Sharing Resources in CLARIN-NL

Jan Odijk
Utrecht University
Trans 10, 3512 JK Utrecht,
The Netherlands
j.odijk@uu.nl

Arjan van Hessen
Twente and Utrecht Universities
Trans 10, 3512 JK Utrecht,
The Netherlands
A.J.vanhessen@uu.nl

Abstract

Sharing resources in a systematic way is essential for conducting high quality scientific research but it imposes requirements on the *documentation*, *visibility*, *referability*, *accessibility*, and *long term preservation* of these resources. Sharing resources only makes sense when others can actually use them, which imposes requirements of *interoperability* on resources. In this paper we describe how the CLARIN-NL project addresses these issues in order to maximize sharing of resources. We submit that the approach taken in CLARIN-NL is an exemplary approach that deserves adoption by other research communities, possibly slightly adapted to their own needs and requirements.

1 Introduction

Sharing resources in a systematic way is essential for conducting high quality research but imposes requirements on the *documentation*, *visibility*, *referability*, *accessibility*, and *long term preservation* of these resources. Sharing resources only makes sense when others can actually use them, which imposes requirements of *interoperability* on resources. We understand the notion *resources* here in a broad sense, including not only data, but also software, including applications and web services. In this paper we describe how the CLARIN-NL project addresses these issues in order to maximize sharing of resources. We submit that the approach taken in CLARIN-NL is an exemplary approach that deserves adoption by other research communities, possibly slightly adapted to their own needs and requirements.

This paper is organized as follows. We first briefly discuss the CLARIN-NL project (§2) and some of the subprojects and activities relevant to sharing resources it undertakes. Next we discuss each of the requirements for optimal sharing, and how they are worked on in the CLARIN-NL project: documentation (§3), visibility (§4), referability (§5), accessibility (§6), long term preservation (§7), and interoperability (§8). We end the paper with our conclusions (§9).

2 The CLARIN-NL Project

The CLARIN-NL project[1] (Odijk 2010) is a national project in the Netherlands that aims to design, construct, validate, and exploit a research infrastructure that is needed to provide a sustainable and persistent eScience working environment for researchers in the Humanities, and Linguistics in particular, who want to make use of language resources and technology for their research. The targeted users include researchers and developers of Human Language Technology (HLT), since they are largely part of the humanities in the Netherlands. The *use* of HLT will play an important role in the CLARIN infrastructure, but this infrastructure is not specifically dedicated to *research* into and *development* of HLT. This is one of the characteristics distinguishing the CLARIN infrastructure from e.g. META-SHARE (Piperidis 2010), the resource exchange facility being constructed in the context of the META-NET project.[2]

Since the targeted users are humanities researchers, the character of the resources differs widely, but their common denominator is that they have a language component. The data resources include dictionaries, text corpora, linguistic databases, audio and video containing speech, in a wide variety of languages, images of historical manuscripts, their transcriptions and annotations. The software resources include lan-

[1] www.clarin.nl/
[2] www.meta-net.eu/

Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, pages 98–106,
Chiang Mai, Thailand, November 12, 2011.

guage technology software for spelling normalization, morphological analysis, lemmatization, PoS-tagging, chunking, parsing, semantic annotation, named entity recognition, sentiment and opinion mining. On the speech side they include speech recognition software for transcribing speech or aligning speech with a transcript, diarisation software for isolating speech from non-speech sounds in an audio file (e.g. as part of an tool for annotating audio/video files created during linguistic field work). They also include a wide range of tools for manually annotating texts, audio and video.

The CLARIN-NL project is part of a Europe-wide enterprise to set up an infrastructure. This was initiated by the just finished CLARIN preparatory project (CLARIN-prep[3]) and is to be continued by a consortium of national projects united at the European level in the so-called CLARIN ERIC[4] expected to start early 2012. The Netherlands played an important role in CLARIN-prep, and the CLARIN ERIC is hosted by the Netherlands.

In the remainder of this section we describe the activities organized by CLARIN-NL that are relevant to the topic of sharing resources.[5]

2.1 Infrastructure implementation

CLARIN-NL will build the infrastructure through so-called CLARIN-Centres. Five organisations have expressed the ambition and the commitment to become such a CLARIN Centre, i.e. INL[6], MPI[7], MI[8], Huygens ING[9] and DANS[10]. They are all organizations that include making resources accessible in their mission. Candidate CLARIN Centres must meet several requirements before they will be recognized as actual CLARIN Centres. Several of these requirements will be described in this paper. A full list can be found in (Roorda et al. 2010).

The CLARIN Centres in the Netherlands work together in a number of projects to *implement the technical infrastructure*. This requires, inter alia, setting up authentication and authorizations systems, several registries, and various other infrastructure services. Especially relevant for sharing resources is the project to implement sophisticated *search facilities* in metadata and data to complement the browsing functionality for which a prototype (the Virtual Language Observatory, VLO[11]) was developed in CLARIN-prep.

