The International Journal of Digital Curation
Volume 8, Issue 2 | 2013

The CLARIN-NL Data Curation Service: Bringing Data to the Foreground

Nelleke Oostdijk, Henk van den Heuvel and Maaske Treurniet,
Radboud University Nijmegen

Abstract

After decades in which a great deal of effort was spent on the creation of resources, there are currently several initiatives worldwide that aim to create an interoperable, sustainable research infrastructure. An integral part of such an infrastructure constitutes the resources (data and tools) which researchers in the various disciplines employ. Whether the infrastructure will be successful in supporting the needs of the research communities it intends to cater for depends on a number of factors. One factor is that resources that are or could be relevant to the wider research community are made visible through this infrastructure and, to the greatest extent possible, accessible and usable. In practice, the durable availability of resources is often not properly regulated within research projects.

CLARIN-NL is directed at creating an interoperable language resources infrastructure for the humanities in the Netherlands. The Data Curation Service was established in order to salvage language resources in this field that are threatened to be lost. In the CLARIN context, a great deal of attention is given to standards, formats and intellectual property rights. Consequently, the Data Curation Service (DCS) has a role as mediator in bringing researchers in the field of humanities and existing data centres closer together.

This article consists of two parts: the first part provides the background to the work of the DCS while the second part illustrates the work of the DCS by describing the actual curation of a collection of language learner data.

Introduction

CLARIN-NL[1] is a programme funded by the Dutch government, which should contribute to the development of a sustainable research infrastructure for the humanities and linguistics in particular (Odijk, 2010). CLARIN-NL is carried out jointly by technical specialists, technology providers and researchers. An effort is made to involve the intended users through various application projects, in which local repositories are integrated and local services set up for prototypical test installations as initial demonstrators. This enables evidence-based contributions to the discussion on standards and best practices for interoperability, and allows users to contribute to the survey of requirements for the infrastructure technology.[2]

From the start of the programme in 2009, CLARIN-NL funding has been available for projects directed at resource curation. These projects have been targeted at rescuing valuable data and educating researchers in using standard formats (see also Thieberger, 2012; Neuroth, Lohmeier and Smith, 2011). As the calls for proposals were less successful in reaching resource producers and owners who were not already aware of and/or participating in CLARIN-NL, the CLARIN-NL Executive Board initiated a pilot project in October 2010 that should investigate the need and possibility for establishing a Data Curation Service (DCS) task force. The idea was that a dedicated team of specialists should be made responsible for curating data residing with humanities researchers, especially those who are reluctant or incapable of undertaking the curation themselves. The data would subsequently be made available to the CLARIN community through one of the CLARIN-NL Data Centres (Odijk, 2010).

The pilot project was carried out between 1 November 2010 and 1 February 2011. The aim was to establish whether there was a sufficient basis to assume that such a service would meet with demand in the field. Moreover, the pilot was intended to develop ideas about the form such a service should take, and also the effort and expertise required. The main findings obtained in the pilot project were summarized in a report (Oostdijk, 2011). As it was found that there was indeed a need and a basis for a DCS task force, in September 2011 CLARIN-NL decided to establish the DCS at the Centre for Language and Speech Technology (CLST) in Nijmegen. The DCS has been operational since January 2012.

The CLARIN-NL DCS

The DCS aims to contribute to the research infrastructure that CLARIN is implementing by salvaging resources and advising on best practices and the use of standards. Set up as a service, the DCS maintains close contact with the research communities and acts as a mediator between these and the CLARIN-NL Data Centres. The DCS prepares resources for archiving at the Data Centres, but does not archive any resources itself. Accordingly, the tasks of the DCS are defined as follows:

[1] CLARIN-NL: http://www.clarin.nl

[2] Cf. CLARIN-NL Long Term Programme 2009-2014 (Odijk, 2009). HYPERLINK \l "Odijk_2009"

 doi:10.2218/ijdc.v8i2.278

1. Curation of resources, especially those presently held by individual researchers or research groups;
2. Assisting in the curation efforts of CLARIN-NL Data Centres, if and when required; and
3. Advising researchers who wish to undertake the curation of their resources themselves.

