The evolving infrastructure for language resources and the role for data scientists

Nelleke Oostdijk, Henk van den Heuvel

CLS / Centre for Language and Speech Technology
Radboud University Nijmegen, The Netherlands

E-mail: {N.Oostdijk|H.vandenHeuvel}@let.ru.nl

Abstract

In the context of ongoing developments as regards the creation of a sustainable, interoperable language resource infrastructure and spreading ideas of the need for open access, not only of research publications but also of the underlying data, various issues present themselves which require that different stakeholders reconsider their positions. In the present paper we relate the experiences from the CLARIN-NL data curation service (DCS) over the two years that it has been operational, and the future role we envisage for expertise centres like the DCS in the evolving infrastructure.

Keywords: language resources; sustainable infrastructure; data curation; research data management

1. Introduction

Over the past few decades the landscape of European scientific research has changed considerably. Since the onset of the digital age the possibilities for data capture and storage but also access and exchange have increased immensely. In fact, what we can observe is nothing short of a landslide that has upset research mores and will undoubtedly bring about further changes before it comes to a halt. In the past, research projects were typically isolated enterprises of individual researchers or research groups who would concern themselves with the collection of the necessary data when the need arose. There was little sharing of resources. In this context standardization efforts were at best local. However, over the years, in a changed research climate, we see that collaborative research replaces the scattered individual efforts of the past. Parallel to this development, resource development and maintenance has caught the attention of the wider research community who has come to realize the potential impact that the sharing and re-use of data and tools has on everyday practice. As resources underlying the research become available, research results can quite easily be validated as research can be replicated and the results obtained verified. Moreover, research may be accelerated as it can continue from the point where previous research stopped.

Through the years we see various initiatives (e.g. the Text Encoding Initiative (TEI; http://www.tei-c.org) and the Expert Advisory Group on Language Engineering Standards (EAGLES; http://www.ilc.cnr.it/EAGLES96/home.html) that aim for the development of standards of various kinds, for example, for text encoding (TEI Guidelines for Electronic Text Encoding and Interchange, (X)CES) and metadata (e.g. ISLE/IMDI). The foundation of ELRA in 1995 can be viewed as a landmark signalling the importance attached to the shared use of resources. The establishment of ELRA/ELDA is a first step towards a sustainable infrastructure for language resources. Since then we have come a long way. In Europe, national governments and research organizations as well as the European Union have put language resources high up on their agendas. Roadmaps have been developed and plans have been implemented to create a set of basic language resources for each of the languages. Driven by a vision of an infrastructure that will include increasingly more as well as more diverse language resources (cf. the Strategic research agenda developed by META and FLaReNET's Strategic Language Resource Agenda (Calzolari et al.)), initiatives such as CLARIN (http://www.clarin.eu), META-SHARE (http://www.meta-share.eu), and EU-DAT (http://www.eudat.eu) are under way that aim for the implementation of an open, sustainable, interoperable infrastructure for language resources.

2. Language resources and LR infrastructure in the Netherlands

In the Netherlands, resource development received a boost from the STEVIN programme (Spyns & Odijk, 2013) in which one of the aims was to fill the remaining gaps in the basic language resources for Dutch. As a result, today Dutch is one of the languages for which there is a fair coverage of basic language resources for a diversity of research areas and applications. The Dutch HLT Centre (TST-Centrale; http://www.tst-centrale.org) was established as a national centre charged with the maintenance and distribution of Dutch language resources. In line with developments we see at the European level, in CLARIN-NL (Odijk 2010) substantial

efforts are made to contribute towards the development of an infrastructure that will support the sharing and re-use of resources, and that will open up new avenues of research as it allows for combining various resources in new and unforeseen ways. Apart from work on the implementation of the technical part of the infrastructure, there are several resource curation and/or demonstration projects which should bring this infrastructure to life and promote its actual use. The data curation service (DCS) hosted at the Centre for Language and Speech Technology in Nijmegen is a centre of expertise set up to assist researchers, especially those without the time, money, or know-how, in preparing their data for delivery to one of the CLARIN centres that operate as hubs in the CLARIN infrastructure (Oostdijk & van den Heuvel, 2012). Data curation involves (where necessary) digitizing data, converting the data so as to conform to CLARIN accepted standards or preferred formats, providing metadata and documentation. The DCS is therefore the intermediary between the researcher and the eventual data centre.