2.2 Data curation projects

CLARIN-NL has set up a range of *data curation* projects, and will set up more in the course of 2011. The goal of a data curation project is to adapt an existing data set in such a way that it becomes properly documented, visible, uniquely referable and accessible via the CLARIN infrastructure. In addition, the format of the resource must be adapted to a standard supported in CLARIN, and the data categories used must be described in a data category registry. In short, these projects are aimed at making it optimally possible and useful to share the resource with other researchers.

In order to speed up the process of data curation and in order to include resources where the owner/researcher does not wish to submit a project proposal or the resource is too small to justify a data curation project, a Data Curation Service is being set up and targeted to start in September 2011.[12]

2.3 Demonstrator projects

CLARIN-NL has also set up a range of *demonstrator* projects. The goal of a demonstrator project is to create a documented web application starting from an existing tool or application that can be used as a demonstrator and function as a showcase of the functionality that CLARIN will offer. Though the main goal is to make a demonstrator, in practice it requires curating the tool or application, so that it becomes properly documented, visible, uniquely referable and accessible via the CLARIN infrastructure, and adapting it to work with CLARIN-supported standards both with regard to formats as well as with regard to the meaning of the data categories used.

In a collaborative project with Flanders the focus is even more on curating the tools and applications. In this project, existing language and speech technology tools for the Dutch language (shared between the Netherlands and Flanders), which were largely developed in the STEVIN programme[13], are turned into web services that

[3] www.clarin.eu
[4] ec.europa.eu/research/infrastructures/index_en.cfm?-pg=eric
[5] See www.clarin.nl/node/76 for a more detailed overview.
[6] Institute for Dutch Lexicology www.inl.nl
[7] Max Planck Institute for Psycholinguistics www.mpi.nl
[8] Meertens Institute www.meertens.knaw.nl/
[9] www.huygensinstituut.knaw.nl/
[10] Data Archiving and Networked Services www.dans.knaw.nl/
[11] www.clarin.eu/vlo/
[12] www.clarin.nl/node/147
[13] taalunieversum.org/taal/technologie/stevin/

can be used in a workflow system. This is only possible if the web services are properly documented, visible, uniquely referable and accessible via the CLARIN infrastructure, and if they operate on formats and work with data categories that are supported in CLARIN.

In short, these projects contribute directly to optimal sharing of tools and applications with other researchers in the CLARIN infrastructure.

2.4 Education and Training

Adapting resources so that they become documented, visible, uniquely referable and accessible, and comply with CLARIN-supported standards both on the formal and on the semantic level is a non-trivial task. The average humanities researcher does not have the knowledge and expertise to carry out such tasks completely independently. Therefore, education, training and support are needed. CLARIN-NL has organized various tutorials and workshops on relevant topics such as metadata and the CLARIN metadata infrastructure, and data categories and data category registries. It has set up a HelpDesk[14] to deal with technical questions on infrastructural matters, including a Frequently Asked Questions section, and appointed infrastructure specialists as second-line support.

3 Documentation

The first step in making a resource suited for sharing is to provide documentation of the resource. Even if a resource is not going to be shared, documenting it is required to guarantee that the resource can still be understood long after its development. So, documentation is a necessity for sharing but requires no or only limited additional effort.

Some parts of the documentation will have to consist of natural language text that is intended for human beings, for example a description of the design decisions in developing the resource. However, other parts of the documentation can be formalized. For example, certain properties of a resource can be systematically assigned to a fixed label (attribute), the possible values of each attribute can be characterized by a type, and in some cases the possible values of an attribute can even be restricted to a finite set or be constrained otherwise (e.g. by a template). In our view, all information of the documentation of a resource that can be formalized should be formalized, since a formalized representation encodes the information in the least ambiguous way (natural language is notorious for its ambiguity), and maximizes the potential for use of this information by software processes. Furthermore, this formalized documentation should be represented in a uniform manner. In CLARIN-NL, we have used and further extended the CLARIN Component-based Metadata Infrastructure (CMDI) originally developed in CLARIN-prep (§3.2).

3.1 Metadata

The term *metadata* is on the one hand very broad. Within a dataset usually a part can be characterized as the "primary data", and metadata then covers all data except the primary data, including annotations, formalized documentation, unformalized documentation, aggregate statistics on the resource, etc. That is such a broad notion, that it may hamper mutual understanding. On the other hand, the name *metadata (*lit. 'data about data'*)* suggests too narrow an interpretation, since we also need documentation (formalized and unformalized) for software. We will therefore try to avoid the term *metadata* here.

