Selected Papers from the CLARIN 2014 Conference

CLARIN
Common Language Resources and Technology Infrastructure

October 24-25, 2014
In Soesterberg (the Netherlands)

Editor: Jan Odijk

Linköping Electronic Conference Proceedings

Selected Papers from the CLARIN 2014 Conference

Jan Odijk (Editor)

October 24-25, 2014

In Soesterberg (The Netherlands)

Published by

Linköping University Electronic Press, Sweden

Linköping Electronic Conference Proceedings # 116

NEALT Series: NEALT Proceedings Series 28

(eISSN 1650-3740 ; ISSN 1650-3686; ISBN 978-91-7685-954-4)

Cover Design: Linda Stokman-Beijer; Photograph: Kontakt der Kontinenten

Preface

Steven Krauwer
CLARIN ERIC Executive Director

We hope that this volume will be the first in a series of publications where members of the CLARIN community share their experiences and their results with their colleagues and with the humanities and social sciences research communities at large.

CLARIN, the Common Language Resources and Technology Infrastructure, is a European Research Infrastructure for the humanities and social sciences, with a specific focus on language in all its forms, and in the many roles it plays in our society and in research, be it as carrier of information, record of the past, means of human expression or object of study.

CLARIN provides a broad range of services, such as access to language data and tools to analyze data, and offers to deposit research data, as well as direct access to knowledge about relevant topics in relation to (research on and with) language resources.

The CLARIN community comprises a variety of groups of people, such as those who build and maintain the infrastructure, those who provide data and tools, and most importantly: those who make use or intend to make use of the CLARIN infrastructure to facilitate and innovate their research. In order to ensure convergence and cross-fertilization between and amongst these groups it is important that they get together and exchange problems and solutions, successes and failures, things yet to be done, and inspiring examples of the capabilities of the infrastructure.

The annual CLARIN conference is one of the places where members of the CLARIN community meet. These proceedings present a selection of the highlights of the 2014 annual conference and we hope that they will not only serve to keep people inside CLARIN informed of what is happening, but that they will also reach a much broader circle of researchers who could benefit from what CLARIN has to offer, or who could contribute to the further development of the CLARIN infrastructure.

Introduction to the CLARIN 2014 Selected Papers

Jan Odijk
Chair Programme Committee

The CLARIN Annual Conference has been organized since 2012. Earlier versions of this conference allowed contributions on invitation only. The CLARIN 2014 Conference is the first one in which there was an open call for contributions and review of the contributions submitted. It is also the first one to result in a publication with Selected Papers from the conference.

The CLARIN Annual Conferences intentionally have an inclusive character: they aim to bring to-gether as many people working on or using CLARIN as possible. For this reason, a light procedure for submissions was adopted. Submissions for this conference are in the form of extended abstracts (2-4 pages), which are evaluated by at least two Programme Committee Members from a different country than the primary author. Presentations and posters elaborating the accepted extended abstracts were presented at the CLARIN 2014 Conference, often in combination with a demonstration of the system or software developed. There were 34 submissions, 29 of which were accepted, which clearly illus-trates the inclusive character of the conference.

The CLARIN 2014 Conference consisted of 13 oral presentations, 16 posters, often combined with a system demonstration, and 7 system demonstrations related to oral presentations. The Programme Committee also accepted 3 posters from the (at that time) new or 'almost' CLARIN members Finland, Sweden, and Slovenia. The keynote speech was given by Jan Rybicki (Jagiellonian University, Poland) on *Visualising Literature: Trees, Maps and Networks*.

The accepted submissions came from 8 countries[1], with Germany (9) and the Netherlands (8) in the lead, and the Czech republic (4) and Denmark (3) following. This distribution is to be expected if one takes into account the start dates and budgets of the various national projects. As to the topics covered, the *construction of the infrastructure* (8) was most prominent, and many topics directly related to this such as *interoperability* and *metadata* were also covered. *Data* and *tools* (4 each) occupied a prominent place. The presence of presentations and posters on the actual *use* of the infrastructure to carry out scientific research (5) shows that CLARIN is beginning to be used by researchers even though it is still under construction. We expect that the proportion of users in the CLARIN conferences to follow will steadily increase.

Many contributions (16) were not bound to any specific scientific discipline. Not surprisingly, *linguistics* was the most dominant scientific discipline (11) in the remaining contributions, but *history, philology, political science,* and *speech recognition* were also present.

Authors of accepted extended abstracts were invited to elaborate their extended abstract into a full paper. The papers submitted were also evaluated by at least two independent Programme Committee members. Eleven full papers were submitted, nine of which were accepted. The papers accepted are included in this volume. A wide range of topics is covered, including *technical infrastructural issues* (metadata (3 papers) and user delegation), *education and training, data curation, data mining* and *corpus exploration* (2 papers) in both *textual* and *multimodal* data, *data curation,* and *user research*. The first authors come from Germany (4 papers), the Netherlands (3 papers), Norway and Poland (1 paper each).

I hope you will enjoy reading the contributions to the *CLARIN 2014 Selected Papers*.

[1] Based on the country of the first author.

Programme Committee

- Koenraad De Smedt, University of Bergen, Norway
- Eva Hajičová, Charles University Prague, Czech Republic
- Erhard Hinrichs, University of Tübingen, Germany
- Bente Maegaard, University of Copenhagen, Denmark
- Karlheinz Mörth, Austrian Academy of Sciences, Austria
- Jan Odijk, Utrecht University, the Netherlands (Chair)
- Maciej Piasecki, Wrocław University of Technology, Poland
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Remco van Veenendaal, Dutch Language Union, The Netherlands/Flanders
- Kadri Vider, University of Tartu, Estland

Table of Contents

Using Data Mining and the CLARIN Infrastructure to Extend Corpus-based Linguistic Research

Thomas Bartz
TU Dortmund University
Department of German Language and Literature
44227 Dortmund, Germany
thomas.bartz@tu-dortmund.de

Christian Pölitz
TU Dortmund University
Artificial Intelligence Group
44227 Dortmund, Germany
christian.poelitz@tu-dortmund.de

Katharina Morik
TU Dortmund University
Artificial Intelligence Group
44227 Dortmund, Germany
katharina.morik@tu-dortmund.de

Angelika Storrer
Mannheim University
Department of German Philology
68131 Mannheim, Germany
astorrer@mail.uni-mannheim.de

Abstract

Large digital corpora of written language, such as those that are held by the CLARIN-D centers, provide excellent possibilities for linguistic research on authentic language data. Nonetheless, the large number of hits that can be retrieved from corpora often leads to challenges in concrete linguistic research settings. This is particularly the case, if the queried word-forms or constructions are (semantically) ambiguous. The joint project called 'Corpus-based Linguistic Research and Analysis Using Data Mining' ("Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining" – 'KobRA') is therefore underway to investigating the benefits and issues of using machine learning technologies in order to perform after-retrieval cleaning and disambiguation tasks automatically. The following article is an overview of the questions, methodologies and current results of the project, specifically in the scope of corpus-based lexicography/historical semantics. In this area, topic models were used in order to partition search result KWIC lists retrieved by querying various corpora for polysemous or homonym words by the individual meanings of these words.

1 Introduction and Project Background

Large digital corpora of written language, such as those that are held by the CLARIN-D centers, provide excellent possibilities for linguistic research on authentic language data (McEnery et al., 2006; Lüdeling and Kytö, 2008; Lüdeling and Kytö, 2009). The size of the corpora allows for remarkable insights into the distribution of notable language usage phenomena with respect to time and/or domain-specific aspects. Not the least thanks to the efforts being done in CLARIN, are analyzing and query tools becoming more and more sophisticated, and thus, enabling researchers to search for word forms or constructions and filter the results with regard to part of speech types or morphosyntactic aspects. Despite these advances, the large number of hits that can be retrieved from corpora often leads to challenges in concrete linguistic research settings. This is particularly the case, if the queried word forms or constructions are (semantically) ambiguous. Researchers in linguistics do not usually examine word forms, but instead the terms representing the relations of word forms and their meanings. It is for this reason that word form-based filtering carried out by the current query tools is insufficient in many cases and leads to an unpredictable number of false positives. Depending on the amount of data, in-

tense manual effort is required for cleaning and disambiguation tasks (Storrer, 2011). Many research questions cannot even be addressed for this reason.

The joint project called 'Corpus-based Linguistic Research and Analysis Using Data Mining' ("Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining" – 'KobRA') is therefore underway to investigating the benefits and issues of using machine learning technologies in order to perform after-retrieval cleaning and disambiguation tasks automatically. To this end, German linguists (Universities of Dortmund and Mannheim), computational linguists (Berlin-Brandenburg Academy of Sciences and Humanities – BBAW, Institute for the German Language – IDS, University of Tübingen) and computer scientists (University of Dortmund) closely collaborate on concrete corpus-based case studies in the fields of lexicography, diachronic linguistics and variational linguistics. The case studies reflect the research actually carried out in these fields and are related to the specific research activities of the project participants. Three major German corpus providers, all CLARIN-D centers (BBAW, IDS, Tübingen University; see above), take part in the project, providing the corpus data and also plan to integrate the project results into the existing infrastructures. The data mining processes, that are made available in RapidMiner (formerly: 'YALE', Mierswa et al., 2006), one of the most widely used data mining environments, operate on search result KWIC lists derived from the corpora. These go beyond a mere search and can be used in order to filter or structure the search results, as well as to simplify the further processing of the data, where necessary, to target specific research questions (e.g. through annotation).

The following article is an overview of the questions, methodologies and current results of the project, specifically in the scope of corpus-based lexicography/historical semantics. In this area, topic models were used in order to partition search result KWIC lists retrieved by querying various corpora for polysemous or homonym words by the individual meanings of these words. The utility and the conditions of these methods are illustrated based on case studies on example words that are of interest to a linguist. German homonyms and polysemes of various parts of speech are presented, that different corpora were queried for. As topic models operate independently from language, it was thought that this method would be suitable for languages other than German as well. Therefore, experiments were also run with English language data.

2 Scope: Research on the Semantic Change of Words

The semantic change of words is interesting for linguistics in two respects: lexicographers and historical semantic researchers. Lexicographers follow the evolution of words in order to construct adequate lexicographic descriptions for example in order to update existing dictionary entries (Storrer, 2011; Engelberg and Lemnitzer, 2009), while researchers in historical semantics explore the possibilities, conditions and consequences of semantic innovations (Fritz, 2012; Fritz 2005; Keller and Kirschbaum 2003). In both cases, the deciding factor in furthering knowledge is the availability of structured text-corpora that allow the use of a word to be tracked over broad time lines and genres. Although comprehensive synchronous and diachronic text corpora with meta data to occurrence dates and text types are available along with accessible retrieval and analysis tools, and especially in the framework of CLARIN, an extensive and automatic semantic annotation of corpora at current technological standards is not yet suitably possible (Storrer, 2011; Rayson and Stevenson, 2008). This means that until now corpus-based exploration of semantic changes of a word have to be manually disambiguated for individual detection. Therefore, the distribution and process of semantic change can presently only be described on the basis of few examples and a relatively small data corpus (Fritz 2005; Keller and Kirschbaum 2003).

3 Data Mining Approach: Disambiguation of KWIC Snippets

Instead of an exhaustive semantic annotation of large text corpora, it appears that it could be more promising to have a subsequent disambiguation of automatically generated KWIC snippets for a searched word retrieved via corpus query. This is also suggested through a series of preliminary results (see Section 4). Already a manual viewing of search results shows that the different meanings of a searched word are most easily recognized through the surrounding words. The usage of a word in a specific meaning evidently corresponds more frequently with occurrences of certain other words or linguistic structures in the environment of this word. Through data mining methods, this latent infor-

mation contained within a search result's context may be used for automatic disambiguation. For that purpose, all occurrences of a relevant word will be placed in context windows of a specific size and with help from word and co-occurrence statistics, distributions of the context words will be determined. These can then be regarded as representations of meanings. As a result, it will be possible to calculate the probability of the relevant word representing a certain meaning for every single context window. A major advantage of such methods that are inductively based on the related words' contexts is that this way unexpected or until now lexicographically unrecorded meanings can now be identified.

4 Related Work: Word Sense Induction (WSI)

The induction of semantic meaning in the area of data mining is already well researched. An early statistical approach was completed by Brown et al. (1991), Navigili (2009) provides a comprehensive overview on the current research. Brody and Lapata (2009) have shown that they obtained the best results with the help of Latent Dirichlet Allocation (LDA; Blei et al., 2003). In addition, they expanded their method to take into consideration various other context features besides the pure word occurrences (e.g. part of speech tags, syntax, etc.). Originally, LDA was used for thematic clustering of document collections. Navigli and Crisafulli (2010) have already shown this to also be useful for the disambiguation of small text snippets, for example when clustering the search results from a web search engine. Rohrdantz et al. (2011) showed the benefits of this method as a basis for the visualization of semantic change of example words from an English newspaper corpus, allowing them to observe the emergence of new meanings and reconstruct their development over time.

The approach proposed in this article differs from these previous works particularly through the application of LDA in search result KWIC snippets derived from queries in large text corpora. The results of a query in a (web) search engine usually correspond to (web) texts, which are closely connected thematically with the searched word. However, search results from a corpus search system are determined through the occurrences of the searched word throughout the corpus, regardless of the thematic relevance of the documents containing the words. In this way, the searched words generally occur in less normal and semantically less clear contexts. The text genre of belles-letters and of newspaper texts often include metaphorical usages. Based on Rohrdantz et al. (2011), KWIC snippets from different text type areas will be used, all of which – apart from one example – are in German.

The benefits and issues of using clustering methods like LDA for the automatization of disambiguation of the search results KWIC snippets derived from corpora are, as of yet, barely researched. In the context of CLARIN-D, there are corpora available to the KobRA project (details about queried corpora see Section 6), which include extensive linguistic (annotations of parts of speech and syntax) and document meta data (examples assigned to text genres and time frames). This is why the project also allows for insights relating to the questions of which attributes may improve the results of clustering methods, such as LDA, and how the KWIC snippets and attributes may ideally be represented for these methods.

5 Evaluated Data Mining Techniques and Environment

The data mining processes evaluated in the KobRA project are implemented as a plug-in in the data mining framework RapidMiner (formerly: 'YALE', Mierswa et al., 2006; see Figure 1). RapidMiner allows one to easily perform large scale data analysis and offers a plethora of methods to import, transform, and analyse textual data as well as to present and visualize the results of the analysis. Besides already available data mining methods for classification and clustering, additional methods were implemented for efficient feature extraction and calculation for large amounts of documents as well as for word sense disambiguation. The plug-in also includes methods to efficiently access linguistic data sources, as well as sophisticated methods to extract linguistic and document features (if available) from KWIC lists. An integrated annotation environment enables linguists to add additional annotations to the KWIC snippets and the words retrieved from the data sources.

For the disambiguation approach described and evaluated in this paper (see Sections 3, 7, 8), we implemented the Latent Dirichlet Allocation method (LDA; Blei et al., 2003; Steyvers et al., 2004; Blei and Lafferty, 2006). LDA models the probability distributions of the words and the snippets from the corpus query result lists. The probability distributions are scattered over a number of what are known as latent topics that correspond to different meanings of a queried word. Based on the words and word

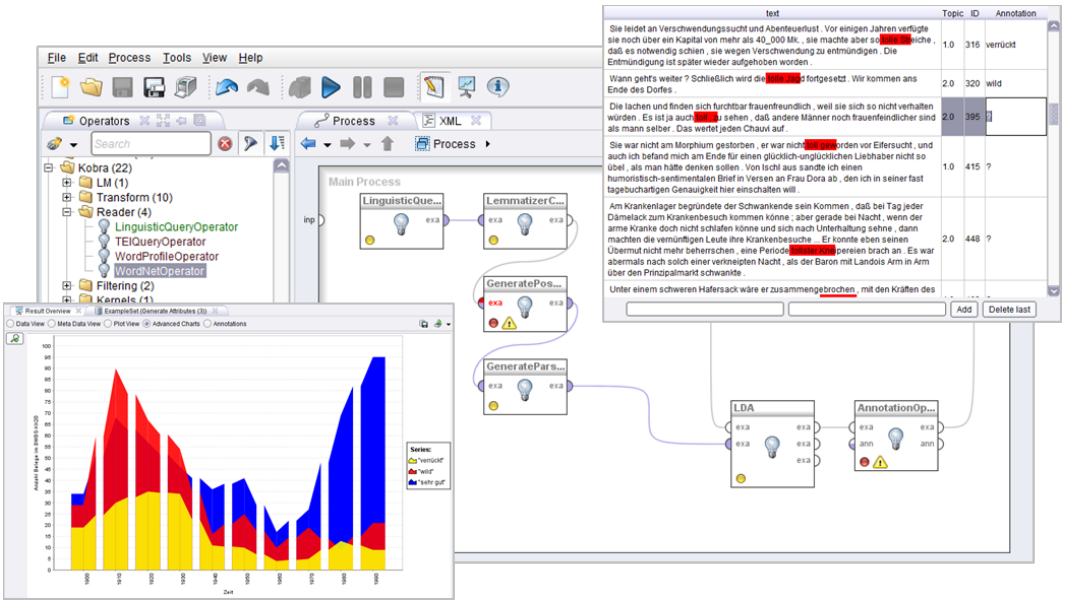


Figure 1: RapidMiner data mining framework and project plug-in.

co-occurrences in the snippets, LDA assigns those words that appear together into the same topics. These topics are then interpreted as meanings.

The probability distributions of the topics for a given word or snippet are multinomial distributions φ respectively θ. These distributions are drawn from a Dirichlet distribution Dir(β) respectively Dir(α) for the meta parameter α and β. The Dirichlet distribution is a distribution over distributions.

The estimation of the distributions is done via a Gibbs sampler as proposed by Griffiths and Steyvers (2004). The Gibbs sampler models the process of assigning a word or snippet to a certain topic based on the topic distributions of all other topics. This is a Markov chain process and converges to the true topic distributions for given words and snippets.

An important aspect, that is investigated, is the possibility to integrate further information into the generation of the topic models. Steyvers et al. (2004), for instance, integrate additional information like authorships of documents in the topic models. We use their approach to integrate information about the text genre classes the query result snippets are attributed to. This enables an additional investigation of how topics, words and snippets distribute over these classes. Moreover, the integration of the publication dates provided with the snippets are of interest to this study. Blei and Lafferty (2006), for example, introduced a dynamic topic model that facilitates analyzing the development of the found topics over time.

6 Words of Interest and Queried Corpora

For the case studies outlined in this article, we queried various corpora for a choice of words that are linguistically interesting, because they recently or over a long period of time have taken on new meanings, or their original meanings have changed. According to the assumed time period of the meaning changes, different corpora were queried. Moreover, we chose example words belonging to different parts of speech. Using this setting, we expect interesting insights in possible corpus- or word class-specific differences in the usefulness of the evaluated data mining techniques. The below examples are the basis for the following outlined experiments. Details about the corpora used can be found subsequently.

- Through the technical innovation of the 20th century, the noun “Platte” had a pronounced differentiation in its range of meanings. Along with the meaning *flaches Werstück* (flat workpiece) or *Teller* (plate), different uses gradually appeared: *fotografische Platte* (photographic plate), *Schallplatte/CD* (gramophone record/compact disk) oder *Festplatte* (hard disk). A search for the lemma of “Platte” in the DWDS core corpus of the 20th century results in 2886 KWIC snippets.

- During the commercial distrubition of the telephone in the 20s and 30s of the 20th century, a new meaning appeared for the verb “anrufen“ besides its original meaning *rufen/bitten* (to cry/to appeal to someone): that of *telefonieren* (to telephone). A search for the verb “anrufen“ in the DWDS core corpus of the 20th century results in 2085 KWIC snippets.
- Since the financial and bank crisis (circa 2007) the noun “Heuschrecke” has a new use, along with its original meaning *Grashüpfer* (locust), to now also describe persons involved in what is known as “Heuschreckenkapitalismus” (locusts capitalism). A search for “Heuschrecke“ in the DWDS newspaper corpus “Die ZEIT” results in 715 KWIC snippets.
- The adjective “zeitnah“ appears to have received a new prototypical meaning in the last 20 to 30 years, *unverzüglich* (prompt), while still retaining its original meaning of *zeitgenössisch* (contemporary)/*zeitkritisch* (critical of the times). A search for “zeitnah“ in the DWDS newspaper corpus “Die ZEIT” results in 597 KWIC snippets.
- The adjective “toll“ has had a remarkable meaning shift in the last century; its original meaning of *irre* (insane) changed to *ausgelassen/wild* (jolly/wild) and to its now positively attributed meaning *sehr gut* (very good). A search for the adjective “toll“ in the Tübingen Treebank of Diachronic German results in 5793 KWIC snippets, and a corresponding search in the DWDS core corpus of the 20th century results in 1745 KWIC snippets.
- The conjunction “da“ (as/because) was almost only used for temporal meaning in early records. Today, it is mostly used causally. A search for the conjunction “da“ in the Tübingen Treebank of Diachronic German results in 123496 KWIC snippets.
- The choice of the English noun “cloud” represents this aricle’s first attempt to use the proposed approach on none German language data. A new meaning appears to have evolved in the last decades with the emergence of large computer networks (clouds), next to the original meaning (*mass of condensed water, smoke, dust or other elements*). A search for “cloud“ in the corpora of the Leipzig Corpora Collection results in 1486 KWIC snippets.

The DWDS core corpus of the 20th century (DWDS-KK), constructed at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), contains approximately 100 million running words, balanced chronologically (over the decades of the 20th century) and by text genre (belles-lettres, newspaper, scientific and functional texts). The newspaper corpus “Die ZEIT” (ZEIT) covers all the issues of the German weekly newspaper “Die ZEIT” from 1946 to 2009, approximately 460 million running words (Klein and Geyken, 2010; Geyken, 2007).

The Tübingen Treebank of Diachronic German (TüBa-D/DC) is a syntactically annotated (constituent parse trees) corpus of selected diachronic language data from the German Gutenberg Project (http://gutenberg.spiegel.de/), a community-driven initiative of volunteers making copyright-free belles-lettres from 1210 to 1930 publicly available via web-interface. TüBa-D/DC that is hosted by the CLARIN-D Center at the Eberhard Karls University of Tübingen contains approximately 250 million running words (Hinrichs and Zastrow, 2012).

The Leipzig Corpora Collection (LCC) consists of corpora in different languages that contain randomly selected sentences from newspaper texts and a web sample (Quasthoff et al., 2006). We used the English corpus with language data from newspapers and the English Wikipedia, covering the time span from 2005 to 2010.

The corpus queries provide KWIC text snippets with occurrences of the investigated words (with including inflected forms). In addition, the publication dates and other document metadata (for the TüBa-D/DC: titles, author names; for the DWDS-KK: text genre classes) are given for each snippet.

7 Experiments and Evaluation

Accounting for our research question for the optimal representation of the KWIC snippets and our selection of example words and corpora (Section 6), eight evaluative treatments of the approach outlined in Section 5 were created. These can be systematically separated into the following aspects:

- **Queried word and part of speech:** noun, verb, adjective, or conjunction

- **Number of meanings:** two or more
- **Queried corpus:** corpus of contemporary German (DWDS-KK, ZEIT) or diachronic corpus (TüBa-D/DC, orthographically normalized)
- **Language of the corpus:** German or English
- **Number of KWIC-Snippets:** More or less than 1000 snippets

In addition, every treatment was tested to check which context size (20, 30, or 40 words) led to the best results for the relevant word. The following Table 1 shows an overview of the evaluative treatments for the outlined data mining techniques in Section 5.

Treatment	Word	Part of Speech	Meanings	Corpus	Language	Snippets	Context		
							20	30	40
1	Platte	noun	5	contemporary	German	> 1000	X	X	X
2	toll	adjective	3	contemporary	German	> 1000	X	X	X
3	anrufen	verb	2	contemporary	German	> 1000	X	X	X
4	Heuschrecke	noun	2	contemporary	German	< 1000	X	X	X
5	zeitnah	adjective	2	contemporary	German	< 1000	X	X	X
6	toll	adjective	2	diachronic	German	> 1000	X	X	X
7	da	conjunction	2	diachronic	German	> 1000	X	X	X
8	cloud	noun	3	contemporary	English	> 1000	X	X	X

Table 1: Evaluation treatments.

For the evaluation purposes, 30 percent of the retrieved KWIC snippets for the queried words were disambiguated manually by two independent annotators. Table 2 shows the obtained inter-annotator-agreement (kappa: Cohen, 1960):

Treatment	Word	Agreement
1	Platte	0.82
2	toll	0.76
3	anrufen	0.97
4	Heuschrecke	0.98
5	zeitnah	0.91
6	toll	0.71
7	da	0.75
8	cloud	0.92

Table 2: Inter-annotator-agreement of the manual disambiguation.

The automatic disambiguation approach was evaluated based on the manually disambiguated data sets. Therefore, topic models (see Section 5) were generated to extract the meanings of the queried words' occurrences and the results were compared to the labels attributed by the annotators. As a measure of reliability for the automatic disambiguation, we use one of the standard measures used to estimate the goodness of a word-sense disambiguation result, the F_1 score. The F_1 score is the weighted average of the disambiguation results' precision and recall in relation to the given annotations. This and further evaluation methods are described by Navigli and Vanella (2013).

8 Evaluation Results

8.1 Reliability of the automatic disambiguation using LDA

The following tables show the results achieved using the above described approach. The Tables 3-8 list the evaluation scores for the investigated treatments:

"Platte"		flat workpiece	plate	photographic plate	gramophone record/compact disk	hard disk
F_1 for context (words)	20	0.800	0.800	0.667	0.287	0.857
	30	0.998	0.875	0.500	0.381	0.988
	40	0.733	0.600	0.750	0.353	0.800

Table 3: Results for treatment 1.

"toll"		insane	jolly/wild	very good
F_1 for context (words)	20	0.519	0.571	0.167
	30	0.714	0.615	0.632
	40	0.625	0.667	0.500

Table 4: Results for treatment 2.

"anrufen"		to cry/to appeal to someone	to telephone
F_1 for context (words)	20	0.727	0.667
	30	0.800	0.800
	40	0.909	0.889

Table 5: Results for treatment 3.

"Heuschrecke"		locust	person
F_1 for context (words)	20	0.857	0.842
	30	0.800	0.933
	40	0.667	0.727

Table 6: Results for treatment 4.

"zeitnah"		prompt	contemporary/critical of the times
F_1 for context (words)	20	0.727	0.667
	30	0.888	0.800
	40	0.895	0.818

Table 7: Results for treatment 5.

"toll"		insane	jolly/wild
F_1 for context (words)	20	0.526	0.571
	30	0.625	0.750
	40	0.556	0.636

Table 8: Results for treatment 6.

"da"		temporal	causally
F_1 for context (words)	20	0.471	0.556
	30	0.353	0.529
	40	0.400	0.611

Table 9: Results for treatment 7.

"cloud"		mass of condensed water, etc.	computer network	name
F_1 for context (words)	20	0.526	0.500	0.471
	30	0.783	0.631	0.615
	40	0.467	0.545	0.684

Table 10: Results for treatment 8.

The results demonstrate that the advised task of automatic disambiguation of KWIC snippets retrieved from corpus queries (see Section 3) yield highly positive outcomes using the approach outlined above (see Section 5). In the best case scenario the average F_1 scores for the reliability of the method is around 0.732. However, depending on the searched word and preferred meaning the values varied in the range between 0,381 and 0,998 (again in the best case scenario). The generality of this method is therefore difficult to hypothesize. Still, according to the above formulated systematization of differences in the treatments (see Section 7) the following trends were established:

Word Form

It appears that the automatic disambiguation of nouns, verbs, and adjectives of the examined examples had, essentially, the same success rates. Similarly good values resulted from the example "Heuschrecke" (see Table 6) as with "zeitnah" (see Table 7) or "anrufen" (see Table 5). Nouns had the highest values (see also Table 3). The finer meaning differences of the conjunction "da" were not satisfactorily recognizable (see Table 9). The method is most promising in terms of content words. This is to be expected because of their function as semantic references. The applicability of this method in relation to grammatical words should be further investigated.

Number of Meanings

It appears, however, that the number of meanings of the examined examples systematically influenced the results. The method revealed lower results for "toll" (see Table 4) and "cloud" (see Table 10) than for examples that had only two meanings. This was also true for single meanings of "Platte" (see Table 3), while for the others the highest values were obtained. In essence, it appears that various meanings are differently identifiable.

Corpus and Language

At first glance, it appears that the chosen corpora (contemporary German vs. diachronic, German vs. English) had relatively similar results with the automatic disambiguation. The snippets' results for "toll" from the DWDS-KK (see Table 4) are comparable to those from TüBa-D/DC (see Table 8); this was also true for the English example "cloud" (see Table 10). In this respect, the evaluative success was expected as the texts from the TüBa-D/DC all lie within the orthographic normative form. More study is needed to determine if this method is also suitable for diachronic corpora with non-normative orthographic language data.

Number of Snippets and Size of the Contexts

Although the number of KWIC snippets used (500-1000 vs. 1000-5000) for each example appeared to have no systematic effect on the results – "zeitnah" (see Table 7) and "Heuschrecke" (see Table 6) were similarly well disambiguated, as were "Platte" (see Table 3), "toll" (see Table 8) or "anrufen" (see Table 5) – it was demonstrated that the method was most useful when the range of the contexts was 30 words before and after the examined word. Yet, it appears that for the verb "anrufen" (see Table 5) the most promising results came from the largest context. A reason for this could be that the verb in its function is more correlated with the sentence as a larger unit, while nouns and adjectives are already specified by their proximal contexts. This is supported by the slightly better results from the primarily adverbial used "zeitnah" (see Table 7) in the treatment with a context of 40 words.

8.2 Application for Research on the Semantic Change of Words

Using the automatic disambiguation easily allows the occurrences of single meanings of examined words to be identified and visualized. From the figures (see Figure 2-6) one can see the benefits of the integration of the query snippet's publication dates into the generation of the topic models: Researchers investigating semantic change are enabled to easily track the use of disambiguated word forms over time:

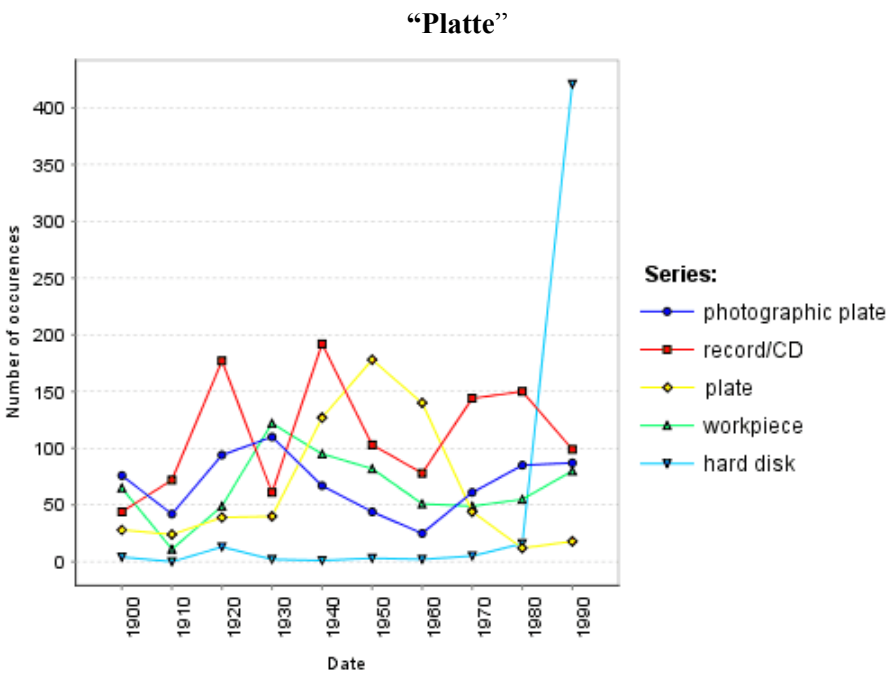


Figure 2: Occurrences of the word "Platte" with the meanings *flat workpiece*, *plate*, *photographic plate*, *gramophone record/compact disk*, *hard disk* in the decades of the 20th century.

The evolution of the meanings of "Platte" is illustrated traceable by Figure 2. The use of the meaning *hard disk* increased dramatically in the 90s, while the other meanings had a more uniform increase in usage in the different phases. The phases of more prevalent usage (e.g. the meaning *plate* in the 40s-60s or the meaning *photographic plate* in the 80s and 90s) are grounds for a more exact studies that would take the underlying KWIC snippets into account. With this in mind, the development of interactive visualisation, which is linked with the corpus base, would further simplify corpus based research on semantic change.

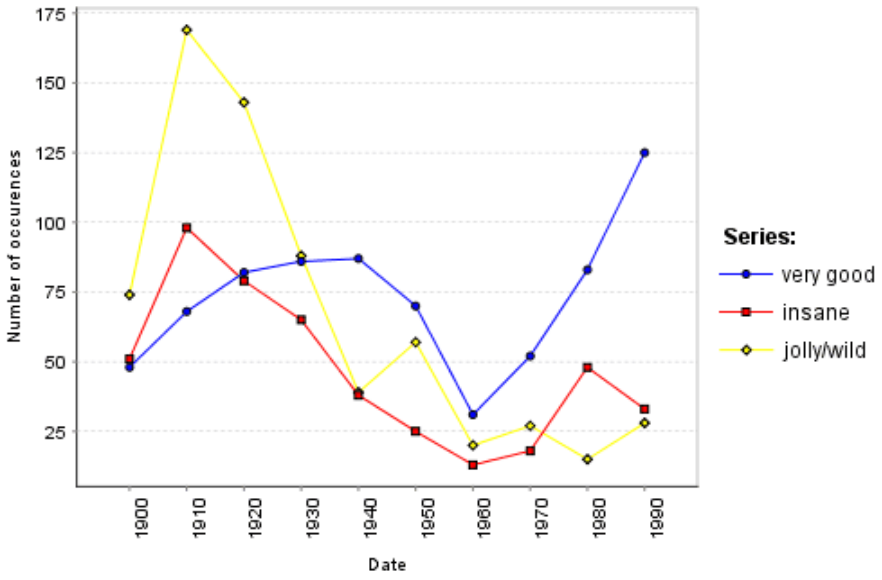


Figure 3: Occurrences of the word "toll" with the meanings *insane, jolly/wild, very good* in the decades of the 20th century.

Figure 3 clearly displays the semantic development of the word “toll“ during the 20th century. To the same degree that the older meanings *insane* and *jolly/wild* dropped in frequency, so did the newer meaning *very good* become more and more prominent.

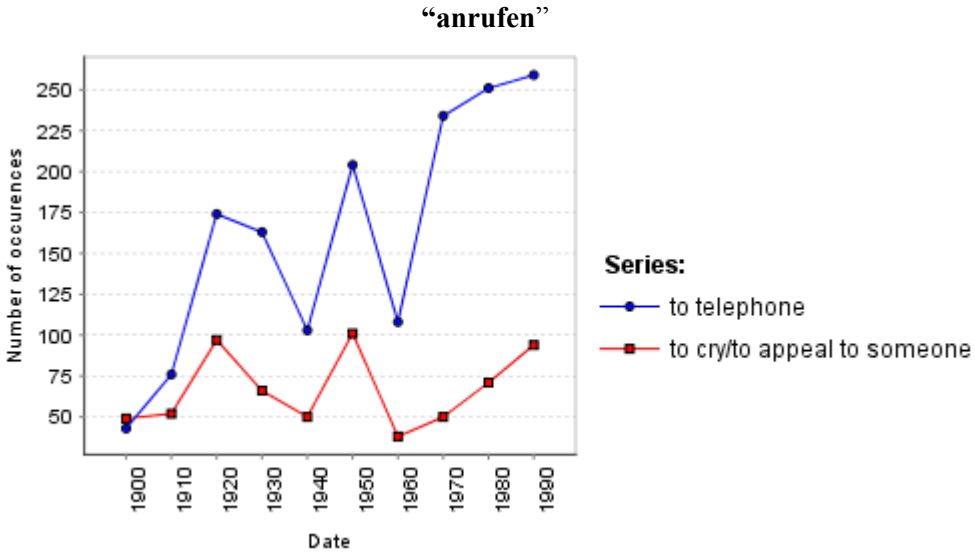


Figure 4: Occurrences of the word “anrufen” with the meanings *to cry/to appeal to someone, to telephone* in the decades of the 20th century.

Figure 4 shows that the strong increase in the use of the word “anrufen“ with the meaning *to telephone* occurred parallel to the commercial spread of telephones. The serrated frequencies that appear for both meanings between 1930 and 1970 could point to an irregularity in the balance of the corpus basis. This, again, underscores the need for a closer investigation of the underlying KWIC snippets.

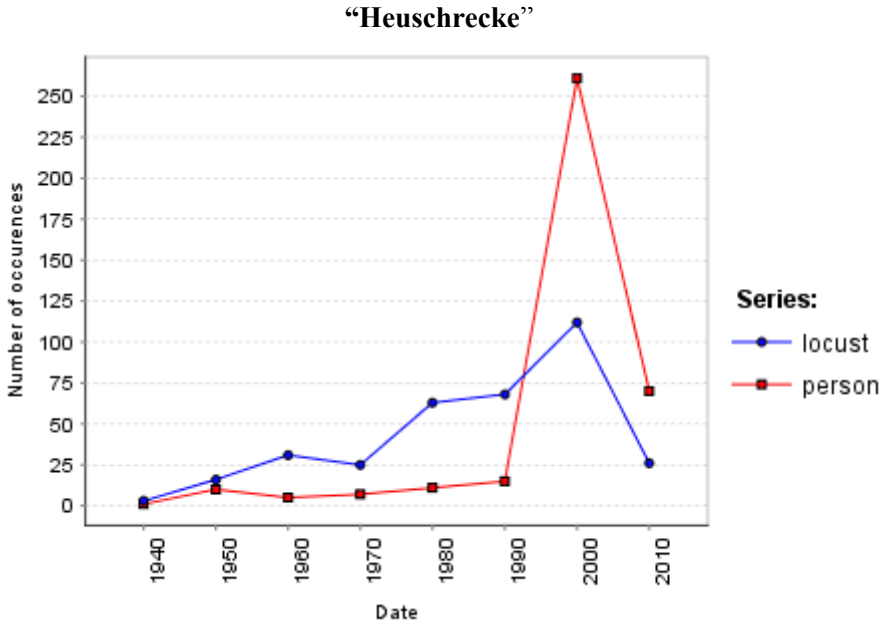


Figure 5: Occurrences of the word “Heuschrecke” with the meanings *locust, person* in the time span 1940-2010.

Figure 5 clearly shows a dramatic increase in the use of “Heuschreke“ with the meaning *person* in the 2000s, in the decades during the international financial and bank crisis. In the decade of the 2010s, there is a noticeable decline in the frequency of this use. However, the markedly smaller amount of records for this decade, in contrast to the others, could explain this discrepancy.

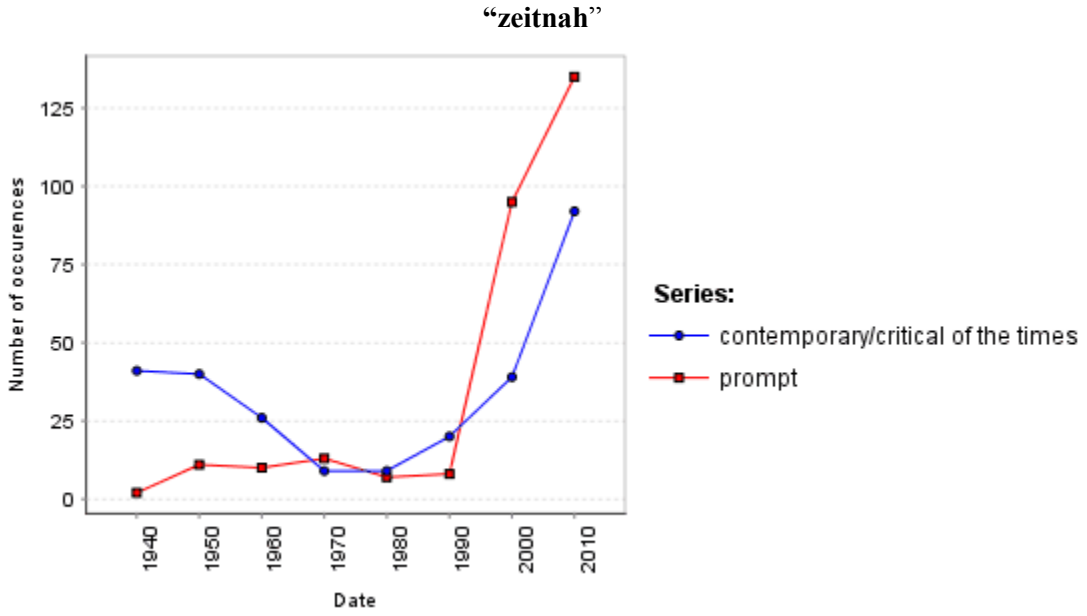


Figure 6: Occurrences of the word "zeitnah" with the meanings *prompt, contemporary/critical of the times* in the time span 1940-2010.

Finally, Figure 6 shows the sudden development, starting in the 2000s, of the meaning *prompt* as a new prototypical meaning for "zeitnah". What is interesting about this, is that at the same time there is a rise in the use of its older meaning *contemporary/critical of the times*. In order to check if this is accurate, or if this is just a cumulation of false meaning associations, it would again be advantageous to have the possibility of directly and interactively accessing the KWIC snippets.

9 Conclusion

The preceding report is a summary of questions, methods and selected results of the joint project called 'Corpus-based Linguistic Research and Analysis Using Data Mining' ("Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining" – 'KobRA') where German linguists, computational linguists and computer scientists closely cooperate in order to investigate benefits and issues of using data mining techniques for the automation of after-retrieval cleaning and disambiguation tasks in the area of corpus-based empirical language research. The methods used and evaluated in this project will be available for research and teaching within the data mining environment RapidMiner and from the CLARIN-D infrastructure.

This article was based mostly on the requirements and issues in the area of corpus-based lexicography/historical semantics. In this area, topic models were used in order to partition KWIC lists retrieved by querying various corpora for a choice of polysemous or homonym words according to the single meanings of the searched words. The reliability of the automatic method was evaluated with help from two independent annotators who manually disambiguated the evaluation data sets.

