Introducing the CLARIN-NL Data Curation Service

Henk van den Heuvel and Nelleke Oostdijk
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

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 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.

From the start of the project (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 Centres (Odijk 2010).

The pilot project was carried out between 1 November 2010 and 1 February 2011. In order 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, 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;
- 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.

On the basis of the report summarizing the main findings, CLARIN-NL in September 2011 decided to establish the Data Curation Service (DCS) at CLST in Nijmegen. 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 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 will form the core of the work to be undertaken by the DCS. The DCS is fully operational since January 2012 and has funding until the end of 2013.

A more elaborated view on the tasks of the DCS is given in the figure below:

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 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 (minor) 8. Convert to a CLARIN preferred format 9. Assign appropriate metadata 10. Ensure semantic interoperability 11. Provide documentation
Task D. Validation	
Actions	12. Validate curated resource
Task E. Archiving	
Actions	13. Transfer to CLARIN Data Centre for long-term storage and maintenance 14. Assign persistent identifier(s) 15. Provide access to content

Our contribution will consist of two parts. In the first part, the main findings of the pilot project will be presented focusing 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 the second part we report on experiences with the curation (cf. the scheme below) by the DCS of various data collections, e.g. the Dutch Bilingual Database (DBD), the Low Educated Second Language and Literacy Acquisition (LESLLA) corpus by Ineke van de Craats, the IPNV interviews with veterans, and dialect dictionaries (Woordenboek Gelderse Dialecten in particular).

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
DCC Curation Lifecycle Model. Retrieved from http://www.dcc.ac.uk/resources/curation-lifecycle-model.
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.

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.