The curation of resources held by individual researchers or research groups form the core of the work undertaken by the DCS. As the DCS receives funding from CLARIN-NL, efforts are directed primarily at language resources stored and used in the Netherlands. The final decision to curate a resource is made by CLARIN-NL's Executive Board, based on a proposal by the DCS.

In the context of the DCS, data curation entails making the data visible (via CMDI metadata), accessible (via a CLARIN centre), uniquely and stably referable (by persistent identifiers), and suited for interoperability with tools and other data, both formally (via standardized data formats) and semantically (via links to data category registries). However, these activities are embedded in a larger context of resource identification and assessment.

The tasks and actions involved in the curation of resources are summarized in Table 1.

Task	Actions
A: Identification and assessment	1. Identify candidate resources; collect information as to: a. The owner/producer b. The type of resource c. The licensing restrictions/conditions d. The size e. The format(s) f. The metadata available g. The nature of enrichment/annotations 2. Assess the desirability of curation 3. Assess the feasibility of successful curation
B: Development of a curation plan	4. Evaluate the content objects and determine: a. What type and degree of format conversion or other preservation actions should be applied b. The appropriate metadata needed for each object type and how it is associated with the objects 5. Estimate cost and lead time 6. Arrange for the necessary expertise to be available
C: Curation	7. Digitize data (limited scale) 8. Convert to a CLARIN preferred format 9. Assign appropriate metadata 10. Ensure semantic interoperability 11. Provide documentation

Task	Actions
D: Validation	12. Validate curated resource
E: Archiving	13. Transfer to CLARIN Data Centre for long-term storage and maintenance
	14. Assign persistent identifier(s)
	15. Provide access to content

Table 1. Tasks and actions in data curation by the DCS.

The tasks and actions show a clear correspondence to those presented as part of the DCC Lifecycle; however, in the context of the DCS, more emphasis is put on the identification and assessment of candidate resources before any curation is carried out.

The identification and assessment of candidate resources may require a great deal of effort, both in terms of the time and the persistence needed for tracking down the resource and whatever relevant information exists. It is a critical step in the curation process, as it should result in a go or no-go for moving ahead with the drawing up of a plan for actually curating the resource. The work undertaken as part of Task A should prevent money and effort going to waste in failed curation efforts. Task A is ideally carried out in close collaboration with the resource owner/producer.

The assessment of a candidate resource considered for curation concerns two aspects:

1. Whether it is desirable to have the resource curated, and
2. Whether indeed successful curation is feasible.

Each of these aspects is discussed in more detailed below. Whether it is desirable to curate a resource (Action A2) is not a question that can be answered straightforwardly, as various factors need to be considered, including:

Relevance to research community. As CLARIN-NL is directed at researchers in the humanities and social sciences, the infrastructure should incorporate the resources that are relevant to these research communities. As the field of Dutch language and speech technology is already very well organized and many resources are available through the HLT Agency[3], the curation of resources of interest to other areas is found to be relatively more urgent. Therefore, in the calls for proposals some priority areas have been identified to solicit project proposals targeted at literary studies, history and political studies, communication and media studies, first and second language acquisition, and historical linguistics.

Uniqueness. It may be argued that priority should be given to resources that are unique in their sort. The extent to which a resource bears resemblance to resources already available should be established, along with the characteristics that set it apart. Only then is there is basis for deciding whether it is interesting enough to be curated. What became apparent with the initial inventory of

[3] HLT-agency (in Dutch *TST-Centrale*): http://tst-centrale.org/

potentially interesting resources is that some resources go under different names, while others that go under the same name are in fact different resources, or at least different versions of a resource. An example is the Eindhoven corpus, which also goes by the name of Corpus Uit den Boogaart, and for which there appear to be different versions (e.g. a Meertens version, a Groningen version, and the HLT Agency's Eindhoven Corpus VU version) without it being clear if, or how exactly, these versions differ.

