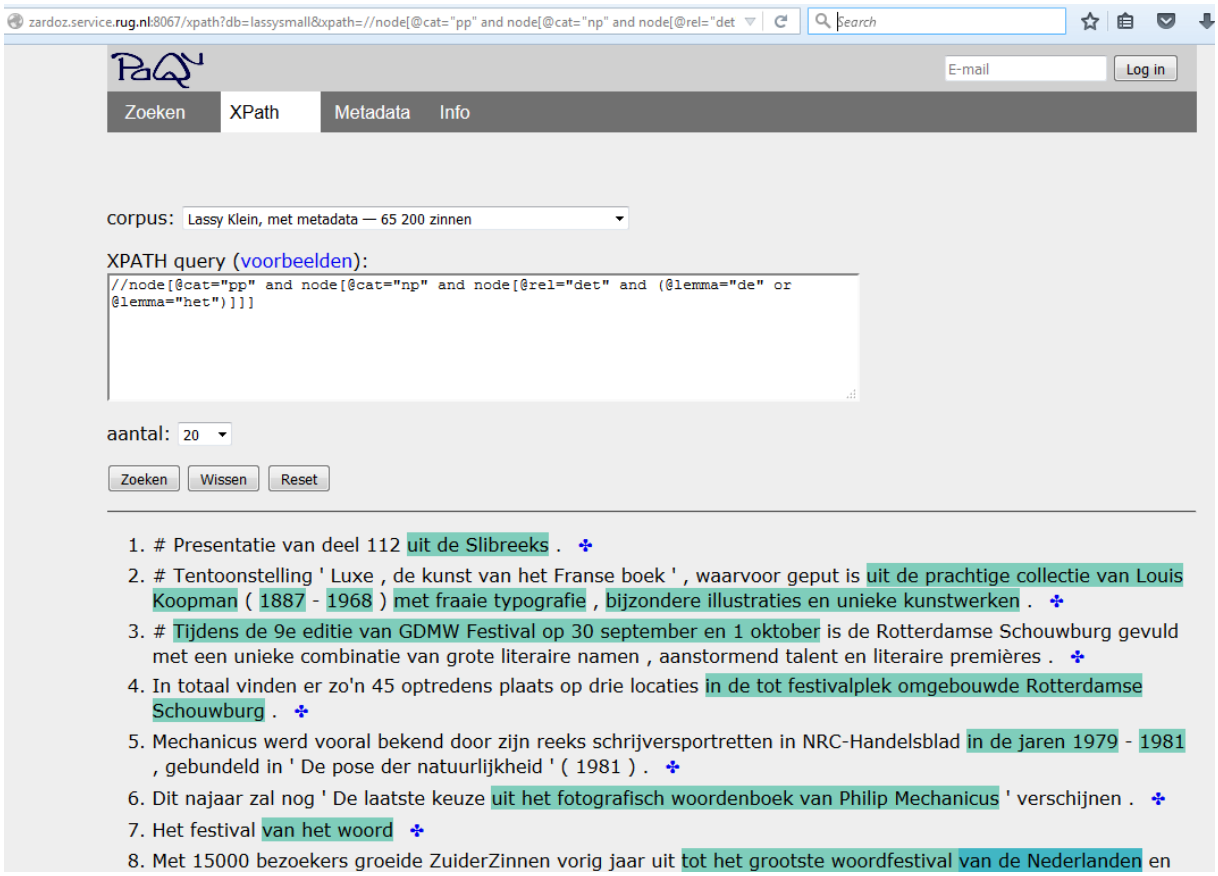



Figure 5: Result of the query in Figure 4.

The researcher can study the examples, download them to his/her own computer, and/or edit the query if the result is not completely satisfactory.

## 6 Evaluation

After completion of approximately 1200 queries that cover the subchapters of the SoD on complementation and modification of adjectives and adpositions, we have learned that creating suitable queries for a given fragment from the SoD requires creativity and careful experimentation, tuning, and documentation (cf. Bouma et al. (2015) for details and statistics). Construction of queries is far from deterministic, that is, different annotators will have different opinions concerning the most suitable query for a given example or phenomenon. In a surprisingly high number of cases, there are mismatches (in constituent structure, in part-of-speech) between the presentation in the SoD and the treebank annotation. While this makes the development of queries harder, it also underlines the value of the current project: by systematically exploring the way various linguistic examples are annotated in the treebank, we provide a starting point for further corpus exploration for users that have a general linguistic interest but who are not necessarily experts on Dutch treebank annotation.

The manually verified treebanks almost always provide sufficient examples of basic word order patterns for queries that are not restricted to a specific adjective or preposition. For queries that search for a specific lexical head or for less frequent word order patterns, the Lassy Large treebank usually has to be used. In that case, users must be prepared to see also a certain number of false hits. However, there are also examples in the SoD that cannot be found in a 700M word corpus. The conclusion that such word orders are not found in the language would be too strong, but it might be a starting point for further research (i.e. *does this construction occur only in certain registers or discourse settings?*) or for an alternative analysis (i.e. *do these cases really involve adjectives?*).

During the process of formulating corpus queries, the student assistants also reported to have run into serious problems:

- There is a certain "mismatch" between the phenomena described by linguists and the phenomena found most often in the wild.
- There is a mismatch between grammar formalisms used by grammarians and the grammar formalisms used by corpus linguists (generative style in the case of Broekhuis, dependency style in the Dutch corpora used).
- The grammars use more semantics than can be handled by/is encoded in the corpora.