Overall, the evaluation gave positive results. The automatic disambiguation performed with similar success for content words such as nouns, verbs or adjectives. It is still to be seen if the usefulness of this method can be extended to grammatical words; for that, more study is needed. The number of meanings of each search word was found to impact the values of the results (less definitions, better results). In most cases it also appeared true that a medium sized context for the relevant word led to the best results. Neither the number of considered KWIC snippets in the range of 500-5000 nor the use of different (orthographically normalized) corpora had any noticeable effect on the results of the automatic disambiguation. More studies are needed to review the performance of this method for diachronic corpora with non-normative orthography.

Using the automatic disambiguation easily allows the occurrences of single definitions of examined words to be identified and visualized. The integration of the query snippet's publication dates enables researchers investigating semantic change to easily track the use of disambiguated word forms over

time. A next step of innovation for this method would be the development and testing of interactive visualizations, which would allow for direct access to the underlying corpus basis.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research ("Bundesministerium für Bildung und Forschung") in the funding line for the eHumanities [01UG1245A-E].

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3* (3), 993-1022.

David M. Blei and John D. Lafferty. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, 113-120.

Samuel Brody and Mirella Lapata. (2009). Bayesian word sense induction. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 103-111.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, 264–270.

Jacob Cohen. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement* 20, 37-46.

Stefan Engelberg and Lothar Lemnitzer. (2009). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Tony McEnery, Richard Xiao, and Yukio Tono. (2006). *Corpus-Based Language Studies – an advanced resource book.* London: Routledge.

Gerd Fritz. (2012). Theories of meaning change – an overview. In C. Maienborn et al. (Eds.), *Semantics. An International Handbook of Natural Language Meaning*. Volume 3. Berlin: de Gruyter, 2625-2651.

Gerd Fritz. (2005). *Einführung in die historische Semantik*. Tübingen: Niemeyer.

Alexander Geyken. (2007). The DWDS corpus. A reference corpus for the German language of the twentieth century. In C. Fellbaum (Ed.), *Idioms and collocations. Corpus-based linguistic and lexicographic studies*. London: Continuum, 23-40.

Thomas L. Griffiths and Mark Steyvers. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences, 101* (Suppl. 1), 5228-5235.

Erhard Hinrichs and Thomas Zastrow. (2012). Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 1622-1627.

Rudi Keller and Ilja Kirschbaum. (2003). *Bedeutungswandel. Eine Einführung*. Berlin: de Gruyter.

Dan Klein & Christopher D. Manning (2003): Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1, ACL '03, pag-es 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wolfgang Klein and Alexander Geyken. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In U. Heid et al. (Eds.), *Lexikographica.* Berlin: de Gruyter, 79-93.

Anke Lüdeling and Merja Kytö. (Eds.). (2008). *Corpus Linguistics. An International Handbook.* Volume 1. Berlin: de Gruyter.

Anke Lüdeling and Merja Kytö. (Eds.). (2009). *Corpus Linguistics. An International Handbook.* Volume 2. Berlin: de Gruyter.

Ingo Mierswa et al. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*.

Roberto Navigli. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys, 41* (2), 10:1-10:69.

Roberto Navigli and Giuseppe Crisafulli. (2010). Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 116-126.

Roberto Navigli and Daniele Vannella. (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, 193-201.

Uwe Quasthoff, Matthias Richter, and Chris Biemann. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, 1799-1802.

Christian Rohrdantz et al. (2011). Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 305-310.

Paul Rayson and Mark Stevenson. (2008). Sense and semantic tagging. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics.* Volume 1. Berlin: de Gruyter, 564-578.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining*, 306–315.

Angelika Storrer. (2011). Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In K. Knapp et al. (Eds.), *Angewandte Linguistik. Ein Lehrbuch.* 3. vollst. überarb. und erw. Aufl. Tübingen: Francke, 216-239.

User Delegation in the CLARIN Infrastructure

Jonathan Blumtritt[1] **Willem Elbers**[2] **Twan Goosen**[2]
Marie Hinrichs[3] **Wei Qiu**[3] **Mischa Sallé**[4]
Menzo Windhouwer[5]
[1] Cologne Center for eHumanities – University of Cologne, [2] CLARIN ERIC,
[3] University of Tübingen, [4] NIKHEF, [5] The Language Archive – Meertens Institute
jonathan.blumtritt@uni-koeln.de, willem@clarin.eu,
twan@clarin.eu, marie.hinrichs@uni-tuebingen.de,
wei.qiu@uni-tuebingen.de, msalle@nikhef.nl,
menzo.windhouwer@meertens.knaw.nl

Abstract

The CLARIN research infrastructure aims to place language resources and services within easy reach of the humanities researchers. One of the measures to make access easy is to allow these researchers to access them using their home institutions credentials. However, the technology used for this makes it hard for services to make delegated call, i.e., a call on behalf of the researcher, to other services. In this paper several use cases, e.g., interaction with a researcher's private workspace or protected resources, show how user delegation would enrich the capabilities of the infrastructure. To enable these use cases various technical solutions have been investigated and some of these have been used in pilot implementations of the use cases. This paper reports on the use cases, the research and the implementation experiences.

1 Introduction

The topic of this paper is the interaction between two of the pillars of the CLARIN research infrastructure:[1] ease of access and integration of services. Ease of access has been implemented by enabling researchers to use their home institution credentials to access resources, tools and services offered by CLARIN on the web. This works well in many cases, but has turned out problematic for the cases where these services themselves need to access other services or resources on behalf of the researcher. To research possible solutions and implement them for a specific use case CLARIN-NL[2] has teamed up with the Dutch BiG Grid project.[3] Last year also a CLARIN-D[4] use case has been solved using the same solution and new CLARIN(-D) use cases are under investigation and in actual development. This paper reports on the results of the research and implementation of these different use cases.

The structure of the paper is as follows: in Section 2 it starts with a description of the problem, the requirements for a good solution, the possible solutions investigated and briefly mentions new development since the research was done. Section 3 then describes in depth the chosen solution and a first implementation thereof. Several use cases in the CLARIN infrastructure would profit from user delegation. These use cases and, where possible, experiences obtained during the implementation are described in Section 4. The paper ends with a description of future work and some conclusions.

[1] http://clarin.eu/content/mission
[2] http://www.clarin.nl/
[3] http://www.biggrid.nl/
[4] http://de.clarin.eu/

2 Shibboleth and User Delegation

Shibboleth[5] is the underlying technology that enables users to use the credentials of their home institute in the CLARIN infrastructure. It is based on the Security Assertion Markup Language (SAML; Cantor, 2012), as a Single Sign-On (SSO) system. Shibboleth is widely used in the research world,[6] providing single sign-on for web applications based on national federations, where the universities and research institutions function as Identity Providers (IdPs). The CLARIN centers that offer services, fulfilling the role of Service Providers (SPs), have grouped together in a CLARIN federation, which makes it administratively easy for the IdPs to deal with the CLARIN SPs.

Its wide support has made Shibboleth a good starting point for CLARIN, but it also has disadvantages. Shibboleth is typically aimed at users logging in and interacting with the SPs via their browser. Although the use cases described in this paper always start out in a browser session, the service invoked needs to invoke another service on behalf of the researcher. Shibboleth does not support this by default. In the next section possible solutions to enable such functionality are described.

2.1 Possible solutions

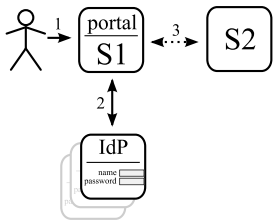
In the research phase of the CLARIN-NL/BiG Grid collaboration many solutions were considered and evaluated against the following requirements (grouped from 3 angles):

1) For the *User*:
   a) Single-Sign-On
   b) Access public and private services from within a portal (and other services)
   c) Transparent use, no required confirmation for every service or service access
2) For *Services*:
   a) Authentication by identity provider
   b) Authorization by service provider
   c) Nested service invocation possible (delegation)
   d) Easy to set up (for researcher)
3) For the *System* as a whole:
   a) Multi-federation authentication using SAML2
   b) REST and possibly SOAP
   c) Using proven technologies
   d) Operational effort minimal
   e) In-line with standards & best practices[7]
   f) Can we start today?

In this section the considered solutions and their evaluations are briefly discussed, for a more extensive discussion see Van Engen and Sallé (2011). In the descriptions and figures S1 indicates the service that calls another service, which is called S2, on behalf of the researcher (represented by the stick figure) authenticated by an IdP. Numbered arrows indicate subsequent requests between the parties involved.

Open

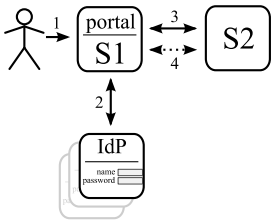
In this simple model all services trust each other. S1 includes the user identity with its request to S2, which accepts this without further checking. This is easy to setup, but does not scale up to the CLARIN infrastructure.



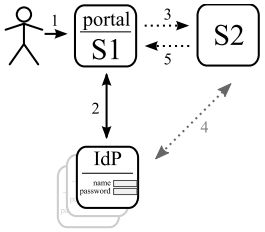
[5] http://www.internet2.edu/shibboleth/

[6] See for example the coverage of research and education identity federations at https://refeds.org/index.html

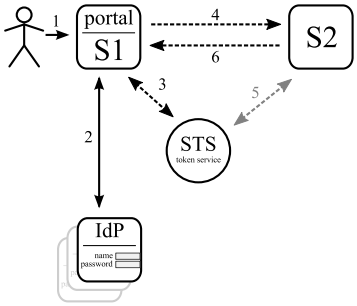
[7] This includes the requirement that the solution should be secure.



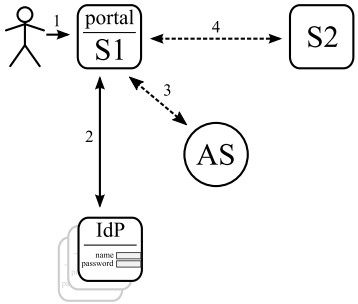
OAuth 1 (Hammer-Lahav, 2010)
This protocol is popular on the Internet and uses delegated security tokens for one site to access another site, e.g., allow LinkedIn to access one's Google address book. When S1 wants to access S2 the researcher's browser will be redirected to S2. There the researcher allows the access, and is redirected back to S1. The drawback is the need for separate confirmation for each combination of services.

SAML ECP (SAML V2.0 Contributors, 2005)


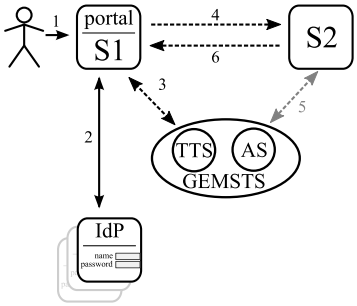
Enhanced Client or Proxy (ECP) is developed to support SAML for programs other than the browser. For Shibboleth, it is actually supported but not enabled by default, while SimpleSAMPLphp[8] does not support delegation via ECP. SAML ECP therefore is not a viable solution: CLARIN cannot force the IdPs to enable ECP and furthermore, since ECP would require a configuration for each AP at each IdP, such a solution does not scale.



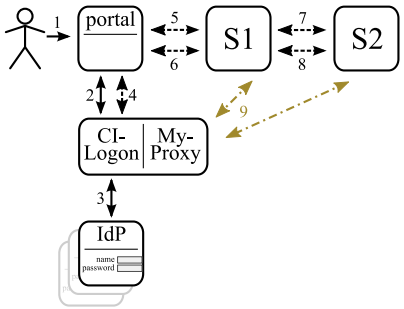
WS-Trust[9]
WS-Trust defines the concept of a security token service for SOAP web services. It is a flexible but rather complex setup, and can also be problematic for REST services.

OAuth 2[10] *(Hardt, 2012)*


This next evolution of OAuth supports more scenarios. As in the WS-Trust case a central service, an Authorization Service (AS), allows S1 to request a security token to pass on to S2, which can check the validity of the token and receive the user identity. Although this solution was fairly new at the time, it was selected as the primary option to be further investigated. It has since then quickly become the de-facto authorization standard on the internet and is replacing OAuth 1.



GEMBus STS
The GEMBus framework[11] is intended as a multi-domain communication environment and provides a number of services, including a security token service. At the time of evaluation GEMBus was alpha software.

X.509 certificates (Cooper, Santesson, Farrell, Boeyen, Housley, & Polk, 2008)


These certificates are the basis of the widely used SSL and TLS protocols. They are based on a public key infrastructure where trusted certificates are signed by trusted certificate authorities (CA). Delegation can be implemented using proxy certificates and is used as such in the 'grid world'. At the cost of additional setup the, much feared, burden of managing the certificate/keypair can be hidden from the user. This solution was selected as the secondary option to be investigated in case the OAuth 2 solution would fail.

[8] https://simplesamlphp.org
[9] http://docs.oasis-open.org/ws-sx/ws-trust/v1.4/ws-trust.html
[10] http://oauth.net/
[11] http://geant3.archive.geant.net/Research/Multidomain_User_Application_Research/Pages/GEMBus.aspx

2.2 Chosen solutions

Eventually two different solutions were chosen for further analysis, since they both could satisfy all the requirements. Firstly a solution based on OAuth 2 was chosen. The only identified risk to this solution at the time was the relative immaturity of OAuth 2 as a protocol and hence also of its implementations, since at the time, most (commercial) internet sites were still using the incompatible predecessor OAuth 1 protocol. This was therefore also the primary reason for choosing a second solution for further investigation. This second option, a solution based on X.509 certificates, which should then be used in such a way that they are hidden from the end-users, also could satisfy all the requirements, and most building blocks were already available at the time. Also this second solution has become more interesting over the past years, in particular in the scientific communities. All the other investigated options showed important shortcomings.

Hence it was decided to start with an OAuth 2 based proof-of-concept implementation, and depending on the experiences from that, to decide whether the X.509-based second option should be implemented as well.

2.3 New developments

Since the research reported on in Section 2.1 and the implementations efforts in the remainder of this paper the EUDAT project[12] has been investigating and developing a solution, named B2ACCESS, that is able to connect the different AAI infrastructures used within different communities, typically providing identity information, to the services offered within the EUDAT infrastructure. The solution provided by the UNITY software[13] supports this integration with different technologies such as SAML, OpenID, username/password and more. This allows for the authentication of the user using their federated identities and mapping these to an EUDAT identity which is then exposed to the EUDAT services in one of three ways: (1) X.509 certificates, (2) OAuth 2 and (3) SAML. Because of this flexibility this solution is very interesting since it allows for different options in the backend. There is support for OAuth 2, which is discussed in depth in this paper, but there is also support for X.509 certificates which might be a good candidate in specific scenarios. Although there is also SAML support, the limitations for the ECP support discussed earlier prevent this from being a viable alternative.

3 Configuring and Running an OAuth 2 Authentication Service

Figure 1 sketches the OAuth 2 delegation workflow in more detail: A user is logged in to Service 1 (S1), which is secured via a Shibboleth SP, using the IdP of his home institution. When the user triggers an action on S1 that requires access to a resource on Service 2 (S2), S1 redirects the user to the AS to collect an access token. Since the AS is also secured via an SP, it sends the user to the Discovery Service (DS) where he selects the IdP for authentication. The AS creates an authorisation code which is sent to S1 via the user. S1 uses it to request an OAuth 2 access token from the same AS. S1 then passes this access token to S2, which checks the validity of the token with the AS and receives user attributes in return (such as the user ID derived from the EPPN (*EduPersonPrincipalName*)). If the token is valid and S2 authorizes the user for the resource (a decision based on the user ID), S2 sends back the response to S1, which can then process it and complete the action triggered by the user. For the lifetime of the initial token, further communication between S1 and S2 can occur without the need to request another token.

In a second report, Van Engen and Sallé (2013) describe how, after attempts to use OAuth 2Lib,[14] a working solution was obtained using the *ndg_oauth* Authorization Server[15] combined with OAuth for Spring Security.[16] The *ndg_oauth* AS is implemented in Python, and for production it is advised to run it via WSGI in an Apache HTTP server. To get it to work for the use cases described below, i.e., to allow S2 to actually receive the user identity, some fixes were needed.

[12] http://www.eudat.eu
[13] http://unity-idm.eu/
[14] http://www.rediris.es/oauth2/
[15] https://github.com/cedadev/ndg_oauth
[16] http://projects.spring.io/spring-security-oauth/docs/oauth2.html

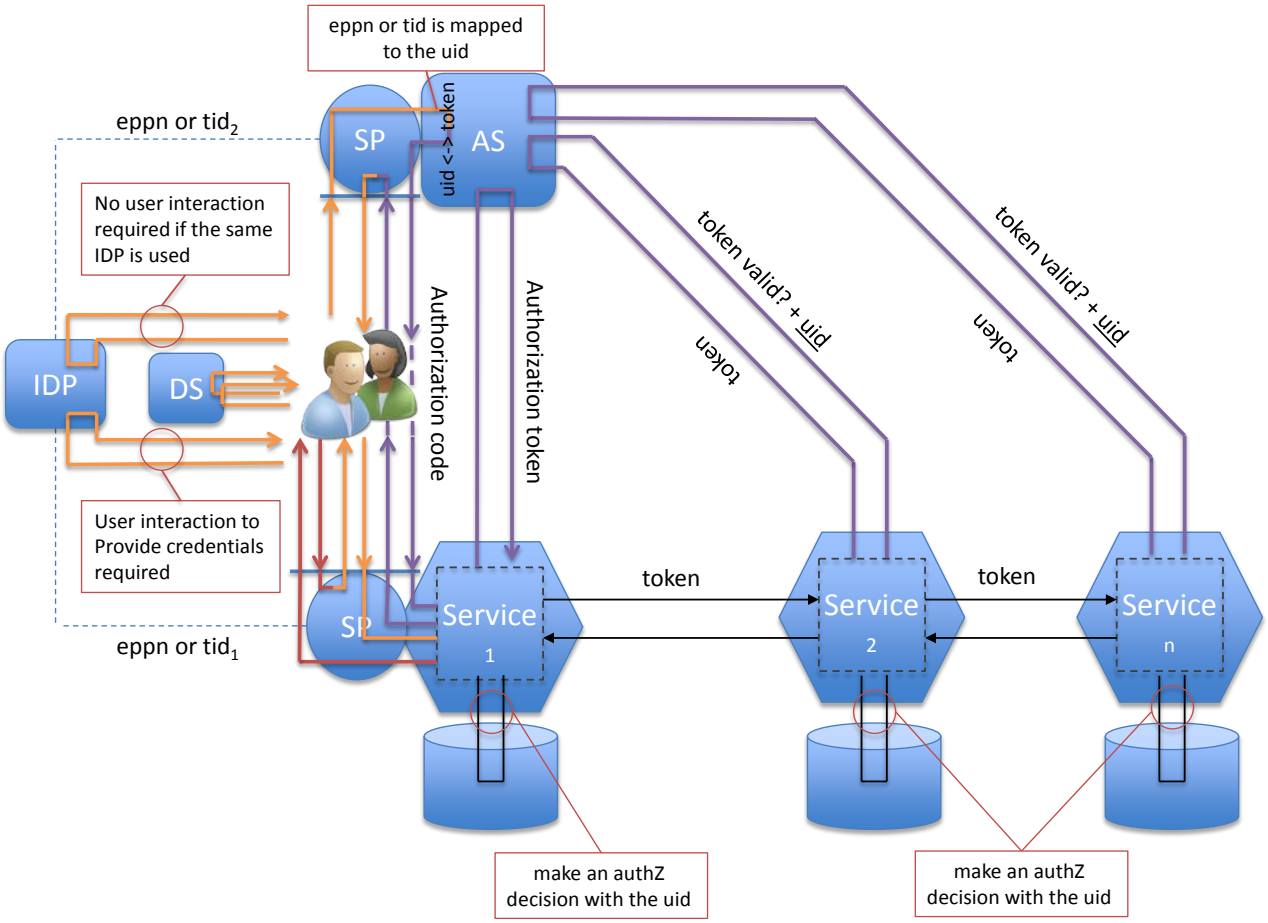


Figure 1. OAuth 2 delegation workflow

Furthermore when later configuration and stability became an issue, the advised WSGI embedding was no longer usable. This was resolved by letting the Apache web server run as a (reverse) proxy in front of an independently running *ndg_oauth* AS. However, the *ndg_oauth* documentation does not cover this, so investigations into the source code were required to achieve this. Documentation covering this setup and the required patches can now be found in the GitHub repository of The Language Archive.[17]

The *ndg_oauth* module is not the only implementation of an OAuth 2 AS. One could, for example, switch to the SURFnet OAuth-Apis AS.[18] The upcoming Section 4.3 reports on some first experiments using this alternative AS.

The solution based on X.509 certificates was not further implemented, but Van Engen and Sallé (2013) state that a smooth transition from OAuth 2 tokens acquired from an AS to certificates acquired from an online CA is possible.

4 CLARIN Use Cases

This section describes a number of cases from the CLARIN infrastructure where integration of services could be extended by means of user delegation. A number of these use cases have already been implemented at a proof-of-concept level. Where applicable, implementation strategies, encountered issues and future perspectives are described.

4.1 CMD Component Registry and ISOcat

This first use case was selected as a pilot because of the availability of development resources within a single institute (the Max Planck Institute for Psycholinguistics) and because the underlying technology

[17] https://github.com/TheLanguageArchive/ndg_oauth
[18] https://github.com/OpenConextApps/apis

stacks of the adapted software components, matches that of the implementation example worked out by Van Engen and Sallé (2013) to a reasonable degree, in particular the client application, which makes use of the Spring framework. Also, the delegation step in this particular use case reflected functionality with (at time of implementation) the potential of real-world application in the production environment.

The Component Registry is part of the Component Metadata (CMD) Infrastructure (Broeder, et al., 2010) implemented by CLARIN. It provides an online editor to metadata modellers to create CMD profiles and components. To enable semantic interoperability, these CMD profiles or components contain references to concept registries. While this use case was developed a prominent registry was the ISOcat Data Category Registry.[19] Within CLARIN, ISOcat has been succeeded by the CLARIN Concept Registry.[20] However, for this paper the experiences to implement the user delegation scenario between the Component Registry and ISOcat are still relevant. The CMD Component Registry editor allowed searching in ISOcat, where the search was initiated by the Component Registry backend, i.e., the backend plays the role of Service 1 and ISOcat that of Service 2 (see Figure 1). Without user delegation only a search for public data categories was possible. Hence the use case is to extend the search for private data categories in the ISOcat users workspace.

To enable this, the Component Registry has been extended with OAuth for Spring Security, providing the following functionality:

1) A method to check if a security token is available in the current session;
2) A method to initiate the request for a security token, i.e., to interact with the *ndg_oauth* AS including logging in and giving permission for delegation;
3) A method to query ISOcat while passing on the security token.

Enabling OAuth for Spring Security required the already present Shibboleth authentication layer to be 'bridged' with Spring Security. This was solved by a simple, though not entirely obvious mapping, involving a custom 'pre-authentication filter' and a dummy 'UserDetailsService'.

On the ISOcat side OAuth for Spring Security could not be used as its implementation is not based on servlet technology. However, this part of the AS interaction is relatively simple. The security token is retrieved from the HTTP header and passed on in a simple check token request to the AS. If the token is valid the identity of the researcher is returned and ISOcat can extend the search to include her workspace.

One implementation issue which still needs to be resolved is the Component Registry's use of frames for the AS interaction. It was pointed out that this hides the URL of the AS and IdP, which makes it hard for the researcher to determine to whom she is providing her credentials.

4.2 CLASS: Cologne Language Archive Services

The CLASS web application[21] implements tools for searching and analysis based on the Poio API,[22] and also provides easy-to-use web interfaces to facilitate field linguists' research. Apart from hosting scripts the main function of the CLASS application is to serve as a gateway to the archives that maintain annotated corpora. The aim is to offer a convenient web-based workflow, which enables the user of the application to access resource files for analysis directly from the repository.

The Cologne use case targets the DoBeS corpus, a core resource hosted by The Language Archive (TLA)[23] at the Max Planck Institute for Psycholinguistics (MPI), a CLARIN center. Most of the collections within the corpus are protected on a personalized level for privacy and ethical reasons. They may only be accessed by the corresponding owner or research group, hence the retrieval of data by external services was unviable in the past. It was soon noticed that this was another case that called for a solution of the delegation issue with the CLASS web application playing the role of S1 and a TLA

[19] http://www.isocat.org/
[20] https://openskos.meertens.knaw.nl/ccr/browser/
[21] http://class.uni-koeln.de/. The CLASS web application was realized as part of the CLARIN-D Curation Projects of Working Group 3, http://de.clarin.eu/en/discipline-specific-working-groups/wg-3-linguistic-fieldwork-anthropology-language - typology/curation-project-1.html.
[22] http://www.poio.eu/
[23] http://tla.mpi.nl/

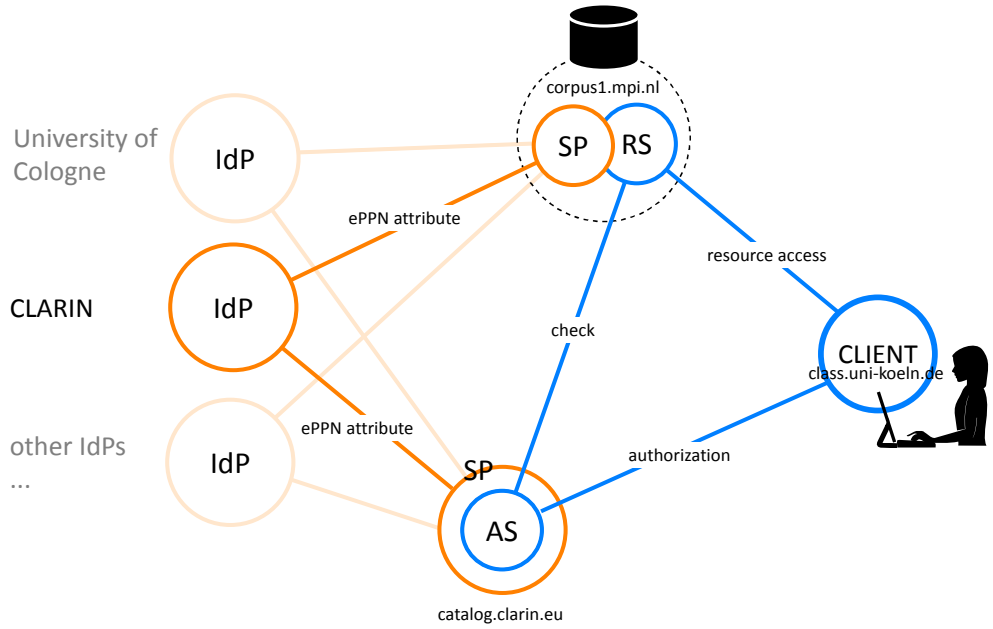


Figure 2. User delegation in the CLASS use case

service that of S2. With the availability of the AS the realization of this layout was possible (see Figure 2).

TLA has implemented a servlet, also known as the TLA Facade Service, which allows delegated access to the resources in the archive. Contrary to ISOcat this servlet can and does use the OAuth for Spring Security. The services provided by the TLA facade are:

1) *accessRights*: receive the access rights (none, read or read/write) the logged-in researcher has for one or more resources;
2) *accessFile*: fetch a specific resource for the logged-in researcher (if she has the right to do so).

The CLASS application uses the *rauth* library[24] written in Python as an OAuth 2 client to talk with the AS and call the TLA facade services. OAuth 2 is specifically designed to reduce complexity on the client side. Tie-ins with common web frameworks are smooth and well documented. Now researchers can run the tools provided by CLASS on resources residing in The Language Archive.

4.3 CLARIN-D ownCloud workspaces

WebLicht[25] is an execution environment for natural language processing pipelines, implemented in CLARIN-D. The online application allows users to construct and execute customized tool chains for text analysis, and subsequently visualize the resulting annotations. OwnCloud[26] is an open-source software system used for file hosting, which provides many features for data sharing and user collaboration. It serves to provide user workspaces, and is deployed and administered at the Forschungszentrum Jülich GmbH (FZJ), a CLARIN-D data center. Currently, in order to save WebLicht results to ownCloud, users must first download the results from WebLicht and then upload to their ownCloud workspace. In this use case, we want to enable users to bypass the download step and directly save results from WebLicht to ownCloud via WebDAV. Both WebLicht and ownCloud are protected by Shibboleth, but behind separate SP's. This scenario exactly demonstrates a user delegation scenario shown in Figure 1, where WebLicht plays the role of Service 1 and ownCloud that of Service 2. This section describes the current state of implementation and further experiments which have been carried out so far.

[24] http://rauth.readthedocs.org
[25] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/
[26] https://owncloud.org/

The CLARIN-D production installation of ownCloud is protected by an SP through a third party plugin called *user_shibboleth*.[27] Some adjustments to the plugin were made by FZJ to make it function with the ownCloud version currently deployed. In the modified version, the IdP is required to release the Persistent-ID and EPPN attributes. The revised ownCloud plugin maps a hashed version of the user's Persistent-ID to an ownCloud user, and the user's shown name is derived from their EPPN.

An environment that mirrors the actual configuration has been created in order to test implementations and perform experiments using various component options. The remainder of this section reports on the work that was done in this test environment.

The first step taken was to adapt the *user_shibboleth* plugin to allow use behind a reverse proxy and to configure the WebLicht SP to pass the HTTP headers to ownCloud. The patches to the plugin can be found on GitHub.[28]

The next step is to add an extra access point to ownCloud to enable it to process requests with valid OAuth 2 access tokens. See Figure 1, where ownCloud acts as a resource server (Service 2). In order to allow access from WebLicht on behalf of a user, the access point must be exposed outside the SP. Only one official plugin for ownCloud is available which offers this functionality - *user_oauth*,[29] and it is not compatible with the deployed version of ownCloud. Furthermore, it relies on several deprecated third party libraries. The CLARIN-D center in Tübingen addressed and solved the problems with the plugin by essentially reimplementing it.[30]

Next, a server (the AS component in Figure 1) is required which:

1) is capable of authenticating users through a Shibboleth IdP
2) supports token introspection compatible with the *user_oauth* implementation, which was done according to a draft specification (Richer, 2013)[31]

Several options are available for the AS component:

- ndg_oauth AS (as described in Section 3)
- php-oauth-as[32]
- SURFnet OAuth-Apis[33]

Since none of the options fulfill all of the requirements out-of-the-box, each one needs to be assessed individually. *ndg_oauth* AS is capable of authenticating users through a Shibboleth IdP, but it is not compatible with user_oauth and the documentation is fairly sparse. *php-oauth-as* seems to be compatible with *user_oauth* and is being actively developed, but its ability to authenticate users via SAML IdP still remains to be investigated. SURFnet *OAuth-Apis* can authenticate users through a Shibboleth IdP, and can be made compatible with *user_oauth* with only minor changes, thanks to its flexible architecture.

SURFnet *OAuth-Apis* was chosen to be evaluated first for various reasons. It is a Spring application fully compatible with the v2-31 version of the OAuth 2 specification. It provides pluggable authentication and user consent handling, which makes customization very easy. This is particularly important because a specification for token introspection has not yet been finalized and customization will be necessary as the specification evolves. Additional advantages are that it has the most extensive documentation and demo applications, is being actively developed, and has a large user community. A demo has been setup using *OAuth-Apis*. In the demo, a client application namely Testlicht[34] is able to access files on ownCloud.

An alternative to adapting the server to meet the requirements of *user_oauth* would be to implement OpenID-Connect[35] on both the server side and *user_oauth* side. OpenID-Connect is in a sense a layer on top of OAuth 2 providing standardized ways to obtain information about the identity behind an

[27] https://github.com/AndreasErgenzinger/user_shibboleth
[28] https://github.com/weblicht/user_shibboleth
[29] https://github.com/owncloud/apps/tree/master/user_oauth
[30] https://github.com/weblicht/user_oauth
[31] http://www.ietf.org/archive/id/draft-richer-oauth-introspection-06.txt
[32] https://github.com/fkooman/php-oauth-as
[33] https://github.com/OAuth-Apis/apis
[34] https://weblicht.sfs.uni-tuebingen.de/testlicht
[35] http://openid.net/specs/openid-connect-core-1_0.html

OAuth 2 token, and it also provides means to restrict the attribute release. Exploring this promising option is left as future work.

4.4 Virtual Collection Registry

The Virtual Collection Registry (Broeder, Van Uytvanck, & Wittenburg, 2010) is an online service developed within CLARIN that allows users to create collections of resources (including metadata documents) from any location and register them in the CLARIN metadata infrastructure. The service assigns a persistent identifier to the collection upon publication so that it can be referenced as a unit.

A stable version of the Virtual Collection Registry (VCR)[46] is currently available. It has a web front end through which users can log in via Shibboleth to create new virtual collections, edit a collection's metadata and existing resource items, or add new items to a collection through a series of forms. In addition, the service exposes a REST service that supports the creation, manipulation and deletion of collections and resource items. It uses the same authentication policy and methods as the web front end, and therefore the potential for usage in other applications is currently limited.

The addition of support for user delegation to the VCR would allow various other applications to be extended with options to add resources, presented in the context of these applications, to one of the user's own collections, or to create a new collection in the user's workspace within the VCR based on a set of resources. An example of such an application is the faceted browser of the Virtual Language Observatory (VLO),[47] in which users can search for metadata records and associated resources. The connection between the VLO and the VCR could consist of an 'add to collection' option available to the user once search results are shown. When the user chooses this option in this scenario, the VLO connects to the VCR's REST service and request the list of the collection that the user has permissions to work on. After selection of a collection, or alternatively the option to create a new collection, the VLO sends the appropriate request including a list of the selected records to the VCR, which in turn applies the requested changes inside the user's workspace. Repositories at CLARIN centres or elsewhere could provide similar options in their repository search and exploration tools. Examples of such tools would be the hierarchical archive browser[48] of The Language Archive or the search engine of the HathiTrust's digital library.[49]

As the VCR REST service is based on the Java servlet and JAX-RS technologies, it is similar to the TLA facade service described above with respect to adding support for authentication through OAuth 2. Notice that this use case is strictly hypothetical and no efforts towards implementing the described support in either the VCR or the VLO have been taken thus far.

5 Future Work and Conclusion

Apart from these first use cases other uses are possible. For example, in addition to accessing archived resources, CLASS tools could also issue delegated calls to protected remote tools, i.e., web services residing on different sites. The same could be done for WebLicht.

Another potential extension is multi-step delegation: the current solution supports single step delegation, i.e., from S1 to S2, but S2 cannot request a security token from the AS to call a next service, S*n*. Support for such multi-step delegation is currently under investigation. The important question to ask here is how S2 could obtain a new token on behalf of the original user. Perhaps S2 should be able to use the original token to authenticate and get a new token. In order to encode the different authorizations involved in this original token, it will be necessary to implement this in the context of OpenID Connect, perhaps as an extension to it. OpenID Connect adds the necessary handles for the required level of fine-grained attribute release. We are not aware of any (full) solution using OpenID Connect for this type of multi-step delegation.

Not all IdPs release sufficient information for the AS to allow identification of the logged-in researcher. Rather than a universally identical user identifier, such as EPPN (*EduPersonPrincipalName*), the IdP might release a EPTID (*EduPersonTemporaryId*). Although the IdP gives the same EPTID each time the researcher accesses a certain SP (so it can use it to identify the return of the researcher),

[46] http://clarin.eu/vcr
[47] http://clarin.eu/vlo
[48] https://tla.mpi.nl/tools/tla-tools/asv/
[49] http://babel.hathitrust.org (an example selection is already available in the VCR at http://hdl.handle.net/11372/VC-1002)

it gives a different EPTID for the same researcher to each different SP. When the AS and S2 thus are hosted at different SPs the EPTID cannot always be used to identify the researcher. Thus researchers with such an IdP are likely to have problems using delegation.

The *ndg_oauth* AS is currently an experimental service at TLA. In the future this or another AS could be a CLARIN service, but to realize this service, the stability and high availability options have to be investigated first. In this respect the experiments in Tübingen with other AS implementations are very relevant.

The developments within the EUDAT project, especially the B2ACCESS service based on UNITY, are a promising development not directly tackling the delegation issue, but offering flexibility in supporting different technologies that have the potential to provide a solution for the delegation problem. Therefore we consider this a valuable solution to look into. As a first step the OAuth 2 based delegation should be integrated and as a second step support for X.509 delegation can be investigated.

As showcased by the various use cases discussed in this paper support for user delegation is a valuable extension of the CLARIN infrastructure, which will allow further and more fluent integration of key infrastructure components. The experiments to implement these use cases have already helped to make the technology more mature and will in the future continue to do so. A production ready implementation will certainly support CLARIN's mission to enable easy access to language resources, services and tools to the community of humanities scholars.

Acknowledgements

The authors like to acknowledge the valuable support of all the experts in the Big Grid/CLARIN project and the current CSNE informal expert group, especially Willem van Engen who implemented an easy to use and understandable *ndg_oath* and OAuth for Spring Security demo setup.[50] We also like to thank Shakila Shayan for the implementation of the TLA facade servlet.

References

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., et al. (2010). A Data Category Registry- and Component-based Metadata Framework. Seventh International Conference on Language Resources and Evaluation. Malta: ELRA.

Broeder, D., Van Uytvanck, D., Wittenburg, P. (Eds.). (2010). Language Resource and Technology Registry Infrastructure (CLARIN Report D2R-5b). Retrieved March 18, 2015 from CLARIN: http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-35

Cantor, S. (Ed.) (2012, May). SAML Version 2.0 Errata 05. Retrieved March 18, 2015 from OASIS: http://docs.oasis-open.org/security/saml/v2.0/sstc-saml-approved-errata-2.0.html

Cooper, D., Santesson, S., Farrell, S., Boeyen, S., Housley, R., & Polk, W. (2008, May). Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. Retrieved June 18, 2014 from Network Working Group: http://tools.ietf.org/html/rfc5280

Hammer-Lahav, E. (2010, April). The OAuth 1.0 Protocol. Retrieved June 18, 2014 from Internet Engineering Task Force (IETF): http://tools.ietf.org/html/rfc5849

Hardt, D. (2012, October). The OAuth 2.0 Authorization Framework. Retrieved September 10, 2014 from Internet Engineering Task Force (IETF): http://tools.ietf.org/html/rfc6749

Richer, J. (2013, May 1). OAuth Token Introspection. Retrieved June 17, 2014 from Internet Engineering Task Force (IETF): http://tools.ietf.org/html/draft-richer-oauth-introspection-04

SAML V2.0 Contributors. (2005). Enhanced Client or Proxy (ECP) Profile. In J. Hughes, S. Cantor, J. Hodges, F. Hirsch, P. Mishra, R. Philpott, et al., Profiles for the OASIS Security Assertion Markup Language (SAML) V2.0 (pp. 21 - 31). OASIS.

Van Engen, W., & Sallé, M. (2011). User Delegation in the CLARIN Metadata Infrastructure: connecting the component registry and ISO-DCR - Part I - Research. CLARIN/BiG Grid. Retrieved March 18, 2015 from NIKHEF: http://wiki.nikhef.nl/grid/images/6/66/Clarin-security_for_web_services-research-report010.pdf

[50] https://github.com/wvengen/oauth2-demo

Van Engen, W., & Sallé, M. (2013). User Delegation in the CLARIN Metadata Infrastructure: connecting the component registry and ISO-DCR - Part II - Implementation. CLARIN/BiG Grid. Retrieved March 18, 2015 from NIKHEF: http://wiki.nikhef.nl/grid/images/1/17/Clarin-security_for_web_services_inplementation.pdf

Sharing Multimodal Data: A Novel Metadata Session Profile for Multimodal Corpora

Farina Freigang[1], Matthias A. Priesters[1,2], Rie Nishio[3], Kirsten Bergmann[1]
[1]Faculty of Technology, Center of Excellence "Cognitive Interaction Technology" (CITEC)
Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany
{firstname.lastname}@uni-bielefeld.de
[2]Human Technology Centre (HumTec), Natural Media Lab
RWTH Aachen University, Theaterplatz 14, 52056 Aachen, Germany
priesters@humtec.rwth-aachen.de
[3]Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Binderstr. 34, 20146 Hamburg, Germany
rie.nishio@sign-lang.uni-hamburg.de

Abstract

In the natural sciences and humanities, scientific data management and in particular the categorisation of data and the publication of (meta)data becomes ever more relevant. A new focus in corpus-based research are *multimodal* data. However, metadata profiles for multimodal data are rare and do not fit the needs of researchers who are searching for particular data. In this paper, we present a novel metadata session profile for describing data collections which contain other modalities beyond text and speech. The profile is based on experiences gained during the work on three different corpora comprising communicative speech-gestural behaviour as well as sign language data. The profile is aimed at creating metadata for individual recording sessions and is technically implemented in the CMDI format. Furthermore, it is designed to be paired with an existing profile for media corpora, which was extended for multimodal data.

Keywords: Metadata profile, multimodal data, multimodal corpora, gesture, sign language, CMDI, ISOcat, CLARIN

1 Introduction

The production of high-quality multimodal corpora is extremely expensive and hence it is of major importance to manage these resources in a way that they are easily searchable and reusable for other researchers. In fact, the reuse of resources is an issue strongly promoted by research funding organizations, for example, by the European Union in terms of their "open data strategy".[1] In the field of corpus linguistics and language resources it is widely agreed that the ever-expanding number and growth of corpora needs *metadata* for the purpose of corpus management. For linguistic resources there already exist a large number of metadata schemes, but so far not much effort has been put into the development of metadata schemes for the particular structure of *multimodal* corpora. This is, at least in parts, due to the fact that multimodal corpora are highly heterogeneous. They might include different modalities or communication channels of natural communicative behaviour such as gestures, facial expressions, body posture or eye gaze for which no standardised coding schemes exist. Moreover, multimodal corpora might comprise multiple synchronous data streams, such as video, audio, time series data (e.g., motion capture or eye tracking) and annotation data. These aspects have previously not been captured by metadata profiles.

