CLARIN-NL Data Curation Service

Nelleke Oostdijk

Contents

Preface

In October 2010 the CLARIN-NL Executive board initiated a 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. [1] 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 Centres.[2]

The project was carried out between 1 November 2010 and 1 February 2011. In order to establish whether there is 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, the following approach was adopted:

- Reading up on
    - various data curation models and frameworks; e.g. through publications of the Digital Curation Centre about their DCC Curation Lifecycle Model, Consultative Committee for Space Data Systems (CCDS) on the Reference Model for an Open Archival Information System (OAIS) and the Research Information Network;
    - data curation policies adopted by other parties (libraries, archives), nationally and internationally (e.g. National Library of the Netherlands, Data Archiving and Networked Services (DANS), British Library, Library of Congress);
- Collecting information about different digital preservation initiatives (e.g. projects such as the InterPARES project) and the recommendations made (e.g. by the NSF-DELOS Working Group on Digital Archiving and Preservation, the RLG/OCLC Working Group on Digital Archive Attributes, the SURF Foundation);
- Charting 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;
- Reviewing the needs and priorities as identified in roadmaps and surveys such as compiled by ELSNET and the Dutch Language Union;

---

[1] The description of the project as it was conceived by the board is included as Appendix A. The project was funded by CLARIN-NL under grant number CLARIN10-025.

[2] See Appendix B and C.

- Consulting the national research database maintained by the Royal Netherlands Academy of Sciences (KNAW) in order to find out which resources feature(d) in current or recent humanities research;
- Formulating criteria for prioritizing resources to be curated;
- Defining the tasks for the DCS task force, identifying people and/or institutes that can contribute to the curation of resources;
- Gathering information as regards tools and data that might be useful in the process of curating resources;
- Consulting various people to fill in gaps in the accumulated information.

The present report summarizes the main findings.

I gratefully acknowledge the contribution of all those who shared with me their views on the subject and provided me with useful information. Special thanks are due to Daan Broeder who supervised the project on behalf of the CLARIN-NL Executive Board.

Nijmegen, 4 February 2011

Nelleke Oostdijk

1. Introduction

CLARIN-NL is a project directed at the development of a sustainable research infrastructure for the humanities and social sciences. 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 researchers have concerned themselves with the collection and annotation of data sets 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, resource allocation, 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 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. [3]

Resource curation is important[4]

- from an economic point of view;
  Curation is needed to prevent loss of resources that only exist in digital form. 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.

[3] Cf. Hedstrom (1997).

[4] While researchers are under obligation to keep the research data safe for a certain period of time (cf. for example the Netherlands Code of Conduct for Scientific Practice which prescribes storage of raw data for at least five years), this is not the prime incentive for setting up a data curation service in the CLARIN-NL context.

- 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.

While ideally curation is part of the day-to-day workflow, with the present state of the art where numerous resources already exist and many are beyond the period of their primary use, in the context of CLARIN-NL the curation of resources is undertaken strictly post-hoc.[5]

In the remainder of this chapter, the situation is described as regards the preservation and curation of digital resources as it presently exists in the Netherlands, including the major stakeholders that are involved and the role foreseen for CLARIN-NL. Before doing so, however, we first introduce the main producers and owners of digital resources as they are used in the fields of humanities and social sciences.

1.1. Digital resources: producers and owners

In the domains targeted by CLARIN-NL there are various institutes and organizations that produce or own digital resources that are of interest. While these include publishers and other commercial enterprises, for CLARIN-NL where the aim is to create a research infrastructure that provides free and open access the most important producers/owners of digital resources are the (non-technical) universities, the Max Planck Institute for Psycholinguistics (MPI), the Institute for Dutch Lexicology (INL) and the three KNAW funded research institutes working in these domains.[6]

Research at the MPI (http://www.mpi.nl) is directed at the study of the psychological, social and biological foundations of language. The MPI has a vast archive which includes many corpora but is especially well-known for its collection of materials from endangered languages. The archive offers researchers the possibility to archive their research data with detailed metadata.

[5] In future, with the CLARIN infrastructure in place, a different situation arises as contacts may be established at a very early moment between resource compilers and those maintaining the infrastructure.

[6] KNAW stands for Koninklijke Nederlandse Academie van Wetenschappen, i.e. the Royal Netherlands Academy of Sciences.

The INL (http://www.inl.nl/) is concerned with the study of the Dutch language (contemporary Dutch as well as Dutch from earlier periods). The institute is responsible for the production of major Dutch dictionaries. To this end it maintains a number of lexicographic data bases and text corpora. Moreover, the institute is the home of the Dutch HLT Centre which is charged with the maintenance and distribution of digital (linguistic) resources.

The orientation of the three KNAW institutes is as follows:

- The Meertens Institute (http://www.meertens.knaw.nl/) studies the diversity in language and culture in the Netherlands. The focus is on contemporary research into factors that play a role in determining social identities in the Dutch society. The institute has produced a large number of databases, many of which are already available online, but also houses many data collections of widely different nature.
- The Huygens ING Institute (http://www.huygensinstituut.knaw.nl/) conducts research in the fields of Dutch letters and the history of Dutch science. It is devoted to the maintenance of Dutch literary heritage, and to providing access to literary and scientific sources for the benefit of researchers and other interested partners. The institute maintains a number of digital resources, including for example the Biographical Portal and the Online Dictionary of Dutch Women.
- The Fryske Akademy (http://www.fryske-akademy.nl/) is concerned with the scientific study of the Frisian language and Frisian culture. The institute has produced the Dictionary of the Frisian language and has compiled various databases. Work is under way to combine these resources to form a single, interoperable database of Frisian spanning the period 1200 to 2010.