3. The CLARIN-NL DCS

In the two years that the CLARIN-NL DCS has now been operational, its focus has been on the curation of data collections residing with and used by individual researchers or research groups in the Netherlands. Candidates for curation were identified and for each it was assessed as to (1) whether it was desirable to have the resource curated and (2) whether successful curation was feasible (a more elaborate description of how these criteria can be operationalized is given in Oostdijk et al. (2013)). On the basis of this assessment a motivated decision could be made as to whether or not to proceed with the curation.

3.1 Resources curated

Most of the data collections targeted by the DCS are collections that were compiled in projects that were already finished and of which many did not receive any follow up, so that in effect the data were at risk of being lost. Curation of such collections can be challenging, especially when they were created in a context where little or no thought was given to the idea of sharing or re-use. Often IPR has not been settled or if it has, the arrangements did not anticipate the distribution or wider use of the data. Typically data formats are diverse, metadata and documentation incomplete. Since settling IPR for already existing collections was deemed problematic, the DCS has refrained from taking on the curation of resources for which any IPR issues remained to be settled.

Among the resources curated by the DCS are the Dutch Bilingual Database (DBD; Oostdijk et al., 2013), LESLLA, a database comprising acquisition data for Dutch as a second language (Sanders et al., 2014), the IPNV database containing interviews with Dutch veterans (Van den Heuvel et al., 2012), and as many as ten dialect dictionaries from various parts of the Netherlands. An overview of curated databases can be found on the DCS website at CLARIN-NL (http://www.clarin.nl/node/414) together with the corresponding curation reports.

These resources were selected for curation for several reasons. They were thought to be of interest to a fairly large number of researchers. Moreover, they represent rather diverse types of resources intended for and used by different user groups and research communities, including dialectologists, researchers interested in language acquisition, and oral historians. Finally, each of these presented a test case for the developing infrastructure in terms of data formats, and metadata/interoperability.

Curation of the resources involved various actions which can be summarized as follows:

- Data collection: obtaining and agreeing upon the complete and final set of data;
- Conversion of data formats into standard formats of CLARIN;
- Anonymization of the data; this was typically done in transcriptions, metadata and file names;
- Finding an appropriate CMDI metadata profile (http://www.clarin.eu/cmdi) and modifying it where needed;
- Filling the metadata profile with the metadata belonging to the database;
- Writing documentation and the curation report;
- Packaging and delivery at a CLARIN data centre. The data centre takes care of adding persistent identifiers and storage of the curated database.

The curation of the resources yielded several beneficial results. Thus we managed to salvage several resources that would otherwise have gone to waste. Obviously, the curated data contributed to filling the infrastructure with a variety of relevant databases. We also found that as researchers were involved in the curation process and they and others from the same research community began to see what possible impact the sharing of research data could have for them, this had a very positive effect and led to increasingly more researchers becoming engaged. An example here is what occurred when we started with the curation of dialect dictionaries. At first, this was limited to four databases but then other researchers came and offered their dialect dictionaries for curation as well. Another result is the feedback/insight we obtained as regards the suitability/usefulness etc. of various standards and formats (LMF, CMDI, ISOCAT, …).

3.2 Lessons learned

Apart from experience in identifying, say formal, problems with adopted formats and implementations, our experiences at the DCS have brought us a number of insights. Firstly, staying in contact with the researcher is of paramount importance for understanding the data. Secondly, the time needed for this interaction should not be underestimated. Substantial efforts are involved in obtaining the data, that is, the final version of the data and documentation accompanying them, especially if more than one researcher has worked on the collection of the

data. Furthermore, interpreting and linking data and metadata should be done involving where possible the researcher, who, understandably, is not at all times available.

Another lesson learned is that IPR issues must be cleared at a very early stage. It is an absolute waste of time and money to enter into a curation enterprise for a database for which an IPR agreement was signed stating that the data may be used for a particular research project and must be destroyed one year after the project end date (to mention just one example).