Urgency. The urgency to curate a resource may arise for a variety of reasons. It may be that the people responsible for the resources are about to disappear or have already disappeared, such as researchers who have completed their PhD research and moved elsewhere, those that have retired or those about to retire. With their departure the risk of data loss is very real. Even when the data can be traced successfully, the knowledge needed to curate them successfully (e.g. knowledge of the content, but also intellectual property rights-related matters) may be lacking. Another cause for urgency may be the limited life of magnetic and optical media, and the fact that the software and devices needed to retrieve the recorded information are disappearing as they are being replaced. Finally, in the context of specific research, certain resources are particularly welcomed if they fill remaining gaps.

Reproducibility of the Resource. When considering the reproducibility of a resource, the first question to be addressed is whether the resource contains primary data (i.e. the original texts, images or recordings), transcriptions, annotations and other forms of enrichment of the primary data, or derived data (e.g. a frequency list or a concordance).

Primary data may be any of a wide range of materials, including data that were collected during field work, while conducting a survey among speakers of a particular language or dialect (including questionnaires and interviews), or while running an experiment in the laboratory (including stimuli), but also a corpus of texts, a grammar or a lexicon that has been compiled. Primary data cannot usually be reproduced, or if they can, reproduction requires an excessive effort. Primary data therefore have high priority.[4]

With transcriptions and annotations, a distinction should be made between enrichments that were obtained either manually/semi-automatically and those that are the result of a fully automatic process. In the first case, recreating these enrichments will appear not be trivial, while at the same time it is unlikely that an identical result can be obtained. Such data should, therefore, be curated. In the case of automatically produced enrichments, these on principle could be reproduced when required, assuming that the tool(s) that is/are needed to do so is/are indeed available.[5] A strong argument in favour of curating the enrichments nevertheless, even when the tools are available, is that the resource

[4] Excepted are data sets that consist of data that have been collected more or less at random (i.e. without a priori formulated design criteria) from the internet and which have no particularly distinctive characteristics. While exact reproduction may not be possible, it can assumed that similar data sets can be produced, if so desired.

[5] In this case, the best strategy would be to consider curating both the data and the tools. However, the curation of tools is complex and serious questions have been raised as to whether it is worth the effort.

with the enrichments is readily usable, whereas users who are left to apply the tools themselves may find it beyond their capabilities to do so efficiently and/or successfully. Even users who do know how to handle the tools may appreciate not having to run complex and time-consuming processes.

Derived data are any kind of data that can be produced on the basis of (a subset of) a primary data set and/or its enrichments. Derived data are not usually to be considered as a prime target for curation, as concordances, frequency lists and such can be generated on demand. However, there may be occasions when the idea of curating derived data may be entertained and actually be given some follow up. This could be the case with derived data that come with a resource (e.g. the various frequency lists with the Spoken Dutch Corpus). It may well be that these data are particularly interesting in their own right for particular user groups (e.g. developers of teaching materials looking for a basic vocabulary list). Curation of derived data must also be considered for complex data sets where it is all but trivial to derive the data one is interested in (e.g. a list of the pronunciation variants of content words in Dutch as spoken by speakers originating from the Netherlands).

Whether a resource that is up for curation can indeed be successfully curated (Action A3) depends on:

The state of the resource. For any resource that is being considered for curation it must be established whether:

- **It can be made available to a wider audience**. Has the resource been cleared for IPR?[6] Should measures be taken to ensure anonymization?
- **It is in digital form**. Is digitization required?

Other questions in this context are:

- Is the resource in a state and form that can still be handled by current hardware and software?
- Is the integrity of the data as yet in tact?
- Upon curation, can the integrity of the data be warranted?
- Is it in a sound state qualitatively?

The availability of documentation. Documentation may take on many different forms. It includes format specifications and descriptions, protocols, annotation guidelines, but also descriptions of the experimental design, the set-up and the stimuli used. The availability of proper (technical and user) documentation is one of the preconditions for curation to be successful, while it is also essential for ensuring that users can use the resource to the full.

The availability of expert knowledge. Expert knowledge of a scientist or the original collector may be indispensable when curation of a resource is to be

[6] In case arrangements have yet to be made, a Creative Commons or similar license is preferred.

undertaken and the conversion of the original form to its projected form is not straightforward.