These problems are not without scientific interest. Annotated corpora deal with types of annotation that are encoded relatively easily without too many errors. This implies, among other things, that a lot of the semantic subtleties discussed in grammars are not addressed in current corpora. Moreover, the language described by grammar (albeit called "descriptive") turns out to be not exactly the same as language as covered by corpora. This holds for the corpora used in the experimental project described in this project – although they deal both with spoken and written language varieties – and it will probably hold for all corpora: "Grammars describe the things grammarians are used to describe", and for good reasons, albeit often biased through discussions (and fashions) in the theoretical literature. Parasitic gaps are a prime example of an extremely rare phenomenon (in the wild) with very serious theoretical consequences (cf. Engdahl 1983, Phillips 2006).

## 7 Concluding remarks

We have described goals and methods of CLARIN-NL15-001, a co-operation project to enrich the on-line Taalportaal (Language Portal) grammatical database with intelligent links that take the form of annotated queries in a number of on-line language corpora and an on-line linguistic morphophonological database. The project contributes to the research infrastructure for linguistics and related scientific disciplines in various ways, since

- It provides users with actual corpus examples for linguistic phenomena described in Taalportaal;
- It points out the existence and usefulness of search interfaces developed in the CLARIN infrastructure such as PaQu, GrETEL and OpenSONAR to linguists;
- By redirecting the user to these front-ends, it stimulates the further use of these applications in the CLARIN infrastructure for modifying queries or submitting new queries. Together with the multiple interfaces of most of these applications, this may also have a significant educational role.

## 8 Acknowledgements

The Taalportaal project is a collaboration of the Meertens Institute, the Fryske Akademy, the Institute of Dutch Lexicology and Leiden University, funded, to a large extent, by the Netherlands Organisation for Scientific Research (NWO Groot 175.010.2009.003).

The project Enriching a grammatical database with intelligent links to linguistic resources (TPC) is a collaboration of Groningen University, Utrecht University, the Institute for Dutch Lexicology (INL), Taalmonsters and the Meertens Institute. It is financed by a grant of CLARIN-NL (NL-15-001: TPC).

## References

Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, & Frank Van Eynde. 2013. Example-Based Treebank Querying with GrETEL – now also for Spoken Dutch. In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. NEALT Proceedings Series 16, pp. 423–428.

R. Harald Baayen, Richard Piepenbrock, & L. Gulikers. 1995. *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Gosse Bouma, Marjo van Koppen, Frank Landsbergen, Jan Odijk, Ton van der Wouden, & Matje van de Camp. 2015. Enriching a Descriptive Grammar with Treebank Queries. In: *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pp. 13–25.

Hans Broekhuis, Norbert Corver, Marcel den Dikken, Evelien Keizer, & Riet Vos. 2012. *Syntax of Dutch*. Amsterdam: Amsterdam University Press, 2012–16 (7 volumes).

Elisabet Engdahl. 1983. Parasitic gaps. *Linguistics and Philosophy* 6, pp. 5–34.

Walter Haeseryn, Kirsten Romijn, Guido Geerts, Jaap de Rooij, & Maarten C. van den Toorn (eds.). 1997. *Algemene Nederlandse Spraakkunst*. Groningen and Deurne: Martinus Nijhoff and Wolters Plantijn. 2nd rev. ed. (2 vols.).

Jack Hoeksema. 2013. Review of: Syntax of Dutch. Noun and Noun Phrases vols. 1 and 2. *Lingua*, 133, pp. 385–390.

Frank Landsbergen, Carole Tiberius, & Roderik Dernison. 2014. Taalportaal: an online grammar of Dutch and Frisian. In Nicoletta Calzolari et al. (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland, pp. 26–31. ELRA.

Gertjan van Noord, Ineke Schuurman, & Vincent Vandeghinste. 2006. Syntactic Annotation of Large Corpora in STEVIN. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06),* pp. 1811–1814. ELRA.

Jan Odijk. 2015. Linguistic Research with PaQU. *Computational Linguistics in The Netherlands Journal* 5, pp. 3–14.

Nelleke Oostdijk, Martin Reynaert, Veronique Hoste, Ineke Schuurman. 2013. The Construction of a 500 Million Word Reference Corpus of Contemporary Written Dutch. In P. Spyns and J. Odijk (eds.): *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Springer Verlag, pp. 219–247.

Colin, Phillips. 2006. The real-time status of island phenomena. *Language* 82, pp. 795–823.

Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra, and Ton van der Wouden. 2003. CGN, an annotated corpus of spoken Dutch. In Anne Abeillé, Silvia Hansen-Schirra, and Hans Uszkoreit (eds.): *Proceedings of 4th International Workshop on Language Resources and Evaluation*, Budapest, pp. 340–347.

Ineke Schuurman. 2015. Concept revival: from ISOcat to CLARIN Concept Registry. *CLARIN News* 7 January 2015. https://www.clarin.eu/news/concept-revival-isocat-clarin-concept-registry.

Ton van der Wouden, Ineke Schuurman, Machteld Schouppe, and Heleen Hoekstra. 2003. Harvesting Dutch trees: Syntactic properties of spoken Dutch. In *Computational Linguistics in the Netherlands 2002. Selected Papers from the Thirteenth CLIN Meeting*, ed. by Tanja Gaustad, pp. 129–141. Amsterdam/New York: Rodopi.