For the (digital) resources that are produced and owned by researchers and research groups associated with the various universities it is impossible to present an overview along similar lines as above: as far as research goes, research topics come and go and there is no exclusive claim on specific topics by specific research groups. Over the past decades many resources were developed and used within the context of a specific research programme or project. As there were hardly any incentives for researchers to share their data or make them available for re-use, the life-cycle of a resource was commonly of limited duration, often expiring as a project ended. However, over the past decade or so, under the influence of a changing academic climate there is a growing interest in sharing and re-using resources as researchers are encouraged to participate in larger collaborative efforts.

Major research funding agencies such as the Netherlands Organisation for Scientific Research (NWO) have started to implement a policy stipulating that research results such as digital resources be deposited with a trusted repository so that they may be re-used.[7] This is a line of action that will ensure the future wider availability of resources, witness the positive effects that are already visible in the form of the many resources that originated from the recent STEVIN programme.

The STEVIN programme was a programme that was initiated by the Dutch Language Union (Nederlandse Taalunie; NTU). Prior to the launch of this programme, the NTU commissioned a survey that was directed at charting the language resource infrastructure for Dutch and identifying the priorities.[8] With the STEVIN programme an important step was taken towards the realization of an effective digital language infrastructure for Dutch, by constructing essential resources that would support the development of NLP applications for Dutch (cf. STEVIN final evaluation report). All deliverables of projects funded within this programme have been deposited with the Dutch HLT Agency where they are available for re-use.

1.2. Digital data preservation

Early, major initiatives directed at preserving digital data have typically been led by archives and libraries who are in charge of digital documents, bibliographic information, catalogues, electronic publications, etc. Thus in the Netherlands, the National Library of the Netherlands (Koninklijke Bibliotheek; KB) is responsible for collecting, cataloguing and preserving publications that are issued in the Netherlands, and keeping them accessible.[9] As more and more publications are becoming available in electronic form, storing them permanently and keeping them accessible has become a task of increasing importance. For the archiving of electronic publications the KB has developed the e-Depot system.[10] More recently, the e-Depot system is also used to store the web archive, thus providing permanent access to an increasing number of websites from the Dutch portion of the World Wide Web.

---

[7] Research collaboration, data sharing and re-use are also addressed by international initiatives such as ELSNET (http://www.elsnet.org/) and FlaReNet (http://www.flarenet.eu/).

[8] See Daelemans & Strik (eds.) (2002), which is sometimes also referred to as the BLARK survey.

[9] See http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html

[10] For more information see http://www.kb.nl/dnp/e-depot/operational/background/index-en.html

While the KB is responsible for publications, Data Archiving and Networked Services (DANS) is the organization that is concerned with the storage and accessibility of research data, more specifically in the fields of humanities and social sciences. Upon its establishment in 2005, the collections and activities of the Steinmetz Archive, the Netherlands Historic Data Archive (Nederlands Historisch Data Archief; NHDA), the Scientific Statistics Agency (Wetenschappelijk Statistische Agentschap; WSA), and the Electronic Depot of the Netherlands Archaeology (EDNA) were transferred to DANS. Funded by the KNAW and NWO, DANS provides various services to researchers and research groups, from archiving to offering access to the files of large data managers. DANS also offers the Data Seal of Approval for archives that meet the criteria for "quality, permanence and accessibility and provides research financers with the guarantee that research results remain accessible for reuse".

The NARCIS portal that is maintained by the KNAW was created with the aim to increase the visibility and findability of scientific research in the Netherlands. [11] NARCIS provides access to scientific information, including (open access) publications from the repositories of the Dutch universities, KNAW, NWO and a number of scientific institutes, and the data sets available through DANS.

1.3. Archiving at DANS

As far as resource curation and archiving is concerned, DANS's ADA service and EASY system deserve some attention.

The ADA service[12]

Between 2000 and 2003 the NHDA carried out the ADA project: Archiveren van Digitaal Academisch Erfgoed (Archiving Digital Academic Heritage). Following the integration of the NHDA in DANS in 2005, ADA developed into a service that is now being offered for the archiving of academic data. Both the results of the original ADA project and the ADA service are discussed in Tjalsma (ed.) (2006).

The ADA project investigated the feasibility of offering digital archiving services to the scientific community, more specifically to the universities and research institutes in the fields of humanities and social sciences. The investigation was directed at a number of aspects, viz. what were the archiving

[11] The acronym NARCIS stands for National Academic Research and Collaborations Information System. See http://www.narcis.nl

[12] The information in this section is derived from Tjalsma (ed.) (2010).

needs of the targeted community, how should archiving services be offered and at what cost, and would such services meet with an uptake by the intended users.

In the ADA project the research data of the Meertens Institute served as a test case. The Meertens was specifically interested in developing a policy for the archiving of especially 'older' research data, i.e. data that were beyond their primary use, so as to prevent data loss. These data comprised research materials that were no longer actively maintained on the institute's server, but were kept on various storage media.

From the experiences gained in the ADA project it is clear that the phase in which the inventory is made of data to be archived and during which all the necessary information is collected is very crucial to the success of the archiving process. It proved to be much more time-consuming than initially anticipated. The involvement of the resource owner/producer is essential as he/she possesses relevant information concerning the content, the manner in which the data were collected, enriched, etc. He/she is also well-qualified to offer ideas on how best to represent the data and make them accessible, and should be able to answer questions as regards copyright and privacy issues. Such information cannot (or at least cannot readily) be obtained in any other manner.

The methodology developed in the ADA project lies at the basis of the ADA service that DANS offers to researchers today. The service is directed at retro-archiving and is considered less appropriate for the archiving of data in on-going projects. The ADA approach is directed primarily at preserving the information content, not at retaining the original formats or structures per se. Rather, where necessary, resources are converted to a standard, non-proprietary format. Emulation software is occasionally used, for example for an initial inspection of data in obsolete formats.