In CLARIN-D, the discipline-specific working group on "Speech and Other Modalities"[2] has initiated a discussion on these issues (cf. Freigang and Bergmann, 2013) which has led to the proposal of a novel

[1] http://ec.europa.eu/digital-agenda/en/open-data-0

[2] Indicated by the group name *SpeechAndOtherModalities* in the CLARIN Component Registry: http://catalog.clarin.eu/ds/ComponentRegistry

metadata session profile for multimodal data: the `MultimodalSessionProfile`.[3] The profile is based on a detailed evaluation of three different multimodal corpora: the Speech and Gesture Alignment ("SaGA") Corpus from Bielefeld University (Lücking et al., 2013), the Dicta-Sign DGS Corpus from University of Hamburg (Matthes et al., 2012) and the Natural Media Motion Capture ("NM-MoCap") Corpus from RWTH Aachen University (Hassemer, 2015). The profile has been developed according to the CMDI[4] standard (Broeder et al., 2012; de Vriend et al., 2013) including unique ISOcat[5] definitions within and for (but by no means exclusively for) the CLARIN infrastructure. It offers a wide variety of corpus descriptions especially designed for, but not limited to, multimodal data. Furthermore, it has been used for the integration (and publication) of the three mentioned corpora into the CLARIN-D infrastructure.[6]

This paper aims to present the new metadata session profile for multimodal data. First, we briefly introduce necessary technical terms in section 2. Subsequently, in section 3, we review existing metadata profiles for multimodal data and discuss why they were not sufficient for the requirements of our corpora. In section 4 we introduce the novel session profile with its modality components and other specifically developed components, and in section 5 the accompanying corpus profile is described. We conclude with a discussion in section 6.

2 Terminology

In this section, we briefly introduce the basic technical terminology pertaining to the CMDI metadata standard insofar as it is necessary for the understanding of the following sections. For more detailed information on CMDI and the underlying infrastructure, see Broeder et al. (2008), Broeder et al. (2012) and de Vriend et al. (2013).

Profile CMDI profiles are templates used to create metadata records (XML files that describe a specific corpus or data set). They are implemented as XSD schemas which define the structure of the actual CMDI records. Profiles consist of components and elements. Profiles are published in the CLARIN Component Registry, which can be used to validate metadata records.

Component A CMDI component is an independent part of a CMDI profile, which groups subordinate components and elements, usually thematically related ones. Components are also published in the Component Registry and can be reused by including existing components into a newly created profile.

Element Elements are the basic building blocks in CMDI, which hold the actual metadata. They are embedded in a component or directly in a profile and have the form of *key–value* pairs. The values are typed and can contain strings, numbers, boolean values, URLs or different date formats. Furthermore, the user can define the values of elements in terms of regular expressions or controlled vocabularies (predefined sets of possible entries).

Cardinality Components and elements in a profile are marked with *number of occurrences* constraints. These constraints specify lower and upper boundaries of how often the component or element can or must occur in an instance of the CMDI profile. Typically, the lower boundary is 0 or 1, and the upper boundary 1 or *unbounded*. In this paper, cardinalities are given in the figures in square brackets (e.g., [0–1]).

Attribute Additional data fields which can be attached to components and elements are called attributes. They can contain additional information about the respective elements or components, for example, attributes can indicate a component's language or link components to other components (cf. section 4.4 and Figure 6).

[3] Monospaced font for designations denotes names of CMDI profiles/components/elements/attributes as they appear in the CLARIN Component Registry.

[4] CMDI: Component Metadata Infrastructure (Broeder et al., 2012; de Vriend et al., 2013); which is compatible with other standards such as Dublin Core (DC), Open Language Archives Community's metadata set (OLAC), and Isle Metadata Initiative (IMDI).

[5] `http://www.isocat.org`

[6] Two corpora were ingested into the repository of the Bavarian Archive for Speech Signals:
`https://clarin.phonetik.uni-muenchen.de/BASRepository/index.php`

3 Related work and problem description

In Freigang and Bergmann (2013), we compared relevant CMDI metadata profiles for multimodal data from the CLARIN Component Registry: `media-corpus-profile` and `media-session-profile` by BAS[7], NaLiDa's[8] `MultimodalCorpus` profile, and `BamdesMultimodalCorpus` used for harvesting purposes by the Harvesting Day initiative. These profiles already included some aspects of multimodality, since simple modality components such as `cmdi-modality` and `ModalityInfo` exist (cf. the modality list in Figure 4), however, various other aspects were missing (for a detailed discussion, see Freigang and Bergmann (2013)). We are not aware of other related work concerning metadata for multimodal corpora.

Basically, we identified two major problems with existing components and profiles. The first problem occurred when generating metadata descriptions for the previously mentioned multimodal corpora (SaGA, Dicta-Sign DGS, NM-MoCap) from existing metadata profiles: the *granularity* in which modality (or multimodal) metadata descriptions were possible was not fine enough. So far, it was not possible to specify, for example, the handedness of an actor[9], the modalities of a stimulus, or that iconic gestures were annotated in the data. Hence, from the corpus user's perspective, it was not possible to search for detailed features of multimodal corpora. Therefore, we focused particularly on the following:

- the development of detailed descriptions of the two modalities *gesture* and *sign language*, among others

- descriptions of how actors relate to different modalities (text, speech, gesture, sign language, etc.): for example, what is their written German language proficiency, or do they gesture a lot, or do they have experience with sign language, or do signers have regular contact to non-signers?

- multimodal descriptions for the study design and the data collection (environment, content, elicitation phase, etc.): for example, did the recording take place in a studio, or what modalities were involved in explaining the task, or did someone gesture in a stimulus video?

- descriptions of the annotation scheme (e.g., according to a gesture or sign language researcher) which is used to analyse the multimodal data.

A second problem with existing metadata profiles was that *technical descriptions* for media files and annotation files were missing. With the recording of multimodal data, novel technical devices typical for gesture or sign language studies are used. For example, one of our reference corpora includes motion capture recordings of gestures. Compared to other profiles, the `media-session-profile` is rather advanced and already includes components for time series data and stereo video (3D) recordings. However, describing a marker setup as used in the NM-MoCap corpus in this metadata structure proved cumbersome and unintuitive. Therefore, a more meaningful way for describing motion capture data among others was one of the requirements for a new profile. Furthermore, descriptions for technologies used in recordings, as for example HD videos, were not elaborate enough and needed extension.

4 Introducing the `MultimodalSessionProfile`

Based on the identified problems, we developed various new components covering different modality aspects and technical descriptions. As discussed in Freigang and Bergmann (2013), there are several

[7] `http://www.phonetik.uni-muenchen.de/Bas`

[8] `http://www.sfs.uni-tuebingen.de/nalida/en`

[9] We chose to use the term *actor* throughout our profiles and in this paper, firstly, because it is most accurate. *Participant* and *subject* are terms which imply an arranged setting such as in studies, which is not always the case since some corpora are collections of data, such as the Dicta-Sign DGS Corpus or a collection of news broadcasts. Secondly, *actor* is the most neutral term available: the terms *speaker* or *signer* would exclude users of signed or spoken languages, respectively. Therefore, we used the term *actor* in newly created components of our session (and corpus) profile. The terms *subject* and *participant* occur rarely and only where components were reused.

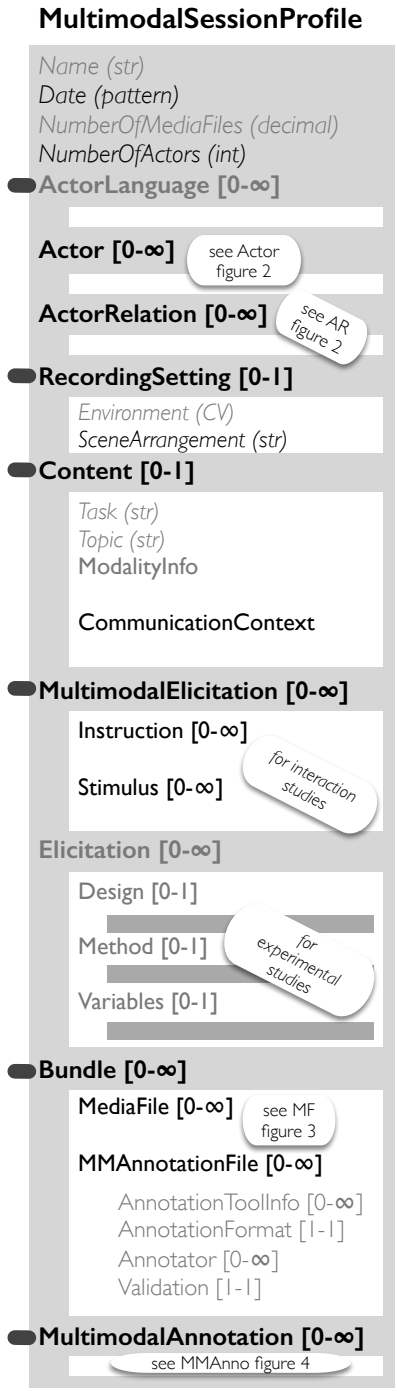


Figure 1: An overview of the session profile with six main thematic parts. Two elicitation components cover interaction studies and experimental studies.

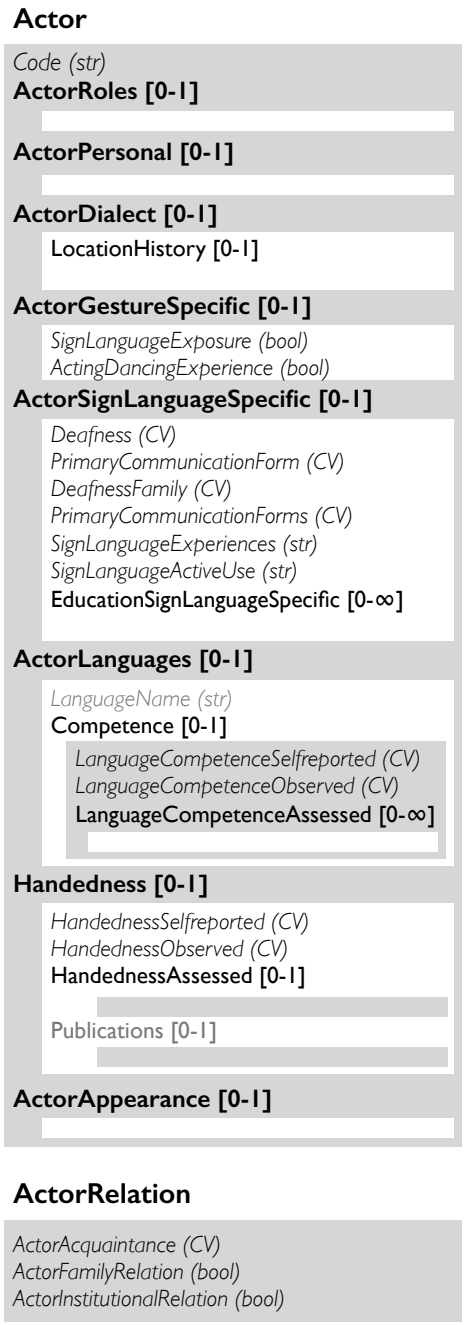


Figure 2: The actor and actor relation components introduced in Figure 1. All components have been newly created for the purpose of a fully multimodal description of actors.

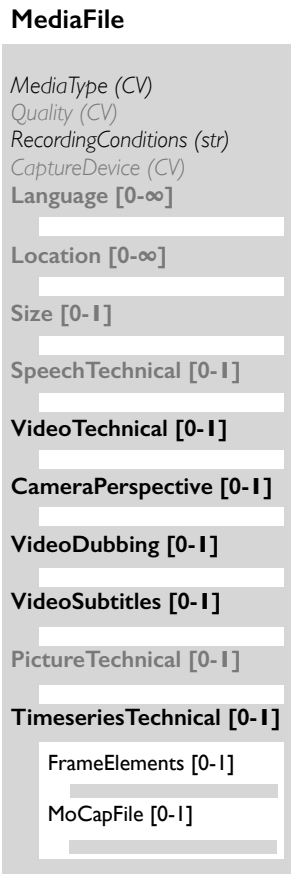


Figure 3: The Media file component introduced in Figure 1. The component is comprised of old and new components.

options of how to realise new metadata components. Many changes were necessary, so that the integration of the new components into an existing profile structure was not feasible. Therefore, we created the novel **MultimodalSessionProfile**[10] in a bottom-up fashion: We designed new components and subsequently compiled them together with relevant existing components from the CLARIN Component Registry into a large profile structure. All figures in this paper illustrate reused components and elements in grey font; all others are newly created. In some cases a component has been newly created which combines (almost) only existing components and elements in a novel fashion. The aim was to group components and elements thematically. Furthermore, the exact names of components and elements may have been changed in this paper for better readability (for the exact designations see the profiles online) and the figures illustrating the metadata components are depicted not to the full extent but are reduced to the most important parts. The profile construction is oriented at `media-session-profile` by BAS and NaLiDa's `MultimodalCorpus` profile, among others. In cooperation with BAS, we also created a multimodal version of the `media-corpus-profile`, discussed in section 5.

4.1 Session Profile Overview

The **MultimodalSessionProfile** (Figure 1) consists of six main thematic parts: metadata descriptions about the *actors*, the *recording setting*, the *content* of a study or a corpus, the *elicitation* methods, the accumulated *data* (media files and linked annotation files), and a description of the *annotation* design (`MultimodalAnnotation`). In some cases, the outline is reminiscent of the temporal development of a data set: for example, when conducting a study, one typically starts with the participants and the study design, continues with the study itself, the recorded data, the post-processing of the data, and the theory behind the used methods. On the top level, the profile contains a number of elements for basic metadata about the session: the date, the time and the place of a recording and the numbers of files and actors involved.

In the following, the main components and the various possibilities of description they provide are discussed. Note that the profile has been designed for flexibility, therefore all components and elements on the profile's top level are optional. If a component has been chosen for description of a data set, some elements or components are obligatory. For example, if `MultimodalAnnotation` has been chosen, we assume that the corpus contains multimodal data that needs to be described and, thus, the `ModalityInfo` component, reused from the META-SHARE metadata profile (Gavrilidou et al., 2012), is mandatory. Or as soon as an actor is involved in a data set it needs to be stated which role she took (cf. section 4.2.1). Since there is a lot of homogeneity among multimodal corpora, flexibility is of major importance to allow users to adapt the metadata profiles according to their needs.

4.2 Modality and multimodal components

Natural communication data include various modality aspects of which only a few are found in metadata descriptions. In order to get the full picture of multimodality, a first step was to define categories for gesture and sign language. In Figure 5, the newly created components are grouped according to three different categories: *speech*, *gesture* and *sign language*. We refined the granularity of the modality metadata descriptions in various ways: to depict the details of performed modalities by the *actors*, the influence of the modalities on the *data collection* design, the *content* of the material, and the *annotations* concerning the various modalities.

In the following, we will give a few examples of how certain components (which will be discussed in more detail below) fall into these categories. We developed category-specific descriptions, namely the `ActorGestureSpecific` and `ActorSignLanguageSpecific` components. Speech-specific descriptions are covered by reused components such as `cmdi-subjectlanguages`. Other components are kept general and allow for speech and sign language descriptions, among others, as does the component `ActorLanguages`. Other components that serve both of these categories are `ActorDialect` with its component `LocationHistory`. The component `Handedness` has been

[10] The `MultimodalSessionProfile` has been published on May 5th 2014 (and edited by CLARIN on July 2nd 2014): `http://catalog.clarin.eu/ds/ComponentRegistry?itemId=clarin.eu:cr1:p_1381926654659`

kept general for the description of both gesture and sign language. Finally, some components fall into all three categories: for example, the MultimodalElicitation component (which is supplemented by the reused Elicitation component for experimental research data) and the Content component comprising the study task, the modalities which are used during the study, and the communication context.

4.2.1 Actors

Providing detailed information about the persons appearing in the corpus material was a major focus of our metadata profile. We realised this in two components: Actor for the properties of an individual person and ActorRelation for capturing relations between multiple persons taking part in the same session (Figure 2). The Actor component comprises multiple subcomponents which allow the metadata user to describe different aspects of participants in various use cases: ActorPersonal contains elements for basic data about participants, such as their name, sex, age, educational or professional status. ActorRoles captures the roles a person takes in a corpus, for example, *experimenter*, *subject* or *confederate*. ActorAppearance describes the actors' physical appearance, insofar it is relevant for the purposes of the recordings (e.g., if the actor is wearing glasses or clothing that could cause problems for image recognition or motion tracking techniques). Besides specifying the language used in the corpus, the component ActorLanguages allows for recording all languages spoken by an actor, including their self-reported or measured proficiency (Competence).

The requirements of researchers recording gesture and sign language corpora include metadata about participants usually not captured by metadata profiles. For theses purposes, we developed the ActorGestureSpecific and the ActorSignLanguageSpecific components. The focus of both types of corpora is set on communicative manual action and thus, our profile includes information about an actor's Handedness (either self-reported or assessed using a test). The gesture component records whether the actor has had previous exposure to sign language, acting or dancing, factors which could influence their gestural behaviour. For corpora including sign language, the personal history of actors, such as their educational background or the location where they grew up, are especially important, as these strongly influence sign language proficiency and the signed dialect (ActorDialect). Furthermore, the sign language component contains detailed elements for describing Deaf actors, such as the use of hearing aids, the deafness status of their family members and their degree of involvement in Deaf culture (e.g., through sign language teaching or using sign language in art). In developing these components, we built upon and extended a set of ISOcat data categories for describing signed language resources compiled and implemented by Crasborn and colleagues (Crasborn and Hanke, 2003a; Crasborn and Hanke, 2003b; Crasborn and Windhouwer, 2012).

4.2.2 Elicitation

The *Elicitation* components provide room for describing the design of the data collection, i.e. the methods which were applied to elicit communicative behaviour from the recorded participants. For studies employing experimental methodologies (e.g., from a psychological background), we included the Elicitation component from the NaLiDa metadata profile. Additionally, we designed a new component for interaction studies (MultimodalElicitation). It is mainly divided into a MultimodalInstruction and a MultimodalStimulus component, each framing a component named InformationChannel, stating how information was given to the actors during the instruction or stimulus phase: which kind of medium was used, how it was physically presented, and which modalities were involved. For stimuli which are common in a certain research community and which have been published, the respective publications can be attached using the documentInfo component, also developed by META-SHARE. Furthermore, it can be specified whether an instruction was recorded beforehand and if a stimulus was accessible to the actor during language or gesture production.

4.2.3 Context and content

For a full picture of the data collection design, information about the content of the language resources and the recording environment is necessary. The profile includes two components for these purposes,

Content and RecordingSetting. The latter includes facts such as the Environment (e.g., *studio* or *field*), the VisualBackground, the Weather or the SceneArrangement (a free description field).

Content offers various ways of specifying what the content of a data set is about, including the Task and Topic elements as well as a modality component and the CommunicationContext component. CommunicationContext provides space for specifics of the conversation, for example the SocialContext (e.g., *family* or *public*), the Channel (e.g., *face to face* or *telephone*), the ConversationType (e.g., *dialogue*), or whether there is an Audience watching the scene.

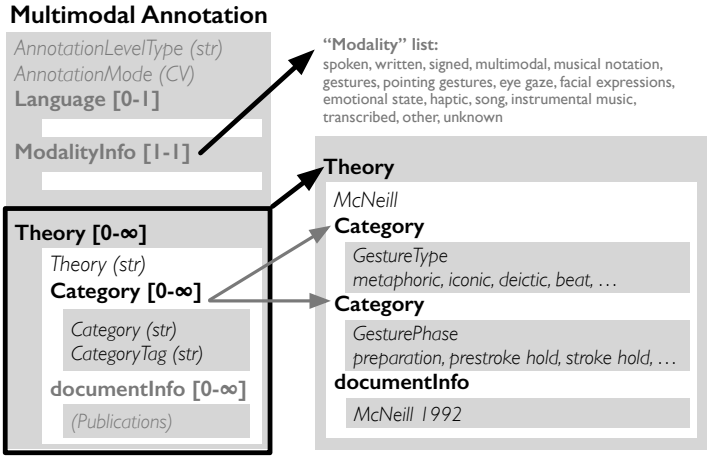


Figure 4: The MultimodalAnnotation component with an exemplary use of the Theory component for gesture categories by McNeill (1992). For the position within the MultimodalSessionProfile, cf. Figure 1.

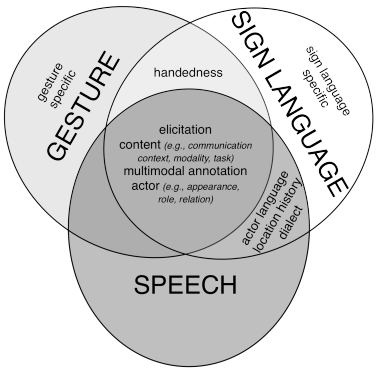


Figure 5: The keywords and names of newly created modality components classified into three main categories. Reused components are not shown.

4.2.4 Multimodal annotations

For multimodal communication data, the categorisation of observed phenomena is usually done by annotating recorded data according to predefined or emergent categories. One crucial aspect of our work was the development of a metadata component which is able to capture various categorisation systems. Our profile records this information in the description of the annotation schemes. In gesture studies, for example, gestures can be classified according to different criteria. One popular method follows McNeill (1992), who distinguishes between *iconic* (resembling the content of speech), *metaphoric* (image of abstract concept), *deictic* (pointing) and *beat* (marking the structure of the utterance) gesture categories. Furthermore, McNeill temporally segments gestures into phases such as *preparation*, *stroke*, *hold* and *retraction*.

In Figure 4, we have sketched how a gesture annotation scheme based on McNeill's categories could be captured with our MultimodalAnnotation component. Several Theory components can be added; each needs to be given a name (e.g., *McNeill*, cf. the example above) and may contain one or several Category components. A Category is also named (e.g., *GestureType* and *GesturePhase*) and contains one or several CategoryTag elements, which represent the individual annotation labels (e.g., *iconic* or *preparation*). Category and CategoryTag may be seen as an annotation category with one or more annotation labels possible. Each Theory component can be enriched with literature references in the documentInfo component. Additional information, for example, explanations about the exact meaning of annotation categories, can be stored in optional Description components. Overall, this Theory component is kept simple in its design and is still flexible enough to cover complex category systems, also those which may be developed in the future. We explicitly encourage metadata creators to use this component to also refer to their own theory or annotation frameworks.

The Theory component appears next to two other components and two elements in the MultimodalAnnotation component. One component is the modality list ModalityInfo mentioned in section 3. It provides modality-related keywords for characterizing the annotations performed on the corpus data. The difference between the Theory and the ModalityInfo components is that the former describes the annotation scheme and the theory behind it in detail, whereas the latter generally lists the specific modalities which were annotated. Thus, the ModalityInfo list allows for quick and shallow modality descriptions, if no particular framework has been used. Furthermore, the metadata creator can specify the AnnotationLevelType (e.g., *part of speech*, *gesture form*, etc.), the AnnotationMode (e.g., *manual*, *automatic*, etc.), and the language of the annotations.

4.3 Technical metadata

Besides capturing information about the recorded data on a conceptual and theoretical level, technical and organizational descriptions of the resulting data files are necessary. These metadata are collected in the Bundle component (Figure 1). Files are grouped in bundles if they belong to the same (usually synchronously recorded) data set. This can mean for example multiple simultaneously recorded video streams, together with motion capture and eye tracking data and the annotation files pertaining to these data.

4.3.1 Media files

With the MediaFile component (Figure 3), the profile includes fine-grained description categories for various types of media data, that is, *video*, *audio*, *image* and *time series* data. Most categories were reused from existing metadata profiles (most notably the media-session-profile), but some components were extended. Among the added features are information about camera perspectives, video dubbing/subtitling and the ability to describe multiple channels of a single video recording (needed for 3D stereo videos). The TimeseriesTechnical component was extended by components for marker sets used in optical motion capture systems and for kinematic data computed from raw motion capture data.

4.3.2 Annotation files

The treatment of annotation files differs from existing profiles in that *annotation files* (labels) are separated from *annotation schemes* (theories), the latter being realised in the Theory/MultimodalAnnotation component (cf. section 4.2.4). There are various established transcription and annotation tools, such as *Praat* (Boersma and Weenink, 2001), *ELAN* (Wittenburg et al., 2006), *Anvil* (Kipp, 2001), *iLex* (Hanke et al., 2010), *EXMARaLDA* (Schmidt, 2002), among others, each of which is based on different annotation file formats. The MultimodalAnnotationFile component (Figure 1) is limited to technical and organisational metadata. Each description instance of an annotation file is linked to the corresponding MultimodalAnnotation component, this way information about an annotation system or scheme only needs to be stored once in each session CMDI file.

4.4 Links between components

In order to better reflect the internal structure of a data set, many components can be linked to each other using attributes (Figure 6). This way, redundancy is kept at a minimum, since each piece of information has to be given only once. Components which can be linked to possess an 'ID' attribute, components which can link to other components possess a 'reference' attribute. The component Actor, for instance, has an attribute ActorID, which can be linked with components such as ActorRelation and MultimodalAnnotationFile through their ActorRef attributes. Those links can capture, for example, that the actor participated in different sessions of the same data set, that two actors are relatives or colleges from work, or that this specific actor in a video file is the same person whose interaction is labelled in an annotation file. Finally, session CMDI files and the corresponding corpus CMDI file are linked to each other.

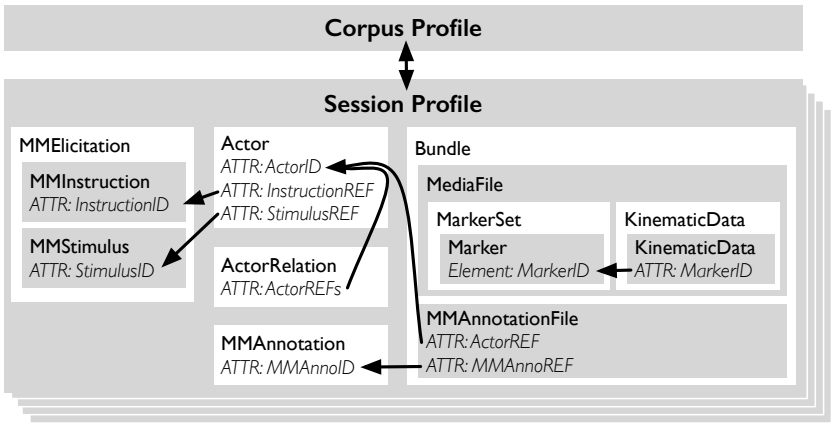


Figure 6: Links between components using attributes.

5 Corpus metadata

The `MultimodalSessionProfile` is designed to describe a single set of contiguous data, usually one recording session as part of a larger corpus. For the description of the corpus as a whole, an enclosing profile is needed, which 'frames' the session data. Therefore, in cooperation with BAS, we extended their `media-corpus-profile` with components for multimodal data (Figure 7; the components in grey font are the original BAS components of the profile). The first version of the profile was mostly geared towards speech corpora containing audio data. The extended version of the `media-corpus-profile`[11] (version 1.1) now contains a `MultimodalCorpus` component capturing information about modalities and an `AnnotationInfo` component with information about the annotated phenomena and the annotation tools and file formats used. The `MultimodalSessionProfile` and the extended version of the `media-corpus-profile` are designed to be used in combination in order to create a complete corpus metadata description. The usage is as follows: for each experiment or sub-study, one session CMDI file is created, whereby one actor can participate in several sub-studies. All sessions that belong to one data set are then linked to a single corpus CMDI file, which describes this data set (for links between components and profiles cf. section 4.4).

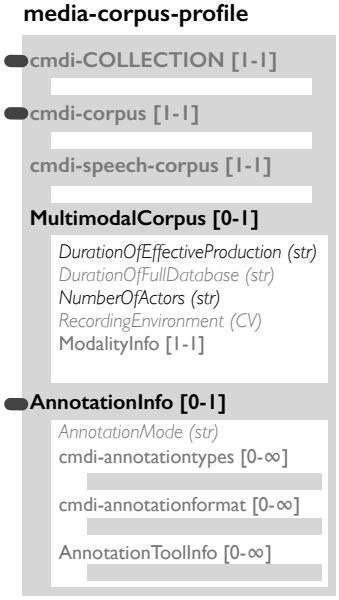


Figure 7: An overview of the enhanced BAS corpus profile with three main thematic parts.

6 Discussion and outlook

In this paper, we presented a metadata profile specifically addressing the needs of researchers working on multimodal communication data, which builds upon and expands earlier profiles. The presentation of our novel scheme and its realization have evoked fruitful discussions at conferences and workshops, both within the CLARIN community and in the relevant research communities. This shows a serious interest in the topic among potential users.

The development of our metadata profile has been driven by the requirements which resulted from work with specific corpora. Nevertheless, we aimed at developing a flexible profile universally applicable to multimodal data, in line with the philosophy behind CMDI: "The CMDI infrastructure encourages reuse of resources [...]. Therefore, metadata that are useful to any researcher [...] is especially valuable

[11] The `media-corpus-profile` has been published on May 5th 2014:
`http://catalog.clarin.eu/ds/ComponentRegistry?itemId=clarin.eu:cr1:p_1387365569699`

and should be focused on first." (de Vriend et al., 2013, 1320) Given the heterogeneity of multimodal corpora, further tests beyond our three corpora with other data collections are necessary to further improve the profile and make it as universally applicable as possible.

To date, a user-friendly tool for the creation of CMDI files based on the `MultimodalSession-Profile` is not available. Some CMDI generation tools exist, as for example ARBIL[12] (Withers, 2012), developed at the Max Planck Institute for Psycholinguistics in Nijmegen, or custom-built CMDI generations scripts. However, these tools are either designed for a particular profile structure or they are not easy to use with complex profiles such as the `MultimodalSessionProfile`. The creation of actual CMDI files remains a challenge, as the profile's size and complexity makes the manual creation of larger numbers of CMDI files infeasible. Technical metadata can easily be extracted automatically from the data itself, but for content metadata, easy-to-use tools for researchers are required and remain future work. Therefore, we highly encourage further tool development for the automatic generation of CMDI files, which would be extremely helpful to create, use and share CMDI files and further improve metadata profiles. In future, such a tool may even be used for flexible and 'on-the-fly' profile creation: Thus, no complete CMDI profiles would need to be prepared as templates, but components from the Component Registry could be combined flexibly by the metadata creator while compiling metadata for a data collection.

Acknowledgments

We thank Florian Schiel, Menzo Windhouwer and Onno Crasborn for their support and cooperation of the multimodal corpus profile. This research was supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication", the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University, and CLARIN-D, the German division of the "Common Language Resources and Technology Infrastructure". Additionally, we thank the anonymous reviewers for their comments and ideas.

References

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

Daan Broeder, Thierry Declerck, Erhard Hinrichs, Stelios Piperidis, Laurent Romary, Nicoletta Calzolari, and Peter Wittenburg. 2008. Foundation of a component-based flexible registry for language resources and technology. In *Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation*, pages 1433–1436.

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a Component Metadata Infrastructure. In *Proceedings of the workshop on Describing LRs with Metadata (LREC 2012)*.

Onno Crasborn and Thomas Hanke. 2003a. Additions to the IMDI metadata set for sign language corpora. Agreements at an ECHO workshop, May 8–9, 2003, Radboud University, Nijmegen. `http://www.ru.nl/publish/pages/522090/signmetadata_oct2003.pdf`.

Onno Crasborn and Thomas Hanke. 2003b. Metadata for sign language corpora. Background document for an ECHO workshop, May 8–9, 2003, Radboud University, Nijmegen. `http://sign-lang.ruhosting.nl/echo/docs/ECHO_Metadata_SL.pdf`.

Onno Crasborn and Menzo Windhouwer. 2012. ISOcat data categories for signed language resources. In Efthimiou, Eleni and Kouroupetroglou, Georgios and Fotinea, Stavroula-Evita, editor, *Gestures in embodied communication and human-computer interaction*, pages 118–128. Springer.

Folkert de Vriend, Daan Broeder, Griet Depoorter, Laura van Eerten, and Dieter van Uytvanck. 2013. Creating & testing CLARIN metadata components. *Language Resources and Evaluation*, 47(4):1315–1326.

[12] `http://tla.mpi.nl/tools/tla-tools/arbil`

Farina Freigang and Kirsten Bergmann. 2013. Towards metadata descriptions for multimodal corpora of natural communication data. In *Proceedings of the workshop on Multimodal Corpora: Beyond Audio and Video, (IVA 2013)*.

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, and Valerie Mapelli. 2012. The META-SHARE metadata schema for the description of language resources. In *Proceedings of LREC 2012, Eighth International Conference on Language Resources and Evaluation*.

Thomas Hanke, Lutz König, Sven Wagner, and Silke Matthes. 2010. DGS Corpus & Dicta-Sign: The Hamburg studio setup. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC 2010)*, pages 106–110.

Julius Hassemer. 2015. *Towards a theory of Gesture Form Analysis: Principles of gesture conceptualisation, with empirical support from motion-capture data.* Ph.D. thesis, RWTH Aachen University.

Michael Kipp. 2001. Anvil – A generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based Analysis of Speech and Gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its Applications. *Journal on Multimodal User Interfaces*, 7(1–2):5–18.

Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. Dicta-Sign – Building a Multilingual Sign Language Corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012)*.

David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, Chicago and London.

Thomas Schmidt. 2002. Exmaralda – ein System zur Diskurstranskription auf dem Computer. *Arbeiten zur Mehrsprachigkeit, Serie B (34). Hamburg: SFB Mehrsprachigkeit*.

Peter Withers. 2012. Metadata Management with Arbil. In *Proceedings of the workshop on Describing LRs with Metadata (LREC 2012)*.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure

Twan Goosen[1] **Menzo Windhouwer**[2] **Oddrun Ohren**[3]
Axel Herold[4] **Thomas Eckart**[5] **Matej Ďurčo**[6]
Oliver Schonefeld[7]

[1]The Language Archive, [2]The Language Archive - Meertens Institute, [3]National Library of Norway, [4]Berlin-Brandenburg Academy of Sciences and Humanities, [5]Leipzig University, [6]Institute for Corpus Linguistics and Text Technology, [7]Institute for the German Language

twan@clarin.eu, menzo.windhouwer@meertens.knaw.nl, oddrun.ohren@nb.no, herold@bbaw.de, teckart@informatik.uni-leipzig.de, matej.durco@oeaw.ac.at, schonefeld@ids-mannheim.de

Abstract

This article reports about the on-going work on a new version of the metadata framework Component Metadata Infrastructure (CMDI), central to the CLARIN infrastructure. Version 1.2 introduces a number of important changes based on the experience gathered in the last five years of intensive use of CMDI by the digital humanities community, addressing problems encountered, but also introducing new functionality. Next to the consolidation of the structure of the model and schema sanity, new means for lifecycle management have been introduced aimed at combatting the observed proliferation of components, new mechanism for use of external vocabularies will contribute to more consistent use of controlled values and cues for tools will allow improved presentation of the metadata records to the human users. The feature set has been frozen and approved, and the infrastructure is now entering a transition phase, in which all the tools and data need to be migrated to the new version.

1 Introduction

Component Metadata Infrastructure (CMDI) has been one of the core pillars of CLARIN since the beginnings of this initiative (for an overview, see Broeder et al., 2012).

It established means for flexible resource descriptions for the domain of language resources with sound provisions for semantic interoperability weaved deeply into the data model and the infrastructure to overcome, in a great extent, the rule of metadata schism it set out to combat. Based on this solid grounding, the infrastructure accommodates a growing collection of metadata records. The development of the joint metadata domain both in number of records and diversity of profiles is proof for the success of the model and the infrastructure as such. Currently, at version 1.1 of the CMDI specification, there are 170 public profiles and over 1,000 public components defined. The CLARIN OAI-PMH harvester[1] periodically collects records from some 60 providers in more than 80 different profiles, almost 1 million as of March 2015.

However, in the first five years of its intensive usage by the CLARIN community naturally a number of design issues have arisen that need further attention. Therefore a dedicated task force consisting of developers and metadata experts from multiple CLARIN centres was established to work towards a successor to CMDI 1.1 based on the existing paradigm. After careful analysis, the task force worked out a proposal for a number of small but important changes and additions to the CMDI model leading to CMDI version 1.2. In April 2014, the Standing Committee for CLARIN Technical Centres approved the proposal, which meant that work on the implementation could begin.

The changes address the following aspects: lifecycle management, structure of the model and schema sanity (namespace issues, consistency of the meta model, attributes, mandatory/optional elements),

[1] http://catalog.clarin.eu/oai-harvester/

use of external vocabularies and cues for tools. They are described in detail in sections 2 and 3. The work on the model is accompanied by a comprehensive transition plan covering the conversion of existing data and adaptation of existing tools using CMDI data, described in section 4. Finally, section 5 details still open issues and further plans for the CMDI model and joint metadata domain.

1.1 Short description of the Component Metadata Infrastructure

It is important to understand that CMDI is not "yet another" static metadata format, but rather a metamodel, a framework allowing for the creation and use of custom schemas. It relies on a modular model of so-called metadata components (Broeder et al., 2010), which can be assembled together, to foster reuse, interoperability and cooperation among metadata modellers. Components are used to group elements and attributes, which can take values, and also other components. They are stored and maintained in the Component Registry.[2] A metadata modeller selects or creates components and combines them into a profile targeted at a specific resource type, a collection of resources or a project, tool or service. A profile serves as blueprint for a schema for metadata records. CLARIN centres offer CMD records, describing their resources, to the joint metadata domain.

Due to the flexibility of this model, the metadata structures can be very specific to an organization, project or resource type. Although structures can thus vary considerably they are still within the domain of metadata for linguistic resources and thus share many key semantics. To establish these shared semantics CMD components, elements and values can be annotated with links to concepts defined in external concept registries.[3] This allows generic tools that operate on all the CMD records in this domain, like the metadata catalogue Virtual Language Observatory (VLO),[4] to overcome differences in terminology as well as structure by operating on this shared semantics layer.

1.2 Related approaches

To position the work on CMDI in the broader landscape and to allow for comparison to the approach adopted by CLARIN we will briefly review a number of alternative approaches taken by similar or related initiatives. In the sister initiative DARIAH[5] no one common solution for resource description and discovery has been adopted yet, however a candidate solution developed within DARIAH-DE (Heinrich & Gradl, 2013) pursues an approach not too different to that of CLARIN: Repositories or collections are registered in the Collections Registry (roughly corresponds to CLARIN's Centre Registry), subsequently harvested via OAI-PMH into the Generic Search, a faceted search engine (corresponds to VLO). The schemas exposed by individual repositories are recorded in the Schema Registry where a mapping (crosswalks) can be defined. The mapping information is used for on-the-fly expansion of the queries. The main difference to the CLARIN approach is how crosswalks or semantic interoperability is achieved, namely via pair-wise mapping between the schemas, whereas in CMDI the concept links serve as pivot points, represent a separate semantic layer to ground the schemas onto, allowing for more efficient mapping (dozens of profiles share the same basic data categories). Furthermore, the DARIAH-DE approach has not yet been adopted on the European level. An alternative proposal within DARIAH is based on Semantic Web technologies: repositories provide lightweight description of their collections via RDFa[6]-annotated web pages, which are being crawled and indexed in a semantic web application.[7] This approach reflects the general tendency especially in the humanities towards adoption of semantic web technologies for resource description. Acknowledging the integrative power of the Linked Data paradigm, the CMDI developer team proposed a complete expression of CMD data (from meta model to the instances) as RDF (Ďurčo & Windhouwer, 2014), which was implemented by CLARIN-NL.[8]

---

[2] http://catalog.clarin.eu/ds/ComponentRegistry/

[3] The primary registry used until recently, the data category registry ISOcat, has been replaced with the CLARIN Concept Registry in December 2014.

[4] http://www.clarin.eu/vlo

[5] https://www.dariah.eu/

[6] Resource Description Framework in Attributes, see http://www.w3.org/TR/xhtml-rdfa-primer/

[7] http://rechercheisidore.fr

[8] https://catalog.clarin.eu/ds/cmd2rdf and https://github.com/TheLanguageArchive/CMD2RDF

The META-SHARE initiative,[9] on the other extreme, imposes one large schema for all resource descriptions with many optional parts and some specialization for the main resource types. Nevertheless it also adopts the basic idea of component-based modelling and concept-based semantic mapping. The principal compatibility has been demonstrated by expressing the META-SHARE schema as *resourceInfo* profiles[10] within the Component Registry.

The European DASISH project[11] delivered a metadata catalogue[12] collecting resource descriptions from the three research infrastructures CLARIN, DARIAH and CESSDA.[13] Given the great disparity of the encountered formats and the goal being a catalogue with a broad coverage but only a small fixed set of facets, individual fields in the schemas were manually mapped to the facets. This work was also strongly inspired by the CMDI approach using concept links for mapping where possible.

2 New CMDI functionality

2.1 Lifecycle Management

There is no definite metadata representation for any given language resource in terms of a single fixed CMDI component or profile. Instead, metadata modellers often encounter situations that make it necessary to adapt or amend existing metadata models. Typically, such situations are caused by needs of data providers that supply more detailed metadata than any of the existing components cater for. To ensure formal and semantic persistence of referenced metadata components, typical applications of CMDI will disallow changes of those components once they are made publically available.

Within the current version of CMDI, there is no possibility to denote the lifecycle status of components, e.g. by marking a component as deprecated and/or superseded by another component. CMDI 1.2 will provide lifecycle management support for components based on four additional header elements: *Status*, *StatusComment*, *Successor* and *DerivedFrom*. These elements appear as direct children of */CMD_ComponentSpec/Header/*.

The mandatory *Status* field is used to record the current lifecycle phase of a component. Allowed values comprise “development”, “production”, and “deprecated”. Infrastructures exploiting the CMDI framework need to ensure that only transitions from “development” to “production” but not vice versa are allowed on the grounds that components should not leave a state that denotes immutability (“production”, “deprecated”) once they reached it.

Each component can optionally be annotated with a *StatusComment*. This field can be used to record the reasons for status changes, reasons for the derivation of a new component from an existing one or other useful information regarding the component’s status in human-readable form.

The optional *Successor* element can be used on deprecated components to specify, if applicable, the URI of the component that should be used instead. Often this will be an updated or improved version of the original component. It is not necessarily a derivative in a technical sense: the successor can be a component created from scratch or another already existing component that represents a different metadata scheme which is meant to replace the scheme in the original component. As the *Successor* field holds exactly one URI, only the direct successor of a component can be specified. Note however, that succession is a transitive relation. Therefore it is possible to construct a complete chain of succession by traversing components via their *Successor* fields.

