Introducing the CLARIN-NL Data Curation Service

Nelleke Oostdijk and Henk van den Heuvel

CLS/CLST (Centre for Language and Speech Technology)
Radboud University Nijmegen
P.O. Box 6500 HD Nijmegen, The Netherlands
E-mail: {n.oostdijk|h.vandenheuvel}@let.ru.nl

Abstract

In this paper we introduce the CLARIN-NL Data Curation Service. We highlight its tasks and its mediating position between researchers and the CLARIN Data Centres. We outline a scenario for successful data curation and stress the need to take notice of the factors that determine the desirability and feasibility of data curation. Finally, we present and discuss an exemplary case that illustrates the relevant issues involved in setting up a data curation plan.

Keywords: data curation, corpus management, sustainable infrastructure.

1. Introduction

Following decades in which a great deal of effort was spent on the creation of resources, currently there are several initiatives worldwide that aim to create an interoperable, sustainable research infrastructure. Examples, more specifically for the arts and humanities, are the US project Bamboo[1] and the European CLARIN initiative.[2] An integral part of such an infrastructure constitute 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 extent possible, accessible and usable.

Over the past decades numerous datasets have been collected and annotated by researchers for use in their own research. Often such data sets sank into oblivion once the research results had been published, while occasionally data were actually lost. With the years it has become apparent that unless appropriate action is undertaken to actively *curate* existing resources, many are at the risk of being lost as individual researchers or research groups often lack the expertise and the means to take the necessary measures to ensure their future availability.

By resource curation we mean the planning, allocation of financial and other means, and application of preservation methods and technologies to ensure that digital information of enduring value remains accessible and usable. It encompasses material that begins its life in digital form as well as material that is converted from traditional analog to digital formats. Digital information must be stored long-term and error-free, with means for retrieval and interpretation, for the entire time span the information is required for; in other words, it must be possible to decode and transform the retrieved files – of texts, charts, images or sound - into usable representations (cf. Hedstrom 1997).

Resource curation is important

- from an economic point of view;
  Curation is needed to prevent loss of resources that were created at substantial efforts and expenses. Loss may occur as a result of media deterioration or digital obsolescence. Costs may incur when resources are lost and resources must be rebuilt. In some cases, resources are unique and cannot be replaced if destroyed or lost.
- in terms of scientific interest;
  Curation grants access to the resources to a wider user community, allowing researchers to share access to data sets and permit replicability in research.
- for reasons of cultural heritage.

The structure of the present paper is as follows. In Section 2 the objectives and the background for setting up the CLARIN-NL data curation service (DCS) are described. Section 3 focuses on the positioning of the DCS in the language resources infrastructure context in the Netherlands and the tasks with which it has been charged. In Section 4 we report on experiences with the curation by the DCS of three data collections, viz. the Dutch Bilingual Database, Roots of Ethnolects and TCULT. Section 5 concludes the paper.

2. Background

In The Netherlands CLARIN-NL[3] (Odijk 2010) received funding from the Dutch government for implementing a programme that would contribute to the development of a sustainable research infrastructure for the humanities and linguistics in particular. CLARIN-NL is carried out jointly by technical specialists, technology providers and researchers. An effort is made to involve the intended

[1] www.projectbamboo.org
[2] www.clarin.eu
[3] www.clarin.nl

users through various application projects in which local repositories are integrated and local services set up for prototypical test installations as initial demonstrators, enabling evidence-based contributions to the discussion on standards and best practices for inter-operability, and to contribute to the survey of requirements for the infrastructure technology.[4]

From the start of the programme (2009), in CLARIN-NL funding has been available for projects directed at resource curation. Although a number of curation projects were undertaken, the calls for proposals have been less successful in reaching resource producers and owners who were not already aware of and/or participating in CLARIN-NL. In October 2010 the CLARIN-NL Executive board Board therefore initiated a pilot project that should investigate the need and possibility for establishing a Data Curation Service (DCS) task force that would salvage valuable corpora and data sets that are at the risk of being lost. 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. In such a scenario curation is carried out with minimal support from the original researcher who created, owns and/or manages the data. 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 a demand in the field and to develop ideas about the form such a service should be take, and also the effort and expertise required.