Although users have shown an interest in the ADA service, it is not all that frequently used. An explanation for this, it would seem, is the lack of financial means.

The EASY system

The EASY system is an electronic archiving system provided and maintained by DANS. The system allows users to store research data "in a permanent and sustainable manner, according to the guidelines of the international Data Seal of Approval. The data are made available to other researchers under specific conditions in accord with the depositor."(https://easy.dans.knaw.nl/dms) Thus, there is the possibility to give researchers free access to the data or to restrict access so that the data can only be accessed after the consent of the depositor. Other forms of restricting the access are also possible, such as allowing

access only after a certain period of time. The EASY system is a light-weight deposit application which expects the users to submit their data themselves and provide minimal metadata (Dublin Core). For researchers wishing to search the archive, the possibilities are limited and only resource bundles (rather than individual files) can be viewed.

1.4. SURF

While it would seem that, with the KB and DANS in place, research data were well taken care of, in actual practice what can be seen is that with the exception of data maintained by the MPI, INL and the KNAW institutes, data that reside within university research institutes or are held by individual researchers or smaller research groups with a few exceptions are not systematically curated. Although over the years researchers and research organizations have begun to take an interest in data sharing and re-use, to date they still do not concern themselves with questions such as *what research data qualify for long-term preservation* and *in what form should they be preserved?* These questions are essentially left for the archives and repositories to decide and organizations such as SURF.

SURF (http:www/surf.nl) is an organization in which universities, institutes of higher education and research institutes in the Netherlands collaborate and which aims at breakthrough innovations in IT. SURF actively promotes open access, open standards and open source. In the recent SURFshare programme (2007-2010) guidelines for the preservation and availability of collaborative data collections were developed.[13] However, the challenge remains to have these guidelines be adopted and see data curation be included as a part of the day-to-day work flow.

1.5. CLARIN data curation

As past experience has shown, there is a variety of reasons why individual researchers but also research groups cannot be expected to take the initiative for the curation of their research data. Data curation in the context of CLARIN is directed at salvaging resources and making them accessible and usable in a sustainable interoperable infrastructure. Researchers should be able to discover and access resources that are possibly relevant to their research by exploring the rich meta data that are provided.

[13] The guidelines are available in the form of three reports, viz. Nauta et al. (eds.) (2010), Russell (ed.) (2010), and Tjalsma & van der Kuil (eds.) (2010).

2. Resources up for curation

In this section the types of sources are identified that qualify for curation and an initial inventory is made of resources that may be worth curating. Criteria are formulated that can be used to set priorities.

2.1. Types of resources

In the widest possible sense of the word, a resource may be understood to refer to any of the following (cf. Nauta et al., eds. 2010):

- Digital surrogates of primary sources, e.g. in case the relevant materials have not been digitized yet
- Research data; these constitute a highly heterogeneous set and include text corpora, databases, transcriptions, annotations, spreadsheets, survey results, collections of hyperlinks, …
- Documentation, e.g. annotation guidelines, transcription protocols
- Bibliographic information
- Publications (other than documentation)
- Simulations
- Tools

In the context of CLARIN the focus will be on research data and documentation. For bibliographic information and publications it may be argued that these are already well taken-care of by various library and archiving services (incl. the KB, and the university and research institute repositories). As regards simulations and tools, it must be observed that their sustainability is at best limited and possibly requires disproportional effort.[14] Therefore, no effort will be put into curating simulations, while tools will only be considered for curation when they are deemed indispensable for computationally less versed researchers to explore and/or exploit the data in a user-friendly way, or if they can be used to facilitate the curation of other resources.

[14] Cf. Hedstrom (1997: 199) on this issue: There are few well developed methods for preserving and migrating software so that it might be used to recreate digital documents that have the "look and feel" of the original sources. Maintaining repositories of obsolete hardware and software has been discussed periodically, but usually dismissed out of hand as too expensive and not demonstrably feasible. This approach deserves more serious consideration as a strategy for maintaining continuing access to certain types of digital materials. Feasibility studies and cost/benefit analyses should be conducted to determine the technological, economic, and commercial feasibility of maintaining selected legacy software systems and performing specialized migrations or, alternatively, of building and maintaining software emulators. Such an approach would support replay of original resources and contribute to the preservation of software as a significant cultural and intellectual resource in its own right.

2.2. Inventory of potentially relevant resources worth curating

In the context of CLARIN –NL the curation effort will be directed primarily at language resources stored and used in the Netherlands.

An initial, list of potentially relevant resources was compiled on the basis of information that was collected from various sources,[15] viz.

- the project descriptions in the national research database (Nederlandse Onderzoeksdatabank; NOD) maintained by the KNAW
- the inventory of resources (data and tools) made in the context of CLARIN-EU (WP5) [16]
- project proposals submitted in response to CLARIN Calls (1 and 2)
- a preliminary version of the CLARIN user survey (Arjan van Hessen and Jenny Audring, July 2010)
- through personal contacts of the author

Since the initial list already contained numerous potential candidates for curation, it was decided to refrain at this point in time from active attempts to identify additional resources and to make a first crude selection on the following grounds:

- it may be assumed that there is currently no immediate need for curation by the CLARIN-NL DCS if
    - a resource is held at a trusted repository, archive or such; this relates to all resources held at ELDA, the LDC, OAI, SIL, TalkBank, CHILDES, etc.;[17]
    - a resource is held at one of the (candidate) CLARIN centres;[18]
    - a resource is already (being) curated in one of the CLARIN-NL funded or associated projects (e.g. TTNWW, or Call 1 or Call 2 projects);
    - a resource is held at an institute abroad; this is the case for a number of resources that were listed in the CLARIN-EU inventory;
    - the resource is as yet under construction;

[15] A first (preliminary) version of this inventory can be found in the form of an Excel file under Data Curation Service on the CLARIN-NL site (http://www.clarin-nl).