With respect to CMDI metadata profiles we have come to the conclusion that it is best to publish a new CMDI profile for each database at project level by selecting and constructing CMDI building blocks from selected other profiles (and introduce one or more new metadata categories). One will never be able to publish an all encompassing CMDI profile covering all databases of a similar type (e.g. second language acquisition), since the variety of encountered metadata is vast, and the overall profile will never be complete.

So far the DCS has focused on existing collections which means that most of its efforts have been directed at trying to make the resources conform to the preferred formats, allowing for their integration in the larger CLARIN infrastructure and the application of various services offered within this infrastructure. Thus one could say that the DCS has been working on a backlog of resources that were created in the past. In the near future, however, we expect the task of the DCS to change in the light of the evolving vision of an infrastructure for language resources.

4. Future perspective

4.1 Developments, stakeholders and their positions

So far data sharing has been much more common in the field of the natural sciences (see e.g. http://www.3tu.nl/datacentrum/en/) than in the humanities In the natural sciences immense data sets have been collected and are commonly shared by everyone, as there is simply too much data for any one research group to deal with. Data sets often serve as reference sets (also e.g. in computational linguistics). By contrast, in the humanities typically we find very many and, in comparison to the data collections used in the field of the natural sciences, quite small and very diverse data collections. These are often created with huge effort and personal involvement on the part of the researcher or research group who took the initiative for the collection. Data sharing then for humanities researchers is not something they necessarily warm to easily, although rationally they may see the idea making sense.

As with the new infrastructure the doors are opening to vast amounts of data, various stakeholders need to reconsider their positions: there is the individual research or research group, the wider research community, funding agencies, universities, university libraries and possibly others.

At the forefront of data salvation, we find the Digital Curation Centre (DCC; http://www.dcc.ac.uk/) established in the UK in 2005 and operational since then. The DCC is very active when it comes to developing and implementing procedures, plans and policies that will support research data management and sharing. They also charted the current policies of UK funding bodies as regards how to warrant data preservation and accessibility, and have found that funding bodies in the UK nowadays increasingly require from grant-holders a data management and sharing plan. An overview of the data policies adopted by various funders shows that what is currently expected as regards a data management plan varies quite a bit. However, a common denominator appears to be that such plans "typically state what data will be created and how, and outline the plans for sharing and preservation, noting what is appropriate given the nature of the data and any restrictions that may need to be applied." Subsequently, the DCC has created a template for a data management plan (Donnelly and Jones, 2009) that researchers may find extremely helpful.

Meanwhile, in the Netherlands, NWO, the national research foundation is developing and implementing a similar policy. While in the past with certain types of grants (e.g. investment grant) it was required that you specify where the resource would be deposited once it was completed. Usually a sentence stating that it would be archived for example with DANS (Data Archiving and Network Services) would suffice. More recently, grant proposals require a paragraph on data management, and in future we undoubtedly can expect to have to submit a full-fledged data management plan.

These days universities strain under the pressure of economic cuts and find themselves in a position where they become more and more dependent on external funding. However, competition is steep and in many cases research grants supplied by funding organizations require matching funds from the universities with which the researchers are affiliated. As financial resources are limited, university bodies have to make choices as regards what they want to allocate the matching funds to. Procedures have been or are being put in place for the vetting of research proposals before they may be submitted. At the same time universities are also addressing issues such as ethics in research and scientific integrity where data sharing for verification purposes finds a legitimate basis.

University libraries are redefining their position and are looking into what their future role could be. Over the past years university libraries already took up the challenge to create and maintain repositories in which the academic publications are collected and made available, usually offering free access in line with the policy of promoting open access publications. More recently the

idea is gaining ground that with the possibilities offered by the technological developments and the infrastructure that is beginning to take on shape, it should be possible not only to have access to the research publications, but also to the data underlying them. The university library appears to be extremely well-suited as one of the points of entrance for researchers looking for literature, existing data or assistance with the management of their research data. Ideally researchers consult the library at the start of their project, and the library can assist in carrying out the data management plan (DMP), e.g. by referring them to an expertise centre like the DCS (see 4.3).