In the pilot project various data curation models and frameworks (such as the DCC Curation Lifecycle Model[5]) were investigated and the data curation policies adopted by other parties (such as libraries and archives, including for example the National Library of the Netherlands and Data Archiving and Networked Services (DANS) were looked into. Moreover, the project charted the role of various stakeholders (e.g. researchers, research institutes but also funding agencies like NWO) and organizations such as SURF and the Dutch Language Union.

As regards the needs and the priorities in curating resources, the roadmaps and surveys compiled by ELSNET and the Dutch Language Union provided pertinent information. Consultation of the national research database maintained by the Royal Netherlands Academy of Sciences (KNAW) served to identify which resources feature(d) in current or recent humanities research. Criteria were formulated for prioritizing resources for curation.

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, CLARIN-NL in September 2011 decided to establish the DCS at CLST in Nijmegen.

3. The CLARIN-NL DCS

3.1 Position and tasks

The DCS has been operational since January 2012. It aims to contribute 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 as a mediator between these and the CLARIN 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. Curation of resources, especially those presently held by individual researchers or research groups
2. Assisting in the curation efforts of CLARIN Data Centres (if and when such is desired)
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.

3.2 Data curation

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

A first step towards curation is the identification and assessment of candidate resources. This 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 there is. 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 failing 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: it should be established whether it is desirable to have the resource curated and whether indeed successful curation is feasible.

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:

[4] Cf. CLARIN-NL Long Term Programme 2009-2014. Retrievable from http://www.clarin.nl/system/files/CLARIN-NL%20Multiyear%20Programme%20090409-2.pdf

[5] http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Task A. Identification and assessment

Actions

1. Identify candidate resources; collect info 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
   etc.
2. Assess the desirability of curation
3. Assess the feasibility of successful curation

Task B. Development of a curation plan

Actions

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

Task C. Curation

Actions

7. Digitize data
8. Convert to a (CLARIN) preferred format
9. Assign appropriate metadata
10. Provide documentation

Task D. Validation

Actions

11. Validate curated resource

Task E. Archiving

Actions

12. Transfer to CLARIN Data Centre for long-term storage and maintenance
13. Assign persistent identifier
14. Provide access to content

Figure 1. Tasks and actions in data curation

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. Seeing that the field of Dutch language and speech technology is already very well organized and many resources are available through the HLT Agency, 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 solliciting 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. To the extent that a resource bears resemblance to resources already available it should be established which are the characteristics that set it apart. Only then is there is basis for deciding whether it is interesting enough to be curated. What became already apparent with the initial inventory of 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 and a Groningen version, while the HLT Agency distributes the Eindhoven Corpus VU version) without it being clear if, and if so 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: Researchers who have completed their PhD research and moved elsewhere, people that have retired or are 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 IPR-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 as 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
- 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 (incl. questionnaires and interviews), or while running an experiment in the laboratory (incl. 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.[6]

[6] 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

With transcriptions etc. a distinction should be made between enrichments that were obtained either manually/semi-automatically or as 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.[7] A strong argument in favour of curating the enrichments nevertheless (i.e. even when the tools are available) is that the resource 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 enrichtments. 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. There may, however, be occasions when the idea of curating derived data may be entertained and actually be given some follow up. This could be 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).

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

<u>The state of the resource</u>

For any resource that is being considered for curation it must be established whether

- it can be made available to a wider audience; questions that need to be addressed here are: Has the resource been cleared for IPR?[8] Should measures be taken to ensure anonymization? Etc.
- it is in digital form; to the extent that resources are not in digital form, digitization is needed.

Other questions in this context are

- is the resource in a state and form that can still be handled by current hard- 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?

<u>The availability of documentation</u>

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.