[16] See http://www.clarin.eu/view_resources

[17] The Language Archive (TLA) links to various archives and mirrors repositories such as CHILDES and TalkBank (see http://corpus1.mpi.nl).

[18] See Appendix C. Resources maintained by individual researchers/research groups are presumed to be the most vulnerable and there seems to be no immediate need to make a special effort for curating resources held at the CLARIN centres as they will do so in due time.

- there is insufficient information as to the nature of the resource or its present whereabouts to follow up on it.

For the remaining items on the list, an effort was made to provide further relevant information. After subjecting this reduced list once more to the selection criteria mentioned above, a list of over 150 resources remains. Additional effort is needed to get hold of all the information that is needed before a motivated decision can be made as to whether curation is desired and feasible.

2.3. Establishing criteria for setting priorities

Bearing in mind that there may be different reasons for curating a resource, viz. reuse, verification, and heritage, in this section a number of criteria are formulated that come into play when considering the question whether a resource should be curated and what priority this should be given (*desirability of curation*), and also whether curation can be expected to yield a satisfactory result (*feasibility for curation*). Apart from these criteria there is the *cost of curation* to be considered.

Desirability of curation

Relevance to research community

CLARIN-NL is directed first and foremost at researchers in the humanities and social sciences. Therefore, 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. This has inspired the identification of some priority areas for the submission of proposals for curation and demonstration projects in Call 1 and 2. These are literary studies, history and political studies, communication and media studies, first and second language acquisition, and historical linguistics.

Uniqueness

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.

Free availability

In some cases arrangements have been made for resources to be distributed by an organization such as ELDA. This means that the resources must be paid for. Examples are CELEX and EuroWordnet. Where contracts with ELDA are on a non-exclusive basis, it may be worthwhile investigating whether there is sufficient interest to pursue curating these resources and making them available in the CLARIN infrastructure for free.

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.[19]

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.[20] 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).

[19] 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.

[20] 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. Cf. note 14.

Feasibility of successful curation

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; questions that need to be addressed here are: Has the resource been cleared for IPR?[21] 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
- it is in a sound state qualitatively?
- …

Availability of technical documentation and user 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 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.

Availability of expert knowledge

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.

Availability of necessary tools, scripts, etc.

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.

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

3. CLARIN-NL data curation service (DCS)

As the mission of CLARIN is to create a sustainable and interoperable research infrastructure, making available resources and providing access are integral parts of what needs to be done in order to make the infrastructure attractive for users to use. While the CLARIN centres will contribute their collections, a select group of researchers submit their resources of their own accord helped by the funding provided through the CLARIN Calls, there are many more resources that could be shared, viz. resources that are "lying around" and are at the risk of being forgotten or lost, where the researchers cannot be expected to take action towards curating them. Seeing that these resources will not find their way to the CLARIN infrastructure without encouragement and active support from the CLARIN community, a data curation service or task force should be set up.

3.1. Tasks of the DCS

The tasks of the DCS are as follows:

1. Curation of resources , especially those presently held by individual researchers or research groups
2. Advising researchers who wish to undertake the curation of their resources themselves

Curation of resources

The curation of resources held by individual researchers or research groups will form the core of the work to be undertaken by the DCS. In the scheme below, an overview is given of the various subtasks involved.

Especially the task of identifying and assessing candidate resources will 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 to be carried out in close collaboration with the resource owner/producer. It is recommended that a small body be formed of representatives of the various user groups. This body should give advice on the relevance of a resource for a specific discipline.

Task B, the development of a curation plan, requires the involvement of the CLARIN centres. Their expertise is indispensable, especially where the content of the resources is concerned. Moreover, it should become clear, at the earliest possible moment in the curation process, at which of the CLARIN centres the resource will be deposited. While devising the curation plan it may become clear that there

are factors that are prohibitive to proceeding with the execution of the curation in which case the process is terminated.

Task	Action
A. Identification and assessment	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
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
	8. Convert to a (CLARIN) preferred format
	9. Assign appropriate metadata
	10. Provide documentation
D. Validation	11. Validate curated resource
E. Archiving	12. Transfer to CLARIN Centre for long-term storage and maintenance
	13. Assign persistent identifier
	14. Provide access to content

Scheme: Tasks and actions in data curation

Task C (curation) is envisaged to be carried out by the DCS, although where appropriate, e.g. in retrieving data from obsolete media, assistance may be sought elsewhere. Task D (validation) should

preferably be carried out by an independent group (e.g. SPEX) with experience in validation.[22] Task E (archiving) is to be carried out by the DCS and the designated CLARIN centre.

Advising researchers who wish to undertake the curation of their resources themselves

Researchers wishing to curate their resources themselves should be able to get advice on how best to proceed. It seems appropriate to have the CLARIN help desk handle also questions as regards curation.

3.2. Staff and expertise required

The staff required for the DCS minimally comprises

- a coordinator (0,3 fte) whose tasks are to coordinate the activities of the DCS, to establish and maintain contacts with resource suppliers and the CLARIN Centres, supervise the identification and assessment of curation candidates and the subsequent planning for and monitoring curation
- an IT-specialist (0,3 fte) who is responsible for data analysis, conversion tests, risk assessment, data migration, conversion
- a documentalist (0,5 fte) who is involved in the identification and assessment of resources, and the creation of metadata profiles, metadata conversion, providing documentation for the curated resources

Additional staff is to be hired as required. This concerns for example research or student assistant(s) for entering metadata, checking the correctness of conversions, clean up processes, etc. It also holds as regards specialized expertise that is needed at some point. The involvement of staff from one of the CLARIN Centres must be considered for tasks requiring specific expertise (e.g. expert knowledge in the field of literary studies). However, it should be noted that such involvement requires timely communication so that it can be included in their planning. Apart from the CLARIN Centres there are various other institutes that can be approached for specific questions or tasks. These include the Frisian Academy (any questions regarding Frisian but also matters having to do with historical/geographical carthography) , the KB (OCR), Sound and Vision (audio and video), TiCC/University of Tilburg (text clean up), CLST/Radboud University (orthographic and phonetic transcription, search and annotation of audio archives).