The URI specified in the optional *DerivedFrom* field allows for the reconstruction of a component’s genesis in relation to other components. Derivation in the context of CMDI is considered in a purely technical sense of copying a component and modifying it independently from the original component. As component editors are free to modify components without restrictions (as long as they are in the “development” state), the *DerivedFrom* relation does neither imply any strict structural or semantic inheritance relation among the components nor is it the inverse relation of succession. Nevertheless, we expect the typical use case for derivation to be the copying of existing components in order to improve them. This is illustrated in Figure **1**. From an existing component C a derived component C’ can

[9] http://www.meta-share.eu/

[10] Altogether 4 resourceInfo profiles were created representing different resource types, reusing most of the components.

[11] http://dasish.eu/

[12] http://ckan.dasish.eu/

[13] http://www.cessda.net/

be forked at any point in time of the original component's lifecycle. After the necessary amendments, C' will eventually enter the "production" state, possibly alongside C for some time. Finally, when C becomes deprecated, it may explicitly instantiate a *Successor* relation with C'.

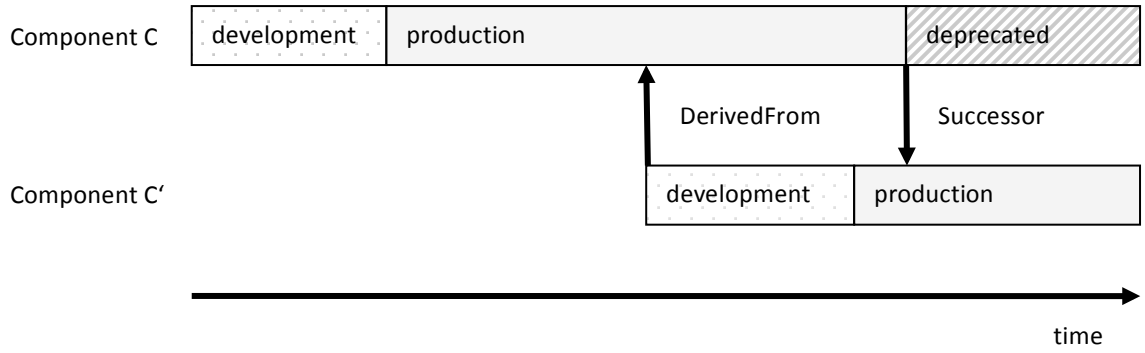


Figure 1: Lifecycle management in CMDI 1.2: lifecycle status and derivation.

It should be pointed out that value assignment to a specific lifecycle status applies only to the component in which it occurs. It is not inherited down through the component structure, nor can any inference on the value of its corresponding element in the child components be drawn safely, at least not in general. For instance, any deprecated component may include child components, which are still in production. Likewise, a deprecated component A containing a deprecated component B may have a successor A' which does not include the successor of B. However, the infrastructure might guide or pose restrictions on transitions. For example, when moving a component into the "production" state, the Component Registry might ask the user to first publish all referenced components that are still in the "development" state and warn the user if any embedded component is marked as deprecated.

The introduction of lifecycle information in components will enable a more sophisticated management of components. For example, as all published components are kept persistently within the Component Registry, the addition of improved versions of components may easily lead to proliferation. Explicit lifecycle management and especially the *Status* field can be used as filtering devices that constrain the users' and modellers' views to a more manageable subset of components. Restricting the selection of available components to those with a "production" status will help users to select the most relevant components. For special tasks such as the development of documentation or component reviews and for ensuring backwards compatibility of the Component Registry with deprecated components, all non-production components will continue to be available within the CLARIN infrastructure.

2.2 Vocabularies

The current version of CMDI requires value domains for elements and attributes to be specified locally in the components. In the cases where value domains are specified as controlled value sets this has several disadvantages. Firstly, updating any value domain is equivalent to updating the containing component. Hence, knowing that many of the value sets are work in progress this may greatly add to the proliferation of components mentioned in section 2.1. Secondly, keeping element value sets as integral parts of components inevitably hampers reuse of components. For example, consider a user looking for a component describing licences, and finding one that is perfectly adequate, except that its element representing the licence name does not list all the licences needed. In such cases, the user has no alternative but creating a new, possibly similar component, making sure that all the needed values are included in the value domain specification. On the other hand, using controlled vocabularies in metadata is in general an effective way to interconnect metadata from various origins, as long as the vocabularies are maintained as shared resources. To this end, CMDI 1.2 will support the use of external vocabularies, thereby increasing the possibility to obtain semantic interoperability across metadata.

Metadata modellers will have the opportunity to associate a vocabulary (identified by its URI) with an element or attribute in their components and profiles. The metadata creator will then be able to pick values from the specified vocabulary or (for open vocabularies) still choose to use a custom value that does not appear in the vocabulary. External vocabularies may be included in component specifications

in one of two ways:

1. Vocabularies may be *imported* verbatim into CMDI components, as enumerated value domains for CMDI elements or attributes. In this case the modeller may choose to import all vocabulary items, or only a subset.
2. Vocabularies may be *referenced* by the component and be used for dynamic lookup and retrieval of values when editing metadata records. Here a non-exclusive (open) use of items from the vocabulary must be assumed.

The above will be facilitated by introducing a new element *Vocabulary* in *ValueScheme* elements, with an optional *enumeration* element for imported, closed vocabularies. Examples are given in Code example 1 and Code example 2 below. At the instance level, an attribute *ValueConceptLink* (in the CMDI namespace) will be allowed on fields that have a vocabulary linked to hold the URI of the selected value, see Code example 3.

```
<Element name="Language" CardinalityMax="1" CardinalityMin="1">
      <ValueScheme>
            <Vocabulary URI="http://openskos.meertens.knaw.nl/iso-639-3"
      ValueProperty="skos:prefLabel" ValueLanguage="en">
                  <enumeration>
                        <item ConceptLink="http://cdb.iso.org/lg/CDB-
                  00138580-001">Dutch</item>
                        <item ConceptLink="http://cdb.iso.org/lg/CDB-
                  00138512-001">French</item>
                        ...
                  </enumeration>
            </Vocabulary>
      </ValueScheme>
</Element>
```

Code example 1: An element in a component specification with a closed external vocabulary

```
<Element name="Institution" CardinalityMax="1" CardinalityMin="1">
      <ValueScheme>
            <Vocabulary
      URI="http://openskos.meertens.knaw.nl/Organisations"
      ValueProperty="skos:notation"> </Vocabulary>
      </ValueScheme>
</Element>
```

Code example 2: An element in a component specification with an open external vocabulary

```
<cmdp:Institution
cmd:ValueConceptLink="http://openskos.meertens.knaw.nl/Organisations/dc0
2b3ea-00d9-433f-a540-9baf94a14be0">Sound and Vision</cmdp:Institution>
```

Code example 3: An element in a metadata record (CMDI instance) with a vocabulary item specified

Note that the two modes of using external vocabularies in CMDI 1.2 have quite distinct implications on the component life cycle as well as metadata management. Importing the vocabulary as enumeration into the component allows for strict schema validation of the values in the instance data, but does not automatically reflect changes in the vocabulary. Updating the local copy will typically be done by deriving a new component from the old one and importing the current version of the external vocabulary into the new component.

On the other hand, referencing a vocabulary allows keeping the list of possible values dynamically up to date, but standard XML validation tools will not be able to handle element values obtained this way. The modeller has to decide based on the expected completeness and change rate of the vocabulary which mode to apply. It is assumed that such decisions will be informed by future usage and experience with the vocabulary service, through which guidance and best practice will emerge. As a

general rule, large and dynamic vocabularies (e.g. an institution or person registry) are typical candidates for referencing, whereas small and stable vocabularies (e.g. small lists of formats or units) might be imported.

The usage of external vocabularies has some impact on the infrastructure. At the model level, the vocabulary facilities are specified to be generic, in the sense that no assumption about specific services is made. On the operational level – as initially supported by the core CMDI infrastructure – it will be designed to support specifically the OpenSKOS-based CLAVAS vocabulary service (Brugman, 2012), through which vocabularies of languages, organisations and value sets extracted from ISOcat are already available. To make the new functionality available for metadata modellers and creators, both Component Registry and existing metadata editors must be updated accordingly. Dedicated validation tools for handling references to external vocabularies would be useful and feasible, but seeing such tools more as part of the vocabulary service than of CMDI as such, there is at this point no plan for providing such tools as part of the upgrade mechanism supplied by the CMDI task force.

2.3 Cues for Tools

Some of the applications in the context of CMDI, especially those directly used by human users, require information that goes beyond formal specification and validation aspects. This includes documentation of meaning and purpose of all content-related elements and hints for improved visualisation of metadata content. Furthermore CMDI 1.2 will provide the basis for a powerful feature that allows automatic derivation of element content.

2.3.1 Improved documentation of CMDI elements

Documentation especially of content-related elements is essential for both metadata creators and human interpreters. CMDI 1.1 already provides an option to document the usage of CMDI elements but lacks this functionality for attributes or components. Therefore CMDI 1.2 expands the existing approach to all kinds of metadata entities. This allows schema creators to document their profiles in all necessary detail. Furthermore, CMDI 1.2 will permit multiple documentation values for different languages, which can be the basis for localised user interfaces. Code example 4 shows the specification of a component that contains an element, which in turn contains an attribute. It has documentation in both English and Dutch for the first two levels.

```
<CMD_Component name="Actor" CardinalityMin="0"
CardinalityMax="unbounded" ComponentId="ex_compid_actor">
  <Documentation xml:lang="en">
   This is a person or entity that plays a role in the resource
  </Documentation>
  <Documentation xml:lang="nl">
   Dit is een persoon of entiteit die een rol speelt in de bron
  </Documentation>
  <CMD_Element name="firstName" ValueScheme="string"
  DisplayPriority="0" CardinalityMax="1">
    <Documentation xml:lang="en">
     This is the given name of a person
    </Documentation>
    <Documentation xml:lang="nl">
     Dit is de voornaam van een persoon
    </Documentation>
    <AttributeList>
      <Attribute name="nickname" Type="string">
        <Documentation xml:lang="nl">
         Bijnaam van een persoon
        </Documentation>
      </Attribute>
    </AttributeList>
  </CMD_Element>
</CMD_Component>
```

Code example 4: New means for documentation of CMD entities

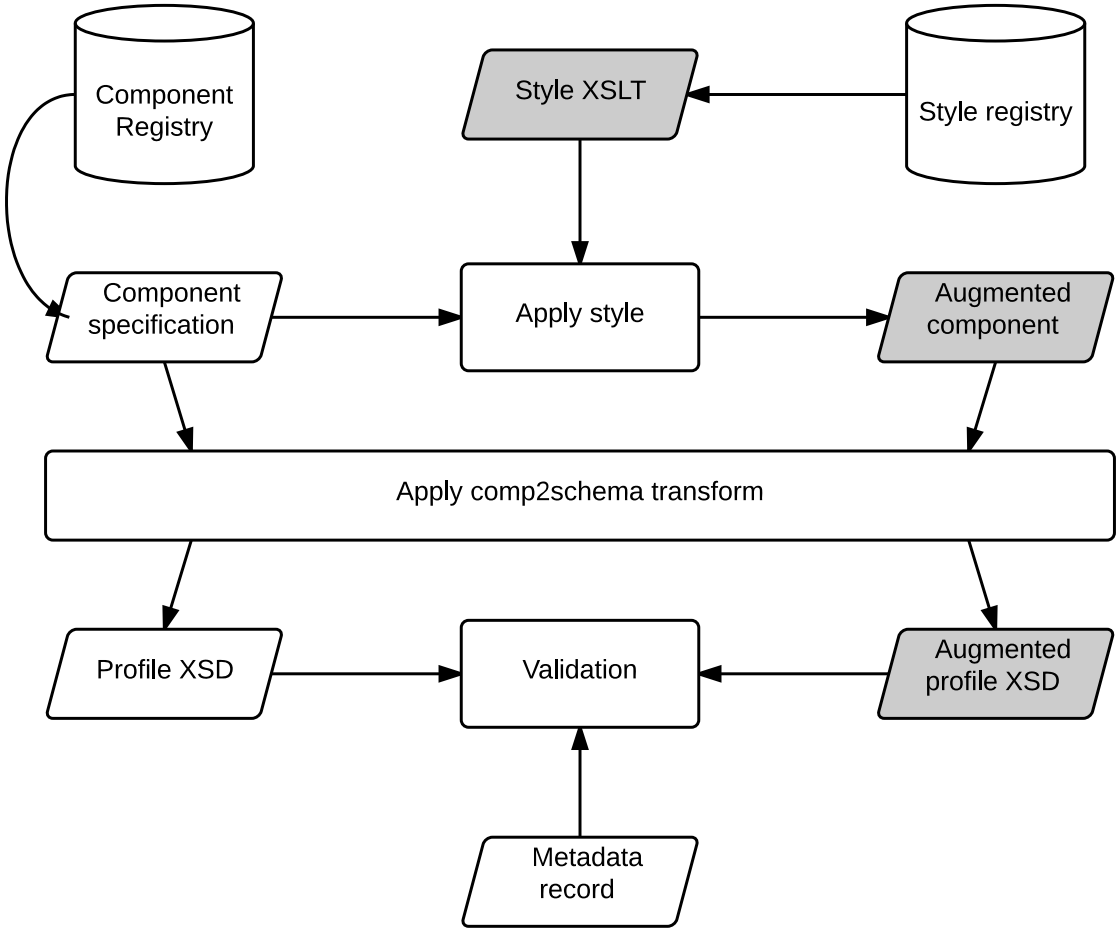


Figure 2: Workflow for visual hints. Shaded components relate to the augmentation of profile schema documents with styling information.

2.3.2 Support for visual hints

Also in the context of user-friendly interfaces extensive changes are introduced to augment metadata profiles with information about how the metadata content should be presented to the user. CMDI 1.1 only provides a very simple approach to specify display priorities for elements. Experiences of recent years showed that this functionality is hardly used and in most cases not even understood by many metadata schema creators. Therefore, this approach is superseded by a new namespace *http://www.clarin.eu/cmdi/cues/display/1.0* for all kinds of display cues. By using an open namespace CMDI 1.2 does not prescribe a closed set of functionality but is completely open for any future extensions that are deemed necessary.

Visual hints that may be useful include:

- Grouping information to allow visual merging of components. This would be especially useful in cases where the content of two components that both contain information about the same issue can be merged and the underlying CMDI structure is not relevant for the end user.
- Selection of elements as representatives of their component. In many cases components contain very extensive information that is relevant in specific contexts but has only minor importance for most users. This could be the case when a component contains very detailed information about a book but only its author and title should be displayed to the user.
- Information about the relevance of a specific element for the whole component that is used as an indication for metadata creators what fields are recommended, optional or even deprecated.

- Explicit visual hints about how an element should be displayed to the end user. This may include suggestions about colour, font size, usage of frames to emphasize specific elements, or the usage of italic or bold letters.

The original component specification is only augmented if necessary. This can be done by means of XSLT transformations according to the workflow as laid out in Figure 2. The current workflow uses CMDI profile specifications that are stored in the Component Registry and converts them via XSLT to XML schemata. These can be used to validate specific metadata record files. This new, supplemental approach extends this workflow by applying XSL transformations provided by a new style registry to enrich the component specification files with additional style attributes. These are also included in the following transformation from component specification to XML schemata. A metadata record that includes style attributes can then be validated against an enriched version of the component specification thus allowing a flexible and expendable workflow without losing the ability to validate a record file against a formal schema. As a consequence of the chosen approach, a tool or a user can decide if display hints are needed at all or may select between different sets of display cues if available.

2.3.3 Value Derivation

A further extension in CMDI 1.2 is the specification of value derivation cues. The experience with CMDI in the last years revealed that a lot of metadata could be automatically derived from other values. The systematic usage of this feature avoids redundancy, helps metadata creators build consistent metadata and allows an explicit definition of relations between elements. Useful applications of this feature may include:

- Definition of duration as the difference of two timestamps.
- Specification of language or country names based on already stated ISO codes.
- Support of keywords like "FileSize" or “CreationDate” that are automatically replaced with their actual value by editor tools.
- Inference of values based on simple regular expressions like the extraction of initials based on already specified first and last name, or the content for a field ‘publication year’ based on a more specific date information.

Similar to the support of visual hints there is no fixed set of allowed rules and keywords. Instead a general framework is specified where most information about relations is defined externally, and the actual derivation is regarded as an optional functionality of applications. Hence it is up to the community what rules, formulas and keywords will establish themselves in the future and what formal structure they will have. Consequently, it is also expected that different tools may support different value calculation methods as there won’t be a central authority that governs a set of allowed values.

In Code example 5, an element holding the age of a file is defined. Its value can be derived from a sibling field *CreationDate*. It assumes a syntax in which a keyword ‘CurrentDate’ exists, as well as a function ‘date’ that in this example takes as its value the path to its sibling element evaluated to its value.

```
<CMD_Element name="AgeOfFile"
          AutoValue="$CurrentDate - date({../CreationDate})"/>
```

Code example 5: Definition of derived values (with hypothetical syntax)

2.4 Attributes in instances

In CMDI 1.1 attributes on instance elements were always optional. The schema for component specifications does not offer a way of expressing the cardinality of an attribute, nor does the Component Registry provide a way of marking an attribute as mandatory. Because of the lack of such an option, it is not possible to closely mimic the constraints of some existing models; the TEI Header (TEI Consortium, 2014), for example, has mandatory attributes. It also poses a needless restriction. In CMDI 1.2, an element 'required' is added to the attribute definition in component specifications in CMDI 1.2 to

allow for both optional and mandatory attributes.

For example, a mandatory attribute could be defined inside an element definition as shown in Code example 6. The profile schema generated from this example would render instances of the *firstName* element without a *nickname* attribute invalid.

```
<Element name="firstName" ValueScheme="string"
Documentation="This is the firstname of a person"
DisplayPriority="0" CardinalityMax="1">
    <!-- provide a nickname attribute for this element -->
    <AttributeList>
          <!-- example of an attribute using a simple type -->
          <Attribute name="nickname" Type="string"
                required="true"/>
    </AttributeList>
</Element>
```

Code example 6: Definition of derived values

3 Fixed CMD functionality

3.1 CMD Namespaces

In CMDI 1.1 a CMD namespace, i.e. *http://www.clarin.eu/cmd/*, was introduced. All CMDI records use this namespace, regardless of the profile, and thus XML Schema. This approach, although simple, has led to problems with the basic assumptions about XML, namespaces and schemas made by tools and standards outside of CLARIN. For example, the metadata harvesting OAI-PMH protocol (Lagoze et al, 2002), which is used by CLARIN but specified by the Open Archive Initiative, demands that only one schema is associated with a metadata prefix. But CMDI metadata comes with many schemas, a different one for each profile. Also tools, such as Xerces2-J,[14] that perform XML Schema validation, assume (backed by the XML Schema recommendation (Thompson et al, 2004)) that a namespace is associated with a unique schema and base their caching strategy on this. In CMDI 1.2 therefore, a general namespace for the CMDI Envelope, and profile specific namespaces for the payload are added. (Code example 7 illustrates the use of these two namespaces.) This allows binding of the CMDI Envelope schema to the OAI-PMH CMDI metadata prefix and also supports caching of profiles specific schemas. In principle this change touches every resource and tool in the infrastructure. Fortunately many of these tools can use various approaches, e.g. wildcards, to ignore the profile specific namespaces when they access arbitrary CMDI records.

[14] http://xerces.apache.org/xerces2-j

```
<cmd:CMD
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:cmd="http://www.clarin.eu/cmd/1"
xmlns:cmdp="http://www.clarin.eu/cmd/1/profiles/clarin.eu:cr1:p_13119
27752306"
CMDVersion="1.2"
xsi:schemaLocation=
"http://www.clarin.eu/cmd/1 http://www.clarin.eu/cmd/1.2/envelop.xsd
http://www.clarin.eu/cmd/1/profiles/clarin.eu:cr1:p_1311927752306
http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/
clarin.eu:cr1:p_1311927752306/1.2/xsd">
    <cmd:Header>
          …
    </cmd:Header>
    <cmd:Components>
          <cmdp:ToolService>
                …
          </cmdp:ToolService>
    </cmd:Components>
</cmd:CMD>
```

Code example 7: Fragments of a CMDI record illustrating the use of the namespace

Another namespace related issue is the potential clash between reserved attributes, i.e. *ref* and *componentId*, and user defined attributes. In CMDI 1.2 reserved attributes are moved to the general CMD namespace, so the user has the freedom to define attributes with arbitrary names. These arbitrary names include the names which were reserved for CMDI attributes in 1.1, as shown for *ref* in Code example 8.

```
<cmdp:name cmd:ref="h42"  ref="http://viaf.org/viaf/113230702"
type="person">Douglas Adams</cmdp:name>
```

Code example 8: The separate namespace for envelope and payload allow usage of the *ref* attribute that was a reserved attribute in CMDI 1.1

3.2 Changes in the CMD Envelope

In CMDI 1.1, *IsPartOfList* with its *IsPartOf* elements can be used to link to collections that the described resources and/or metadata are part of. However, the nature of the (implicit) subject of an *IsPartOf* statement has been unclear. While its current position within the Resources element may indicate that any *IsPartOf* relation applies to *all* resources referenced in *ResourceProxyList*, its mere name 'IsPartOf' indicates a single subject.

In CMDI 1.2, this issue will be resolved by moving *IsPartOfList* to the envelope top level alongside Resources, and restricting the semantic of *IsPartOf* to express a partitive relationship between the described resource as a whole and some collection or larger resource. See Code example 9 for an illustration.

```
<CMD xmlns="http://www.clarin.eu/cmd/1">
      <Header>
            <MdProfile>clarin.eu:cr1:p_1345561703673</MdProfile> ...
      </Header>
      <Resources>...</Resources>
      <IsPartOfList>
            <IsPartOf>http://infra.clarin.eu/example/mycollection.cmdi
            </IsPartOf>
      </IsPartOfList> ...
</CMD>
```

Code example 9: Usage of *IsPartOfList* in a metadata record (CMDI instance)

Other relationships between resources than *IsPartOf* can, broadly speaking, be expressed in one of two ways in the CMDI framework; either using components and elements, or as *ResourceRelation* elements within the *Resource* section of the CMDI envelope. *ResourceRelations* in CMDI 1.1 contain simply a *RelationType* element giving a name for the relation, together with elements *Ref1* and *Ref2* pointing to the related resources.

Existing data shows that the latter method has been very little used. There seems to be a general feeling that the current *ResourceRelation* is too simplistic and underspecified to convey the intended information. Although no fundamental change will be performed in CMDI 1.2, the intention is to clarify the semantics of the current specification, all the while keeping the door open for expressivity extension at a later date.

In CMDI 1.2, *ResourceRelation* elements should always contain exactly two Resource elements (replacing *Res1* and *Res2*), explicitly constraining relationships to be binary. In these elements, a mandatory *ref* attribute (indicating a resource listed in the same CMDI record) and an optional *Role* element with an optional *ConceptLink* attribute is added. Moreover, *RelationType* is extended with an optional *ConceptLink*. The new scheme is illustrated in Code example 10, in which the (fictitious) *ConceptLinks* refer to the CLARIN Concept Registry.[15] This way, both relationship direction as well as semantic marking of both relation type and resource roles may be defined by metadata creators.

```
<ResourceRelationList>
      <ResourceRelation>
            <RelationType
      ConceptLink="http://hdl.handle.net/11459/CCR_C-2318_bfda5ab9-
      a429-c2e5-8f08-7c8dfca8245a">annotates</RelationType>
            <Resource ref="rp1">
                  <Role ConceptLink="http://hdl.handle.net/11459/CCR_C-
            346_bfda5ab9-a430-c2e6-8f08-7c8dfca8245a">annotation</Role>
            </Resource>
            <Resource ref="rp2">
                  <Role Conceptlink="http://hdl.handle.net/11459/CCR_C-
            417_bfda5ab9-a429-c2e6-8f08-7c8dfca8245a">annotated</Role>
            </Resource>
      </ResourceRelation>
</ResourceRelationList>
```

Code example 10: Example of *ResourceRelationList* in a metadata record (CMDI instance)

3.3 Component Schema Cleanup

Since the development of CMDI started, multiple developers have worked on the schema that governs how CMDI profiles and components are specified in XML. Different modelling strategies have been applied leading to a mixed bag, e.g. most properties of CMDI elements are specified via XML attributes while similar properties are specified in XML elements for CMDI attributes, as is showcased in Code example 11 (left hand side). In CMDI 1.2 (example on the right hand side) these different approaches are cleaned up by going back to the original approach of using XML attributes whenever applicable.

[15] http://www.clarin.eu/conceptregistry

```
<CMD_Element
Multilingual="true"
CardinalityMax="1"
CardinalityMin="1"
ValueScheme="string"
name="Description">
<AttributeList>
    <Attribute>
      <Name>LanguageID</Name>
      <Type>string</Type>
    </Attribute>
 </AttributeList>
</CMD_Element>
```

```
<CMD_Element
Multilingual="true"
CardinalityMax="1"
CardinalityMin="1"
ValueScheme="string"
name="Description">
<AttributeList>
    <Attribute
    name="LanguageID"
    type="string"/>
</AttributeList>
</CMD_Element>
```

Code example 11: Comparison of the element and attribute definition in CMDI 1.1. and 1.2

4 Migration from CMDI 1.1 to 1.2

Centres should upgrade their data and tools if they wish to benefit from the changes in CMDI 1.2 and good integration with the infrastructure as other centres are upgrading as well. New tools and future versions of existing tools may support CMDI 1.2 only and may not be applicable to unconverted metadata (although conversion can always be performed on the fly, either transparently by the tool or as a pre-processing step by the client).

CMDI 1.1 will be phased out in the future, but initially the core infrastructure components will support both version 1.1 and 1.2, allowing centres to migrate at their own pace. Centres may choose to keep supporting both versions after upgrading, for example by performing on the fly transformations. Migrating to CMDI 1.2 is an active migration process requiring varying degrees of effort from the centres depending on the specifics of the repository and/or tools maintained by the centre involved. The CMDI task force will supply a ready-to-use upgrade mechanism, based on Extensible Stylesheet Language Transformations (XSLT) stylesheets, that will allow centres to convert their metadata records from CMDI 1.1 to 1.2, either one time statically (individually or in batch) or dynamically on the fly. Figure 3 shows a schematic overview of the various aspects of the migration from CMDI 1.1 to CMDI 1.2.

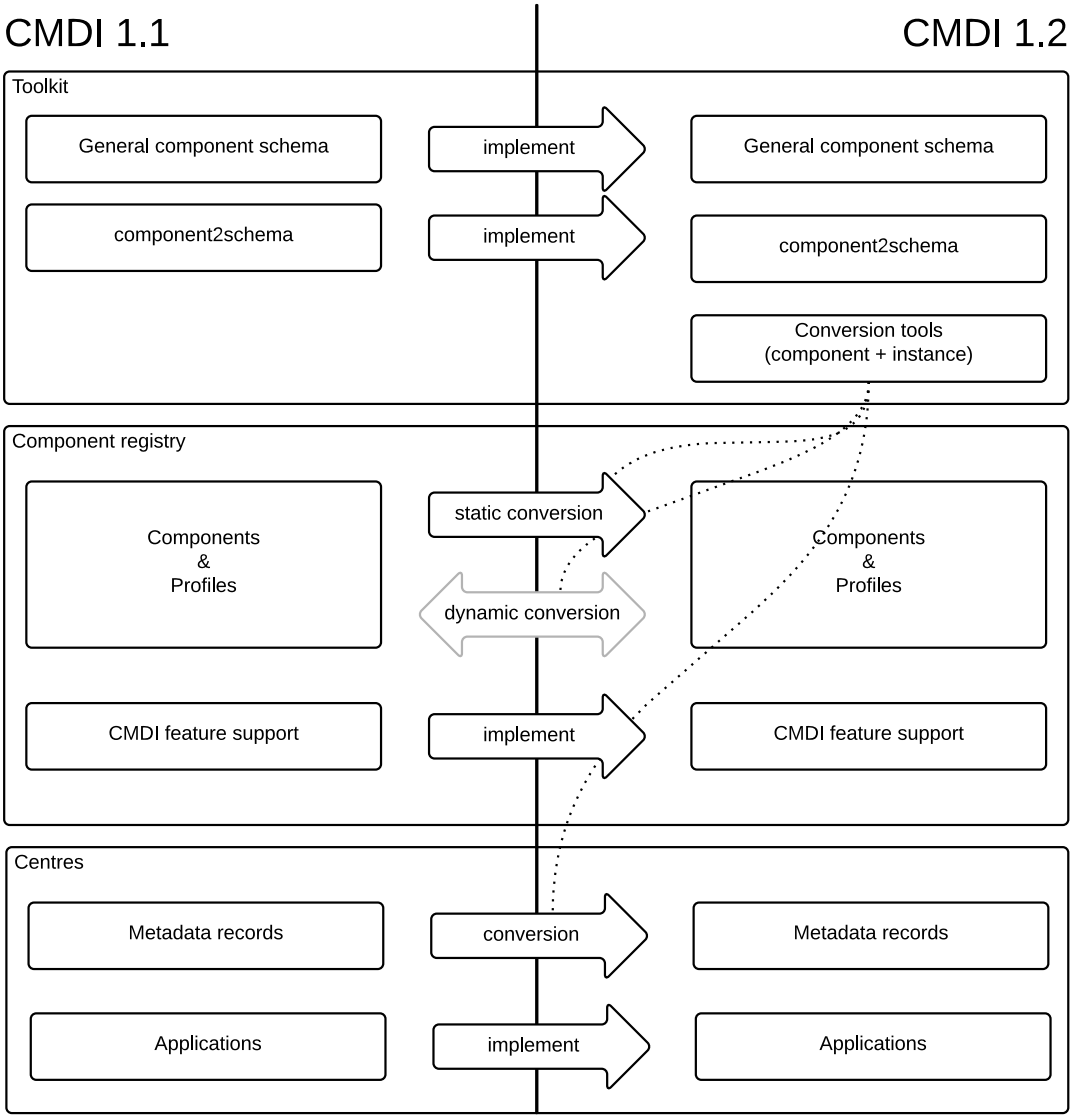


Figure 3: Conversion between CMDI 1.1 and CMDI 1.2: general overview of the toolkit (top) upgrade and the necessary upgrade steps in subsequent parts (below) of the infrastructure.

4.1 CMDI Toolkit and Component Registry

The CMDI toolkit comprises the definitions (in the form of XML Schema Definition (XSD) and XSLT documents) that define the language for the specification of metadata components and profiles as well as the structure of metadata instances in relation to profiles. The task force will produce a new version of this toolkit, which then provides the essential components for creating CMDI 1.2 metadata.

The Component Registry is built on top of this toolkit and will be the first infrastructure component to be adapted to support CMDI 1.2. All existing components and profiles stored in the Component Registry will be statically converted to CMDI 1.2 using an XSLT stylesheet that is part of the toolkit. These components and profiles will become available at a new location in the Component Registry's web service. CMDI 1.1 versions of all components and profiles will be generated on-the-fly by applying a downgrade XSLT and can be requested by tools and users at their current locations. Therefore, the Component Registry will remain compatible with existing infrastructure components. An analysis has shown out that converting existing components and profiles (i.e. those that were present in the registry before the conversion to 1.2) back to CMDI 1.1 after the upgrade can be carried out losslessly, therefore the validity of existing metadata instances is not affected.

Components and profiles that will be created after CMDI 1.2 support has been added to the Compo-

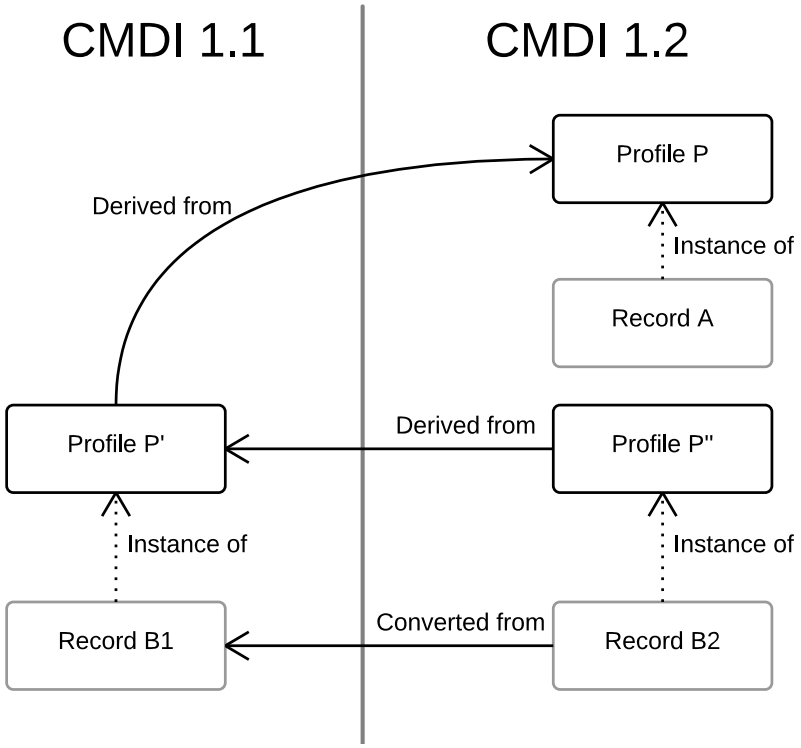


Figure 4: Workflow for a CMDI 1.1 record that is created on basis of a CMDI 1.2 profile and later converted to a CMDI 1.2 instance

nent Registry cannot be lossless converted to 1.1 in all cases, as they may make use of one or more of the newly added features. This poses no problem, as there are no pre-existing instances based on these specifications.

A scenario that needs to be supported by the infrastructure is depicted schematically in Figure **4**. In this scenario, CMDI 1.1 metadata gets created based on profiles that are based on 'native' CMDI 1.2 specifications. If such metadata eventually gets converted to CMDI 1.2 it will not necessarily be valid to the original CMDI 1.2 specification. For example, a CMDI 1.2 profile schema (*P*) might define a mandatory attribute, an option not available in CMDI 1.1. Therefore, the 'dumbed down' profile schema (*P'*) will allow omission of this attribute in instance records (such as *B1* in the diagram). To allow for such a scenario without rendering the metadata invalid when upgrading (yielding *B2*), the Component Registry will also provide a 'dumbed down' CMDI 1.2 version (*P"*) of each profile, which in fact will be the result of applying the specification upgrade script to the result of the specification downgrade script applied to the original specification. This version of the profile schema will be available through a separate call, which will perform the chained conversion on the fly. When upgrading a CMDI 1.1 metadata record, its schema location reference should be set to this version of the schema in case the profile is based on a 'native' CMDI 1.2 specification; in other cases, the original CMDI 1.2 version of the schema should be referenced, allowing usage of new CMDI 1.2 features in the instance. This is not an issue if the output of the conversion is either transient or not subject to change.

5 Conversion of CMD Records

The task force will provide an XSLT stylesheet for upgrading metadata records from CMDI 1.1 to CMDI 1.2. Upgrading a record entails transforming the schema reference into a reference to the schema based on the CMDI 1.2 version of its profile (in some cases this should be the 'dumbed down' version, see above) and applying all required changes to make the document compliant with the CMDI 1.2 specification (see sections 2 and 3). No information will get lost in the upgrade process, and the component structure will not change.

In some exceptional cases, an automated transformation cannot be carried out. Specifically, if no profile reference is present in the original record or multiple *ref* attribute values are found on a single

element (both of which are schema valid in CMDI 1.1). If such a case is encountered during transformation, the stylesheet will yield an error and the owner of the record will have to adapt the record manually.

A method for converting (downgrading) CMDI 1.2 records to CMDI 1.1 will not be provided by the task force, as there is no generally applicable way of doing so without potentially losing information. In cases where centres or individuals do wish to perform such a conversion, a conversion targeting specific profiles should in generally be quite straightforward. A reason for doing so could be the desire to apply a tool that only supports CMDI 1.1 to a native CMDI 1.2 record.

5.1 Tools, Services and Repositories

Since the Component Registry will keep supporting CMDI 1.1, the need to upgrade other tools, services and repositories hosted and maintained by the centres will not be pressing immediately in most cases. Centres will probably not be inclined to permanently switch to CMDI 1.2 before the majority of relevant tools supports it. On the other hand, the development and adaptation of tools will be driven by the availability of metadata. Adding support for CMDI 1.2 to central tools and services that deal with a broad variety of metadata sources and types, such as the Virtual Language Observatory, will be most urgent. As soon as some support exists in the exploitation stack, it makes sense for repositories to start providing CMDI 1.2 metadata. In some cases this can be achieved by simply applying (additional) transformations. In other cases, however, this will depend on more thorough modifications in the metadata creation pipeline, including editors and content management systems, especially if the new features of CMDI 1.2 are to be harnessed. Centres that generate CMDI on the fly, based on a separate primary data source such as a relational database, have the choice to keep providing both CMDI 1.1 and CMDI 1.2 alongside each other.

Based on the namespace URIs OAI endpoints are able to provide different versions of the CMDI records. The *http://www.clarin.eu/cmd* namespace URI corresponds to CMDI 1.1. While any higher minor version of CMDI 1.x will use the *http://www.clarin.eu/cmd/1* and with the next major version change the *http://www.clarin.eu/cmd/2* URI will be used. This scheme does require future minor versions within a major version to be compatible with each other.

6 Roadmap

Work on the implementation has begun mid 2014, starting with the creation of a new version of the toolkit. Once this has been completed, the Component Registry software stack (REST service and front end web application) will be updated, followed by the migration of all registered components and profiles. After this, the remainder of the infrastructure can be migrated in a distributed fashion. CMDI 1.1 can be formally deprecated once a significant share of the existing records has been migrated and all relevant tools have been adapted. CMDI 1.1 will keep being supported at the core infrastructure level even after deprecation, as will CMDI 1.2 after its eventual succession.

There are a number of tasks related to CMDI 1.2, some of which are currently being worked on, and some of which are planned for after or in parallel to the implementation of CMDI 1.2. First of all, the CMDI task force has initiated the process of writing an extensive and formal specification of CMDI. Such a specification does not exist for CMDI 1.1. Members of the task force have started working on this specification and expect to finish the document in the second half of 2015. In addition to this formal description of the technical scope of CMDI, a document describing *best practices,* targeted primarily at the metadata modeller, is under development.[16]

There is on-going work - coordinated by the CLARIN Metadata Curation task force - on evaluating the quality of the metadata records in the joint metadata domain (cf. Trippel et al., 2014). The main goal is to provide a service that examines individual records or whole collections, performing a number of basic checks (schema validation, "dead links", etc.), and optionally normalisation of values based on controlled vocabularies, producing a curation report that lists encountered issues. The checks will especially also cover the specifics of the CMD versions, to support the data provider in the transition period. Once completed, this service will be integrated into the basic workflow for harvesting the

[16] At time of writing, a draft version of the CMDI best practice guide is available at http://www.clarin.eu/content/cmdi-best-practice-guide

metadata and filling the VLO.

6.1 Open issues

The CMDI task force has decided to leave a number of known shortcomings and potential improvements unaddressed in CMDI 1.2. Rather, these specific issues should be investigated further so that, if feasible, a reliable and non-controversial solution can be incorporated in a future version of CMDI. This section briefly describes four salient ones.

Metadata record versioning information
Most metadata records are subject to change over the course of their lifespan due to for example content fixes, extension, or adaptation to external circumstances. Sometimes a change is applied impromptu, so that a newer version overwrites an existing one, while in other cases a versioning policy is in place that ensures that older versions remain available and each new version gets a distinct identifier. The same applies to resources. In either case, it's often desirable to encode versioning information close to the versioned item. Ways of doing this within CMDI records can be thought of, but an investigation of use cases and ways of representing such information, ideally based on existing common practices, has to be carried out in order to derive one or more appropriate candidate solutions.

Recursive component definitions
Any component or profile specification in any existing version of CMDI, including version 1.2, can be modelled as a tree. The Component Registry does not accept component specifications that hold a reference resulting in a cycle. Therefore no CMDI schema can be derived that allows for arbitrary depth of nesting. To illustrate, one cannot model a component *A* such that it contains a component *B* which in turn specifies component *A* as a descendant. XSD does allow for such circular references, and in fact some existing metadata schemata contain them. For example, the schema for MODS defines an element 'relatedItem' within 'mods', which can hold the same child elements as a 'mods' elements, including 'relatedItem' (Gartner, 2003). However, in CLARIN it is strongly suggested best practice to use semantically explicitly specified concepts for metadata elements. This is in strong contrast to very general concepts such as 'relatedItem' where the position within the metadata tree crucially contributes to the semantics of a given metadata element. The best practice approach strongly reduces the need to exploit the structure of the metadata and therefore reduces the need for recursive use of components. Moreover, introducing the possibility of circular references in component specifications would require a number of fundamental changes in tools that process CMDI records on basis of component specifications or profile schema files.

Nillable fields
Element types in CMDI are derived from XSD types. An option on types that is available in XSD, but not adopted in CMDI, is nillability. While many of the potential use cases for *nil* values can be covered by omitting optional fields, or leaving a string element blank, there are also cases where there is no proper alternative. For example, a modeller might decide that date information should be mandatory, but also want to support cases where date is undefined. Leaving a mandatory date element empty renders an instance document invalid with respect to the schema, so that would not be a proper solution.

The need for workarounds, such as representing dates and booleans as strings or making fields optional where they should not, can be removed by allowing, on selected elements, for the XML-standard attribute *xsi:nil*[17] (which takes the value true to indicate a non-value in the record). It can be combined with a specification of the semantics of *nil* in its particular context (e.g. whether it represents 'unknown' or 'unspecified'), either in the record or in the component specification. This would also prevent metadata creators from entering bogus information to force validity. However, before such support can be added, the methods for defining the exact semantics of *nil* values need to be decided on. Furthermore, the effects on profile schema generation from the component specification need to be investigated.

[17] http://www.w3.org/TR/xmlschema-1/#xsi_nil

Resource proxy constraints

Finally, the task force has discussed but not yet designed or implemented, ways of controlling, via the component specification, the range of resource proxy types and allowed reference points to these proxies. By means of such a facility, a metadata modeller could for instance specify that an instantiation of an 'audio recording' profile should only contain resource proxies with 'audio' media types.[18] Similarly, profiles intended to be used for metadata collections could restrict resource proxies to those of the 'metadata' type.

As an example of controlling the resource reference points, a multimedia session profile might require a reference for each audio or video resource proxy from a 'technical details' section and to a text file proxy from a 'transcription' section.