<u>The availability of expert knowledge</u>

Expert knowledge of a scientist or the original collector may be indispensable when curation of a resource is to be undertaken and conversion of the original form to its projected form is not straightforward.

<u>The availability of the necessary tools, scripts, etc.</u>

To the extent that specific tools etc. 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.

4. 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, viz. the Dutch Bilingual Database, Roots of Ethnolects and TCULT. They form an interesting and representative case for a number of reasons:

- the data over time have been produced and held at various locations, some data is presumed missing but chances are that these may yet be retrieved
- 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) which –to the extent that

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.

[7] The best strategy therefore 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.

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

they do not conform to one of the CLARIN preferred formats– should be converted, thus providing an ideal test case for applying available tools
- two CLARIN-NL data centres (the Meertens Institute and The Language Archive at the MPI) are involved as targeted archiving centres.

4.1 Description of the resources

The Dutch Bilingualism 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, Arabic, Berber and Turkish speakers . At the basis of the collection is the research project TCULT[9] (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 sets collected in the period 1985 – 2005. The current version of the DBD has IMDI[10] metadata files and is made accessible by the MPI at http://corpus1.mpi.nl/ds/imdi_browser/. 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, PDF or EAF (Elan). 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).

Roots of Ethnolects is a well-structured collection of 168 audio recordings of Dutch, Arabic, Berber and Turkish. The data are stored at the Meertens Institute together with metadata (IMDI). For a number of recordings transcripts are available (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 are directed towards the TCULT/DBD collections.

IPR

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

4.2 Development of a 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, we made a list of metadata considered relevant and of preferred formats.

On the basis of our findings we will establish a curation plan (Action B in Figure 1). This plan addresses the following issues:

Restoring data:

Missing data are identified at two levels: A. missing files in currently available recordings (e.g .either audio files or transcription files); B. missing sessions that must/can be added to the collection. Whether data can be added to the curated corpus depends on the effort needed so as to make the material accessible (e.g. AD-conversion of audio, or scanning transcriptions available on paper).

Restructuring the corpus

The structure of the corpus in its present condition is very inconsistent. At the first level the data are divided according to language which is well justifiable. However, below this layer the structure becomes very diffuse. Subdirectories are introduced with names referring to a variety of metadata, e.g. collector, informant, city of recording. A more consistent approach is to put the session names as sublayer directly below the language directory since all the other information in subdirectory names is already part of the metadata of the recording session.

Converting the metadata to CMDI

Within CLARIN, CMDI[12] is the preferred format for metadata, IMDI can be considered as a predecessor format of CMDI. CMDI categories should be ISOcat[13] categories or be related to these (Windhouwer et al., 2010). For curation this involves a number of tasks:
- establish a list with relevant metadata categories for this corpus
- establish a mapping list showing in which IMDI fields these metadata occur, and, where appropriate, including additional metadata
- establish a mapping list of corresponding CMDI and ISOcat metadata categories

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 which are not present in the DBD, but are considered relevant for this type of corpus.

Converting data formats

The following conversion steps for DBD data have been identified:
- the conversion of MP3 to WAV
- the digitization of retrieved audio recordings (if so required)
- the conversion of DBD transcripts currently in TXT or PDF to CHAT or EAF format

[9] http://ebookbrowse.com/tcult-pdf-d68190469
[10] http://www.mpi.nl/isle/
[11] See the Conditions_for_use_DBD (under node DBD at http://corpus1.mpi.nl/ds/imdi_browser/).
[12] http://www.clarin.eu/cmdi
[13] http://www.isocat.org/

- converting the metadata (IMDI) to CMDI. Although the data are already stored at the MPI and the Meertens Institute and also the curation of ‘core’ IMDI to CMDI is available, the extensions to IMDI developed in the DBD project (the IMDI DBD profile) which were used for both the DBD and Roots of Ethnolects data remain to be converted.