[22] As one of ELDA's validation centres, SPEX has ample experience in the validation of resources.

3.3. Role for the CLARIN Centres

The picture that emerges from talks with representatives of the CLARIN Centres is mixed.[23] While they subscribe to the need for actively pursuing the curation of resources and recognize that there is a role for them (cf. Section 3.1), there appears to be some reluctance to participate in the curation of resources other than the ones that are already part of their collections. The reasons for this are that

- institutes such as Meertens, INL, MPI and Huijgens are evaluated on the basis of their scientific achievements; curation by the research institutes is usually done in the context of a current research programme or project, curation by itself is considered to be not very prestigious
- curation is not the most exciting/sexy activity;
- personnel costs are not fully covered by the current CLARIN rates;
- by employing additional staff the institute runs the risk of being held responsible for any unemployment pay that is due;
- additional work must be planned for;
- institutes have a preference for certain research data (e.g. the Huygens Intistute has a special interest in literary and historical data, while the MPI focuses more specifically on acquisition data and endangered languages)
- the quality of the resource may not be up to the standards normally upheld by the CLARIN Centre.

The DCS and the CLARIN Centres should make clear arrangements so that the DCS can benefit from the expertise available at the CLARIN Centres without unnecessarily burdening them.

3.4. Planning for curation

Most of the resources that are mentioned in the overview of data collections that was produced as part of the user survey for one reason or another do not (or at least do not presently) qualify for curation by the DCS. Some resources are already or about to be deposited with a trusted archive (CHILDES, HLT Agency, MPI), or curation is already undertaken in a CLARIN-funded project. Other resources are still in the process of being built or are being planned. For a number of resources IPR and privacy issues need to be resolved before curation can be considered. Thus from the resources listed in the overview, only the following items remain for which there are no immediate hindrances on the curation path:

[23] Questions that were addressed in the discussions with representatives of the CLARIN Centres are listed in Appendix D.

- Corpus of metaphors (Gerard Steen, VU)
- Child language data, incl. data from SLI children (Fred Weerman, UvA)
- Word lists and grammar of Papua languages (Wilco van den Heuvel, VU)
- A collection of some 100,000 records comprising idioms, collocations, text fragments (Jack Hoeksema, RUG)

For these resources additional information must be collected in order to be able to decide on a course of action.

Additional information is also needed for most of the resources in the inventory that remain potential candidates. At present it is not deemed possible to judge whether curation is desirable and/or feasible, nor is it possible to estimate the cost involved.

Resources for which a proposal was submitted in one of the CLARIN Calls but which did not (as yet) receive any funding obviously form an exception.[24] Here the detailed information that is required *is* available. Moreover, as is apparent from the fact that in each case a proposal exists, there is an interest for curating these resources. Thus it is recommended that curation of be undertaken of (1) the LESLLA corpus, (2) the DBD, Roots of Ethnolects, and TCULT data, and (3) the IPNV corpus.

LESLLA Corpus:[25]

The LESLLA corpus contains speech of low/educated learners of Dutch as a second language. The corpus comprises some 24,000 utterances. Besides audio (.wav) files there are orthographic transcriptions (PRAAT TextGrid files), and for all 15 learners there are also metadata available. IPR is not a problem: all participants were paid and have signed a form stating that the recordings can be used freely for scientific purposes.

Curation of the LESLLA Corpus involves making the corpus CLARIN-compliant, better searchable and accessible through different platforms (PRAAT, ELAN). A suitable CMDI metadata schema will be created to accommodate all available relevant metadata. Where necessary new concepts will be created in the ISOcat. It will be investigated to what extent it is possible to harmonize with other similar corpora (e.g. second language learner, ESF corpora) at the metadata level. The metadata available in suitable format

[24] Some of the proposals that were submitted that did not receive funding involved combined projects with a curation and a demonstration component.

[25] The description below is based on the LESLLA++ proposal originally submitted in CLARIN Call 2.

will be converted to CMDI. In addition new metadata records will be created by hand using the CMDI editor.

The total effort required is estimated to be 4 months (2 PM * € 5,249; 2 PM * € 3,200).

DBD, Roots of Ethnolects and TCULT data:[26]

The Dutch Bilingualism Database (DBD) comprises recordings of Dutch, Sranan, Sarnami, Papiamentu, Arabic, Berber and Turkish speakers. With the recordings transcripts (ELAN) are available as well as metadata (IMDI). The data are stored at the MPI. Roots of Ethnolects is a collection of 168 recordings of Dutch, Arabic, Berber and Turkish. Metadata are available in IMDI, while there are selected transcripts in ELAN. The TCULT data comprise recordings and transcripts in Word. Both the Roots of Ethnolects and the TCULT data are stored at the Meertens Institute.

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) remain to be converted. The additional effort will make all metadata of the above resources available in a coherent format.

The total effort required is estimated to be 11 months (6 PM * € 5,249; 5 PM * € 3,200).

IPNV Corpus:[27]

The IPNV Corpus comprises a collection of 1000 2,5-hour interviews, 250 of which were already curated in the INTER-VIEWS project. The recordings are in .wav format. With all of them extensive metadata are available in an electronic, but not yet standardized form. For each interview also an extensive textual summary is available (XML).