Such a specification mechanism provides the modeller with some control over the resources coupled to metadata documents, which is lacking from current CMDI implementations. A level of validity (in addition to schema validity) could be derived from this aspect of the specification in relation to a metadata instance. For this reason, this proposal needs to be worked out further before it can be integrated centrally into the CMDI framework. The 'cues for tools' extension mechanism described in this paper could be used to add comparable functionality on the level of user guidance, depending on support by editors and other tools.

7 Conclusion

After 5 years of intensive usage by the community the CMDI task force has reflected the gathered experience in a new minor version of CMDI – CMDI 1.2 – with a number of fixes and improvements. The proposal being finalized and approved, the work now concentrates on a smooth transition of the infrastructure and the data. It is hoped that the CMDI community will largely and successfully adopt CMDI 1.2 and provide the support required to implement these and other enhancements in the future.

References

Broeder, D. Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C (2010, May). A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*, pages 43–47, Valletta, Malta

Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T., and Trippel, T. (2012). CMDI: a Component Metadata Infrastructure. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme*.

Brugman, H., and Lindeman, M. (2012). Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme*.

Ďurčo, M. & Windhouwer, M. (2014). From CLARIN Component Metadata to Linked Open Data. In *Proceedings of the third Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing,* LREC 2014 Workshop.

Gartner, R. (2003). MODS: Metadata Object Description Schema. *JISC Techwatch report TSW*, 03-06.

Gavrilidou, M.; Labropoulou, P.; Desipri, E.; Giannopoulou, I.; Hamon, O. & Arranz, V. (2012). The META-SHARE Metadata Schema: Principles, Features, Implementation and Conversion from other Schemas. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme*. LREC 2012, Istanbul.

Henrich, A., & Gradl, T. (2013). DARIAH (-DE): Digital Research Infrastructure for the Arts and Humanities-Concepts and Perspectives. International Journal of Humanities and Arts Computing, 7(supplement), 47-58.

[18] As defined by the *mimetype* attribute in the Resource Proxy, referring to Media Types, formerly known as MIME types; see http://www.iana.org/assignments/media-types/media-types.xhtml.

Lagoze, C., Van de Sompel, H., Nelson, M., and Warner, S. (2002). *The Open Archives Initiative Protocol for Metadata Harvesting*. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm. Accessed on 20 June 2014.

TEI Consortium, eds. (2014). *Guidelines for Electronic Text Encoding and Interchange*. 20 January 2014. http://www.tei-c.org/P5/. Accessed on 20 June 2014.

Thomson, H.S., Beech, D., Maloney, M., and Mendelsohn, N. (2004). *XML Schema Part 1: Structures Second Edition*. http://www.w3.org/TR/xmlschema-1/. Accessed on 20 June 2014.

Trippel T., Broeder, D., Durco, M. and Ohren, O. (2014) Towards automatic quality assessment of component metadata. In *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC). Reykjavik, Iceland, 26-31 May, 2014.* Pages 3851-3856.

Windhouwer, M., Goosen, T., Schonefeld O, Ohren, O., Eckart, T., Herold, A., Misutka, J., Frankhauser P., Schiel, F., Eckart, K., et al. (2014). *CMDI 1.2 changes - executive summary*. Technical Report CE 2014-0318, CLARIN ERIC, http://www.clarin.eu/content/cmdi-12-changes-executive-summary.

Data curations by the Dutch Data Curation Service:

Overview and future perspective

Henk van den Heuvel
CLST, Radboud University
Erasmusplein 1, 6525 HT Nijmegen
The Netherlands
h.vandenheuvel@let.ru.nl

Nelleke Oostdijk
CLST, Radboud University
Erasmusplein 1, 6525 HT Nijmegen
The Netherlands
n.oostdijk@let.ru.nl

Eric Sanders
CLST, Radboud University
Erasmusplein 1, 6525 HT Nijmegen
The Netherlands
e.sanders@let.ru.nl

Vanja de Lint
CLST, Radboud University
Erasmusplein 1, 6525 HT Nijmegen
The Netherlands
v.delint@let.ru.nl

Abstract

Data curation comprises activities such as digitizing data (where necessary), converting the data so as to conform to accepted standard formats, (re)shaping metadata and adding documentation. In this contribution we present the motivation for a data curation service (DCS) in the CLARIN-NL project, and the activities the DCS employed during the past years in curating a variety of resources, including dialect dictionaries, speech databases for language acquisition and interview data. In the second part, we present a view on how in the future data curation is best addressed as an integral part of research data management and what could be the role for an expertise centre like the DCS in this context. We envisage and advocate a shift in the future in which data management becomes an integral part of the overall research data management plan (DMP) right from the start of a project. For researchers the university libraries are a natural entry point for data management issues. The data expertise centres can be installed as back offices for consultancy and data curation tasks.

1 Introduction*

In line with developments we see at the European level (e.g. Calzolari et al., 2014), in the CLARIN-NL project (Odijk, 2014; Odijk 2010) substantial efforts have been made to contribute towards the development of an infrastructure that supports the sharing and re-use of resources, and that opens up new avenues of research as it allows for combining various resources in new and unforeseen ways. Apart from work on the implementation of the technical part of the infrastructure, there have been several resource curation and/or demonstration projects which should bring this infrastructure to life and promote its actual use.[1] The Data Curation Service (DCS) hosted at the Centre for Language and Speech Technology in Nijmegen was originally set up as a centre of expertise which aimed to assist researchers, especially those without the time, money, or know-how, in preparing their data for delivery to one of the CLARIN centres that operate as hubs in the CLARIN infrastructure (Oostdijk & van den Heuvel, 2012). Data curation involves digitizing data (where necessary), converting the data so as to conform to

[1] For an overview of resources that were created within the CLARIN NL project and that are now part of the CLARIN NL infrastructure, or that were created by other projects but are essential for functioning of the CLARIN (NL) infrastructure, we refer to the CLARIN NL Portal pages (CLAPOP): https://dev.clarin.nl

CLARIN accepted standards or preferred formats, (re)shaping metadata and adding documentation. The DCS typically has served as intermediary between the researcher and the eventual data centre.

In this contribution we first give an overview of the data curation efforts the DCS has been involved in, which at once shows the diversity of the language resources at stake and the various issues we came up against. In the second part, we present a view on how in the *future* data curation is best addressed as an integral part of research data management and what could be the role for an expertise centre like the DCS in this context.

2 Data Curation

In the two years that the CLARIN-NL DCS has been operational, its focus has been on the curation of data collections residing with and used by individual researchers or research groups in the Netherlands. Candidates for curation were identified and for each it was assessed as to (1) whether it would be *desirable* to have the resource curated and (2) whether successful curation would be *feasible*. A more elaborate description of how these criteria can be operationalized is given in Oostdijk et al. (2013).

Most of the data collections targeted by the DCS were collections that were compiled in projects that were already finished and of which many did not receive any follow up, so that in effect the data were at risk of being lost. Curation of such collections can be challenging, especially when they were created in a context where little or no thought was given to the idea of sharing or re-use. Often IPR has not been settled or if it has, the arrangements did not anticipate the distribution or wider use of the data. Typically data formats are diverse, metadata and documentation incomplete. Since settling IPR for already existing collections was deemed problematic, the DCS has refrained from taking on the curation of resources for which any IPR issues remained to be settled.

Thus, curation of resources as undertaken by the DCS involved a number of actions. We combine this overview with a report on a number of experiences and lessons learned.

Data collection

Upon identifying a resource that was in need of curating, the first step in the curation process was directed at establishing what constituted the complete and final set of data. Especially with data that came into existence in the course of research projects where at the start of the project not much thought was given to what would happen to the resource once the project ended, we found that datasets were not always well-defined in the sense that data collection within the project did not necessarily follow a strict plan: some of the data planned were not realized whereas apparently other unplanned data were found and subsequently included. The time needed for data collection should not be under-estimated. Substantial efforts were sometimes involved in obtaining the data, that is, the final version of the data and the accompanying documentation, especially if more than one researcher was involved in the project. Furthermore, interpreting and linking data and metadata should be done involving where possibly the researcher, who, understandably, is not at all times available.

IPR check

Since the DCS restricted itself to resources for which IPR supposedly had been settled, the IPR check was directed at making sure that the data could be incorporated in the CLARIN infrastructure. Depending on the IPR this incorporation could take on a variety of forms ranging from showing (e.g. in the Virtual Language Observatory) the mere existence of a resource via its metadata to making it completely accessible and downloadable for end-users.

Format conversion

The curation of existing resources often required that the formats that had been used be converted into the standard formats adopted in CLARIN. This logically followed from the fact that many resources had been created before the current standards had been established. Moreover, the list of accepted standards evolved over time. For instance, Praat[2] transcription files were not among the standard formats at the start of CLARIN-NL, but were accepted in later stages.

[2] http://www.fon.hum.uva.nl/Praat/

Anonymization
Occasionally, it proved necessary to anonymize the data. Anonymization was typically done in transcriptions, metadata and file names. It appeared too difficult to implement a single anonymization methodology for all data-sets since particular types of data may require and/or make possible different approaches while occasionally individual researchers had clear preferences for one approach or the other.

Providing metadata (CMDI-compliant)
As CMDI is the current standard for metadata in CLARIN, the metadata available with all resources should be CMDI-compliant.[3] In terms of curation this entailed that an appropriate CMDI metadata profile had to be identified and modified where necessary. Subsequently, this profile had to be filled with the metadata pertaining to the resource at hand. With respect to CMDI metadata profiles we came to the conclusion that it is best to publish a new CMDI profile for each database at project level by selecting and constructing CMDI building blocks from selected other profiles (and introduce one or more new metadata categories) and not at database type level. One will never be able to publish an all encompassing CMDI profile covering all databases of a similar type (e.g. second language acquisition), since the variety of encountered metadata is vast, and the overall profile will never be complete.

Documentation
With each curated resource two types of documentation were to be made available: (1) documentation describing the design, collection, annotation etc. of the resource, preferably with reference to the research context in which it was produced, and (2) a curation report in which the various steps taken in the curation process were documented and accounted for.

Packaging and delivery
Once the curation process had been completed, the resource was delivered to a CLARIN data centre. The data centre would then take care of adding persistent identifiers and storage of the curated resource.

3 Curated language resources

The DCS has curated a variety of resources. In this section we report on their curation while grouping them into different categories following the language resources typology presented in Gavrilidou et al. (2012):

1. Lexical resources: Dialect databases
2. Multimodal and multilingual corpora: Language acquisition databases
3. Oral/spoken corpora: IPNV interviews

3.1 Lexical resources: Dialect databases

Over the years various projects have undertaken the description of Dutch dialects. This has resulted in an extensive collection of books (dictionaries) covering a wide range of regional and local dialects. These dictionaries are unique instruments for research into variation linguistics, which is currently a field of study that is attracting a lot of interest. The dictionaries have been compiled on the basis of oral and written surveys in which thousands of informants have taken part and the analyses of the collected material by dozens of dialectologists. Most of these dictionaries have been completed, and the researchers and other people involved are retired. The digital files are in different formats and are located at many different institutions; sometimes they are kept by individuals. These files are thus fragmented and are seriously at risk of remaining inaccessible for others. If nothing would be done, they might eventually be lost all together. By bringing the files together and curating them into standard formats, they become accessible to a large group of users. This, we expect, will enable researchers to formulate new research questions since the different datasets can now be studied and consulted individually but also in

[3] For more information on CMDI, see http://www.clarin.eu/CMDI

comparison to the other datasets. Thus a range of dialect dictionaries for which the IPR had been cleared were offered for curation by the DCS.

The dialect databases originally came in various formats including exports of MySQL, MS Access, and FileMaker Pro. None of these formats is an accepted CLARIN format. The LMF format, however, is. LMF stands for Lexical Markup Framework and is an XML standard which is typically suited to capture hierarchical lexicon structures (Francopoulo, 2013). We departed from a first LMF model used in the COAVA project[4] and made an extended version of this. Our LMF model is based on three head features associated with Lexical Entry, viz.

- Form
- Sense
- Location

Two further head features are Definition and Context (both positioned under Sense). Each individual feature is linked to an ISOcat[5] data category (cf. Windhouwer & Wright, 2013) as shown in Table 1. Only Form Keyword is mandatory.

LMF feature	**Corresponding ISOcat element**
Form Keyword=	278 keyword
Form Representation aggregatedKeyword=	278 keyword
Form Representation lexvariant=	5585 lexical variant
Form Representation morphologicalvariant=	5758 morphological variant (new, defined by DCS)
Form Representation grammaticalInformation=	2303 grammatical unit
Form Representation dialectform=	1851 geographical variant
Form Representation standardizedform=	1851 geographical variant
Form Representation phoneticform=	1837 phonetic form
Sense lemma-id=	288 lemma identifier
Sense lemma=	286 lemma
Sense meaning=	464 sense
Definition definition=	168 definition
Definition sourcelist= Definition sourcebook=	5759 source list (new, defined by DCS) 471 source
Definition sourcelistnumber= Definition sourcebookpage=	5760 source list number (new, defined by DCS) 4126 pages
Context timecoverage=	3664 Time coverage
Context example=	3778 example
Context comment=	4342 Comment

[4] http://www.meertens.knaw.nl/coavasite/
[5] http://www.isocat.org/

Location place=	3759 source
Location area	3814 region
Location subarea=	3814 region
Location informant-id=	3597 speaker id
Location kloeke=	3651 Kloeke geo-reference

Table1: LMF features in the LMF model for dialect databases and corresponding ISOcat elements

We were able to capture all dialect databases in this framework. The databases were converted into Excel which was considered the intermediary format. Excel files can be converted and imported by tools that are typically used by dialectologists. Care was taken that all data was encoded using UTF-8. The databases were exported as tab-separated text files and converted to LMF by means of a Perl script. This script is a generic script based on a mapping of field headers to corresponding LMF features which has to be defined in the header of the script. Phonetic transcriptions (as found to occur in the WBD, i.e. the Dictionary of the Brabant Dialects, and the WLD, i.e. the Dictionary of the Limburgian Dialects)[6] were preserved in SIL IPA.

Metadata for each lexical database was entered in the WND profile,[7] a CMDI profile created for the COAVA project (Cornips et al. 2011).

In this way the following dialect databases were curated:

- WLD and WBD part III (Dutch dialect dictionaries from Brabant and Limburg)
- Woordenboek Gelderse Dialecten, Rivierengebied
- Woordenboek Gelderse Dialecten, Veluwe
- Melis-van Delst (2011) Bikse Praot. Prinsenbeeks Dialectwoordenboek. (Dialect dictionary of the town Prinsenbeek in Brabant)
- Swanenberg, A.P.C. (2011). Brabants-Nederlands Nederlands-Brabants: Handwoordenboek. (Dictionary Brabantic-Dutch, Dutch-Brabantic)
- Panken, P.N. (1850) Kempensch taaleigen. (Dialect dictionary of the town Bergeijk in Brabant)
- Hendriks, W. (2005) Nittersels Wóórdenbuukske. Dialect van de Acht Zaligheden. (Dialect dictionary of the town Netersel in Brabant)
- Laat, G. de (2011) Zoo prôte wèij in Nuejne mi mekaâr. (Dialect dictionary of the town Nuenen in Brabant)
- Bergh, N. van den, et al. (2007) Um nie te vergeete. Schaijks dialectboekje. (Dialect dictionary of the town Schaijk in Brabant)

All curated databases were transferred to the Meertens Institute where they were assigned persistent identifiers and stored.

3.2 Multimodal and multilingual corpora: Language acquisition databases

LESLLA

The LESLLA corpus was collected between 2003-2005 in the framework of the research project *Stagnation in L2 acquisition: under the spell of the L1?* sponsored by NWO (the Dutch Organisation for Scientific Research). The corpus contains valuable data for studying low-educated second language and literacy acquisition, but had been lying idly on the shelf ever since the project came to an end.

[6] For the WBD and WLD see http://dialect.ruhosting.nl/wbd/index.htm and http://dialect.ruhosting.nl/wld/index.htm respectively.

[7] See http://catalog.clarin.eu/ds/ComponentRegistry/#

The main research question in the project was to what extent the first language impeded the acquisition of the second language in the tutored context of a language course. The 15 participants in the original study had to carry out five tasks which all involved spoken language but varied from strictly controlled to semi-spontaneous. The recordings took place in three cycles of about 6 months each. In each cycle the same tasks were repeated by each participant. The recordings of one cycle were done in three separate sessions (in order to avoid an overload for the participant). Thus there were 9 recording sessions per participant over a period of 1.5 years.

The data was stored on 135 DVDs in Praat[8] collection format, which is a text-based format with both the speech signal and the annotation. The files were split into MS riff wave files and Praat TextGrids. The TextGrids were converted to ELAN[9] transcription files by using ELAN's export function. The database was restructured into sessions with the structure Task/L1/Speaker/Cycle. All files were renamed in the same structure, using a fixed format in such a way that each file could be uniquely identified by its name. As only first names were used in the database there was no need for anonymization.

The metadata profile for LESLLA was adapted from the DBD (see below).[10] The metadata was stored in an MS Excel file and CMDI files were created using a Python conversion script.

LESLLA is available through one of the CLARIN data centres, viz. the Max Planck Institute in Nijmegen. It can be accessed via:

https://corpus1.mpi.nl/ds/asv/?openpath=node:2102153

A full description of the database and its curation can be found in the documentation that comes with the curated database and also in Sanders, Van de Craats, & De Lint (2014).

DBD/TCULT

The Dutch Bilingual Database (DBD) is a rather substantial collection of data (over 1,500 sessions[11]) from a number of projects and research programmes that were directed at investigating multilingualism. It comprises data originating from Dutch, Sranan, Sarnami, Papiamentu, Arabic Berber and Turkish speakers. At the basis of the collection lies the research project TCULT (1998-2002) in which intercultural language contacts in the Dutch city of Utrecht were studied. Many more bilingual datasets collected over the period 1985 – 2005 were later added to the database.

The DBD corpus was stored at the Max Planck Institute with metadata in IMDI format. During the curation process, missing CHAT[12] files (i.e. files that belonged to the database but had not before been included), were added. Because all data was already in CLARIN approved format, there was no need for any data conversion.

A new DBD metadata profile was set up in CMDI, based on the existing IMDI profile. A shell script was created to convert the IMDI files to CMDI files. Where necessary, information was made consistent and missing information (e.g. about file sizes) was added. New ISOcat elements were introduced that were submitted to the ISO committee for formal approval.

Documentation on the DBD can be found in the PhD theses by the various researchers who originally collected and interpreted the data. The curation has been described in the curation report.

The data has been made available through one of the CLARIN data centres, viz. the Max Planck Institute in Nijmegen. It can be accessed via

https://corpus1.mpi.nl/ds/asv/?openpath=node:2102153/

3.3 Oral/spoken corpora: IPNV interviews

The IPNV Corpus is a corpus originally compiled by the Veteraneninstituut (VI). It comprises a collection of more than 1,100 (recorded) interviews with veterans who were involved in wars and other military actions that the Dutch military forces took part in. The average duration of an interview is 2.5 hours. Most interviews are with veterans of World War II, the decolonization wars with Indonesia and New

[8] http://www.Praat.org
[9] https://tla.mpi.nl/tools/tla-tools/elan/
[10] The CMDI profile can be found at
http://catalog.clarin.eu/ds/ComponentRegistry/?item=clar%20in.eu:cr1:p_1375880372947#/
[11] In this context 'session' is used to denote an audio file recorded with one informant at a specific point in time.
[12] http://childes.psy.cmu.edu/

Guinea, the UN action in Korea, the UN observe mission in Lebanon, UN missions in Cambodia and former Yugoslavia, and the NATO missions in Iraq and Afghanistan. Some 100 interviews are with veterans who were involved in small-scale observation, monitoring and humanitarian missions.

In the INTER-VIEWs project[13] 246 of the interviews were curated: the audio recordings (in riff wav format) of the interviews were transferred to DANS[14] and the metadata were made available in CMDI/ISOcat format adopting the profile *OralHistoryInterview* in CLARIN's component registry. The data and metadata can be accessed through the DANS EASY system.

For the remaining interviews all recordings are in wav format as well. They have also been transferred to DANS by the Veteraneninstituut. For these data, some metadata (at least covering Dublin Core categories) is available. The Veteraneninstituut has provided additional metadata (in an MS Access database) such that the metadata are comparable (and thus compatible) with the metadata for the 246 interviews that were curated in the INTER-VIEWs project (Van den Heuvel et al., 2012).

Around 950 interviews were curated (including an update of the 246 previously curated interviews). All corresponding CMDI metadata files were delivered to DANS. DANS has been authorised to publish various aspects of the metadata in accordance with their agreement (Convenant) with the Veteraneninstituut.

4 Future perspective

4.1 General and reusable workflows

Funded by CLARIN NL the DCS has served as an expertise centre that was charged with and focused on the curation of existing collections. This explains why most of its efforts so far have been directed towards attempts to try and make these resources conform to the (CLARIN) preferred formats, allowing for their integration in the larger CLARIN infrastructure and the application of various services offered within this infrastructure. Thus one could say that the DCS has been working on a backlog of resources that were created in the past. From our experience we have learned that the diversity in data is enormous, even in our own linguistic field of research, which makes it hard and partly impossible to devise efficient generalized procedures and tools for data curation. Still, curation efforts such as for example those pertaining to the curation of the various dialect dictionaries can be looked on as quite successful, as they have shown that certain existing manual workflows can at least partly be automated, offering a significant speed-up in corpus ingestion and annotation. This generalization will be further explored and extended in a new CLARIN-NL project: *CARE* which stands for Curation of Regional dialect dictionaries.

4.2 From posthoc to frontline

When we turn to look at the future, we advocate that data expertise centres such as the DCS shift their attention towards a point much earlier in the lifecycle of a resource, preferably even to the point where researchers are still in the first stages of proposal writing. Much is to be won if data curation is to become an integral part of the overall research data management plan right from the very start, rather than that it has been so far where curation came into view well after the resource was created and used (once, in the context of a specific research project), that is, at a time when the resource was at risk of vanishing all together. Current developments show that various stakeholders (individual researchers, research groups, the wider research community, but also for example the various funding agencies) are becoming increasingly aware of the vested interest they have in data sharing and preservation. More and more researchers are subscribing to the idea that research involving data requires a data management plan (DMP). Funding agencies have begun implementing a policy where a DMP is a prerequisite for being eligible for funding. Research plans should describe not only what kind of resource will be created (with attention for the design, data collection and annotation, formats, IPR, etc.), but also how it is envisaged that the resource can be stored and made accessible for other researchers and beyond the lifetime of the research project in which the resource was created.

[13] Project funded by CLARIN-NL under grant number CLARIN09-015.

[14] DANS (Data Archiving and Networked Services) is one of the Dutch CLARIN centres. See also http://www.dans.knaw.nl

4.3 The DCS of the future

Ideally, researchers can be held responsible for the data from the point of creation up to the point where the resource can be delivered to a data centre where the resource can be persistently stored and accessed via web portals containing aggregated metadata. However, the effort required for making data available to the wider research community should be proportionate, i.e. it should be born in mind that the core business of the researcher is to conduct research, and can only devote limited time and effort to data curation. Therefore, it is not to be expected that (all) researchers can carry out the complete data preparation of their resources up to inclusion in the data centres themselves. Expertise centres like the DCS will therefore remain indispensable in the years to come.

Part of the funding for setting up and maintaining such data expertise centres will need to come from national or international funding bodies such as NWO in the Netherlands including resource infrastructure programs such as CLARIAH.[15] As observed above, research proposals in the future can be expected to be required to contain a data management plan specifying the design of the resource, procedures for data acquisition, data formats, ethic and legal arrangements, etc. The set-up and execution of such a plan can be (partly) subcontracted to one of the data expertise centres whose role it will be to offer various services to researchers developing and implementing their data management plans. In the expertise centres, data scientists, technical staff, and documentalists should be available. At a local level, one can imagine that for example within universities, the university library will act as a front office where researchers can turn to with their questions. These questions will typically pertain to data-sets in all stages of development: planned (DMP needed!), under construction, or completed. The expertise centre will then operate as a back-office.

Thus, in future the principal tasks of the data expertise centre will be

- to assist researchers in drawing up data management plans;
- to advise on licenses both for data acquisition and for data use by the end-users;
- to provide information on standards and best practices, guidelines, etc.;
- (where necessary) to convert data and metadata in standard formats;
- to give support to researchers as regards delivery of the resource to the repository with which the data will be archived.

Where relevant, the centre will refer researchers to other (national or international) centres of expertise, for example for having their resources validated.

Since the diversity of data is immense, we recommend that such expertise centres are organized according to scientific discipline or subdiscipline.

5 Conclusion

So far the DCS has focused on existing data collections which means that most of its efforts have been directed at trying to make the resources conform to CLARIN preferred formats, allowing for their integration in the larger CLARIN infrastructure and the application of various services offered within this infrastructure. We have shown that even in our own field of research, linguistics, there is a wide variety of language resources requiring tailor-made curation solutions which makes it difficult to create generic data and metadata conversion procedures that can be used as ready-made, off-the-shelf procedures that fit other datasets. This being said for resources developed in the past, we envisage a more promising perspective for the future if data curation is to become an integral part of the overall research DMP right from the start. It is here where procedures and guidelines can be developed to maximize uniformity in database design, data formats and perhaps even metadata categories, thus advancing efficient data management and avoiding time-consuming posthoc curation labour.

[15] http://www.clariah.nl/en/

We do not believe that the full data management cycle can or should be completely left in the hands of the researchers since it is not their primary task. For this reason we advocate the lasting support of data expertise centres funded by national and/or international funding entities. These data centres need this funding for continuity and visibility of their work, and to guide researchers in setting up their DMPs. The actual implementation of the DMP could (also) be funded by allocating part of the budget in the research proposal to data management support by a data expertise centre.

For researchers the university libraries are a natural entry point for posing questions regarding data management. The expertise centres can be installed as back offices for consultancy and data curation tasks.

For the Netherlands efforts directed at data curation will be undertaken within the framework of the CLARIAH project in which data curation is one of the pillars of WP3.

References

Calzolari, N.; Quochi, V. and Soria, C. (2014) *The Strategic Language Resource Agenda*. Retrieved from: http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf. Retrieval date: 20 March 2014.

Francopoulo, G. (2013). *LMF Lexical Markup Framework.* Chapter 3. Wiley-ISTE. ISBN: 978-1848214309.

Gavrilidou, M.; Labropoulou, P.; Desipri, E.; Piperidis, S.; Papageorgiou, H.; Monachini, M.; Frontini F.; Declerck, T.; Francopoulo, G.; Arranz, V. and Mapelli, V. (2012). The META-SHARE Meta Schema for the description of language resources. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2012*, Istanbul, Turkey.

Odijk, J. (2010). The CLARIN-NL project. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2010*, pp. 48-53. Valletta, Malta.

Odijk, J. (2014). CLARIN-NL: Major results. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2014*, pp. 2187-2193. Reykjavik, Iceland.

Oostdijk, N. and Van den Heuvel, H. (2012). Introducing the CLARIN-NL Data Curation Service. In *Proceedings of the Workshop Challenges in the management of large corpora*. *LREC2012*, Istanbul, 22 May 2012. http://www.lrec-conf.org/proceedings/lrec2012/index.html. Retrieval date: 20 March 2014.

Oostdijk, N.; Van den Heuvel, H. and Treurniet, M. (2013). The CLARIN-NL Data Curation Service: Bringing Data to the Foreground. *The International Journal of Digital Curation*, Vol. 8, Issue 2, 134-145.

Oostdijk, N. and Van den Heuvel, H.(2014). The Evolving Infrastructure for Language Resources and the Role for Data Scientists. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2014, Reykjavik.*

Sanders, E.; Van de Craats, I. and De Lint, V. (2014). The Dutch LESLLA Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2014, Reykjavik.*

Van den Heuvel, H.; Sanders, E.; Rutten, R. and Scagliola, S. (2012). An Oral History Annotation Tool for INTERVIEWs. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2012*, Istanbul, Turkey.

Windhouwer, M. and Wright, S.E. (2013). LMF and the Data Category Registration: Principles and application. In: G. Francopoulo (ed.): *LMF Lexical Markup Framework.* Chapter 3. Wiley-ISTE. ISBN: 978-1848214309.

User Required? On the Value of User Research in the Digital Humanities

Max Kemman
University of Luxembourg
www.maxkemman.nl
max.kemman@uni.lu

Martijn Kleppe
Erasmus University Rotterdam
www.martijnkleppe.nl
kleppe@eshcc.eur.nl

Abstract

Although computational tools play an increasingly important role in the humanities, adoption of tools by scholars does not always reach its potential. One approach to this problem is user research to uncover the needs of the users. However, it is uncertain whether such user requirements can be generalized to a wider group of humanities scholars, and whether users are able to explicate their requirements for methodological innovation. We ask what the role of user research is in the Digital Humanities by discussing gathered user requirements for two projects. We categorized the requirements as within- or out-of-scope of the projects' goals, and found a tension between the specificity of humanities' research methods, and generalizability for a broader applicable tool. With the out-of-scope requirements we are able to map the wider research workflow, showing DH tools will most likely take a spot in the wider workflow, and that it is infeasible to create a tool for the entire workflow that is generic enough for a larger user group. However, the within-scope requirements led to features that were sufficiently generic for the tool to be adopted, also for unintended purposes. These insights show user research has a clear benefit for DH projects.

1 Introduction

The development of tools plays an important role in the Digital Humanities. With the increasing quantities of digitised as well as born-digital source material, computational tools have become necessary for exploring, analysing and enriching this material. While many tools have been and are being developed, adoption by the target audience, i.e., humanities scholars, does not always reach its potential (Edwards, 2012; Gibbs & Owens, 2012; Warwick et al., 2007). In projects where the research data is published within a tool, this can result in neither the tool nor the research data being fully used by other scholars. One partial solution to this problem is to publish research data separately from the tool, as advocated by Borgman (2012), and Kansa et al. (2010).

Furthermore, in order to create tools that will be adopted by scholars, development should take into account the practices and conventions adhered to in subdisciplines of the humanities (Bradley, 2005; Kemman et al., 2014b). One approach is to focus on the users, actively involving them during development and evaluation of designs, known as *user-centred (systems) design* (Gulliksen et al., 2003). To achieve this, *user research* is performed (Warwick, 2012), for which one of the tasks is to uncover the needs and wishes of the user group, commonly referred to as *user requirements* (e.g., Sweetnam et al., 2012).

There is however an ongoing debate whether such user requirements can be sufficiently generalized to a wider group of humanities scholars. On one end of this debate, we see the suggestion that research contains generic tasks called *scholarly primitives*, defined as "*basic functions common to scholarly ac-*

tivities across disciplines, over time, and independent of theoretical orientation" (Unsworth, 2000). Unsworth presented a non-exhaustive list of primitives, summarized by Martin Weller as follows (Weller, 2011):

1. discovering – knowledge either through archives or research;
2. annotating – adding layers of interpretation;
3. comparing – for example, texts across languages, data sets;
4. referring – referencing and acknowledging;
5. sampling – selecting appropriate samples;
6. illustrating – clarifying, elucidating, explaining; and
7. representing – publishing or communicating.

Tools can be developed to support these primitives, and are thus applicable to a broad community of scholars. For the development of an infrastructure for the arts and humanities, the scholarly primitives have been combined into *discovering*, *collecting*, *comparing*, and *delivering* (Anderson et al., 2010; Blanke & Hedges, 2013). The idea is to create a user-centric infrastructure to support the entire research process with primary source material.

At the other end of the debate, we see the suggestion that scholarly practices are very specific and that a "`*one size fits all' approach would be a disastrous underestimation of the specific needs of humanities research*" (van Zundert, 2012). Van Zundert suggests that insofar methodological innovation is desired, generalization and standardization might be detrimental.

Whether user research enables targeted users to explicate their requirements for methodological innovation is furthermore met with scepticism in literature. Although interviews are regularly used as a method for gathering user requirements (Benyon et al., 2005), users supposedly do not know what they want, and cannot predict their own future behaviour (Nielsen, 2001). Moreover, innovation is said to be driven by focusing on new technology, even though people do not yet need such technology, nor have a clear use case for it (Norman, 2010). Nevertheless, in the wider Human-Computer Interaction literature, user research is regarded crucial during development (Hofmann & Lehner, 2001). Following from the above discussion, we ask what the role of user research is in the Digital Humanities. Our research question is: *what is the added value of user research for developing tools aimed at digital research methods?*

To address this question, we will discuss results from user research for gathering user requirements for two Digital Humanities projects we coordinated; PoliMedia and Oral History Today. In these projects, we held interviews with scholars to inform development. We will show user requirements that were within- or out-of-scope, where the scope is determined by feasibility and the project goal, and examine how many user requirements were common to multiple participants. By doing so we aim to provide insight into the added value of user research for these two case studies.

This paper is structured as follows: first, we will introduce the research projects and their goals. Second, we will explain how scholars were involved in these projects to voice their needs and wishes. Third, we will review the user requirements that were collected and whether these were determined to be within- or out-of-scope. Fourth, we will discuss how our findings relate to the literature. Finally, we will discuss what we learned from the user requirements, and what the added value was of user research.

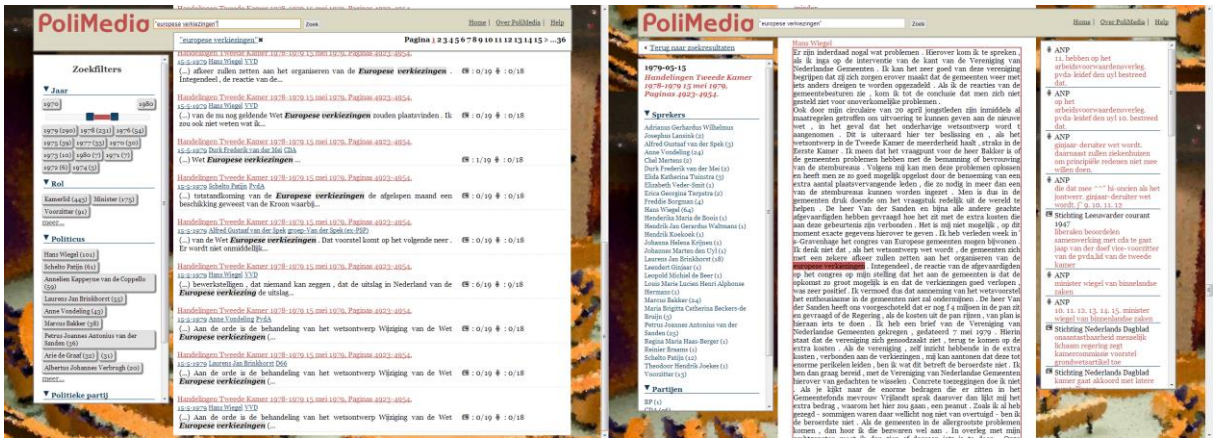


Figure 1: PoliMedia user interface. Left: search results page. Right: debate page, with on the right links to media items.

2 User requirements for PoliMedia and Oral History Today

The PoliMedia project[1] aimed to facilitate a digital research method for large-scale cross-media analysis of the coverage of political debates (Kleppe et al., 2014). Investigating how political debates are covered in the media required scholars to explore three distinct collections: 1) the minutes of the Dutch parliament, 2) Dutch newspapers and 3) Dutch radio bulletins. Additionally, a fourth dataset of interest is the Dutch television broadcasts, but due to a lack of links found between the proceedings and television broadcasts, this dataset was dropped from implementation, although it was included in the interviews. In order to present a dataset with as much overlap of these three collections as possible, we set the timeframe from 1945-1995.

Although access to the collections has already improved with digitization, each collection still required scholars to learn and use three different user interfaces, as well as redo searches for the same subject in each system.[2] To better facilitate such research, PoliMedia provides a search user interface where scholars can explore the minutes of the Dutch parliament with integrated links to media coverage, see figure 1.

For each speech in the parliament, information was extracted to represent the speech; the speaker, the date, important terms (i.e., named entities) from its content and important terms from the description of the debate wherein this speech was held. This information was then used to query the archives of the newspapers and radio bulletins, and links were created to items that correspond to the query (Juric et al., 2013). The debates and links were then represented as RDF, a Semantic Web standard (Juric et al., 2012). By employing Semantic Web technology, information about entities (such as people, places, subjects) can be aggregated from multiple collections to gain a broader perspective. The scope of the project could thus be described as follows: *automatically creating links between debates of the Dutch parliament to media items, made available in a search user interface in which debates of the Dutch parliament can be explored.*

The Oral History Today project[3] aimed at facilitating a digital research method for exploring and searching of aggregated, heterogeneous oral history content (Kemman et al., 2014b). Discovering interesting oral history interviews is a difficult task, as many small collections are available at many different locations: sometimes digitized, sometimes annotated by archivists, and sometimes available through an online portal. To better facilitate this process, Oral History Today provides a search user interface where scholars can search through over fifty oral history collections containing over four thousand interviews, enabling scholars to discover interviews across several collections, see figure 2. The collections were aggregated in a previous project (Ordelman & De Jong, 2011), and are hosted by DANS (DANS, 2012), where the collections were annotated to fit this archive's schema. The metadata was then indexed and made searchable through a search user interface with a focus on usability. Since Google is immensely

[1] http://www.polimedia.nl

[2] Shortly after the PoliMedia project, the Dutch National Library launched a new search system that integrates the newspapers and radio bulletins, Delpher (http://www.delpher.nl).

[3] http://zoeken.verteldverleden.org

popular among scholars (Kemman et al., 2014a), the search system was designed to be like 'a Google for oral history interviews', i.e. the system would provide a simple search bar and a high recall of results ranked by relevance. This was extended with several filtering and ranking features. The scope of the project could thus be described as follows: *a search user interface similar to Google but including advanced filter options, in which oral history interviews and collections can be searched and explored to discover topics across a multitude of collections.*

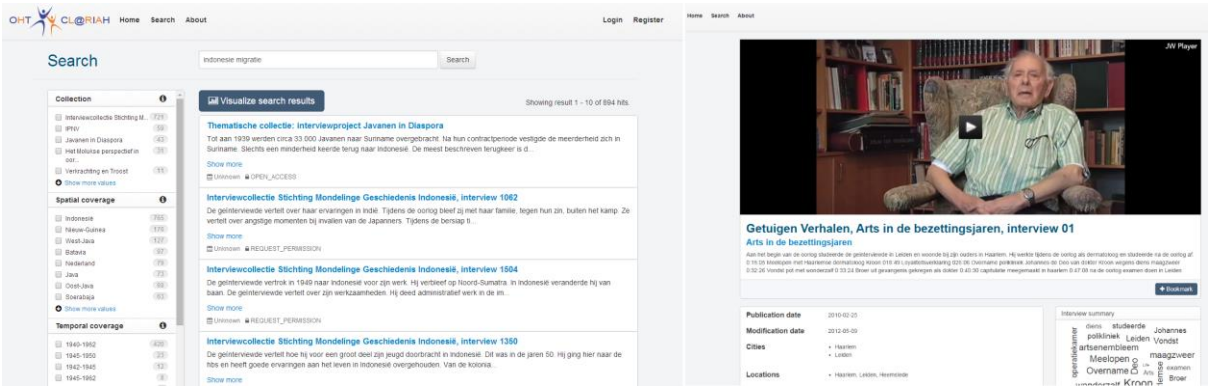


Figure 2: Oral History Today user interface. Left: search results page. Right: interview page.

3 Methods

In the PoliMedia project, before development commenced, we held semi-structured face-to-face interviews with five scholars. Interviewees were invited from our own network and represented both qualitative and quantitative methods. The interviewees worked at different universities. There were no further selection criteria regarding demographics. One interview was with two scholars simultaneously, and is treated as a single interviewee, thus leading to four interviewees in our data. Interviewees talked about their research questions, methods and requirements for cross-media analyses. Questions were specifically related to their general research problems and approaches, which databases and search engines scholars used, what they liked or disliked about these, and asking feedback on a verbal description of the PoliMedia plans.

In the Oral History Today project, we held semi-structured interviews with fifteen scholars via Skype. Interviewees were selected from our own network as well as via the oral history working group of the Dutch Research Institute and Graduate School for Cultural History.[4] We selected interviewees in all stages of careers from project assistants to PhD Candidates to Professors. There were no selection criteria regarding other demographics. All interviewees were given a monetary reward for their participation. Interviewees talked about their research questions, methods and requirements for a federated search engine for oral history collections. Questions were specifically related to how they performed Oral History research, which collections they used, and asking feedback on a rudimentary search user interface that was created before the interviews, particularly regarding their first actions in the interface, how they explored collections, how they did more directed searches, and how they evaluated interviews. Interviewees were mainly knowledgeable in employing the oral history method; less than half of the interviewees created or reused oral history collections.

After each interview, the interviewer summarized this information into functional requests, which was then sent back via e-mail to the interviewee for approval, allowing edits where needed. These functional requests were then categorized into user requirements by the interviewer, where similar statements were combined. These user requirements were finally discussed by the project team to classify them as within- or out-of-scope, determined by feasibility and the project goal. The within-scope requirements were then prioritized for development.

For PoliMedia, after developing the user interface, 24 scholars evaluated the usability of the portal (Kemman et al., 2013). Feedback voiced during this evaluation led to an improved final version of the search interface.

[4] http://www.huizingainstituut.nl/werkgroep-oral-history/

For Oral History Today, after an update of the search interface, five scholars were interviewed via Skype to explore the collections, try search questions of their own interest and provide feedback. The results of these evaluations were then considered for the next update; we repeated this process a second time leading to the final version of the search interface.

In this paper we report the user requirements that we gathered and classified for the first round of interviews for both projects.

4 Results[5]

4.1 PoliMedia

The interviews for PoliMedia led to 39 user requirements. A total of 21 requirements were deemed within-scope of the project, and were related to functionality such as:

- gaining insight into contextual information (e.g., `Function of actors`, `Party of actors`, or `Type of programme` (news, talk show, late night, etc.)),
- the frequency of terms (e.g., `Mathematical queries`, `Frequency of searched, related, and important terms in documents`, `Comparing/sorting search results by frequency of terms`),
- search operators (`Boolean operators` and `Google search operators` (esp. the combination of a string with quotation marks)), and
- analysis of the debates (e.g., `Length of document per actor`, `Ability to export non-formatted text`).

18 requirements were deemed out-of-scope. These requirements were related to computational analysis of the sources with advanced techniques:

- image processing of newspaper pages (e.g., `Size of headers`, `Number of columns on a page`, `Presence and size of photographs`),
- audio-visual processing of television programmes (e.g., `Length of talk`, `Presence of music`, `Use of filming techniques`), and
- linguistic analyses of debates (e.g., `Speech functions`, `Type of speech fragments` (interruptions, questions, jokes, etc.)) and of newspaper articles (`Genre` (report, comment, letter to the editor, etc.)).