The availability of the necessary tools, scripts, etc. To the extent that specific tools are necessary for the curation of a resource, they should be available or it should be possible to develop them without disproportional effort.

After a candidate resource has been properly assessed, the next step in the curation process is to develop a curation plan. The plan should specify what actions are necessary to preserve the data and accompanying metadata. This may involve digitization and conversion to CLARIN preferred formats. From a very early stage in the curation process, the designated CLARIN Data Centre that will eventually store and maintain the curated resource is involved. Elements of such a curation plan are addressed in more detail below.

To the Foreground: The Case of the Dutch Bilingual Database, Roots of Ethnolects and TCULT Collections

In this section we report on our experiences in the DCS with the curation of three data collections: the Dutch Bilingual Database, Roots of Ethnolects and TCULT. They form an interesting and representative case for a number of reasons:

1. Over time, the data have been produced and held at various locations. Some data is presumed missing, but chances are that these may yet be retrieved.
2. There are several types of data (audio recordings, transcripts, images and descriptions of materials used to elicit the data and protocols/descriptions of the task), metadata and formats (wav, mp3, mp4, jpg, mpeg, txt, pdf, chat, imdi, eaf) which – to the extent that they do not conform to one of the CLARIN preferred formats– should be converted, thus providing an ideal test case for applying available tools.
3. Metadata of the recordings be converted to the preferred CLARIN standards involving ISOcat approved elements.
4. Two CLARIN-NL Data Centres (the Meertens Institute and The Language Archive at the Max Planck Institute) are involved as targeted archiving centres.

Description of the Resources

The Dutch Bilingual Database (DBD) is a rather substantial collection of data (over 1,500 sessions) from a number of projects and research programmes that were directed at investigating multilingualism, and comprises data originating from Dutch, Sranan, Sarnami, Papiamentu, Moroccan(-Arabic), Berber and Turkish speakers. At at the the basis of the collection is the research project TCULT[7] (1998-2002) in which intercultural language contacts in the Dutch city of Utrecht were studied. DBD established a first curation of the TCULT data and added many more bilingual data

[7] TCULT: http://ebookbrowse.com/tcult-pdf-d68190469

sets collected in the period 1985-2005. The current version of the DBD has IMDI[8] metadata files and is made accessible by the MPI[9]. The audio and text data are stored at the MPI and at the Meertens Institute. The DBD data consists of audio recordings, most of which are in .wav format while some are in .mp3. Most transcripts that are available are in .chat (Childes), some in .txt, or .eaf (Elan[10] format). In addition, there are summaries of some sessions in .pdf. Metadata are available in .imdi. In a number of cases, additional metadata are available in .txt or .pdf format. Occasionally, additional materials are available. These include descriptions (.pdf) or images (.jpg; .pdf) of the pictures books or cartoons used to elicit the data and protocols/descriptions of the tasks involved (.pdf), and covers of PhD theses in which the data are described and analysed (.pdf).

Roots of Ethnolects is a well-structured collection of 168 audio recordings of Dutch, Moroccan Arabic, Berber and Turkish. The data are stored at the Meertens Institute together with metadata (.imdi). Transcripts are available for a number of recordings (.eaf). In addition, protocols are available describing how the data were collected. Since this collection is well shaped and complete, the main efforts of curation at the DCS were directed towards the TCULT/DBD collections.

Permission for use of the DBD (including the TCULT) data was obtained from the subjects under the condition that they will be anonymized.[11] For the Roots of Ethnolects, data subjects gave their consent for the data to be used freely.

Development and Execution of the Curation Plan

We conducted interviews with the researchers involved in the TCULT/DBD collections. Based on what we learned from these sessions, we made an inventory of extra data that should be available, created a list of metadata considered relevant and specified the preferred formats. On the basis of our findings we drew up a curation plan (Action B in Table 1). Below we report on the elements of this plan and our experiences during execution.

Step 1: Restoring data

Missing data were identified at two levels:

1. Missing files in currently available recordings (e.g. either audio files or transcription files);
2. Missing sessions that should/could be added to the collection.