Curation will involve the integration of the rich metadata in the CMDI metadata profile developed in the INTER-VIEWS project. New metadata categories will be defined, and registered in ISOcat, when needed.

The total effort required is estimated to be 3 months (2 PM * € 5,249; 1 PM * € 3,200).

---

[26] The description below is based on the MULTINED proposal originally submitted in CLARIN Call 2.

[27] The description below is based on the AUDITIONS proposal originally submitted in CLARIN Call 2.

3.5. Standards and tools that can be used in the curation process

In Kemps-Snijders et al. (2009) an overview is given of standards for language resources with recommendations for CLARIN. The Digital Formats web site[28] of the Library of Congress provides information about the sustainability of digital formats for a number of content categories, viz. sound, textual, still image, moving image, and web archive. The web site is particularly useful in that it discusses the quality and functionality factors appropriate for the evaluation of the various formats and also gives an illustrative view of a curator's identification of significant characteristics of the content and preferred formats.

An overview of tools can be found on the site of CLARIN-EU (http://www.clarin.eu/view_tools). Relevant for curation are tools created by the MPI such as ARBIL and ELAN, and also tools that are available through CLARIN-NL projects such as TICCLops, Adelheid, and AAM-LR. Moreover, on the websites of the Digital Curation Centre[29] and the National Digital Information Infrastructure and Preservation Program[30] information can be found on various tools and services that may be employed in the preservation/curation of digital data. To what extent any of these tools and services are useful in the context of CLARIN-NL, e.g. because they address specific needs presently uncatered for, is as yet unclear. It may well be that the need for such additional tools only arises when the curation of particular resources is being considered. While it is suggested that a technical specialist go over the tools and services listed on these websites in order to see how they might support the work of the DCS, the active search for other tools and services is best undertaken when the need arises.

[28] See http://www.digitalpreservation.gov/formats/

[29] See http://www.dcc.ac.uk

[30] See http://www.digitalpreservation.gov/partners/resources/tools/index.html

4. Concluding remarks

The picture that has emerged from the present investigation is one that shows that there is clearly a need for a data curation service (or task force) that actively pursues the curation of resources. The DCS should direct its efforts primarily towards the curation of resources held by individual researchers or research groups which cannot be expected to undertake the curation themselves. As is apparent from the initial inventory, there are plenty of resources that may be worth curating. In the conversations with various researchers, many other resources were hinted at, indicating that we have only begun to uncover what must be an overwhelming amount of data that has accumulated over the years. It was suggested that one should also check what data have been deposited at university libraries. However, more specific and detailed information is needed about the individual resources before it can be decided which resources should be given priority and what effort is required from the DCS exactly. Only then will it be possible to give an estimate of the staff and budget required.

As regards the role of the CLARIN Centres, this remains a delicate matter: On the hand, the CLARIN centres want to be involved while the DCS can benefit from the expertise they have to offer; on the other hand, the curation of resources that are not their own is not necessarily a priority for the CLARIN centres. The role envisioned for the CLARIN centres is something that urgently needs to be discussed and clarified in talks between the CLARIN Executive Board and the CLARIN Centres.

There are presently two initiatives that CLARIN may want to follow up and possibly engage in. One concerns the curation of resources held at the Leiden University Centre for Linguistics (LULC), the other is an initiative for a project in which pathological data will be curated. As regards the LULC resources, last autumn there have been initial talks between DANS and a representative from the LULC, Prof. Maarten Mous, who is investigating the possibilities for salvaging these resources. Alternatives currently under consideration are archiving these resources with DANS, or with support of SURF archiving them with the university library. As for the pathological data, a group of researchers from the universities of Amsterdam, Leiden, and Nijmegen are considering a grant proposal for a project for setting up an “open multimedia data archive for Dutch pathological language data”.[31]

[31] The current working title is “Vulnerability in Acquisition: Language Impairments in Dutch: Creating a VALID Data Archive”.

References

Brown, A. 2003. *Digital preservation guidance note 2: Selecting storage media for long-term preservation.* The National Archives. Retrieved from http://www.nationalarchives.gov.uk/documents/selecting_storage_media.pdf (12 November 2010)

Daelemans, W. & H. Strik (eds.). 2002. *Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen. Een rapport in opdracht van de Nederlandse Taalunie.* Retrieved from http://taalunieversum.org/taal/technologie/stevin/documenten/batavo.pdf (1 November 2010).

CLARIN Resources overview. Retrieved from http://www.clarin.eu/view_resources (1 November 2010).

CLARIN Standards for LRT (v-6). Retrieved from http://www.clarin.eu/recommendations (2 November 2010).

DCC Curation Lifecycle Model. Retrieved from http://www.dcc.ac.uk/resources/curation-lifecycle-model (4 October 2010).

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 (16 November 2010).

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 (10 November 2010)

Graaf, M. van der. 2010. *Organisatorische aspecten duurzame opslag en beschikbaar stelling onderzoeksdata.* SURFshare programma, Stichting SURF. Retrieved from http://www.surffoundation.nl/nl/publicaties/Pages/Organisatorischeaspectenduurzameopslagenbeschikbaarstellingonderzoeksdata.aspx (7 December 2010).

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 (4 October 2010).

Hessen, A. van. 2010. *CLARIN Gebruikersonderzoek. Overzicht Data- en Toolscollecties* (v. 6 July 2010 and 15-11-2010).

Kemps-Snijders, M. N. Bel, P. Wittenburg, D. Broeder, D. van Uytvanck, L. Romary, E. Hinrichs & G. Budin. 2009. *Standards for LRT*. Retrieved from http://www.clarin.nl/system/files/Standards%20for%20LRT-v6.pdf (16 November 2010).