It will be investigated to what extent it is feasible to convert all transcripts in CHAT to EAF format, using the tools provided by Brian McWhinney. The idea here would be that researchers working with both DBD and Roots of Ethnolects data could have all the data in a single format (EAF), if they chose to do so. Where transcripts exist in the CHAT format, both the CHAT and the EAF are to be maintained, which means that attention must be given also to the synchronization of these versions.

Once the individual tasks have been detailed, the corresponding personnel effort and costs must be estimated and included in the curation plan. After approval of the plan, the necessary personnel with the required expertise must be made available for the envisaged tasks. These activities need not per se be executed by DCS staff, but may be outsourced to third parties, such as CLARIN-NL Data Centres.

5. Conclusion

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. Apart from the curation of this and other resources, an important task will be the identification and assessment of resources that qualify for curation in the near future.

Acknowledgement

The research for this paper was funded by CLARIN-NL (http://www.clarin.nl) under grant numbers CLARIN-NL-10-025 and CLARIN-NL-11-005.

References

Duranti, L. The long-term preservation of accurate and authentic digital data: the InterPARES project. In *Data Science Journal*, Volume 4, 25 October 2005: 106-118.

ELSNET’s HLT Roadmap. http://elsnet.dfki.de/ (November 2010).

Hedstrom, M. 1997. Digital preservation: a time bomb for Digital Libraries. In *Computers and the humanities*, 31(3): 189-202. Retrieved from http://www.uky.edu/~kiernan/DL/hedstrom.html.

Hedstrom, M., S. Ross, K. Ashley, B. Christensen-Dalsgaard, W. Duff, H. Gladney, C. Huc, A. Kenney, R. Moore & E. Neuhold. 2003. *Invest to Save. Report and recommendations of the NSF-DELOS working Group on digital archiving and preservation*. Prepared for National Science Foundation (NSF) Digital Library Initiative & The European Union under the Fifth Framework Programme by the Network of Excellence for Digital Libraries (DELOS). Retrieved from http://delos-noe.iei.pi.cnr.it/ activities/ internationalforum/Joint-WGs/digitalarchiving/ Digitalarchiving.pdf

Gray, J., A. Szalay, A. Thakar, C. Stroughton, J. vanden Berg. 2002. *Online Scientific Data Curation, Publication and Archiving*. Technical Report MSR-TR-2002-74. Redmond, Microsoft Research. Retrieved from http://research.microsoft.com/apps/pubs/default.aspx?id=64568.

Nauta, G.-J., R. Grim, I. Angevaare, H. Tjalsma, A. van Nispen & A. van der Kuil (eds.). 2010. *Data curation in arts and media research*. Stichting SURF. Retrieved from http://www.surffoundation.nl/nl/publicaties/Pages/StudieDataCurationinArtsandMediaResearch.aspx.

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

Oostdijk, N. 2011. *CLARIN-NL Data Curation Service*. CLARIN-NL internal publication. Retrievable from …

Russell, K. (ed.). 2010. *IISH Guidelines for preserving research data. A framework for preserving collaborative data collections for future research*. Stichting SURF. Retrieved from http://www.surffoundation.nl/nl/publicaties/Pages/StudieIISHGuidelinesforpreservingresearchdata.aspx.

Tjalsma, H. & A. van der Kuil (eds.). 2010. *Selection of research data. Guidelines for appraising and selecting research data.* A report by DANS and 3TU.Datacentrum. Stichting SURF. Retrieved from http://www.surffoundation.nl/nl/publicaties/Pages/StudieSelectionofResearchData.aspx.

Trusted digital repositories: Attributes and responsibilities. An RLG-OCLC report. 2002. RLG, Mountain View, CA. Retrieved from http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf.

Windhouwer M, Wright SE, Kemps-Snijders M (2010) *Referencing ISOcat data categories*. In Budin G, Declerck T, Romary L, Wittenburg P (eds) Proceedings of the LREC 2010 LRT standards workshop, Malta, http://www.lrecconf.org/proceedings/lrec2010/workshops/W4.pdf