The project scope, as described above, focused on creating links between collections, and developing a search user interface to explore the proceedings to which media items are linked. The computational analysis of these items then is clearly out-of-scope. Moreover, such tasks are far from trivial considering the size of the collections: eight million pages from newspapers, 1.8 million radio bulletins (Delpher, n.d.), 2.4 million pages of parliamentary proceedings (Staten-Generaal Digitaal, n.d.), and 2500 hours of television material (Academia.nl, n.d.). Finally, computer vision tasks such as the classification of filming techniques are research problems not yet solved.

27 requirements were unique, i.e., voiced by a single interviewee. The most common requirements were the inclusion of `Media output about subject before debate`, `Names of actors (people) involved`, and `Location in the newspaper` (page number, location on page), each mentioned by three interviewees. The first two were deemed within-scope, while the third was deemed out-of-scope due to required image processing as described above.

Some user requirements that we had not considered before the interviews, but that were considered within scope and made a big impact on our thinking about the tool:

[5] All user requirements are available open access via Kemman, M., Kleppe, M. (2014): User Requirements for Two Digital Humanities Projects: PoliMedia and Oral History Today [dataset]. fig**share**. http://dx.doi.org/10.6084/m9.figshare.1170077

- `Function of actors` (e.g., minister, member of parliament, but also show host, interviewer, etc.) – voiced in one interview.
- `Party of actors` (e.g., VVD, PvdA, but also Greenpeace or other lobby groups) – voiced in two interviews
- `Media output about subject before debate` – voiced in three interviews.

The first two requirements could be addressed without too much difficulty, since this information was already part of the dataset. Making this information available at the front-end for interaction introduced the opportunity to explore the proceedings on the level of the speaker's role (in our implementation, as member of government or of parliament). The third requirement introduced a different perspective on the interaction between politics and media than was envisioned. Not only do newspapers report on what happens in parliament, parliament discusses events in society by referring to newspaper reports. Newspaper articles regularly set the stage for parliamentary debates. Unfortunately, due to technical reasons, it was ultimately not implemented.

4.2 Oral History Today

The interviews for Oral History Today led to 75 user requirements. A total of 33 user requirements were deemed within-scope of the project, and were related to:

- more instructions and clearer details of functionality and collections (`Support page describing interviews and search technology`, `Description of project (within which collection was created) and how collection came to be`, `Organization behind collection (management/creation)`),
- more advanced searching with filters (e.g., `Locations`, `Collection`, `Topics`, `Year Event`, `Access conditions`),
- navigation within the search user interface (e.g., `Navigate from interview to interview collection`, `Clicking a topic should result in all interviews with the same topic`, `Links between related interviews`), and
- workspaces (`Search trail` (i.e., a history of queries), `Bookmark functionality for interview`).

42 user requirements were deemed out-of-scope. These requirements were mainly related to:

- features of the search technology (e.g., `Boolean operators`, `Search explicitly for broad or narrow terms`, `Detect synonyms of search terms`), and
- additional metadata on the interviewee (e.g., `Age/Year of birth`, `Gender`, `Religion`, `Community of experience`, `Social class`), interviewer (e.g., `Age`, `Gender`) and the interview (e.g., `Research question underlying interview`, `Location of interview`, `Description of interview per 10 minutes`).

On first sight, such requirements might appear well within the scope of the project. The decision to categorize them as out-of-scope was mainly due to limitations of what we had available. The requests regarding search technology were dependent on the search technology provider we had chosen before the start of the project. The search technology that we used focused on high recall with relevance ranking, i.e., adding more search terms broadened the result set but improved the search results ranking. This conflicted with the search behaviour we observed from the interviewees who aimed at reducing the search result set until it became a manageable set that could be assessed interview by interview. This wish for precision is also reflected in the wish for more metadata to assess relevance and the broader context of the interview. However, since we used a dataset created in a previous project, we could only use the metadata that was made available then. We cannot provide information we do not have ourselves, and enriching the metadata was out-of-scope.

34 user requirements were unique, i.e., voiced by a single interviewee. The most common requirement was a filter for `Year event`, voiced by ten interviewees, and was deemed within-scope.

Some user requirements that we had not considered before the interviews, but that were considered within scope and made a big impact on the final tool were the following:

- `Description of project (within which collection was created) and how collection came to be` – voiced in five interviews.
- `Organization behind collection (management/creation)` – voiced in one interview.
- `Distinguish facets between relating to content or general` (where content relates to the contents of the oral history interviews, e.g., Year or Location, while general is about the interview files, e.g., Open Access, or Audio/Video)– voiced in one interview.

What is interesting about the first two requirements is how oral history interviews are understood within the context of their collection. While we started with the assumption of a keyword search bar, we learned that a significant portion of the interviewees wanted to browse and view the interviews in the context of their collections. Our observations showed that while half the interviewees (8/15) started by typing interesting terms into the search bar, the other half (7/15) started by browsing the collections. Knowing this, we introduced more fine-grained exploration of the collections, and navigation controls to move from an interview to a collection page. The third requirement described gave us input to further consider the search filters we provided; what type of filter is this, and how should the search filters thus be presented? Considering such questions ultimately led to a better search filter interface than we initially provided.

5 Discussion

What can we learn from the gathered user requirements? The user requirements show that our users, the humanities scholars, are very aware of what they want, agreeing with the findings of Warwick (2012). In PoliMedia many user requirements reflect the research methods of the interviewed scholars, who would like their heuristic process simplified, i.e., the discovery of primary and secondary sources for investigation. Automatic analysis was perceived as helpful for this process to easily discover e.g., debate sentiments, framing of topics by media, and topic importance. In Oral History Today the requirements reflect the fine-grained control oral historians desire during their heuristic process: being able to find interviews related to a specific place, time and event. Additionally, insight into the background of both the interviewee and interviewer is desired to properly understand the interview.

Still, to some extent, our results agree with the criticisms of asking users about their requirements (Nielsen, 2001; Norman, 2010). First, in the case of PoliMedia only three uniquely voiced user requirements, out of 39, were related to the project's technological goal of linking debates and media items and publishing these as RDF:

- `External linking to databases about persons` (e.g., www.parlement.com).
- `Function of actors` (e.g., minister of defence, member of parliament, show host, interviewer).
- `Search on committee`.

Second, in the case of Oral History Today the user requirements are based on current, rather than future practices, and even show a distrust of potential innovations. The idea of a simple Google-like search bar and high recall ranked by relevance did not appear to match the desire for high precision resulting in manageable sets, reminiscent of the *"perfect thirty-item"* online search identified by Bates (1984). Interviewees explained they could not trust the search ranking in a way to be confident search results further down the list would not have to be looked into, regardless of the performance of the ranking mechanism. This seems to show a tension between a need for completeness of search results, while at the same time keeping the number of search results manageable. Potential innovations in the discovery of oral history interviews are deemed undesirable, despite the proven utility of other search engines with high recall ranked by relevance (cf. Kemman et al., 2014a).

Finally, our results agree to a large extent with the suggestion that the humanities are too specific for generic innovations (van Zundert, 2012). A large number of user requirements we found were unique, underscoring the specificity of humanities research.

6 Conclusions

What is the added value of user research for the development of tools aimed at digital research methods? In our investigation of user requirements for two Digital Humanities projects, we found scholars have a clear idea how they perform their research, and how tools could simplify some steps in the process of discovering and analysing sources. On the other hand, we hardly see scholars immediately embrace the full potential of the projects' goals in their user requirements: i.e., semantic web technology in the case of PoliMedia, and simple Google-like searching in the case of Oral History Today. Whether this means that scholars are unaware of how such facilities might help them, or whether scholars are aware that such goals do not match with their methods, remains an open question. To answer this question requires a deeper understanding of how (digital) technology is adopted by scholars. A study of how historians adopt digital technology and how it affects their practices is the topic of PhD research by Kemman currently in progress.[6]

Alternatively, perhaps the scope we chose was already too much tied to specific requirements dependent of the researcher, i.e., the linking between different collections is perhaps already a specific rather than a generic research method.

The findings and such questions seem to confirm the criticism that interviewing users for their requirements might not be the most effective method to advance methodological innovations. Instead, alternative approaches such as observations might give more insights into practices. Another promising approach is to move beyond the list of user requirements, and emphasize *participatory design* as a negotiation between users and developers (Muller, 2003), e.g., as done in the HistoGraph project (Novak et al., 2014).

Still, despite the specificities of the user requirements, we also find that the tools contain generic features. For example, we were happy to find an article in a Dutch newspaper in which the author stated to often use PoliMedia. This author however mainly used the tool for its search and filter options, without using the linked media coverage (Sanders, 2014). The within-scope user requirements helped to improve the tool to be used even for purposes not specifically intended.

The out-of-scope user requirements on the other hand provide hints of what the wider research workflow consists of for the different participants. That is, after finding the related media items with PoliMedia, scholars want to analyse these media items, or annotate it with their observations. With Oral History Today, we see that after finding an interesting video, scholars want to contextualize it and come to a full understanding of the interview. User researchers should thus keep in mind that the tool will most likely take a spot in a wider research workflow, and that it is infeasible to create a tool for the entire workflow that is generic enough to be applicable to a larger user group. In this sense our conclusions are in opposition with the ambitions of e.g., Blanke and Hedges (2013). Our findings instead suggest to focus on a single task within the workflow, which is reminiscent of the old adage *do one thing and do it well* (McIlroy et al., 1978). This proposition is compatible with Van Zundert's suggestion of light-weight tools for specific humanities tasks (van Zundert, 2012). To some extent it seems compatible with Unsworth's suggestion of tools for specific *scholarly primitives* (Unsworth, 2000), in that the research workflow is split up in a set of primitives. However, what we have learned from our user research is that the requirements for a tool related to certain tasks are related to what tasks come further down the workflow. As such, these tasks are not true primitives since their implementation is dependent of the rest of the workflow. To what extent certain tasks are generalizable is a question that requires further user research.

These insights furthermore lead us to conclude that in order to enable a workflow with multiple tools, Digital Humanities projects should separate the tool and the data. Even when the tool would not be compatible with a specific scholar's research methods, the data should still be usable. In PoliMedia, not only was a tool created, but also a dataset, which was made available via a SPARQL-endpoint.[7] Alter-

[6] For more information about this PhD research and for future updates, see http://www.maxkemman.nl/category/phd-thesis/
[7] http://data.polimedia.nl

native approaches are an API or a downloadable dataset. This introduces a new continuum in the innovation of digital research methods, namely that from developing tools *for* scholars, via developing tools *with* scholars, to scholars developing tools. An in-depth discussion of this continuum is beyond the scope of this paper, but in receiving feedback on our published datasets we do observe that many humanities scholars have difficulty using data without an accompanying tool. Data reuse is not as simple an undertaking as one might hope (Borgman, 2015; Edmond & Garnett, 2015).

We note that there is a tension between the specificity of humanities' research methods, and generalizability for a broader applicable tool. Our findings suggest however that user research has a clear benefit for Digital Humanities projects: first, the out-of-scope user requirements give insight into the tool's compatibility with existing research practices. Second, the user requirements that were within-scope led to usable features that were sufficiently generic for the tool to be adopted, also for purposes for which it was not specifically intended. User research thus proved useful for the development of tools to be compatible with specific research methods of scholars, taking a place in a wider research workflow.

Acknowledgements

The PoliMedia and Oral History Today projects were funded respectively by CLARIN-NL (http://www.clarin.nl) and CLARIAH (http://www.clariah.nl). We would like to thank Joris van Zundert for his helpful feedback on an earlier version of this paper, as well as Andreas Fickers who also inspired the title of this paper.

References

Academia.nl. (n.d.). Hulp. Retrieved February 10, 2015, from http://www.academia.nl/faq/28341#t28345n1847

Sheila Anderson, Tobias Blanke, & Stuart Dunn. (2010). Methodological commons: arts and humanities e-Science fundamentals. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *368*, 3779–3796. http://doi.org/10.1098/rsta.2010.0156

Marcia J. Bates. (1984). The Fallacy of the Perfect Thirty-Item Online Search. *RQ*, *24*(1), 43–50.

David Benyon, Phil Turner, & Susan Turner. (2005). *Designing interactive systems: People, activities, contexts, technologies*. Pearson Education.

Tobias Blanke, & Mark Hedges. (2013). Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems*, *29*(2), 654–661. http://doi.org/10.1016/j.future.2011.06.006

Christine L. Borgman. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. http://doi.org/10.1002/asi.22634

Christine L. Borgman. (2015). *Big Data, Little Data, No Data* (ebook). MIT Press.

John Bradley. (2005). What You (Fore)see is What You Get : Thinking About Usage Paradigms for Computer Assisted Text Analysis. *TEXT Technology*, *14*(2).

DANS. (2012). Thematische collectie: Oral History. DANS. http://doi.org/10.17026/dans-z3c-f26d

Delpher. (n.d.). Collecties. Retrieved February 10, 2015, from http://www.delpher.nl/nl/platform/pages?title=collecties

Jennifer Edmond, & Vicky Garnett. (2015). APIs and Researchers: The Emperor's New Clothes? *International Journal of Digital Curation*, *10*(1), 287–297. http://doi.org/10.2218/ijdc.v10i1.369

Charlie Edwards. (2012). The Digital Humanities and Its Users. In Matthew K. Gold (Ed.), *Debates in the Digital Humanities* (online). University of Minnesota Press.

Fred Gibbs, & Trevor Owens. (2012). Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs. *DHQ: Digital Humanities Quarterly*, *6*(2).

Jan Gulliksen, Bengt Göransson, Inger Boivie, Stefan Blomkvist, Jenny Persson, & Åsa Cajander. (2003). Key principles for user-centred systems design. *Behaviour & Information Technology*, *22*(6), 397–409. http://doi.org/10.1080/01449290310001624329

Hubert F. Hofmann, & Franz Lehner. (2001). Requirements engineering as a success factor in software projects. *IEEE Software*, *18*(August), 58–66. http://doi.org/10.1109/MS.2001.936219

Damir Juric, Laura Hollink, & Geert-jan Houben. (2013). Discovering links between political debates and media. In *The 13th International Conference on Web Engineering (ICWE'13)*. Aalborg, Denmark. http://doi.org/10.1007/978-3-642-39200-9_30

Damir Juric, Laura Hollink, & GJ Houben. (2012). Bringing parliamentary debates to the Semantic Web. In *DeRiVE workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*.

Eric C. Kansa, Sarah Whitcher Kansa, Margie M. Burton, & Cindy Stankowski. (2010). Googling the Grey: Open Data, Web Services, and Semantics. *Archaeologies*, *6*, 301–326. http://doi.org/10.1007/s11759-010-9146-4

Max Kemman, Martijn Kleppe, & Jim Maarseveen. (2013). Eye Tracking the Use of a Collapsible Facets Panel in a Search Interface. In *Research and Advanced Technology for Digital Libraries* (pp. 405–408). Springer-Verlag Berlin Heidelberg. http://doi.org/10.1007/978-3-642-40501-3_47

Max Kemman, Martijn Kleppe, & Stef Scagliola. (2014a). Just Google It. In Clare Mills, Michael Pidd, & Esther Ward (Eds.), *Proceedings of the Digital Humanities Congress 2012*. Sheffield, UK: HRI Online Publications.

Max Kemman, Stef Scagliola, Franciska de Jong, & Roeland Ordelman. (2014b). Talking with Scholars: Developing a Research Environment for Oral History Collections. In Łukasz Bolikowski, Vittore Casarosa, Paula Goodale, Nikos Houssos, Paolo Manghi, & Jochen Schirrwagen (Eds.), *Theory and Practice of Digital Libraries -- TPDL 2013 Selected Workshops* (Vol. 416, pp. 197–201). Springer International Publishing. http://doi.org/10.1007/978-3-319-08425-1_22

Martijn Kleppe, Laura Hollink, Max Kemman, Damir Juric, Henri Beunders, Jaap Blom, Johan Oomen, & Geert-Jan Houben. (2014). PoliMedia - Analysing Media Coverage of Political Debates By Automatically Generated Links to Radio & Newspaper Items. In Mathieu D'Aquin, Stefan Dietze, Hendrik Drachsler, Marieke Guy, & Eelco Herder (Eds.), *Proceedings of the LinkedUp Veni Competition on Linked and Open Data for Education*. Geneva, Switzerland: CEUR-WS.

M. D. McIlroy, E. N. Pinson, & B. A. Tague. (1978). UNIX Time-Sharing System: Foreword. *Bell System Technical Journal*, *57*(6), 1899–1904. http://doi.org/10.1002/j.1538-7305.1978.tb02135.x

Michael J. Muller. (2003). Participatory Design: The Third Space in HCI. In Andrew Sears & Julie A. Jacko (Eds.), *Human-Computer Interaction: Development Process* (pp. 1051–1068). Routledge.

Jakob Nielsen. (2001). First Rule of Usability? Don't Listen to Users. Retrieved July 1, 2014, from http://www.nngroup.com/articles/first-rule-of-usability-dont-listen-to-users/

Donald A. Norman. (2010). The way I see it: Technology first, needs last. *Interactions*, *17*(2), 38. http://doi.org/10.1145/1699775.1699784

Jasminko Novak, Isabel Micheel, Mark Melenhorst, Lars Wieneke, Marten During, Javier Garcia Moron, Chiara Pasini, Marco Tagliasacchi, & Piero Fraternali. (2014). HistoGraph -- A Visualization Tool for Collaborative Analysis of Networks from Historical Social Multimedia Collections. In *2014 18th International Conference on Information Visualisation* (pp. 241–250). IEEE. http://doi.org/10.1109/IV.2014.47

Roeland Ordelman, & Franciska De Jong. (2011). Distributed Access to Oral History collections: Fitting Access Technology to the needs of Collection Owners and Researchers. In *Digital Humanities 2011*. Stanford, USA.

Ewoud Sanders. (2014, April 14). Voor al haar mantelzorgen. *NRC Handelsblad*.

Staten-Generaal Digitaal. (n.d.). Digitalisering. Retrieved February 10, 2015, from http://www.statengeneraaldigitaal.nl/overdezesite?document=digitalisering

Mark S. Sweetnam, Maristella Agosti, Nicola Orio, Chiara Ponchia, Christina M. Steiner, Eva-Catherine Hillemann, Micheál Ó Siochrú, & Séamus Lawless. (2012). User Needs for Enhanced Engagement with Cultural Heritage Collections. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7489 LNCS, pp. 64–75). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-33290-6_8

John Unsworth. (2000). Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? In *Symposium on Humanities Computing: Formal Methods, Experimental Practice*. King's College, London.

Joris van Zundert. (2012). If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities. *Historical Social Research / Historische Sozialforschung*, *37*(3), 165–186.

Claire Warwick. (2012). Studying users in digital humanities. In Claire Warwick, Melissa Terras, & Julianne Nyhan (Eds.), *Digital humanities in practice* (pp. 1–21). Facet Publishing.

Claire Warwick, M. Terras, Paul Huntington, & N. Pappa. (2007). If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data. *Literary and Linguistic Computing*, *23*(1), 85–102. http://doi.org/10.1093/llc/fqm045

Martin Weller. (2011). *The Digital Scholar* (online ed). Bloomsbury Academic. http://doi.org/10.5040/9781849666275

TeLeMaCo—A tool for the dissemination of teaching and learning materials

Hannah Kermes and **Jörg Knappen** and **José Manuel Martínez** and **Elke Teich** and **Mihaela Vela**
Fachrichtung 4.6, Universität des Saarlandes, Campus A2.2, D-66123 Saarbrücken, Germany
{h.kermes, j.knappen, j.martinez, e.teich, m.vela}
@mx.uni-saarland.de

Abstract

This paper presents TeLeMaCo, a collaborative portal for training and teaching materials relevant in linguistics and digital humanities hosted at the CLARIN-D centre at Saarland University in Saarbrücken. The portal is easy to use both for casual users who search for teaching and training material and for community members who want to contribute descriptions of their materials. We collect structured metadata of the described resources to provide advanced search and to integrate them in the wider CLARIN framework of resources.

1 Introduction

For language resources and tools, there is a growing number of repositories, and there are platforms to search the metadata of many collections at once, like the Virtual Language Observatory (VLO) (Van Uytvanck et al., 2012). Also there are web services and chaining tools such as, e.g., WebLicht (Hinrichs et al., 2010) that provide facilities for text and speech processing and support processing pipelines for a wide variety of scientific tasks.

However, the documentation and teaching materials remain scattered over many places, including institutional web pages, YouTube channels, or software repositories like sourceforge[1] or Bitbucket[2]. A common interface to access and search those teaching and learning materials was lacking when we started the design of our service.

We developed the **Te**aching and **Le**arning **Ma**terial **Co**llection (TeLeMaCo[3]) to overcome this situation. Our approach is community driven as we collect descriptions of relevant materials contributed from all over the world in our service.

TeLeMaCo offers search and access to a wide range of teaching and learning materials, including the following

- technical documentation (e.g., quick starts, tutorials, or full manuals),
- learning material for self study (e.g., YouTube videos and screen casts),
- short teaching modules (2–4 hours) that can be integrated in existing courses,
- full courses covering a broader spectrum of language resources and tools or focusing on specific topics of application of language resources and tools,
- reference materials (e.g., specialised dictionaries).

TeLeMaCo brings together materials stored at different institutions and locations through a unified interface and it provides access to materials published in different languages (currently mostly German

[1] http://sourceforge.net/
[2] https://bitbucket.org/
[3] http://fedora.clarin-d.uni-saarland.de/hub/

and English, but also French). For some tools, e.g., EXMARaLDA (Schmidt and Wörner, 2009), there is already a comprehensive coverage of the available materials.

In Section 2 of this paper we present some ideas behind TeLeMaCo. In Section 3 we describe TeLeMaCo as seen from a casual user searching for learning and teaching materials. In Section 4 we explain how to add materials to the TeLeMaCo service. Section 5 is about the metadata we store and in Section 6 we talk about some additional benefits of TeLeMaCo. We conclude with an evaluation of TeLeMaCo (Section 7) and a comparison to three other services for the dissemination of teaching and learning materials (Section 8).

2 Ideas behind TeLeMaCo

2.1 A social network

The idea behind TeLeMaCo is that the user community can easily contribute descriptions of teaching and learning materials. We see it as a colloborative and ongoing effort, not as the work of a small dedicated project which is finished at some fixed date.

TeLeMaCo allows the users to choose their own tags for the teaching and learning materials, creating some kind of folksonomy (Vander Wal, 2007). We expect some alignment of the chosen keywords because the user can easily see already existing keywords and we have implemented the autosuggest feature of tags in the service.

TeLeMaCo is not a wiki; every contributor to TeLeMaCo "owns" the descriptions he/she has added and he/she's the only one (except for the administrators) who can change them.

Another kind of user interaction is given by the feedback system we have implemented. Any user can give feedback on the usefulness of the material (and of its description) by answering four simple questions (see Figure 2). TeLeMaCo displays the aggregated feedback score.

We also allow the user to report on quality issues like stale links or inappropriate content to the portal administrator.

2.2 Implementation considerations

TeLeMaCo ist implemented using the well-established and sustainable LAMP (Linux, Apache httpd, mysql, Python) software stack with Django as a web framework. All tasks are automated to minimise administrator interaction with the system.

All local dependencies are stored in a small configuration file. TeLeMaCo can be migrated to another machine with minimal effort.

We have designed the URLs of the description pages (see Figure 1 for an example) to be plain bookmarkable URLs, no forms are involved. The same holds for other primary pages, like the browsing interface, the *what's new* page, or the (currently experimental) i*metadata* pages. These pages are easy to index by search engines and get good rankings.

2.3 Embedding into the CLARIN infrastructure

TeLeMaCo is a complementary part of the CLARIN infrastructure. It does not copy features of existing services like WebLicht or the VLO; it supplements them. It uses CMDI aware metadata, and an integration with the VLO using an OAI provider is planned. In addition, TeLeMaCo is integrated into the Helpdesk of CLARIN-D.

3 Searching and browsing TeLeMaCo

TeLeMaCo offers several facilities to search for materials: Simple and advanced search as well as browsing keywords and authors.

Simple search does a substring query over author, title, keyword and description, this way optimising the recall.

Advanced search allows querying specific fields, e.g., to get only materials in a certain language.

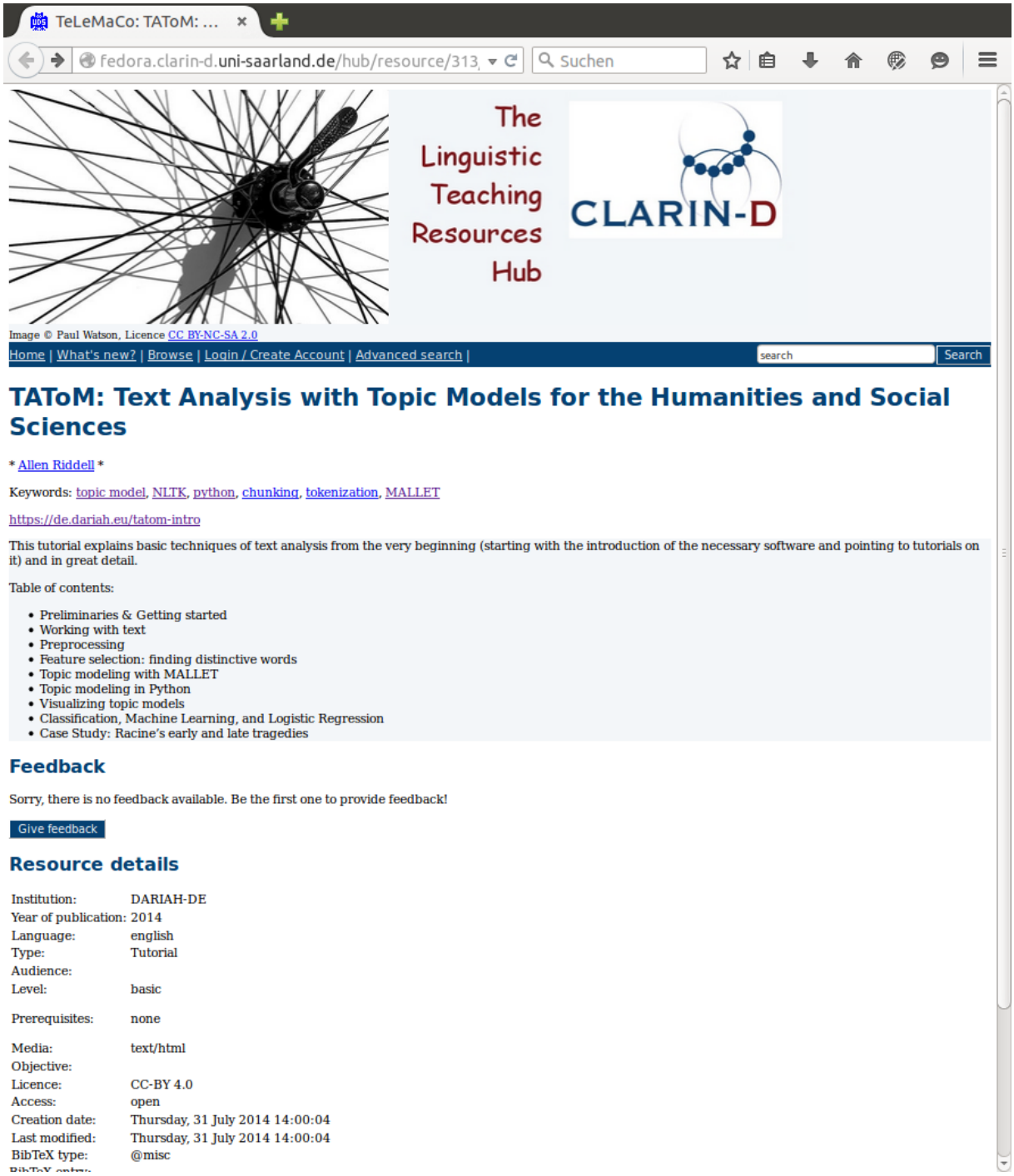


Figure 1: Display page of a sample material in TeLeMaCo

Was the described material relevant for me?— No, not at all . . . Yes, very much
I think, the level of the described material is . . . —Basic . . . Expert
I reached the objectives given by studying the material.— No, not at all . . . Completely
I think, the prerequisites are—Grossly wrong . . . Accurately stated

Figure 2: The four feedback questions and their answer ranges

Alignieren, Aligning, Analysis, ANNIS, annotation, annotation management, annotation of speech data, annotation panel, Annotieren, AntConc, Artificial Intelligence, Äußerungsliste, automatic segmentation, bash, Bayes, character encoding, chunking, CLARIN-D, COCA, collocation, Coma, combinatorics, command line, Concordance, corpus, corpus analysis, corpus annotation, corpus linguistics, Corpus Manager, corpus search, corpus workbench, corrections, COSMAS II, CQPweb, creation of speech data, data mining, Demo Corpus, DeReKo, dictionary, digital editing, Dublin Core, entropy, errors, estimator, etree, EXAKT, EXMARaLDA, Fehler, formant analysis, fraktur, GATE, geolocation, Gesprächslinguistik, glossary, HIAT, hidden Markov model, historical text, Importieren, Importing, infrastructure, intensity analysis, Introduction, IPA, Java, JavaScript, Korpus erstellen, Korpuslinguistik, Korrigieren, language resources, latent Dirichlet allocation, LaTeX, LaTeX2e, LDA, lemmatization, lexicography, lexicon, linguistic resources, linguistics, Linux, lxml, Machine Learning, Maintenance, MALLET, manual annotation, markup language, MAUS, Merging, metadata, Metadaten, MMAX2, MOSES, named entity recognition, NER, NLP, NLTK, OAI-PMH, OCR, Open Archives Initiative, parser, part of speech, Partitur Editor, personal name, phonetic alphabet, PHP, pitch analysis, Praat, python, quality management, R, RDF, Regular Expressions, Reguläre Ausdrücke, relation, resource description framework, Rhetorical Structure Theory, RST, scientific writing, sed, segmentation, segmentation errors, Segmentieren, Segmentierung, sentence splitting, sound, spectral analysis, Splitting, Spoken Language, Stanford, statistical analysis, statistical machine translation, statistics, Statistik, Struktur, STTS, Stuttgart, Support, tagger, tagging, tagset, teaching, TEI, teilen, TeLeMaCo, test theory, text encoding, Textile, Textlinguistik, TIGERSearch, TIGERXML, tokenization, tools, topic model, Transcription, Transkription, treebank, Tübingen, typesetting, typography, UAM corpus tool, UNIX, utterance list, verbinden, Video, video annotation, Visualisierung, W3C, WebAnno, WebLicht, WebMAUS, Weka, Windows, word cloud, written data, XML, XPath,

Figure 3: Available keywords in the browse interface (as of 2015-03-09)

The browse page[4] allows the user to select a keyword or an author. A list of known keywords is found in Figure 3.

4 Adding material to TeLeMaCo

Adding material to TeLeMaCo is easy and should not take longer than five minutes. The following fields can be filled (only two of them—marked with a star—are compulsory):

Title* The title of the material. This field must be filled.

Keywords Some keywords describing the resource. The user can chose the keywords freely. There can be any number of keywords.

Author The author(s) of the material. Note, that you can add materials to the portal that other people have created.

Language The language in which the material is written. We offer currently a maintained list of admissible languages.

Institution The institution that makes the material available. The portal uses this information to create a BIBTEX entry for the material.

Year The year when the material was published.

Objective What can be learned from the material?

Audience Who are the intended users for the described material?

[4] http://fedoara.clarin-d.uni-saarland.de/hub/browse/

input field	Dublin Core term	notes
access	accessRights	
BIBTEX entry	bibliographicCitation	generated by TeLeMaCo
author	contributor	
institution	creator	
media	format	recommendation: MIME type
	identifier	generated by TeLeMaCo
language	language	
url	relation	
licence	rights	
keywords	subject	
title	title	
type	type	we don't use the DCMI Type Vocabulary

Table 1: Correspondence between input fields and Dublin Core terms.

Prerequisites What is needed to make use of the material? It is possible to link to other descriptions in the portal using the Textile[5] markup language.

Level The level of the material, described by a closed vocabulary containing the values "not specified", "basic", "intermediate", "advanced", and "expert".

Type The type or genre of the material. The values can be freely chosen by the user, popular choices include quickstart and tutorial.

Media The media of the material, given as a MIME type.

URL* The location where the material can be accessed. This field must be filled.

Licence The licence under which the material can be reused.

Access There are only two admissible values: "open" and "academic". We do not support commercial items.

Description A free text describing the material.

BIBTEX type A type for creating the BIBTEX entry for the material, usually misc. It is possible to manually post-edit the generated BIBTEX entry.

5 Metadata

Most of the input fields of TeLeMaCo have a natural correspondence to Dublin Core terms (DCMI Usage Board, 2012), as detailed in Table 1. Some more Dublin Core concepts are automatically filled by TeLeMaCo; this includes biliographicCitation and identifier. For the Dublin Core term type, DCMI suggests a coarse grained controlled vocabulary. We don't follow this suggestion. Instead, we allow the users to enter a type at their deliberation, frequently used types include *Tutorial* and *Quickstart*.

For the pedagogical metadata we consider a mapping to the concepts of the Learning Resource Metadata Initiative (LRMI) (Association of Educational Publishers and Creative Commons, 2014).

We have a preliminary implementation[6] of CMDI (Broeder et al., 2011) metadata for the materials registered in TeLeMaCo. We plan to make these metadata harvestable via an OAI-PHM interface.

[5] `http://txstyle.org/`

[6] See, e.g., `https://fedora.clarin-d.uni-saarland.de/hub/cmdi/313/`

6 Additional benefits

The contents of TeLeMaCo are crawled and indexed by the big search engines (Google, MSN, Yahoo, Yandex, Baidu). This has two effects to materials added to TeLeMaCo: Some users will find the TeLeMaCo display page in the search engine of their choice and go on to the wanted material, and the page rank of the original page is boosted (leading users directly from the search engine to the material).

The display page of a teaching or learning material[7] has internal weblinks to the authors and keywords weaving a web of related resources. This allows the user to navigate to other materials for the same tool, the same task, or by the same author.

7 Evaluation

After a phase of internal testing within the CLARIN-D project, TeLeMaCo went public in September 2013 and was announced at GSCL 2013 (Amoia et al., 2013). Since then, a steady trickle of descriptions has been added to TeLeMaCo, now holding a total of 145 materials. Most of the contributions still come from members of the CLARIN-D project, but we start seeing submissions from other places, too.

We see in the logs that users from all over the world start consulting TeLeMaCo for teaching and learning materials. It was a surprise for us to see calls to a specific description directly without prior searching or browsing. These hits are coming from users being directed to TeLeMaCo by a search engine.

The assignment of keywords by the contributors works reasonably well as can be seen in Figure 3. The autosuggest feature helps in selecting already existing keywords. Some contributors have chosen German keywords for materials in German language. We currently do not attempt to merge the different languages.

Since September 2014 TeLeMaCo is listed in the large directory at LINSE (*Linguistik-Server Essen*)[8]. LINSE is a German language portal to all kinds of materials and services around linguistics.

8 Comparison with other services

We are aware of two collections of teaching resources for the documentation of endangered languages. The E-MELD School of Best Practices[9] is a project supported by the LINGUIST list. Resources are added by the project team, and although there was little activity since 2007 the project is still alive. There are short descriptions of the materials in free text format and there is a search function.

The Resource Network for Language Documentation (RNLD)[10] hosts a more up-todate list of resources for language documentation. Materials are described in free text format. They provide a full text search over the whole website.

The project DARIAH-DE[11] has launched a service called *Schulungsmaterial-Sammlung*[12] in July 2014. The target group are Digital Humanitites. The interface to this service is in German, materials are added by the staff members only. The materials have short textual descriptions and come with the following annotations: institution, media, title, tools, didactic type, discipline, language, date, keywords (up to three) choosen from the closed TaDiRAH (Perkins et al., 2014) vocabulary, and licence. There is a full text search over all the fields available.

9 Conclusion

We think that TeLeMaCo fills a gap in the existing ecosystem of language infrastructures by providing easy and quick access to teaching and learning materials. Descriptions of materials can be provided by everyone after registration at the service, avoiding a bottleneck in extending the service.

[7] e.g., `https://fedora.clarin-d.uni-saarland.de/hub/resource/313/`
[8] `http://www.linse.uni-due.de`
[9] `http://emeld.org/school/index.html`
[10] `http://www.rnld.org/resources`
[11] `http://www.dariah-de.eu`
[12] `https://de.dariah.eu/schulungsmaterial`

TeLeMaCo provides structured metadata of the resources that can be further integrated in the CLARIN infrastructure.

Both tools and available documentations benefit from being added to TeLeMaCo. They are not only findable through TeLeMaCo itself, but also their visibility in search engines is improved.

References

Marilisa Amoia, Hannah Kermes, Jörg Knappen, José Manuel Martínez Martínez, Elke Teich, and Mihaela Vela. 2013. TeLeMaCo—a collaborative repository for training and teaching materials in linguistics. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2013)*, Darmstadt, Germany.

Association of Educational Publishers and Creative Commons. 2014. LRMI Specification Version 1.1. `http://dublincore.org/dcx/lrmi-terms/1.1/2014-10-24/`.

Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to XML interoperability—the component metadata infrastructure (CMDI). In *Balisage: The Markup Conference 2011*, volume 7.

DCMI Usage Board. 2012. DCMI Metadata Terms. `http://dublincore.org/documents/2012/06/14/dcmi-terms/`.

Marie Hinrichs, Thomas Zastrow, and Erhard Hinrichs. 2010. Weblicht: Web-based LRT services in a distributed escience infrastructure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valetta, Malta.

Jody Perkins, Quinn Dombrowski, Luise Borek, and Christof Schöch. 2014. Building bridges to the future of a distributed network: From DiRT categories to TaDiRAH, a methods taxonomy for digital humanities. In *International Conference on Dublin Core and Metadata Applications 2014*, pages 181–183.

Thomas Schmidt and Kai Wörner. 2009. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19(4):565–582.

Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Thomas Vander Wal. 2007. Folksonomy Coinage and Definition. `http://www.vanderwal.net/folksonomy.html`.

COMEDI: A component metadata editor

Gunn Inger Lyse*, Paul Meurer†, Koenraad De Smedt‡
University of Bergen*‡, Uni Research Computing†
Bergen, Norway
E-mail: gunn.lyse@uib.no*, paul.meurer@uni.no†, desmedt@uib.no‡

Abstract

The flexibility of component metadata (CMDI) brings about a need for editing tools which are equally flexible. Moreover, such tools should be as user friendly as possible in order to lower the threshold for beginners and to promote efficiency even for advanced users. The current paper presents COMEDI, a new web-based editor which handles any CMDI profile. We evaluate currently existing metadata editors and argue that the COMEDI editor is the first one to combine a good level of user-friendliness with sufficient support for CMDI. COMEDI also offers up to date support for CLARIN features such as current license types.

1 Introduction

Digital language data and tools (hereafter for short called 'resources') benefit from good metadata to promote their visibility, reusability and durability (Trippel et al., 2014; Piperidis et al., 2014; Dima et al., 2012; Lyse et al., 2012; Soria et al., 2012; Wittenburg et al., 2010). CLARIN[1] aims to provide researchers with improved access and added value to existing resources for their reuse. To this end, structured documentation of resources is a key factor, enabling researchers to find existing resources and judge their relevance for the intended purposes.

However, the process of creating and managing metadata is a challenge for the less experienced and time consuming even for the experienced. The quality and completeness of metadata will suffer if the tools for metadata creation and management are cumbersome or have missing functionalities (Withers, 2012). Therefore, it is crucial to offer tools that lower the threshold for filling in and editing metadata in a format that CLARIN requires.

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 XML-based CMDI format attempts to strike a balance between flexibility and stability. The basic building blocks of CMDI are *components*, which 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 certain 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.[2]

In CLARIN, CMDI is the recommended standard for metadata, and certified CLARIN B-centres are required to offer component based metadata for harvesting via the Open Archives Iniative Metadata Harvesting Protocol (OAI-PMH).[3]

The flexibility and power of component metadata makes it desirable to develop and deploy suitable editing tools which are equally flexible and powerful. Moreover, such tools should be as user friendly as possible in order to lower the treshold for beginners and to promote efficiency even for advanced users.

[1] http://www.clarin.eu. Websites cited in this paper were consulted on June 30, 2015.
[2] http://catalog.clarin.eu/ds/ComponentRegistry/
[3] http://www.clarin.eu/node/3577

In this paper, we evaluate the currently available metadata editors and argue that these do not sufficiently support the full power of CMDI in combination with an adequate level of user-friendliness. We then present COMEDI (COmponent Metadata EDItor), which was designed and implemented in the CLARINO project, the Norwegian part of CLARIN.

COMEDI is a web service which can start from any metadata schema conforming to CMDI. Among its features are an intuitive web interface, cloning of information from existing metadata files and validation. It supports all features of the Resources section, such as defining resource proxies which can be referred to in components. COMEDI also offers up to date support for CLARIN features such as current license types.

2 Existing metadata editors

We propose the following as some desirable features for the design and evaluation of component metadata editors:

- handling of any registered CMDI profile;
- import and export of CMDI files;
- fully online web interface to remote processing, saving and storage;
- navigation and editing with menus as well as keyboard shortcuts;
- reduction of repetitive typing tasks, e.g. by cloning of information pieces;
- controlled vocabulary where needed;
- validation of input;
- authentication and authorization of users, with management of shared access;
- support for up to date CLARIN features such as license types.

2.1 Arbil

Arbil[4] is a metadata editor, browser and organizer tool for CMDI, IMDI and similar metadata formats (Withers, 2012). Arbil has been characterized as 'the reference implementation for a CMDI editor' (Dima et al., 2012), but it is also acknowledged that this tool is primarily directed towards archivists or librarians rather than non-expert researchers (ibid.); the latter group may therefore find the learning curve in Arbil rather steep.

Arbil must be installed on the user's system and can be used offline as well as via a webstart. It allows users to create and edit metadata displaying the underlying XML code as plain table structures. Arbil offers several facilities to reduce the burden of repetitive typing tasks, such as bulk editing of metadata via copy and paste into multiple fields of multiple rows. It displays trees of metadata in its user interface. Frequently used sections of metadata can be collected from the *Favourites* directory and be reused for new metadata, thus reducing the amount of repetitive data entries.