Lord, P., A. Macdonald, L. Lyon, D. Giaretta. *From Data Deluge to Data Curation*. The Digital Archiving Consultancy Limited and the Digital Curation Centre. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf (4 October 2010)

Martin, S., D. Golding, P. Noakes, H. van Wijngaarden, J. Pijpers, A. Lampers, J. van der Hoeven, R. Altenhöner, J. Kett, T. Steinke & S. Brygfjeld. 2010. Long-term preservation services. A description of LTP services in a Digital Library environment. Retrieved from http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_Long_Term_Preservation_Services_2010-08-05.pdf (10 November 2010).

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 (8 December 2010).

NDIIPP Partner Tools and services inventory. Retrieved from http://www.digitalpreservation.gov/partners/resources/tools/index.html (16 November 2010).

Research Information Network. 2008. *Stewardship of digital research data: a framework of principles and guidelines*. Retrieved from www.rin.ac.uk/system/files/attachments/Stewardship-data-guidelines.pdf (8 December 2010).

Russell, K. (ed.). 2010. *IISH Guidelines for preserving research data. A framework forpreserving collaborative data collections for future research*. Stichting SURF. Retrieved from http://www.surffoundation.nl/nl/publicaties/Pages/StudieIISHGuidelinesforpreservingresearchdata.aspx (8 December 2010).

Sustainability of digital formats. Planning for the Library of Congress Collections. Retrieved from http://www.digitalpreservation.gov/formats/ (19 November 2010).

Tjalsma, H. (ed.). 2006. *Archiveren van Digitaal Academisch Erfgoed. Een verslag als voorbeeld*. DANS studies in digital archiving 2. Den Haag. Retrieved from http://www.knaw.nl/publicaties/pdf/20061011.pdf (4 January 2011)

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 (8 December 2010).

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 (12 November 2010).

Appendix A. CLARIN-NL Data Curation Service

The text below constitutes the original description of the task the author was charged with.

A plan must be created to work out the set-up of a CLARIN-NL Data Curation Service. This service will make it possible to carry out curation for data residing at humanities researchers by a dedicated team of specialists and with minimal support from the original researcher who created/owns/manages the data.

We observe that many of these researchers do not initiate a project to carry out such curation (even though such a project could be funded by CLARIN-NL), for a variety of reasons, for example

- *The resource is (judged to be) rather small*
- *The resources is (judged to be) rather idiosyncratic*
- *The relevant expertise is not available*
- *Such a project is given low priority relative to real research projects*
- *No added value for resource curation is perceived by the researcher*
- *"it is available to everyone because I put it on my website"*
- *Etc.*

The idea is that this dedicated team optimally utilizes any existing tools in the curation of the data, and actively selects data to be curated (of course, permission of the researcher will be required).

Data curation involves:

- *Converting the data into a standard format supported in CLARIN*
- *Creating metadata for the data conforming to CMDI*
- *Mapping data categories used in the resource and its metadata onto ISOCAT data categories, and/or extending ISOCAT with new data categories. Extend RELCAT (as soon as it is available or a format has been fixed for it) to establish relations of new data categories with closely related (but different) existing ones*
- *Storage of the data and its metadata on a server of a CLARIN-Centre, so that they become accessible to the whole research community*
- *Assignment of permanent IDs to the data and the metadata (by the relevant CLARIN-centre)*

The plan must

- *Identify existing tools and data that can be used to facilitate the data curation. For example, CMDI and the associated tools, any existing converters, XML checkers, tools and expertise for corpus clean-up (e.g. in Tilburg), etc.*
- *Identify relevant expertise centres and/or persons that could/should be part of the dedicated team*
- *Identify and prioritize resources that qualify for curation. Here the overview being created by the user survey will be especially helpful. Priorities can be assigned in function of the relevance of the data to the research community but also in function of the technical possibilities and limitations that hold. Initially data for which there are no IPR-problems for academic (research) use should be selected.*

- *Make an estimate of the required knowledge, expertise, tools, effort, lead time and budget for curation of the highest priority data.*

Though we have already seen that many existing resources are not available in digitized form and therefore require a digitalization effort, the focus should be on resources that are digital in nature, though small-scale digitization is not completely excluded.

Appendix B. CLARIN members in the Netherlands

Overview, retrieved from the CLARIN-NL site (http://www.clarin.nl/node/7).

Name	Acronym	Location
Utrecht Institute for Linguistics OTS	UIL-OTS	Utrecht
Landelijke Onderzoeksschool Taalkunde	LOT	Utrecht
Max-Planck-Institute for Psycholinguistics	MPI	Nijmegen
Meertens Instituut (KNAW)	Meertens	Amsterdam
Huygens Instituut (KNAW)	Huygens	Den Haag
Data Archiving and Networked Services (KNAW)	DANS	Den Haag
Fryske Akademy (KNAW)	FA	Leeuwarden
Digitale Bibliotheek voor de Nederlandse Letteren	DBNL	Leiden
Instituut voor Nederlandse Lexicologie	INL	Leiden
Centre for Language and Speech Technology	CLST	Nijmegen
Centre for Language Studies	CLS	Nijmegen
Amsterdam Center for Language and Communication	ACLC	Amsterdam
Center for Language and Cognition	CLG	Groningen
Centre for Linguistics	LUCL	Leiden
Tilburg Centre for Creative Computing	TiCC	Tilburg
Human Media Interaction Group	HMI	Twente
Katholiek Documentatie Centrum	KDC	Nijmegen
Koninklijke Bibliotheek	KB	Den Haag
Veteraneninstituut	VI	Doorn
Taal en Communicatie Vrije Universiteit	VU	Amsterdam
Instituut voor Beeld en Geluid	BG	Hilversum
Nederlands Instituut voor Oorlogsdocumentatie	NIOD	Amsterdam
Aletta (Instituut voor Vrouwengeschiedenis)	Aletta	Amsterdam

