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.