Arbil allows the user to type metadata in any order. It warns the user when a metadata field is missing or is not in the required format but allows the user to continue editing and to save locally, even with errors.

When exporting the metadata, all metadata files are checked for inconsistencies and if necessary, warnings are given. Only at the point of pushing the metadata into the remote archive will the user be blocked if they have not correctly completed all the required fields.

In many metadata sets, the number of fields required to describe the data and its context can be extensive; this can make it difficult for a user to see their relevant information at a glance. The table columns in Arbil are therefore customizable, so that only those relevant to a particular user need to be displayed.

Despite its many useful features, Arbil has some drawbacks, among which are its steep learning curve and the local setup and local saving. This makes Arbil less user-friendly for an occasional user, such as the average metadata provider within CLARIN. Moreover, users report that they experience Arbil as responding very slowly.

[4] `http://tla.mpi.nl/tools/tla-tools/arbil/`

2.2 ProFormA

ProFormA[5] is a web-based CMDI editor (Dima et al., 2012). Through a web interface, the user selects a CMDI profile to start from, and the editor displays the profile as a plain online form hiding the XML code. The user may create new files or upload and edit existing files; records can be downloaded as XML files. The ProFormA editor works on local copies of CMDI records, either created in the tool or uploaded by the user. Since ProFormA only allows local saves during a working session, the user must make sure to export an XML copy from ProFormA before closing a session.

ProFormA allows the user to select and upload any CMDI profile in the Component Registry by typing its profile ID into a field in the editor's start page. The CMDI NaLiDa profiles can be selected directly from a drop-down menu in the editor's start page. An existing CMDI file may be edited by pointing to its URL or by uploading the file to ProFormA. In the case of new files, the user selects a CMDI profile and types a filename for the metadata file, which is then opened as a web form. The file is displayed as a simple online form. At the top of the web form, all components in the relevant profile are listed, allowing the user to navigate between components by clicking. Below the component menu, all elements in the selected component are listed consecutively.

Although ProFormA accepts any CMDI profile, some weaknesses became apparent when testing with a profile other than the NaLiDa profiles. As a test, we uploaded the META-SHARE profile *resourceInfo* for lexical resources.[6] Some display errors related to cardinality, i.e. the number of instances of a component or element, were observed. First, ProFormA does not display the required number of occurrences (the cardinality) of an element or a component, even though this information is available in the profile specification in the Component Registry. Second, we found that in the case of elements that may occur arbitrarily many times (cardinality $0 - \infty$), only one instance could be created; conversely, the editor *did* allow arbitrarily many instances of elements where the CMDI profile only allows at most one item. The screenshot from ProFormA in Figure 1 illustrates that it is possible to create more than one instance of the element *metadataLastDateUpdated*, which, according to the META-SHARE profile *resourceInfo* for lexical resources, has cardinality 1.

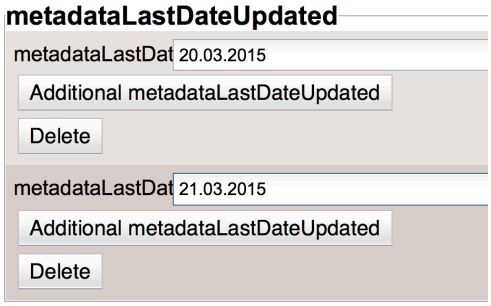


Figure 1: Screenshot from ProFormA illustrating that it is possible to create more than one instance of an element with cardinality [0-1].

Moreover, the hierarchical structure is not always correctly displayed. Sometimes the elements of a subcomponent are displayed without a subcomponent title, which leaves the contents of the subcomponent completely out of context. For example, inside the *distributionInfo* component of the META-SHARE profile *resourceInfo* for lexical resources, a subcomponent *licenceInfo* follows after the element for describing userNature. However, the editor does not display the name *licenceInfo*, so that its contents (person role, surname, given name etc.) do not have an appropriate context (see Figure 2). This is particularly unfortunate since ProFormA also omits the documentation text that ususally accompanies elements and components in the CMDI profile in order to explain the intention of that element.

There are also limitations regarding controlled vocabulary. For some elements (e.g. for *resourceName*) there is an accompanying language element, since a resource may have one English name and another in

[5] `http://www.sfs.uni-tuebingen.de/nalida/proforma/web/`

[6] META-SHARE v3.0 – lexical/conceptual resources; ID: clarin.eu:cr1:p_1355150532312

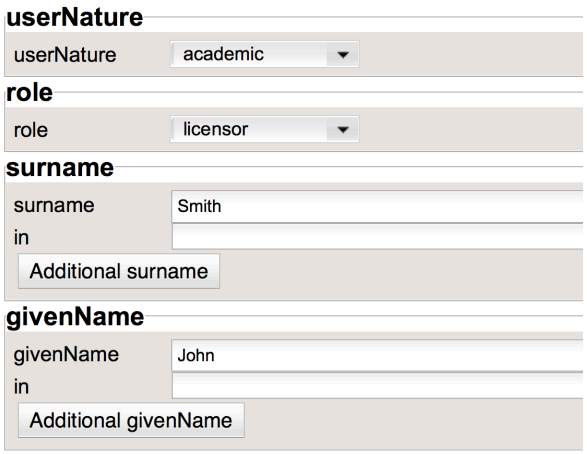


Figure 2: Screenshot from ProFormA illustrating problems with displaying the hierarchical structure of a component correctly.

the native language. ProFormA allows the user to choose language names from a drop-down menu which only lists four languages: Dutch, English, French, and German. It is possible to type other language names manually, but we found that ProFormA does not validate whether the input provided by the user is a correct language name, despite the claim (Dima et al., 2012) that ProFormA validates content provided by the user. Any string, even ill-formed, will pass. Also, ProFormA currently does not support editing persistent identifiers or editing of the header of CMDI files, even though this is an important functionality for archives and repositories.

In conclusion, ProFormA currently appears to be of limited use. It does not provide the full choice of CMDI profiles in the Component Registry and fails to provide some relevant information and functionality to the user.

2.3 META-SHARE

The META-SHARE model is tailored to describe language resources and tools relevant for language technology research and development, and thus has not been designed to support the full CMDI framework. META-SHARE is intended as an open infrastructure for sharing language resources in Europe.[7] It has been created in the META-NET project in cooperation with the META-NORD, META4U and CESAR projects. META-SHARE was deployed Europe-wide in early 2013 and allows anyone to search for language resources online, based on metadata.

META-SHARE offers metadata schemata for four basic resource types: corpus, lexical/conceptual resource, tool/services and language description. Provided that these schemata meet the user's needs, metadata can be created or uploaded, stored, edited, searched and downloaded using an online META-SHARE editor which is implemented as a web-based form. The META-SHARE model has been mapped and incorporated into the Component Registry of CMDI (Piperidis et al., 2014), facilitating a conversion from META-SHARE metadata to CMDI metadata. For this conversion, XSL stylesheets have been written.[8] There is ongoing work to implement and embed an identifier (corresponding to a persistent identifier), using the International Standard Language Resource Number (ISLRN) (Choukri et al., 2012).

The META-SHARE system is constructed as an integrated online solution allowing anyone to search in metadata and allowing authorized users to edit metadata. It is a platform independent, open-source solution, implemented using the Python-based framework Django, which is maintained by active open source communities. META-SHARE is itself open source, released under a BSD licence.[9]

If desired, stored META-SHARE metadata records can be published directly in its online public search interface. Alternatively a metadata record may be tagged as internal (visible and accessible only to au-

[7] http://www.meta-share.eu, http://www.meta-share.org
[8] https://github.com/metashare/META-SHARE/tree/master/misc/tools/CMDIConverters
[9] https://github.com/metashare/META-SHARE

thorized users) or as ingested (visible and accessible in the web editor but not published), or it may be downloaded as an XML file. Resource owners can define editing groups and manage their membership, thereby allowing easy cooperation between several individuals.

The META-SHARE editor supports differing degrees of descriptive detail at two levels: the minimal schema contains obligatory elements (e.g. *Resource name* and a *resourceType* classification), and the maximal schema contains optional information. The minimal schema provides the basic elements for the description of a resource which are therefore obligatory, whereas the maximal schema offers the option to add more detailed information on each resource. Values can be free text or from a limited vocabulary. For example, originally the user could enter license names at will, but this has now become a closed list of license names (plus an option 'other' if the user cannot find the relevant license in the drop-down list).

META-SHARE also has mappings and links to ISO and DC whenever relevant, and also offers certain autofill possibilities. For example, when language is specified, the user can type either the language name (in which case the editor makes autocompletion suggestions) or the ISO code. In either case, the editor fills in the corresponding field for the user.

It is not possible to make an intermediate save without first filling in all obligatory elements. This has the unfortunate side effect that the metadata creator may type 'dummy' information if the relevant information is not available at that time, thus creating a backlog for the metadata creator at a later stage. Even worse, if the user is unable to correct the source of an error message (for the unexperienced user error messages may sometimes be unclear), all data that have been typed so far (even correctly filled in fields) will be lost.

The editor offers autofill (e.g. for today's date) and features controlled vocabularies extensively through drop-down lists (e.g. *Linguality type*) and autocompletion (e.g. language codes), which promotes consistent and correctly typed metadata. The amount of repetitive typing is greatly reduced by storing person info, institutional info and research project info as separate objects in a database which can be pointed at by multiple metadata records. The next time the same object occurs (e.g. the same person), its complete information can simply be selected from this database. If information about an object must be changed later (e.g. changing the e-mail of a person object), the change is automatically updated in all metadata records pointing to this object, provided that the record is stored in the META-SHARE repository. The META-SHARE editor does not support, however, cloning entire records or chunks of information, such as an entire component. Cloning would be desirable in cases where multiple metadata records have a lot of similar information, e.g. if we are creating ten metadata records describing ten parts of a collection, with only minimal differences. In META-SHARE this can only be achieved outside the editor: first, create a record in the editor and making sure it contains all the information you need to duplicate; then download it in XML to your personal computer; create duplicates of your file; then edit each file in XML (and risk syntax errors) or upload each duplicate to META-SHARE and edit it further there. Either way, it is a time-consuming procedure.

A challenge with large metadata schemata is how to portion out elements, components and subcomponents, and META-SHARE appears as very complex and at times confusing in this respect. Subcomponents are usually shrunk initially, but the editor misses a uniform display both for shrunk and expanded components and for how to expand components.[10] For example, main components such as the *Corpus Text* component can be found by clicking via the left-side menu (Figure 3). Whereas the first of the main components, *Administrative Information*, will appear on the same page, the others will appear in a pop-up window.

As illustrated in Figure 4, the component *Distribution* (inside the main component *Administrative Information*) appears in a framed, multiple-line box with documentation text below the component name. The component opens in a pop-up window by clicking a green plus sign. By contrast, the component *Annotations* (within the 'main' component *Corpus text > Recommended*) has a different display: the component name appears in a one-line, grey-coloured box, without documentation text. The component name is a grey font, which makes it even harder to find, especially since it is surrounded on the same page by components that appear as already opened, multiple-line boxes with a blue header line (cf. Figure 5,

[10] See for instance the discussion at: `https://github.com/metashare/META-SHARE/issues/315`.

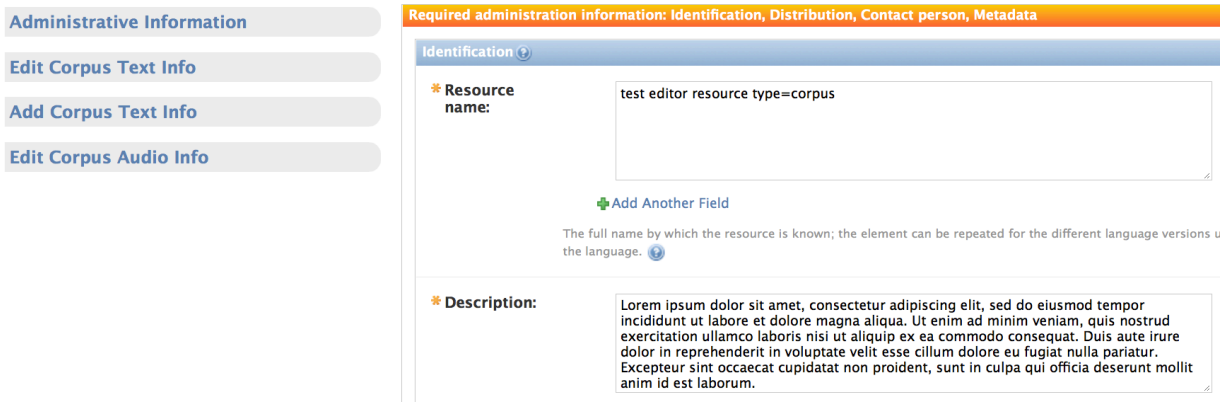


Figure 3: Screenshot from META-SHARE illustrating the heterogenous display for shrunk and expanded components and for how to expand components. Main components are clickable from the left-side menu.

where the component *Annotations* is hidden in the middle of the screenshot). In this case, the component can be opened by clicking on a word *Vis* "Show" after the name,[11] requiring the user to recognize a cue which is quite different from the plus sign used for the same purpose elsewhere. Upon clicking, the component is displayed on the same page instead of as a pop-up window, thus presenting a completely different solution than that in Figures 3 and 4.

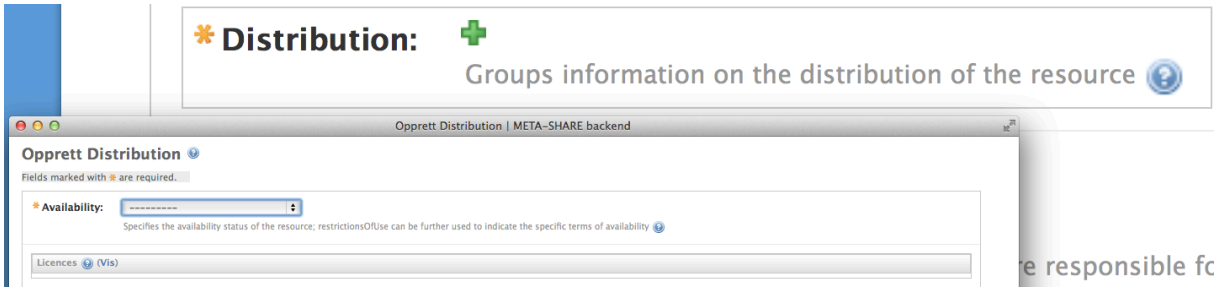


Figure 4: Screenshot from META-SHARE illustrating the heterogenous display for shrunk and expanded components as well as for the expansion of components. In this case, a subcomponent is displayed in a pop-up-window in front of the main page upon clicking the green plus sign.

In conclusion, the META-SHARE editor has some user-friendly features that support efficient and correct editing; however, like ProFormA, it has not been designed to fully support CMDI profiles, which is a non-trivial drawback. The interface shows some lack of consistency and there is also a potential for other improvements.

2.4 Other editors

General purpose XML editors such as Oxygen[12] are challenging to be used by non-experts, since their effective use requires some insight in XML technologies (Dima et al., 2012). The IMDI-editor[13] seems to be more or less replaced by Arbil and will not be further discussed in this paper. CLARIN-D has created the HTML5 web app CMDI Maker,[14] which is now part of the CLARIN infrastructure. It allows users to load files that are part of a resource, to which the researcher can add metadata. CMDI Maker creates IMDI records that may be exported and subsequently uploaded as CMDI with an IMDI profile via Arbil. This editor allows cloning of components for person-related data. It seems limited to a specific set of profiles,

[11] Some elements in the interface have been localized to Norwegian.
[12] http://www.oxygenxml.com/
[13] https://tla.mpi.nl/tools/tla-tools/older-tools/imdi-editor/
[14] http://class.uni-koeln.de/cmdi_maker/

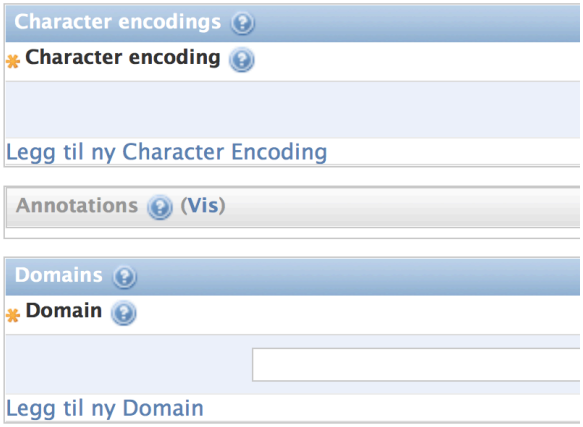


Figure 5: Screenshot from META-SHARE illustrating the heterogenous display for shrunk and expanded components and for how to expand components. In this case, a subcomponent will be opened in the same window after clicking the word *Vis* "Show".

and would therefore not be widely usable.

3 COMEDI

3.1 Overview

The COmponent Metadata EDItor (COMEDI) is a fully web-based editor which handles any CMDI-compatible profile.[15] In COMEDI, one can create a metadata record from scratch, or upload, edit and download any CMDI XML file. It may be used simply as an editor for filling in metadata (and downloading the resulting XML file), but it also functions as a full metadata server for storing, searching, viewing and managing metadata, with user group management for controlling the access rights to individual metadata records. A metadata record in COMEDI can be exported as a CMDI XML file, and is harvestable with OAI-PMH. In accordance with the OAI-PMH standard, all metadata can also be harvested in Dublin Core format.

The developer's instance of COMEDI is currently integrated in a web framework at the emerging national CLARIN type B center at the University of Bergen, Norway. The system can also be deployed as a stand-alone web service which can be installed on individual servers, and the software is available under a BSD license. A stand-alone instance has been installed at the National Library of Norway, with the intention to provide a national metadata registry that will be harvesting from metadata providers in Norway. Another European CLARIN centre has decided to use the local developer's instance, since it offers adequate facilities to define user groups for the relevant centre to have their own 'workspace' within COMEDI. Moreover, they do not intend to store their data on the installation, only to create metadata and download the resulting XML files.

3.2 Authentication and authorization

The contribution of metadata by unidentified users is generally undesirable; moreover, user identification makes it possible to define user groups with shared editing rights to sets of metadata records. Therefore, users of the editor are authenticated via login. The COMEDI editor provides an Authentication and Authorization Infrastructure (AAI) based on the DiscoJuice IdP discovery service. If this AAI is properly set up at the installation, users can login to COMEDI through the CLARIN IdP, the eduGAIN interconnection of IdP federations, or OpenIdP (a self-registration service offered by the Norwegian academic identity provider).

A user management system has been implemented that operates on two levels: the user and the group level. Authorization with differing degrees of rights can be given to authenticated users on an individual

[15] At present, version 1.1 is fully supported, while support for CMDI 1.2 is planned.

basis, such as administration of other users, creating persistent identifiers, etc. When a metadata record is created, it is owned by the creator, who by default has sole write access to the record. Other users can request write access to the record by clicking a button on the metadata page. The owner, who receives an email notification, may then grant or deny access to the record.

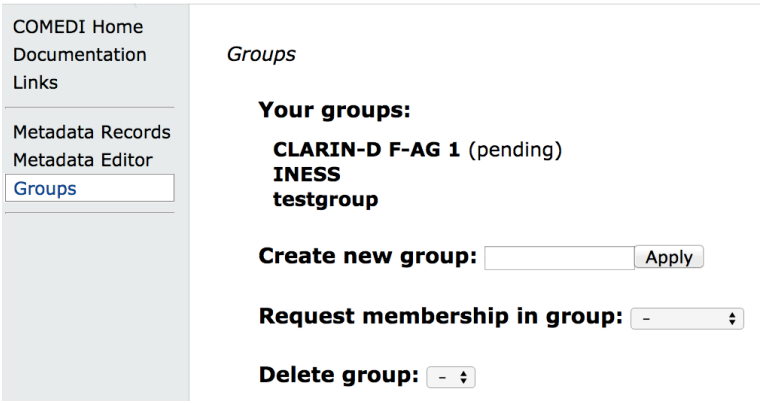


Figure 6: With the user management system, a logged in user may define user groups, apply for membership in existing groups and grant other user access to a group.

In addition, any user can define a user group, initially with that user as its sole member. A group is created by clicking on *Groups* on the left-side main menu of the editor (cf. Figure 6). Again, other users can apply for membership in the group. Subsequently, the group creator, upon receiving an email notification, can grant or deny membership. When a group has been established, a group member can connect an owned record to that group, thus giving all members of the group write access to that record. These operations are reversible. If necessary, a user can be removed from a group, and a group can be deleted.

3.3 Creating, uploading, cloning and searching for metadata records

The left-side menu of the main page has, among other things, a link to a documentation page explaining the functioning of the editor (as can be seen on the left side of the screenshot in Figure 6). By clicking on the *Metadata records* page via the left-side menu, the user finds the main page for starting to work in the editor. At the top of the start page, there are four dialog boxes allowing the user to create a new metadata record, upload a metadata file, clone a record, or search for and select existing metadata in the editor's database for viewing and editing. Each of these options will be described in the following.

Similarly to the solution in ProFormA, the user can create new records by typing the CMDI profile ID (cf. Figure 7), or by choosing among a the set of profiles already registered in the database in a drop-down menu. The user then has to choose a new record identifier which will be used internally for naming the file (e.g. in Figure 7 the name 'demo-corpus' was entered). By clicking the "Go" button, a new, empty record is created based on the selected CMDI profile.

Like in Arbil, COMEDI allows you to use both published and unpublished profiles in the Component Registry, by entering enter the profile ID into the box for choosing a profile. Since profiles published in the Component Registry cannot be modified, it is very convenient to be able to test a metadata profile on real metadata before publishing it in the Component Registry.

Having created a metadata file in COMEDI, the user will find at the top of the page some information about the profile that was used as a blueprint for this metadata file: its profile ID, name, description and the date of fetching it from the Component Registry. There is also a button for refetching the profile, in case the profile is still unpublished and may have been modified in the Component Registry since it was last fetched.

Existing metadata can be uploaded in two ways: via a Web form, or programmatically, using a POST request to a dedicated URL. When the Web form is used (see Figure 8), a CMDI record file is uploaded

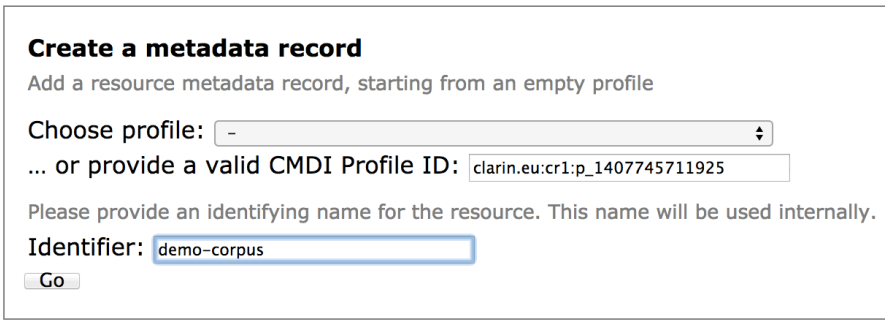


Figure 7: Screenshot illustrating the creation of a new metadata record based on a CMDI profile referred to by its ID.

from a local computer. As COMEDI-internal identifier the name of the file is used; information about the used profile is extracted from the CMDI header.

Figure 8: Screenshot illustrating the dialogue box for uploading CMDI-conformant metadata to COMEDI.

If there are many CMDI files to be uploaded a scriptable mechanism is certainly preferable. To this end, there is a URL that accepts POST requests with a file upload, and an optional parameter that specifies the user group the metadata should be attached to. A slight complication lies in the fact that users who deposit metadata to COMEDI have to be authenticated. Since there is no easy automated way of doing federated authentication, a workaround has been implemented that makes use of the browser-based authentication built into COMEDI: when a user logs into COMEDI a session token is generated that is stored in a cookie. This session token is also exposed on the COMEDI website, and can be used for authentication when programmatically uploading metadata, using the parameter *session-index*. An example shell command for uploading metadata using the Unix *curl* utility could look like the following request, which returns a JSON object signalling either success or failure:

```
curl -F "file=@tiger-treebank.xml"
        "http://clarino.uib.no/comedi/upload?group=CLU&session-index=_a19a23c4"
```

As a last resort, when a service heavily relies on being able to upload metadata, where it is not feasible to manually insert the session index, a permanent access code for uploading can be given upon request.

A metadata record in COMEDI can be accessed at a dedicated URL as a CMDI XML file. Additionally, there is an OAI-PMH endpoint that allows harvesting all published CMDI records. A metadata record has a COMEDI-internal status attached to it which tells whether the record is still unfinished, ready to be published, or finalized. Via the OAI-PMH endpoint, only the finalized records are published and available for harvesting. The OAI-PMH endpoint also makes use of the user groups; they are encoded as OAI-PMH sets. Thus, to harvest all metadata created in the INESS user group one use the following:

```
http://clarino.uib.no/comedi/oai?verb=ListRecords&MetadataPrefix=cmdi&set=INESS
```

Metadata records describing similar resources tend to have much information in common. To ease the creation of very similar metadata records, COMEDI offers the possibility to clone an existing CMDI record.

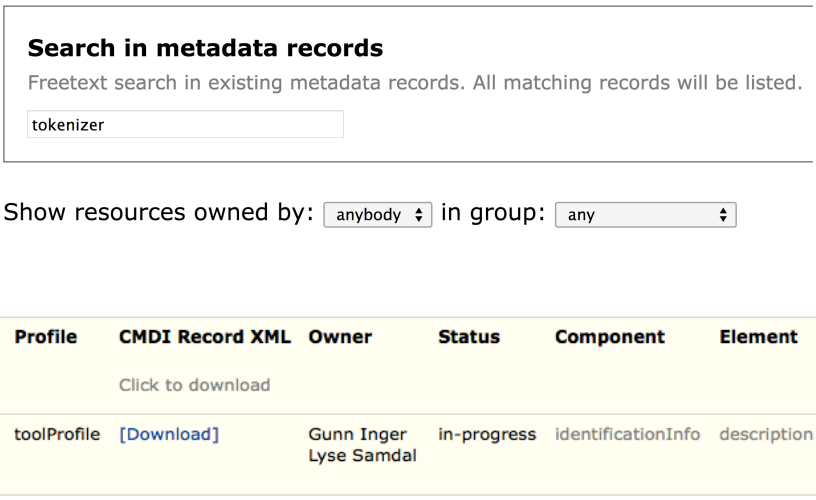


Figure 9: Two screenshots illustrating the dialogue box for searching for content in metadata files in COMEDI (top) and an example of a search result (bottom).

A new CMDI record will be created that is an exact copy of the original record, except that a new identifier has to be supplied, and some fields, like *Self link* and other persistent identfiers relating to the original resource will be cleared, and the header fields *Creator* and *Creation Date* will be adapted.

Below the three dialog boxes for creating, cloning and uploading metadata, there is a box for searching in existing metadata records that are stored in the COMEDI database (cf. Figure 9). Currently this is implemented as a simple string search, and all hits are returned as a list specifying the element and component in which the search string was found. Existing metadata records are listed by resource name and identifier. For instance, Figure 9 shows a resource with name *NLTK tokenize punkt* as specified in the metadata, and below it, the identifier *nltk-tokenize-punkt* is shown. By clicking on the identifier the user may inspect or edit the metadata.

3.4 Component display and navigation

A full metadata schema may appear overwhelming, so the ability to navigate efficiently and selectively show or hide elements to the user is therefore beneficial.

COMEDI displays one top level component at a time, while keeping a menu at the top of the page where the user may switch to another component by clicking on it or by using keyboard shortcuts. The component menu can be seen as the top line menu in Figure 10, where the currently chosen component, *Contact person*, is displayed in boldface[16].

COMEDI offers advanced navigation functionality with consistent shrinking and expansion. All navigation can be done with keyboard shortcuts alone or by mouse-clicks, such as navigating from one component or element to another, switching between edit and view mode, editing content, adding or removing components and elements, showing and hiding subcomponents, or switching between top-level components. The editor also offers basic tab navigation. This range of interaction modes should accommodate both occasional and regular users. An on-line wiki-type documentation is provided.

To accommodate the need for going into detail while typing metadata, and at the same time keeping an overview, COMEDI has two display modes, *view* mode (Figure 10) and *edit* mode (Figure 11). In view mode, the editor displays only the necessary information; specifically, it displays obligatory elements and any other user-provided content. Missing obligatory elements, or user-provided content that does not validate correctly, are marked in red. For instance, Figure 10 shows the *Contact person* component in view mode, where the obligatory e-mail address is missing. Thus, by switching to view mode, the user can easily review the contents filled in thus far and check for missing, obligatory information.

[16]The example uses the META-SHARE corpus profile, ID: clarin.eu:cr1:p_1361876010571

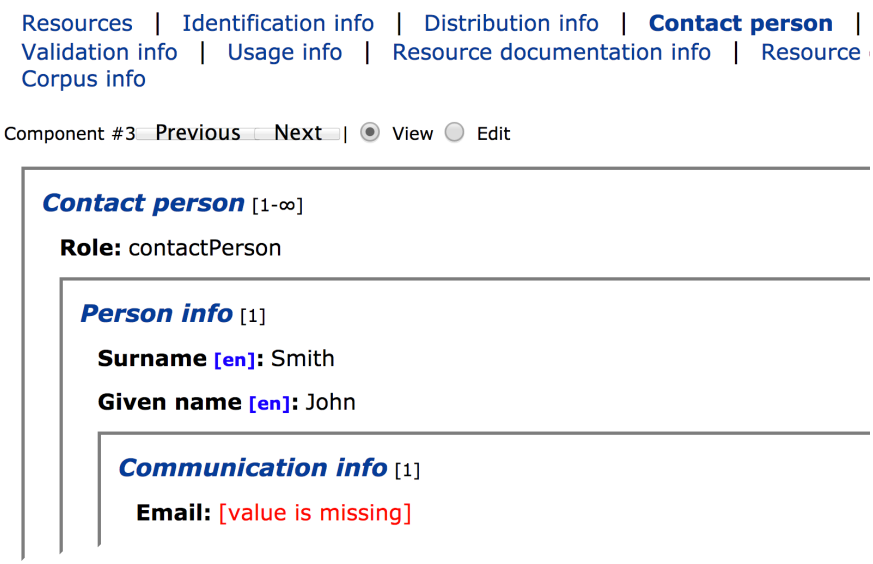


Figure 10: Displaying one top-level component at a time (view mode).

In edit mode, metadata information is displayed in an appropriate level of enhanced detail. In both modes, the components and their elements are shown hierarchically as boxes containing boxes (cf. Figure 10). Akin to the profile display in the Component Registry, components can be expanded and shrunk to adjust the level of detail. Initially, all but the top-level components are shrunk. Optional uninstantiated components are hidden in view mode, but visible in gray in edit mode. Obligatory components and elements are by default open in edit mode whereas optional ones are initially shrunk. Optional components and elements that are not instantiated are also gray.

To illustrate the edit mode, Figure 11 shows the *Communication info* subcomponent inside the *Contact info* component illustrated in Figure 10. In edit mode, each component and each element is displayed along with its documentation from the Component Registry. Adjacent to the component or element name, inside square brackets, the user can see the number of instantiations of a component or element (the left-hand number), along with its allowed minimum and maximum, i.e. its cardinality (the right-hand number). For instance, the element *Given name* in Figure 10 has been instantiated once and may occur zero times or once, hence we see: [1/0-1] in Figure 11. Instances of elements and components can be created or deleted using [+] and [−] buttons, if allowed by the component definition. To prevent accidental deletion of metadata, the user will be asked to confirm before deleting any data.

The user may enter metadata in any order desired. As opposed to META-SHARE, a reliable save functionality is available regardless of whether all user content is valid according to the profile specification. Moreover, after every edit operation, the current metadata is stored in the server database; no explicit save command is necessary. In addition, dated snapshots can be stored at any time, and if necessary, a user can revert to a previous snapshot. Furthermore, through the web interface, some ordinary editing functionality is provided by all modern browsers, such as spell checking and editing functions, including undo, within a form field. COMEDI supports Unicode character input and right-to-left scripts.

Validation, controlled vocabulary and the automatic insertion of metadata are indispensable tools to reduce the amount of inconsistencies and errors in the metadata, while also saving time for the metadata creator. The COMEDI editor validates input according to the *ValueScheme* specification in the element definition and displays an error message on invalid input, for instance for language names, correctly typed `s` and date. Support for vocabulary services like OpenSKOS is planned in the transition to CMDI version 1.2. If a value is invalid, an error message is displayed below the value (Figure 11). Some fields, like metadata creation date and last change date, are filled in automatically.

COMEDI allows easy cloning of components from existing metadata records stored in the repository, thus greatly reducing the work burden of repetitive typing tasks. When clicking on *select component*

Figure 11: Edit mode: validation of language code on the fly.

next to the component name (in edit mode), a list of existing components with the same *ComponentId* appears (Figure 12) and the user can scroll through the list. Since the list of existing components may be extensive, the list may be filtered by entering a search string. In Figure 12, the full list of existing *Funding project*s was narrowed down to 3 existing components containing the string *META-*. The illustrated existing component is the first of three, (denoted at the top of the component as *1/3*), and this first listed component instance appears in 14 metadata records (denoted by *(14) at the top of the component)*. Upon selecting one of the items, its content is copied into the component in focus. To make this feature safe to use, a component can only be filled with new content if it is empty; otherwise, it has to be cleared first. If a component already has content, the user can clear it by clicking *Remove content*. Similarly to Arbil, useful components can be marked as favorites in the component selection box; they will then appear first next time the user evokes the component list for this component type. Cloned contents may be edited as desired. Such editing will not affect the contents of the original, and the edited version will in that case simply appear as a new item in the list of components.

Figure 12: Cloning components: selecting an existing component for cloning.

COMEDI supports all features of the Resources section. Resource proxies can be defined, and they can be referred to in components. This is for instance useful to express that certain parts of the metadata, such as a license or a contact person, only refers to parts of the resource. Users can choose the naming of the *id* values themselves, but uniqueness is checked in COMEDI. The Resources section of the metadata record can be edited in much the same way as ordinary components.

Figure 13 illustrates an example where the user is creating metadata for a treebank and wants to define one contact person in the metadata for questions about the treebank (named in the example with the proxy ID 'iness-nob') and another for questions about a search interface (named with the proxy ID 'corpuscle'). To achieve this, the user must first define the needed resource parts in the Resources section. Next, the user must create two instances of the *Contact* component. As illustrated in Figure 13, every component instance gets a clickable *ref* ('reference') section to the right of its name. When clicking on this reference section (in edit mode), a list for selecting proxy IDs appears. The user may then select, for example, that the first *Contact* instance should refer to 'iness-nob' whereas the second instance should refer to 'corpuscle'.

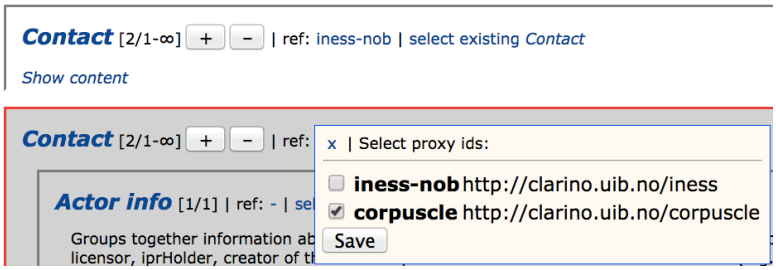


Figure 13: Resource proxies can be defined and subsequently referred to in components.

3.5 Integration of persistent identifiers

In CLARIN, persistent identifiers (PIDs) should be based on the *handle* system offered by the Handle.Net organization. If an installation of COMEDI has access to a local handle server with a dedicated prefix, the handling of handles can be tightly integrated into the system. This is at present done in two places: in the metadata self-link, and in PID elements. A self-link handle can be registered automatically with one click (if appropriate and desired); this handle will point at the URL of the metadata record. For profiles that use the *IdentificationInfo* component, a handle can be registered automatically when the component's PID element is created, and in case there are instantiated URL elements in that component, the newly created handle will point at them. Vice versa, when a handle already exists, newly created URL elements are automatically connected to it. When a URL element is deleted again, it will also be removed from the handle in the handle system, and when a PID element is deleted, the corresponding handle is deleted from the handle server, thus avoiding dangling handles. In addition, the PIDs can also be inserted manually if needed.

The handles that are created in the system are by default EPIC-compatible[17]; they consist of the handle prefix (a number identifying the handle owner) and a suffix of 12 hexadecimal digits plus a checksum digit; they are devoid of semantics (with the exception that the creation time is coded into the suffix). A typical handle thus looks like the following:

```
hdl:11495/D8B8-3AA2-3332-1
```

Handle.Net handles support a flexible handle template mechanism, which allows one to add extensions to the base handle. Only the base handle is registered and resolved, whereas the extension is attached to the resolved URL, possibly in a transformed shape.

In COMEDI, the extension is attached verbatim to the resolved URL and thus retains the full flexibility of URL parameters.

[17]EPIC, the European Persistent Identifier Consortium, is a handle service based on Handle.Net; see `http://www.pidconsortium.eu`

As an example, the handle

```
hdl:11495/D8B8-3AA2-3332-1@version=1.1
```

would resolve to a request for version 1.1 of a resource:

```
http://clarino.uib.no/corpuscle/landing-page
    ?identifier=bul-treebank&version=1.1
```

3.6 Support for CLARIN features

COMEDI also offers support for CLARIN license information.[18] Among other things, CLARIN has introduced a classification system where licenses are placed in one of the three main categories PUB (publicly available), ACA (available for persons with an academic affiliation) and RES (restricted availability on a case-by-case basis). A major challenge for the non-legal expert is to provide consistent and correct license information in metadata. The CLARINO project, which is implementing the Norwegian contribution to the CLARIN infrastructure, has classified existing licences with respect to user Category, license family and conditions of use, and the CLARIN Legal Issues Committee (CLIC) has quality-checked the license table. This license table is a good candidate to be published as external vocabulary in the CLARIN Vocabulary Service in OpenSKOS (CLAVAS) which will be supported in CMDI 1.2. With a CMDI profile that uses CLARINO's license component, the user only needs to fill in the license family and the license name. COMEDI will automatically add information from the license table related to the given license, such as the correct user category, a license URL and conditions of use, as illustrated in Figure 14.

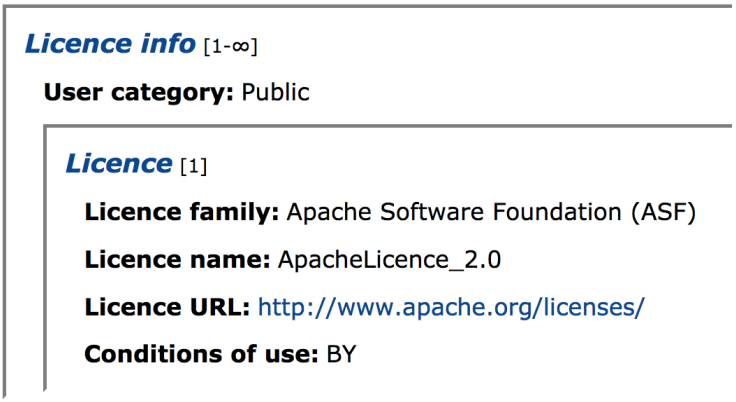


Figure 14: With the CLARINO license component in COMEDI, the user only supplies the license name, and the editor automatically fills in related metadata such as conditions of use.

3.7 User evaluation

Even though a clean web interface and features such as validation and cloning are known to add to user-friendliness, the actual user experience is an empirical issue. As a pilot study, COMEDI was tested on a researcher of linguistics with high technical skills in general, but without previous experience with metadata. The researcher was asked to fill in metadata for a lexical resource that he knows well. Overall, the test person found that the threshold for using COMEDI was low: getting started was easy and it was easy to keep track of the editing process thanks to the top-level component menu at the top of the page and the effortless shift between view and edit mode. The user identified some weaknesses which were easily improved in a subsequent version of COMEDI. For instance, the researcher missed an autosave function when navigating from edit to view mode. This has been implemented. Also, it proved confusing that the action of clicking on a title causes different things to happen depending on whether it is an element title (sending the user to the CLARIN Concept Registry Browser (CCR) page documenting that element) or a component title (shrinking or enlarging the component by clicking on it). To avoid this confusion, the

[18] For information on the CLARIN license classification, see `http://www.clarin.eu/content/license-categories`.

link to the CCR is instead now available as an explicit link next to the title. As the COMEDI user base is currently expanding, we are responding to user feedback to continuously improve the usability.

3.8 Changes to profiles

A particular feedback which we received referred to rare cases when nodes were not processed. While it is always recommended to create metadata based on a stable CMDI profile, it may happen in certain circumstances that a profile needs to be changed, and consequently metadata may lose elements or components upon importing. Such changes are hopefully rare, but should not be overlooked. The consequences of such changes may be very complex and unforseeable in scope, so that automatic adaptation would be intricate. The strategy adopted in COMEDI for coping with such situations is therefore limited to detecting when a change in a profile has occured, thus leaving it up to the user to fix the metadata. Concretely, COMEDI shows an warning if nodes have not been processed, as exemplified in Figure 15.

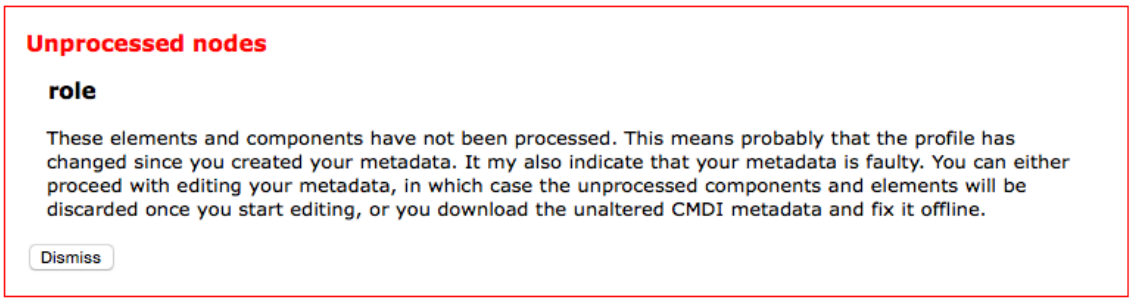


Figure 15: Warning that nodes have not been processed when a profile has been changed.