Researchers, obviously, are essential stakeholders in these developments. Their work is at the basis of any DMP. This means that whenever they envisage the creation of a resource, their plans should describe not only what kind of resource will be created (with attention for the design, data collection and annotation, formats, IPR, etc.), but also how they envisage the resource can be stored and made accessible for others. To require this extra effort from researchers will fail if researchers do not see the benefit of data sharing. Their benefit may reside in the principle of scientific integrity (replication and verification of research), but also in more tangible results such as the first right of publication for the individual researcher or research group responsible for the creation of the resource, or in the official assignment of an ISBN number associated with the data set so that the data set itself counts as a publication.

In this new data landscape ELRA/ELDA as traditional stakeholder should find a role for themselves, too. For ELRA it is important to include these language resources into its Universal Catalogue (http://www.elra.info/Universal-Catalogue.html), so that the resources can be retrieved via the ELRA search portal. As we are dealing typically with academic resources funded by public money, it seems implausible that ELRA can set a price on re-use of such resources by academic members. However, for commercial parties ELRA could negotiate with the database owners (maybe through the libraries) on licenses for commercial use. In this way ELRA can fulfil its broker role for both academic and commercial parties.

4.3 Expertise centres

Ideally, researchers can be held responsible for the data from the point of creation up to the point where the resource can be delivered to a data center where the resource can be persistently stored and accessed via webportals containing aggregated metadata. It is important to keep in mind that the effort required for making data available to the wider research community should be proportionate, i.e. it should be born in mind that the core business of the researcher is to conduct research, and can only devote limited time and effort to data curation. Therefore, it is not to be expected that (all) researchers can carry out the complete data preparation of their resources up to inclusion in the data centres themselves. Expertise centres like the DCS will remain indispensable in the years to come.

Part of the funding for setting up and maintaining such expertise centres will need to come from national or international funding bodies via (granted) research proposals. As observed above, research proposals in the future will be required to contain a data management plan specifying the design of the resource, procedures for data acquisition, data formats, ethic and legal arrangements, etc. The set-up and execution of such a plan can be (partly) subcontracted to one of the expertise centres which will offer various services to researchers developing and implementing their data management plans. In the expertise centres, data scientists, technical staff, and documentalists are available. In our view the university library will act as a front office where researchers can turn to with their questions.

The expertise centre will act as a back-office and

- assist researchers in drawing up data management plan;
- advise on licenses both for data acquisition and for data use by the end-users;
- provide information on standards and best practices, guidelines, etc.;
- give support to researchers as regards delivery of the resource to the repository with which the data will be archived.

Where relevant, the centre will refer researchers to other (national or international) centres of expertise, for example for having their resources validated.

5. Acknowledgment

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.

6. References

Calzolari, N., V. Quochi & C. Soria. *The Strategic Language Resource Agenda*. http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf. Retrieval date: 20 March 2014.

Donnelly, M. & S. Jones. *Template for a Data Management Plan*. Digital Curation Centre, 2009.

Meta Technology Council. 2012. *The META_NET Strategic Research Agenda for Multilingual Europe* (Version 0.9). Retrieved from http://www.meta-net.eu/vision/reports/meta-net-sra-version_0.9.pdf. Retrieval date: 20 March 2014.

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. & H. van den Heuvel. 2012. Introducing the CLARIN-NL Data Curation Service. In *Proceedings of the Workshop Challenges in the management of large corpora*. *LREC2012,* Istanbul, 22 May 2012. http://www.lrec-conf.org/proceedings/lrec2012/index.html. Retrieval date: 20 March 2014.

Oostdijk, N., H. van den Heuvel & M. Treurniet. 2013. The CLARIN-NL Data Curation Service: Bringing Data to the Foreground. *The International Journal of Digital Curation,* Vol. 8, Issue 2, 134-145.

Sanders, E., Van de Craats, I. & V. de Lint. 2014. The Dutch LESLLA Corpus In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC-2014, Reykjavik.*

Spyns, P. & J. Odijk. 2013. *Essential Speech and Language Technology for Dutch. Results by the STEVIN programme.* Springer.

Van den Heuvel, H., Sanders, E., Rutten, R. & S. Scagliola. 2012. An Oral History Annotation Tool for INTER-VIEWs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC-2012*, Istanbul, Turkey.