DBD contained 40-50 sessions with missing audio files and 40-50 (other) sessions with missing transcripts, as well as around 450 sessions that are part of a subcorpus that does not contain transcripts at all. We retrieved 32 of the missing transcripts from the CHILDES database.[12] These were copied into the DBD at the correct location.

[8] IMDI: http://www.mpi.nl/isle/

[9] See: http://corpus1.mpi.nl/ds/imdi_browser/

[10] Elan: http://www.lat-mpi.eu/tools/elan

[11] See the conditions for use at: http://corpus1.mpi.nl/ds/imdi_browser/. Click on the node DBD in the left panel, then on Condition_for_use_DBD.pdf.

[12] Please note that CHILDES itself does not contain audio recordings.

From an inventory it became clear that a handful of recordings in Meertens' TCULT list were not present in the DBD. Inquiries regarding the missing sessions were made at the Meertens archive. However, the additional sessions that we hoped to obtain could not be retrieved in the archives of Meertens.

Step 2: Setting up a metadata profile

Within CLARIN, cmdi[13] is the preferred format for metadata, imdi can be considered as a predecessor format of .cmdi. Cmdi categories should be ISOcat[14] categories or be related to these (Windhouwer et al., 2010). For curation this involves a number of tasks:

- Establishing a list with relevant metadata categories for this corpus;
- Establishing a mapping list showing in which imdi fields these metadata occur and, where appropriate, including additional metadata. Key-value pairs that are current in imdi should be replaced by new cmdi categories;
- Establishing a mapping list of corresponding cmdi and ISOcat metadata elements;
- Creating and defining new ISOcat metadata elements for metadata not yet included in ISOcat, and submiting these for approval by ISO.

Upon curation of this resource, the DCS simultaneously develops a cmdi profile for bilingual speech corpora which can be applied to other similar corpora. This implies that the profile includes metadata categories that are not present in the DBD, but are considered relevant for this type of data.

We found that the introduction and approval of metadata elements in ISOcat is a laborious and time-consuming procedure (cf. Zinn, Hopperman and Trippel, 2012). However, this is a necessary step in assuring uniformity and semantic interoperability for metadata categories.

Step 3: Converting data formats

The following conversion steps for DBD data were identified:

- The conversion of ten .mp3 files to .wav
- The conversion of DBD transcripts currently in .txt or .pdf to .chat or .eaf format. We consulted Brian McWhinney (developer of the .chat format) and Han Sloetjes (expert on .eaf). Unfortunately, most transcripts in DBD do not contain the time annotations necessary to do a proper line up of media files and transcription files. In cases where time annotation are present, they are not useful when the files are converted to .eaf. Time stamping then has to be added manually, partly by speakers who master the languages involved.
  Conversion from .chat to .eaf can be done automatically with a chat2eaf command, but without the time stamps it is questionable to what extent this conversion will be of added value to future users of DBD. Since .chat

[13] See CLARIN Component Metadata: http://www.clarin.eu/cmdi

[14] ISOcat: http://www.isocat.org/

and .eaf files are both CLARIN-standard formats, conversion is not strictly needed. Adding time stamps and converting .chat files to .eaf files would be meaningful, but requires a disproportional amount of manual labour. Therefore conversion of transcript files was not carried out.

- Converting the metadata (imdi) to cmdi. This conversion was done using the imdi2cmdi xslt scheme, provided by the Max Planck Institute. Since we decided to transform the imdi keys to new ISOcat cmdi categories, the xslt scheme needed to be adapted. Some of the imdi keys contained information that would fit in an existing field, so this information was replaced. Moreover, the data category labels were renamed according to the DBD cmdi profile.

It is relevant to note that no files were discarded during the conversion activities, since all original files were kept.

Step 4: Documentation

Two types of documentation were written: documentation on methods used and actions carried out for the curation (the curation plan was the basis for this document), and user documentation (in the form of a readme file) describing the background, the content, the structure and the format of the DBD.