3.9 Implementation

COMEDI is written in Common Lisp, like the other advanced tools in the emerging CLARINO centre. The advantages of Common Lisp are, among other things, a very high level, extensible language allowing rapid development and seamless integration of components, since even the web server is written in the same language. The metadata are stored in a relational database.

Central to all operations performed on metadata in COMEDI are the CMDI profiles. When needed, a profile is fetched by COMEDI from the Component Registry in the XML-based CMDI Component Specification Language (CCSL). Schema descriptions of metadata are derived from the CCSL format.

The main idea in the implementation of COMEDI is to keep the profile, as a description of possible metadata, tightly connected to the metadata as a valid instantiation of the profile. In concrete terms, both the profile and the (complete or emerging) metadata are aspects of the same in-memory tree representation.

To start with, the profile (in CCSL format) is parsed into a DOM tree using a DOM parser. The DOM parser is however modified in such a way that it adds specific wrapper nodes around the CMD_Component and CMD_Element nodes of the profile: first, each CMD_Component is wrapped into a CMD_Component_Wrapper. Its cardinality attribute is set to 0 or 1 (or even higher), in accordance with the value of the component's CardinalityMin. Then, each CMD_Element is wrapped into a CMD_Element_Wrapper. If the element's CardinalityMin is 1, a CMD_Element_Realization is appended as a child node of the wrapper after the CMD_Element node.

This is the state when the metadata is still empty, or, more precisely, minimal. All components that have to be instantiated have been given cardinality 1 in the wrapper, and all elements that have to be instantiated are represented as a CMD_Element_Realization, but their values are empty.

Editing of the metadata is reflected in changes in the DOM tree: When an instantiation of an element is added in the editor, a new CMD_Element_Realization node is created. When a component is added in the editor, the wrapper's cardinality is increased, and unless the new cardinality equals 1, the CMD_Component node itself is cloned and added as a new child node to the wrapper. Editing of values

simply results in changing the element realization's value attribute. New values are immediately validated against the CMD_Element. Basically the same operations are executed when existing metadata is read.

Starting from the unified DOM tree representation of both the profile and the metadata, the HTML and Javascript/AJAX code of the editor can be generated in a quite straightforward way: The DOM tree is serialized to XML, and appropriate XSL and CSS stylesheets create the HTML code. Since all component and element information of the profile is available in the XML, the stylesheets can create the necessary buttons and input elements to manipulate the metadata, and specifically, will only create those buttons and elements that are in accordance with the profile and result in admissible manipulations.

4 Conclusion and future work

Since CMDI has been adopted as a metadata standard across the whole CLARIN infrastructure, good support for handling this fairly new metadata format may be of great importance. We presume that our work on COMEDI fills a current gap and hope it will be useful to researchers throughout the whole CLARIN infrastructure.

As we have seen in Section 2, previously existing editors have their individual strengths, but they also exhibit weaknesses, in particular concerning user-friendliness and coverage of the full CMDI specification. The development of COMEDI is well motivated since it offers several advantages over existing editors, above all a clear but highly functional web interface abstracting away from technical details in an elegant manner while still keeping the internal structure of the metadata explicit, thus helping to produce metadata faster and more consistently (component cloning, validation of user input). It also features advanced navigation through keyboard shortcuts.

The editor is currently fully functional but will benefit from further testing and user feedback for continued development. Among other things, we foresee the need to improve the search possibilities in metadata. Whereas simple cloning has been implemented, a similar but distinct feature is the availability of instantiated components that can be pointed at from different places, in the sense of structure sharing. When such a component's content is changed in one place, the changes will be reflected in all metadata records referring to it. This feature, which is provided in the META-SHARE editor (cf. Section 2.3), is particularly useful for components describing person or institutional info and the like, where changes should be propagated *passim*. A problem with implementing this is how to assign editing rights in a safe way, to avoid uncontrolled overwriting of existing information that may be shared by several users.

The system can be installed in as many centers as desired, but since metadata creation and management does not require very large computing resources, a few installations may suffice to cover CLARIN-wide needs.

COMEDI is available in the public domain under a BSD license.

5 Acknowledgements

The research reported in this paper has received support from the Research Council of Norway through the CLARINO project.

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).

[Choukri et al.2012] Khalid Choukri, Victoria Arranz, Olivier Hamon, and Jungyeul Park. 2012. Using the international standard language resource number: Practical and technical aspects. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Dima et al.2012] Emanuel Dima, Christina Hoppermann, Erhard Hinrichs, Thorsten Trippel, and Claus Zinn. 2012. A metadata editor to support the description of linguistic resources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Lyse et al.2012] Gunn Inger Lyse, Carla Parra Escartín, and Koenraad De Smedt. 2012. Applying Current Metadata Initiatives: The META-NORD Experience. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR (Workshop at LREC'2012)*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 20–27.

[Piperidis et al.2014] Stelios Piperidis, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo Del Gratta, Bernardo Magnini, and Christian Girardi. 2014. Meta-share: One year after. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Soria et al.2012] Claudia Soria, Núria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, and Nicoletta Calzolari. 2012. The flarenet strategic language resource agenda. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Trippel et al.2014] Thorsten Trippel, Daan Broeder, Matej Durco, and Oddrun Ohren. 2014. Towards automatic quality assessment of component metadata. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Withers2012] Peter Withers. 2012. Metadata Management with Arbil. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR (Workshop at LREC'2012)*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) 21.-27. May, pages 72–76.

[Wittenburg et al.2010] Peter Wittenburg, Nuria Bel, Lars Borin, Gerhard Budin, Nicoletta Calzolari, Eva Hajicova, Kimmo Koskenniemi, Lothar Lemnitzer, Bente Maegaard, Maciej Piasecki, Jean-Marie Pierrel, Stelios Piperidis, Inguna Skadina, Dan Tufis, Remco van Veenendaal, Tamas Váradi, and Martin Wynne. 2010. Resource and service centres as the backbone for a sustainable service infrastructure. 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).

Spokes – a search and exploration service for conversational corpus data

Piotr Pęzik
University of Lodz
Corpus & Computational Linguistics Laboratory
pezik@uni.lodz.pl

Abstract

Spokes is an online service for conversational corpus data search and exploration, currently developed as part of CLARIN-PL – the Polish CLARIN infrastructure. This paper describes the data sets currently available through Spokes, the architecture of the service and the data and metadata search functionality it provides to its users. We also introduce some of the more experimental features which have been developed to facilitate more advanced research on multimodal conversational corpora.

1 Introduction

Open-access speech corpora and speech databases are still rare and undersized, if at all available, for most languages. Many such resources have been made available simply as collections of annotated transcription and media files, which can be downloaded and processed by their prospective users, e.g. (Du Bois et al., 2000), (Coleman et al., 2012). For other corpora, dedicated, web-based tools have been developed, which make it possible not only to browse selected transcriptions and play the associated recordings, but also to search and retrieve text spans matching corpus queries, cf. (Johannessen et al., 2009), (Gasch, 2010), (Freitas and Santos, 2008). These tools vary widely with respect to their general functionality, query syntax and the range of supported export formats. For example, while some online tools only support basic searching for exact strings occurring in the transcriptions (Douglas, 2003), others make it possible to directly search and display time-aligned phonetic transcriptions[1].

The PELCRA Conversational Corpus (PELCRA CC) contains over 2.2 million words of casual Polish spoken data collected since 1999 in a number of research projects (Waliński and Pęzik, 2007), (Pęzik, 2012). The most recent set of samples was acquired and added to the corpus in the CLARIN-PL project. In contrast to other speech databases and spoken corpora available for Polish, the PELCRA CC includes mostly transcriptions of *in vivo* recordings of casual conversations, many of which were taken surreptitiously in everyday situations by trained acquisition agents. Although prior and ex post facto permissions were granted by the recorded speakers to process and distribute the transcriptions for research purposes, many of the speakers did not realise their conversations were being recorded at the exact time of recording[2]. This in turn makes this corpus particularly useful for casual spoken discourse studies as well as for the development of formal models of casual Polish speech (Pęzik, 2012).

Although parts of the PELCRA CC corpus have previously been released in raw source formats under open-source licenses, its full research potential has remained dormant for many potential users such as linguists and spoken discourse analysts from domains other than linguistics. This was mainly due to the technical difficulties related to exploring large quantities of casual conversational data. Many researchers simply lack the technical expertise needed to process XML-encoded transcription files in order to extract relevant samples of texts. Also, due to their sheer size, the sound files available for the transcriptions have proved problematic for users who need to identify and analyse large sets of audio samples.

[1] See, for example, the Spock system (http://spock.iltec.pt/) developed for the CORP-ORAL corpus (Freitas and Santos, 2008).

[2] By contrast, the term *in vitro* speech corpora can be used to describe corpora which contain mainly data from scheduled interviews arranged specifically for the purpose of corpus acquisition

To address the need for a centralized, user-friendly tool which would make this data more useful and more available to both technical and non-technical users, we have developed Spokes – a web-based service providing search and analysis functionality with GUI and programmatic access. Once the first version of Spokes for Polish conversational data was released, we proceeded to develop an experimental version of the service for the spoken component of the British National Corpus (BNC) in order to test those search features which require phone-level time-alignment of spoken data. Both of these versions of Spokes are discussed in this paper.

2 Annotation

As mentioned above, Spokes has so far been used for two PELCRA CC and spoken BNC data. The general nature of speaker and conversation metadata as well as the linguistic annotations is similar in both corpora and they are stored and searched using similar backend models. From the point of view of Spokes development, the most important difference between these two corpora is related to the level of phonetic annotation available as it has significant implications on how the data is searched and accessed.

2.1 PELCRA CC

The original PELCRA CC recordings were transcribed orthographically, anonymized and aligned manually at the level of utterances with ELAN (Wittenburg et al., 2006). In addition to basic demographic metadata about the conversations (such as place of recording, date, register) and speakers (such as age, sex, education), the transcriptions were automatically part-of-speech tagged and lemmatized. Using the manual time-alignments, it was also possible to extract the corresponding fragments of the recordings, process them and add pitch pattern annotation to individual utterance spans. The pitch properties for the data shown in Listing 1 were extracted with Praat (Boersma, 2002) and they include observed strength, intensity and frequency values for each time point.

Listing 1: Automatic pitch annotation in PELCRA CC.

```
<audio-segment id="Ekz6a">
   <pch s="0.778" i="0.171"
      t="0.230">164.648</pch>
   <pch s="0.899" i="0.150"
      t="0.240">164.273</pch>
   <pch s="0.915" i="0.135"
      t="0.250">164.214</pch>
  <pch s="0.936" i="0.176"
      t="0.260">163.977</pch>
  <pch s="0.960" i="0.199"
      t="0.270">163.405</pch>
  <pch s="0.934" i="0.203"
      t="0.280">161.971</pch>
  ...
</audio-segment>
```

Since the quality of some of the audio recordings is poor due to the conditions in which they were taken, each conversation was additionally rated and annotated for its overall acoustic quality. The ELAN-annotated transcriptions are transferred to a relational database for management and further processing. The audio recordings are stored in wav files, which are currently approximately 69 gigabytes in size, and accessed on demand by a file retrieval mechanism which is separate from the two other backends.

2.2 Spoken BNC

The spoken component of the BNC contains both text metadata such as source, text types and classification codes as well as speaker metadata including sex, social class, occupation and dialect codes. As a

result of a joint project of the British Library Sound Archive and the Oxford University Phonetics Laboratory, most of the original recordings from the spoken component of the BNC were recently digitized from cassette tapes and made available with the time-aligned transcriptions (Coleman et al., 2012). In order to test the flexibility of the solution described in this paper we have transferred this release of spoken BNC data to a separate instance of Spokes. In contrast to the PELCRA CC data, which is only manually time-aligned at the level of utterances, the alignment available in the BNC relates individual phone units to time offsets. The phonetic transcription is available in the SAMPA (Wells and others, 1997) and IPA alphabets. Needless to say, this level of alignment opens up the possibility of supporting more sophisticated phonetic queries against this data, some of which are discussed below.

3 Architecture

A basic overview of the current Spokes architecture is presented in Fig. 1. The three main tiers of Spokes are the search and storage backends, the REST API service and the Web application. We believe that this separation has several advantages. First of all, all these three modules are separated and they can be distributed and hosted independently. Secondly, access to the backend modules is always mediated by the REST API which means that the backends do not need to be directly exposed to third party users. Finally, because the dedicated Spokes web application uses the same REST API service which is available for other (programmatic) clients, the web application developers were the first users to thoroughly test the API. In the process of developing the web application, a number of new API methods of serving and accessing the data were developed and made publicly available.

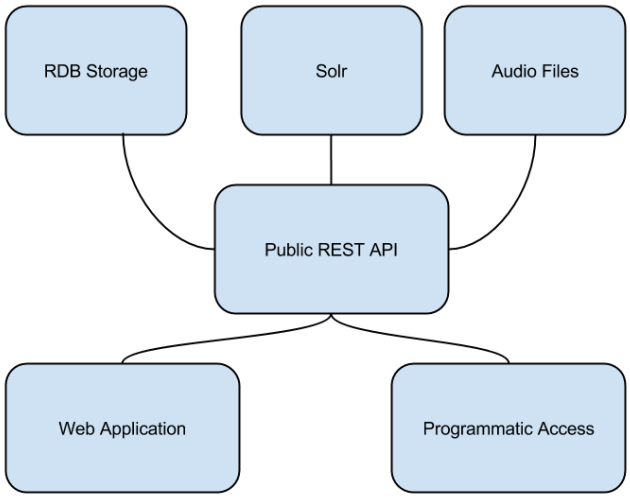


Figure 1: An overview of the Spokes architecture.

3.1 Backends

The three backend modules used in Spokes serve distinct purposes. The main function of the relational database backend[3] is to store and manage the data in a highly normalized model. It is also used for relatively standard data retrieval operations which are well supported in the SQL syntax, such as joins and aggregations. The Solr backend is used to provide very fast full text search capabilities and aggregated views of data matched by corpus queries. The main Solr core used in Spokes contains a flat index of all

[3]We are currently using PostgreSQL 9.2 for the RDB backend.

utterances, which in Solr parlance, are called "documents". In some cases, separate backend modules are used in tandem for different phases of a complex query. For example, while Solr is always used to fetch concordance spans, the RDB backend is used to fetch pitch data for the matching utterances which are then visualised in the front-end web application (see Fig. 5)[4] .

3.1.1 RDB

The XML-encoded datasets were transferred and normalized in the the RDB backend. For example, in the BNC data model, utterances, word tokens and phone segments are stored in separate tables in order to represent the information available in the original annotation. As a result, with a relatively simple SQL query, word and utterance durations can be aggregated from the durations of their constituent phones. The RDB storage has proved to be highly suitable for data management tasks such as batch updates and versioning or validation routines including detecting duplicated, missing and erroneous data values which are less straightforward to detect in the original XML-format of the transcriptions. Additionally, the RDB backend is used to directly support those search functions of Spokes which involve complex metadata aggregation and joining.

3.1.2 Solr

Apache Solr, which is based on Apache Lucene, is a general-purpose search technology known for its maturity, speed, scalability and advanced full-text search capabilities. As explained below, with some modifications to its standard index structure and query syntax, Solr can also be used to support positional queries on part-of-speech tagged corpus data. By customizing this technology, we took advantage of its highly performant "faceting" functionality which opens up interesting possibilities of query-based corpus metadata aggregation.

3.2 REST API

In addition to the Spokes web-application specific methods, the REST API exposes full metadata and partial data access methods. All of its methods and resources are documented using mashery/iodocs [5], which makes it convenient for API users to learn and test them interactively.

4 Search and exploration

4.1 Metadata browsing

Basic data retrieval and metadata browsing are the most obvious functionalities of Spokes. As shown in Fig. 2, users can browse and filter full transcription metadata. The PELCRA CC transcriptions can also be viewed, played, and downloaded with full utterance-level metadata about the individual speakers who took part in the conversations. Similar browsers are available for word frequency and formulaic sequence lists extracted from the two corpora.

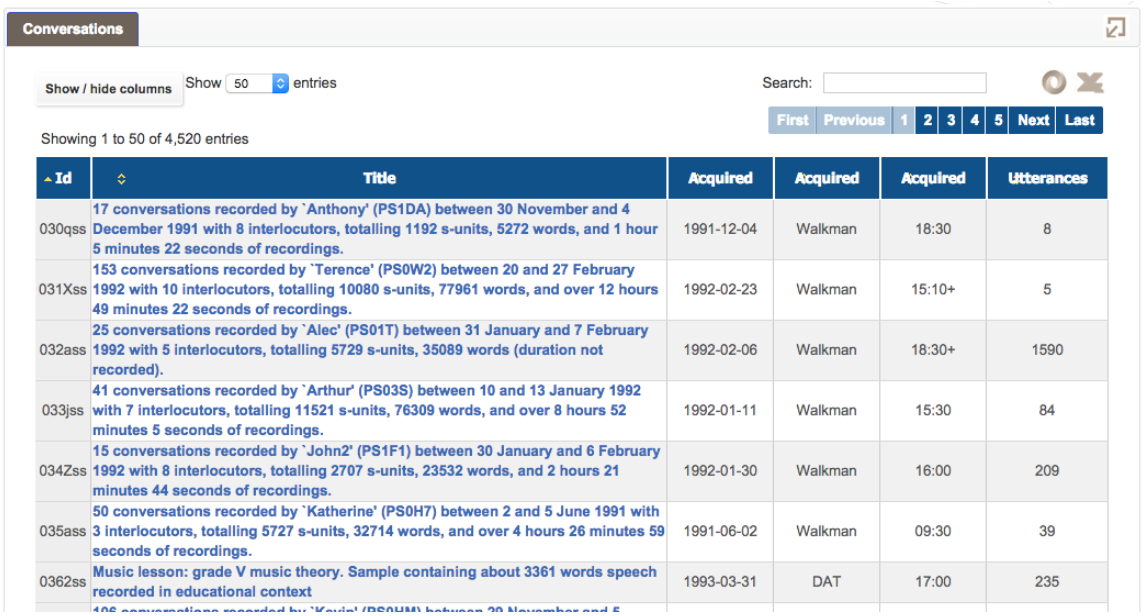
4.2 Corpus query syntax

In order to support positional concordance queries for annotated token sequences, we developed a dedicated text analyser and a query syntax for the Solr backend. The query syntax, called SlopeQ, is illustrated in Table 1. Apart from simple queries for surface and lemmatised terms, it supports regular expressions, part-of-speech terms, variant negation and slop-factor (proximity) operators.

The two proximity operators supported by Spokes can be particularly useful when searching for spans of word tokens which are often discontinued by discourse markers and hesitation devices in casual conversational data. For example, the query `(an|a way of <pos=v.+>)=2` will match sequences such as *Yeah, I mean there could be **a way of sort of coming** together.* It should be noted that regular expressions are only matched against single tokens which are listed in the inverted index of the corpus. Although it is possible to specify that a token matching a regular expression such as `<lemma=.*>` is required at a certain position, it is more convenient to use the slop factor syntax in order to match loosely

[4] All of the visualisations provided by Spokes can be downloaded in bitmap (PNG, JPEG) and vector (PDF and SVG) formats.

[5] See https://github.com/mashery/iodocs.



Id	Title	Acquired	Acquired	Acquired	Utterances
030qss	17 conversations recorded by `Anthony' (PS1DA) between 30 November and 4 December 1991 with 8 interlocutors, totalling 1192 s-units, 5272 words, and 1 hour 5 minutes 22 seconds of recordings.	1991-12-04	Walkman	18:30	8
031Xss	153 conversations recorded by `Terence' (PS0W2) between 20 and 27 February 1992 with 10 interlocutors, totalling 10080 s-units, 77961 words, and over 12 hours 49 minutes 22 seconds of recordings.	1992-02-23	Walkman	15:10+	5
032ass	25 conversations recorded by `Alec' (PS01T) between 31 January and 7 February 1992 with 5 interlocutors, totalling 5729 s-units, 35089 words (duration not recorded).	1992-02-06	Walkman	18:30+	1590
033jss	41 conversations recorded by `Arthur' (PS03S) between 10 and 13 January 1992 with 7 interlocutors, totalling 11521 s-units, 76309 words, and over 8 hours 52 minutes 5 seconds of recordings.	1992-01-11	Walkman	15:30	84
034Zss	15 conversations recorded by `John2' (PS1F1) between 30 January and 6 February 1992 with 8 interlocutors, totalling 2707 s-units, 23532 words, and 2 hours 21 minutes 44 seconds of recordings.	1992-01-30	Walkman	16:00	209
035ass	50 conversations recorded by `Katherine' (PS0H7) between 2 and 5 June 1991 with 3 interlocutors, totalling 5727 s-units, 32714 words, and over 4 hours 26 minutes 59 seconds of recordings.	1991-06-02	Walkman	09:30	39
0362ss	Music lesson: grade V music theory. Sample containing about 3361 words speech recorded in educational context	1993-03-31	DAT	17:00	235

Figure 2: A transcription metadata browser.

defined phrases which contain two or more obligatory tokens. Additionally, the slop factor queries are noticeably faster than their regular expression equivalents in large corpora.

4.3 Metadata query syntax

Spokes allows users to run metadata queries which are formulated in the Solr Extended DisMax syntax[8]. Currently these queries are always appended as filters to an obligatory concordance query. For example, the following DisMax filter: `speaker_age:[0 TO 10]` can be appended to the SlopeQ query "mamo" (Pol. "mom" in the vocative case) in order to make sure that only concordances of this word found in utterances of speakers up to 10 years old are returned. Some of these filtering criteria can also be set using the graphical controls of the web application user interface.

4.4 Concordance grouping

Another search option of Spokes based on Solr is the ability to group concordances matched by a SlopeQ query by one or more metadata field values. For example, it is possible to define the maximum number of concordance results per speaker or text identifier. This in turn may serve as a basic way of sampling the results of queries which are likely to match many spans per conversation. The range of the matching concordances can be specified as well. It is thus possible to sample and group results from different sections of the corpus. It is also possible to implement hierarchical grouping of concordances, which would make it possible to specify the maximum size of samples matching a conjunction of metadata field values. For example, the maximum number of utterances with a unique combination of speaker and text identifiers or any other metadata field value stored in the Solr backend could be specified to be fetched in the concordances matching a query.

4.5 Facets

One of the Solr-based features of Spokes is its facet generation mechanism. For any concordance query, the Solr backend automatically runs a full aggregative query which collects counts of distinct metadata values found in all the matching documents. In other words, even if the user chooses to fetch, say, only 20 matching concordance spans from the index, a full report about the number of matching results found

[6]The pipe operator is always interpreted as a token boundary.
[7]The negated variant is marked with "!" and subtracted from the set of alternatives specified on the same position.
[8]See https://wiki.apache.org/solr/ExtendedDisMax.

#	Query	Returns text spans containing
1	mamo	A single surface token
2	wiesz co	A sequence of surface tokens
3	<lemma=palić>	All variants of a single lemma token
4	<lemma=mieć> <lemma=szansa>	A sequence of immediately adjacent lemma tokens
5	słuchaj <lemma=ja>	A sequence of surface and lemma tokens
6	tultutaj	Variants separated by the pipe operator[6]
7	<lemma=facet>|<lemma=koleś>	Lemma variants
8	bardzo|strasznie dużo	Surface token variants in a sequence
9	(ta kobita)=1	Sequence separated by zero or one unspecified tokens
10	(<lemma=jechać> tam)=2	A lemma and a surface token separated by up to 2 tokens
11	(<lemma=jechać> tam)∼2	As above except that the tokens may occur in any order
12	(<lemma=dać> do zrozumienia)=2	3 obligatory tokens separated by up to 2 unspecified ones
13	<lemma=prosić>|!proszę	Any form of *prosić* except for *proszę*[7]
14	t.* bab.*	Tokens matching regular expressions
15	szykow.+|przygotow.+	Variants with regular expressions
16	<lemma=p.+biec>	Lemmas matching a regular expression
17	<tag=subst:pl:.*>	Any plural noun
18	<tag=subst:.+:inst:.+>	Any noun in the instrumental case
19	<lemma=zdać pos=verb:sg:.*>	Singular forms of the verb “zdać”
20	<tag=adj:.+> <lemma=temat>	Sequences of adjectives preceding the noun “temat”
21	(<lemma=słuchać> <tag=.*:gen:.*>)=1	Lemma followed by any genitives with a slop factor

Table 1: SlopeQ 2 query syntax

in the different sections of the corpus is returned. Fig. 3 shows an example facet-based report for the concordance query *aye*. The available facets are listed in the left panel and they can be visualised in the middle panel as piechart graphs or bar plots. The one hundred most frequent values of each facet are also listed in the right panel, where they can be selected and deselected as filtering criteria.

4.6 Collocations

Another feature of Spokes which is aimed at helping users digest large sets of concordance results is the positional collocation extraction module[9]. The module can be used to aggregate a frequency list of the most frequent tokens co-occuring with the spans matched by any SlopeQ query. Fig. 4 shows an example concordance query for all inflections of the adjective *dobry* which has been transformed into a collocation query. Users need to specify the maximum number of contexts from which collocates will be extracted as well as the allowed part-of-speech tags and positions of potential collocates. The resulting list of frequent collocates contains 35 combinations which occurred four or more times in the concordance results. Each of the positionally defined collocates in this list is presented as a hyperlink to its full concordance. It is possible to extract potential collocates from sets of up to 100 000 matching spans in a single query.

4.7 Data export

Although we have tried to make the web application as powerful and easy to use as possible, it is nevertheless possible to envisage non-technical users who will still want to download thousands of results in order to further process and analyse them. For example, some researchers may want to filter and annotate all the instances of a specific linguistic feature which may be difficult to describe with the query syntax supported by Spokes. To address this need, we make it possible to download up to 100 000 results per

[9]The term “positional collocation extraction” is used here to refer to extraction methods which only rely on aggregating co-occurrences of words in a predefined window rather than explicitly encoded syntactic relations between them, cf. (Evert, 2004).

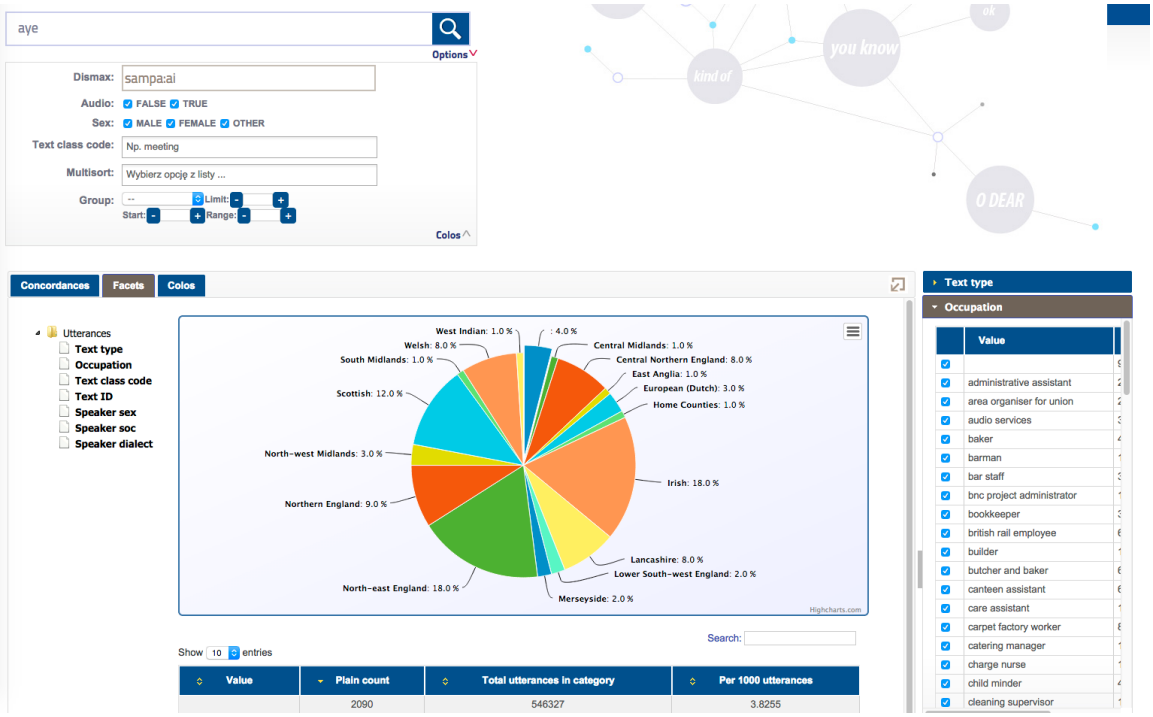


Figure 3: Interactive search facets.

single query in the form of an MS Excel spreadsheet (which can also be processed in OpenOffice). Apart from metadata-annotated listings of concordance lines, such spreadsheets also contain complete sets of facets extracted for a given query.

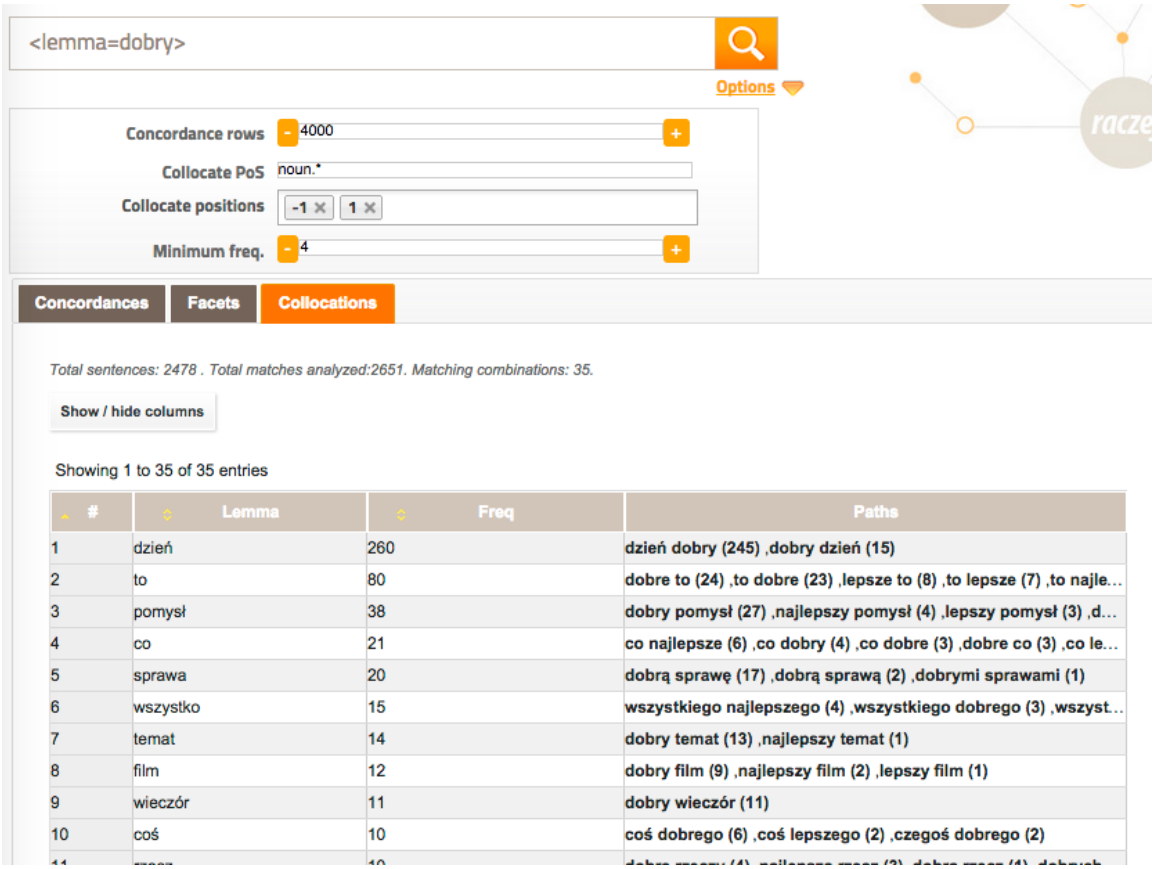
4.8 Searching phonetic annotation

The spoken BNC data indexed in Spokes can be searched using SAMPA- and IPA-encoded query terms which correspond to the phonetic transcriptions of word tokens. Such queries can be combined by means of logical operators with conditions specified for other index fields. For example, the SlopeQ query `row` can be combined with the Sampa query `r@U` or the equivalent IPA query rəʊ to return spans where the corresponding pronunciation of the word *row* was recognized.

Users can display pitch annotations (such as the f0 values shown in Listing 1) for any concordance query, which returns spans aligned with the time offsets of the utterances. This is illustrated in Fig. 5, which shows a pitch contour for an utterance matching a concordance query. Thanks to the availability of phone and word level time alignment in the spoken BNC data, it is possible to generate similar contours for the exact spans (and not just utterances) matching concordance queries. This functionality is also an example of how the separate backend modules are combined to serve different types of data.

5 Experimental features

One of the experimental features of Spokes, which may be particularly useful in spoken discourse analysis makes it possible to carry out on-the-fly analyses of the duration of spans matching users' corpus queries. For example, users who type in a corpus query matching the word "right", which happens to be both polysemous as a lexical item and multifunctional as a discourse marking device in conversational English. may want to order the resulting concordances by their observed duration whenever such time-alignment is available for a transcribed span. The purpose of generating such a ranking would be to check whether certain meanings or functions of "right" are marked by longer or shorter durations. For instance, one working hypothesis here could be that instances of "right" as a turn-opening discourse marker may be characterized by significantly higher average or median duration values than instances of "right" as an adjective pre-modifing heads of noun phrases.



#	Lemma	Freq	Paths
1	dzień	260	dzień dobry (245) ,dobry dzień (15)
2	to	80	dobre to (24) ,to dobre (23) ,lepsze to (8) ,to lepsze (7) ,to najle...
3	pomysł	38	dobry pomysł (27) ,najlepszy pomysł (4) ,lepszy pomysł (3) ,d...
4	co	21	co najlepsze (6) ,co dobry (4) ,co dobre (3) ,dobre co (3) ,co le...
5	sprawa	20	dobrą sprawę (17) ,dobrą sprawą (2) ,dobrymi sprawami (1)
6	wszystko	15	wszystkiego najlepszego (4) ,wszystkiego dobrego (3) ,wszyst...
7	temat	14	dobry temat (13) ,najlepszy temat (1)
8	film	12	dobry film (9) ,najlepszy film (2) ,lepszy film (1)
9	wieczór	11	dobry wieczór (11)
10	coś	10	coś dobrego (6) ,coś lepszego (2) ,czegoś dobrego (2)

Figure 4: Collocation extraction in Spokes.

This type of analysis is made possible by combining information available in two backend modules. Table 2 shows a simplified representation of word token values stored in the RDB backend. The last column in this view shows the duration of a given word derived from the relative offsets shown in columns 3 and 4 extracted from the original BNC transcriptions. The lists of concordances retrieved from the Solr index for any corpus query contains the same identifiers of the word tokens occurring in the matching spans. It is therefore possible to use those identifiers to join the database records and sort or pass them to an aggregate function.

The result of this operation is not only a duration-sorted list of concordances, but also a summary of descriptive statistics for the sample of concordance spans retrieved. For example, Fig. 6 shows a standard box plot for the 6433 time-aligned instances of the word “right” found in the BNC data indexed by Spokes. The median duration of “right” is 190 ms with a mean of 216.4 ms and standard deviation of 132.8. The box plot reveals number of outliers with the maximum value of 2.2 seconds, which could be rather long but genuine instances of the word or simply misalignments. A similar analysis can be carried out for any concordance results matching multiword unit spans which can be specified in the SlopeQ syntax.

6 Challenges and planned developments

As described above, the analysis of the distribution of duration as a prosodic feature is fairly straightforward to implement. A much more challenging extension of Spokes which we are currently developing is motivated by the need to enable automatic identification and classification of pitch patterns of concordance spans. There is a considerable amount of research into recurrent discourse-functional lexical sequences which seem to exhibit regular prosodic characteristics such as “stereotyped” intonation contours (Bolinger, 1986). One hypothesis formulated in such studies is that “intonation conveys information about the intentional as well as the attentional structure of discourse” (Hirschberg and Pier-

Figure 5: Pitch annotations for utterances and matched concordance spans.

Id	Word	Start	Stop	Duration
45799700	right	538673	539013	340
46153393	right	811603	811733	130
49168719	right	1999793	2000043	250
46154674	right	2296643	2296823	180
45802240	right	2651043	2651133	90
47388854	right	372829	373159	330
46155247	right	2306773	2306863	90
47388909	right	375643	375733	90
45802580	right	2659473	2659613	140
45803157	right	195935	196295	360

Table 2: Time offsets for retrieved concordance spans.

rehumbert, 1986). For example, one of the functions of “hello” is to express surprise or irritation at what has just been said or done. This function, as opposed to its “greeting” function, seems to be prosodically stereotyped in that it is marked by a rising pitch contour of this word.

We are currently investigating the possibility of using time-aligned conversational data to facilitate such analysis in Spokes. As already mentioned, pitch annotations for every utterance indexed in the Solr backend are stored and can be retrieved from the RDB backend. In the case of the BNC data, they can also be mapped to word tokens matched in concordance spans. This in turn makes it possible to extract and analyze thousands of automatically recognized pitch patterns for any concordance query. Such results could simply be presented to the user as shown in Fig. 5 for further inspection. Users of Praat may actually prefer to download the audio snippet provided by Spokes and perform their analysis offline.

At the same time, it may also be possible to use different methods of automatic detection and categorization of prosodic events (**?**) to provide an automatic classification of pitch contours observed in the concordance spans retrieved for a query. The technique we have experimented with so far involves producing a distance matrix based on dynamic time warping similarity values (cf. (Müller, 2007), which is computed for all pairs of pitch contours observed in the concordance spans. The resulting distance matrix is then used to produce a dendrogram showing clusters of concordance spans with “similar” pitch contours.

This feature of Spokes is still highly experimental and it requires careful validation. The choice of a reliable measure of comparing and clustering highly volatile prosodic signals remains a challenge. Also,

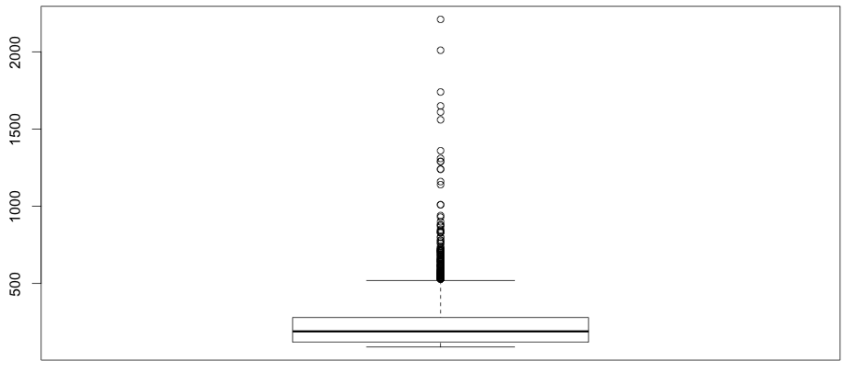


Figure 6: A box plot generated for the 6433 time-aligned instances of “right” found in the spoken BNC data.

the poor audio quality of many of the original recordings also makes it difficult to compare pitch contours for different instances of the same word forms.

7 Availability

Current versions of the Spokes for PELCRA CC and Spokes for BNC web application services are publicly available at `http://spokes.clarin-pl.eu`[10] and `http://pelcra.clarin-pl.eu/SpokesBNC`. The help pages of these applications provide up-to-date links to the REST API. The entire Spokes data will also be available through a Federated Content Search endpoint as part of the CLARIN-PL resource center.

Acknowledgements

The work described in this paper has been financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education. The relational database backend and the REST service was implemented mainly by Łukasz Dróżdż. Paweł Wilk and Paweł Kowalczyk are the main developers of the Web application. Finally, we extend thanks to Danijel Koržinek for sharing his expertise in developing speech signal analysis modules used in Spokes.

References

[Boersma2002] Paul Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

[Bolinger1986] Dwight Bolinger. 1986. *Intonation and its parts: Melody in spoken English*. Stanford University Press.

[Coleman et al.2012] John Coleman, Ladan Baghai-Ravary, John Pybus, and Sergio Grau. 2012. Audio BNC: the audio edition of the Spoken British National Corpus.

[Douglas2003] Fiona M Douglas. 2003. The scottish corpus of texts and speech: Problems of corpus design. *Literary and linguistic computing*, 18(1):23–37.

[Du Bois et al.2000] John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000. Santa Barbara corpus of spoken American English.

[10] See also `http://hdl.handle.net/11321/47`.

[Evert2004] Stefan Evert. 2004. *The statistics of word cooccurrences*. Ph.D. thesis, PhD Dissertation, Stuttgart University.

[Freitas and Santos2008] Tiago Freitas and Fabíola Santos. 2008. Corp-oral: Spontaneous speech corpus for european portuguese. In *Proceedings of LREC*.

[Gasch2010] Joachim Gasch. 2010. Dgd 2.0: A web-based navigation platform for the visualization, presentation and retrieval of german speech corpora. *Sprache und Datenverarbeitung*, 34(1):27–38.

[Hirschberg and Pierrehumbert1986] Julia Hirschberg and Janet Pierrehumbert. 1986. The intonational structuring of discourse. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, pages 136–144. Association for Computational Linguistics.

[Johannessen et al.2009] Janne Bondi Johannessen, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The nordic dialect corpus-an advanced research tool. In *Proceedings of the 17th Nordic conference of computational linguistics NODALIDA 2009. NEALT proceedings series*, volume 4, pages 73–80.

[Müller2007] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

[Pęzik2012] Piotr Pęzik. 2012. Język mówiony w NKJP. In Adam Przepiórkowski, Mirosław Bańko, Rafał Górski, and Barbara Lewandowska-Tomaszczyk, editors, *Narodowy Korpus Języka Polskiego*, pages 37–47. Wydawnictwo Naukowe PWN, Warszawa.

[Waliński and Pęzik2007] Jacek Waliński and Piotr Pęzik. 2007. Web access interface to the PELCRA referential corpus of polish. pages 65–86. Lang.

[Wells and others1997] John C Wells et al. 1997. Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.

[Wittenburg et al.2006] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006.

CLARIN Conferences until 2014

Year	Place	From	To
2012	Sofia, Bulgaria	24 October	27 October
2013	Prague, Czech Republic	21 October	22 October
2014	Soesterberg, the Netherlands	23 October	25 October