Appendix C. CLARIN Centres in the Netherlands

(Source: CLARIN-NL site, http://www.clarin.nl/node/130)

INL

Coordination:	Jan Theo Bakker
Technical Matters:	Wim Kok
Postal Address:	Matthias de Vrieshof 2-3, 2311 BZ Leiden, The Netherlands
Tel:	+31 71 5141648
Fax:	+31 71 5272115
e-mail:	Jan Theo Bakker and Wim Kok

Meertens Instituut

Coordination:	Marc Kemps-Snijders
Technical Matters:	Jan Pieter Kunst
Postal Address:	Postbus 94264, 1090 GG Amsterdam, The Netherlands
Tel:	+31 20 4628500
Fax:	+31 20 462 85 55
e-mail:	marc.kemps.snijders@meertens.knaw.nl and jan.pieter.kunst@meertens.knaw.nl

MPI

Name:	Daan Broeder
Postal Address:	PO Box 310, 6500 AH Nijmegen, The Netherlands
Tel:	+31-24-3521103
Fax:	+31-24-3521213
e-mail:	daan.broeder@mpi.nl

DANS

Name:	Dirk Roorda
Postal Address:	P.O Box 93067 2509 AB Den Haag, The Netherlands
Tel:	+31-70 3494450
Fax:	+31-70 3494451
e-mail:	dirk.roorda@dans.knaw.nl

Huygens Instituut

Name:	Karina van Dalen
Postal Address:	P.O Box 90754 2509 LT Den Haag, The Netherlands
Tel:	+31-70 331 58 00
Fax:	+31-70 382 05 46
e-mail:	karina.van.dalen@huygensinstituut.knaw.nl

Appendix D. Interview questions

Questions that were addressed in the interviews with representatives of (candidate) CLARIN-NL Centres and other institutes that might play a role in the data curation effort.

1. Welk soort resources worden door het instituut beheerd?
   Tekst, spraak, multimodal?
2. Wat is tot dusver het beleid t.a.v. curatie? Wat wordt er gecureerd en op welk niveau?
   bv
    a. is curatie vnl. gericht op het garanderen van toegankelijkheid voor de eigen onderzoekers binnen het eigen instituut of eventueel ook naar buiten waarbij de eigen praktijken voorop staan (evt. idiosyncratische keuzes t.a.v. formaten en tools)
        i. Gaat het daarbij om de curatie van data en tools, of alleen data? (Tools: welke tools evt. wel, welke zeker niet; Data: alleen primaire data of ook afgeleide data?)
        ii. Slaagt men er inderdaad in alle resources veilig te stellen? Zo niet, wat gaat er dan mis?
        iii. Is er sprake van enigerlei prioritering? Zo ja, waar is die dan op gebaseerd?
    b. of omvat curatie ook, of misschien wel juist, het overgaan op CLARIN standaarden en best practices? Zo ja,
        i. Wat zijn daarbij de stappen die gezet moeten worden?
        ii. Wat is de effort die daarmee gemoeid is?
        iii. Beschikt het instituut over alle benodigde know how en faciliteiten? Zo nee,
            1. Welke bottlenecks zijn er?
            2. Hoe tracht men die op te lossen? (bv samenwerking met andere instituten, uitbesteding …)
        iv. Gaat het in principe om alle resources?
            1. Zo ja, wat bepaalt de volgorde waarin resources worden gecureerd?
            2. Zo nee, welke niet en waarom niet?
3. In het kader van de CLARIN Calls is het X projectvoorstel ingediend dat beoogde Y te cureren.
    a. Wat is de reden geweest om dit voorstel in te dienen?
    b. Wat heeft de keuze voor de te cureren resource(s) bepaald?
    c. Wat zijn de ervaringen? Bv
        i. curatie verloopt spoedig, levert weinig verrassingen (vanwege bekendheid met de resource)
        ii. bepaalde aspecten zijn meer problematisch/tijdrovend (of anderszins?) dan aanvankelijk gedacht

    welke conclusies verbindt men hieraan?
4. In het kader van CLARIN-NL wordt overwogen om een data curatie service op te zetten:
   This service will make it possible to carry out curation for data residing at humanities researchers by a dedicated team of specialists and with minimal support from the original researcher who created/owns/manages the data.
   Data curatie wordt in deze context geduid als
   (1) het converteren naar een format dat door CLARIN wordt ondersteund
   (2) het creëren van metadata conform CMDI
   (3) de opslag van data en de bijbehorende metadata op een server van een CLARIN-Centre

(4) toewijzing van permanent ID (PID)s aan de data en metadata (door het betrokken CLARIN-centre)

a. Bent u voorstander van dit idee?
b. Gegeven het feit dat het hier gaat om resources van anderen
    i. is het reëel om te verwachten de curatie met minimale betrokkenheid van de oorspronkelijke eigenaar/producent verloopt?
    ii. wat is een absolute vereiste om curatie succesvol te kunnen laten verlopen?
c. Welke rol zou dit instituut of specifieke personen hierin kunnen spelen? Bv gegeven
    i. expertise
    ii. faciliteiten
    iii. capaciteit
d. Heeft u nog specifieke suggesties? Bv t.a.v.
    i. De opzet/inrichting van een data curatie service
    ii. Welke resources met voorrang gecureerd zouden moeten worden (bv vooral/uitsluitende primaire data, data uit specifieke disciplines, data die veel gebruikt worden, …)?
    iii. Hoe een reële inschatting te maken van wat de (diverse aspecten van) curatie zouden moeten/mogen kosten?

5. Andere opmerkingen?