Step 5: Assignment of Persistent Identifiers (PIDs)

The curated DBD has been archived at the Max Planck Institute (MPI) in Nijmegen – one of CLARIN-NL's data centres. The MPI uses the Handle system and has attributed PIDs to each audio file, transcription file and metadata file in the DBD.

Conclusions

Researchers who possess valuable data that are on the verge of oblivion should be stimulated and guided to make these available and accessible to the research community and (where relevant) to the wider public. In this paper we have introduced the CLARIN-NL Data Curation Service that has been established for exactly this purpose. Of course, the main task of the DCS is curating resources. However, before starting any curation the DCS has a clear obligation to assess the desirability and feasibility of the curation of a data resource. We have outlined the leading considerations underlying such a decision. Upon a positive decision, another relevant preparatory action is setting up a curation plan for the resource. We have illustrated this in our work on the DBD/TCULT database.

Since the DCS was put into service we have had several requests for data curation. These were mainly a result of our publicity campaign among researchers. A variety of data sets were offered, not all of which qualified for curation. For example, we were informed of a substantial corpus of police and court interrogations, which holds material that is valuable and difficult to obtain. Unfortunately, the intellectual property rights (IPR) were not settled in an acceptable manner, which would make curation pointless. For other data that was offered IPR is well arranged. This concerns such diverse material as two dialect dictionaries, some 1,000 interviews with veterans of missions of the Dutch military (part of which was curated earlier, see Van den Heuvel

et al., 2012), and the LESLLA corpus, which contains speech of 15 low-educated learners of Dutch as a second language.[15]

An important lesson we learnt since the DCS has come into operation is that the position of the DCS as middleman between researchers and data centres is a very time consuming one. Considerable time is needed to clear out the IPR status of relevant resources with the researchers. Furthermore, explaining the idea and sense of data curation to researchers, and, once convinced, the interplay of requesting and obtaining (additional) materials and information requires a substantial time investment. Such investments pay off in the end (mostly), but they should not be underestimated.

Acknowledgements

The research for this paper was funded by CLARIN-NL under grant numbers CLARIN-NL-10-025 and CLARIN-NL-11-005.

References

Neuroth, H., Lohmeier, F. & Smith, K.M. (2011). TextGrid – Virtual environment for the humanities. *The International Journal of Digital Curation, 2*(6), 222-231. doi:10.2218/ijdc.v6i2.198

Odijk, J. (2009). *CLARIN-NL long term programme 2009-2014.* Retrieved from http://www.clarin.nl/file/18763

Odijk, J. (2010). The CLARIN-NL project. Paper presented at the Seventh International Conference on Language Resources and Evaluation. Valletta, Malta.

Oostdijk, N. (2011). *CLARIN-NL Data Curation Service.* CLARIN-NL internal publication. Retrieved from http://www.clarin.nl/page/about/147

Thieberger, N. (2012). Using language documentation data in a broader context. In Seifart, F., Haig, G., Himmelmann, N., Jung, D., Margetts, D. & Trilsbeek, P. (Eds), Potentials of Language Documentation: Methods, Analyses, and Utilization. *Language Documentation and Conservation Special Publication No. 3.* Retrieved from http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4527/18thieberger.pdf?sequence=1

Van den Heuvel, H., Sanders, E., Rutten, R., Scagliola, S. & Witkamp, P. (2012). An oral history annotation tool for INTER-VIEWs. Paper presented at LREC2012, Istanbul. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/320_Paper.pdf

[15] An overview of resources curated by the DCS and the corresponding curation reports can be found at: http://www.clarin.nl/node/147.

Windhouwer, M., Wright, S.E. & Kemps-Snijders, M. (2010). Referencing ISOcat data categories. In Budin, G., Declerck, T., Romary, L., Wittenburg, P. (Eds), *In Proceedings of the LREC 2010 LRT standards workshop, Malta*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/workshops/W4.pdf

Zinn, C., Hoppermann, C. & Trippel, T. (2012). The ISOcat registry reloaded. In E. Simperl, Ph. Cimiano, A. Polleres, O. Corcho & V. Presutti (Eds), *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference.* Heraklion, Crete. doi:10.1007/978-3-642-30284-8_26