We assume that each dataset contains a set of "primary data" and a set of additional data with information on the primary data (which we will call *annotations*). Certain pieces of primary data and annotations form a natural unit (following in part from the nature of the data and/or the purposes of the data). We will call such a unit a *resource*. Multiple resources can be organized in composite resources recursively. A description of a resource is called a *resource description*. CMDI has mainly been developed for the formalized parts of resource descriptions. The term *resource description* is also more appropriate than the term *metadata* for resources that consist of software (e.g. applications, web services, command line tools, etc.)

3.2 CMDI

CMDI [15] is a flexible metadata infrastructure which enables the researcher to use a component-based approach to resource descriptions. Because it is component-based, it does not require a single rigid scheme, something that is not feasible given the wide variety of resources CLARIN-NL has to deal with. The meaning of the resource description elements and its values is encoded by linking the data categories used to

[14] trac.clarin.nl/trac

[15] www.clarin.eu/cmdi

a data category registry, which will be discussed in more detail in §8.2.

CMDI enables the researcher to make a resource description *profile* for a class of resources. Such a profile is composed of *components* recursively. This makes it possible to define small components that can be reused easily and provides the required flexibility for making resource descriptions while at the same time maximizing uniformity where this is possible. CMDI provides editors for components, profiles and resource descriptions, and a registry for storing new instances of such objects and finding existing ones for reuse.

CMDI has metadata elements that correspond to the Dublin Core[16] metadata elements also in use by OLAC[17] and is therefore fully compatible with Dublin Core but it allows for much more fine-grained metadata descriptions.

Providing flexibility entails the danger that different researchers will diverge in making resource descriptions even when there is no reason to do so, e.g. because they are working on resource descriptions independently. In order to prevent this and to offer maximum opportunities for reuse of profiles and components, CLARIN-NL started a project with a small team of specialists to make initial components and profiles for a wide variety of resources in the Netherlands. The researchers in the data curation and demonstrator projects, which started later, could therefore maximally reuse components and profiles created by this specialist team and optimally profit from the knowledge and expertise gained by this team. Unfortunately, such components and profiles were made only for data, not for software. So, a set of components and profiles that can be reused for describing software is urgently needed, as was clear from several data curation and demonstrator projects. A project to do exactly that is therefore planned for 2011.

Creating resource descriptions in accordance with CMDI for each relevant resource was a requirement for data curation and demonstrator projects. Therefore, an initial obstacle for sharing these resources in the CLARIN infrastructure has been overcome. The Data Curation Service will increase the number of resources with proper resource descriptions, and we already noticed that research projects unrelated to CLARIN-NL as well as several data providers are willing to provide CLARIN compatible resource descriptions for data they produce and/or make available.

4 Visibility

All resources and resource descriptions dealt with in a CLARIN-NL project must be stored on a server of a CLARIN-centre. CLARIN-centres are obliged to make the resource descriptions for these resources and for resources they have available from other sources available for harvesting (using a standardized protocol, OAI-PMH[18]). In the CLARIN infrastructure all resource descriptions are harvested regularly and made available via a central CLARIN portal. This ensures the *visibility* of the resources and the resource descriptions. Researchers only have to visit the CLARIN portal to find the resources they are looking for and are not dependent anymore of knowledge about resources via informal contacts, accidental encounters or effort-wasting search actions via the web or systematic visits of the catalogues of resource distribution centres.

The CLARIN portal will offer various opportunities for finding the resources one is interested in. This includes browsing facilities with faceted browsing, of which a first prototype developed in the CLARIN preparatory project is available (VLO, see above). It also includes facilities to search in the resource descriptions, not only with a Google-style string search but also with structured search that takes into account the resource description XML syntax and the semantics of the resource description elements and their values. It also includes search in the actual resources. However, the actual resources will be distributed over the various CLARIN-centres. Searching in the actual resources will therefore be carried out via *federated search*. Results of search queries can be collected and stored as a *Virtual Collection*, to which new, possibly more refined search queries can be applied.

Many CLARIN-supported standard formats for written resources consist of tagged text (e.g. XML). Searching in many (tagged) textual resources is generally not possible with computers in a reasonable amount of time. This problem will not disappear when computers are increasing in capacity every two years (as *Moore's Law*[19] appears to implicate), since (1) many problems are inherently intractable and solutions can only be approximated, and (2) the amount of data grows at least as fast and very likely orders of

[16] http://dublincore.org/
[17] http://www.language-archives.org/
[18] www.openarchives.org/OAI/openarchivesprotocol.html
[19] en.wikipedia.org/wiki/Moore's_law

magnitude faster. So even though Moore's Law may be true, it is also true and much more relevant that computers are slow[20] and getting slower every two years.

Fortunately, smart people have found smart ways to avoid the computer's slowness to a significant extent by a range of techniques. However, this requires storing the information contained in the tagged textual data in special formats in database systems (e.g. relational databases) and/or adding various indexes. In the central portal, the resource descriptions harvested from the various CLARIN centres will therefore also be stored in a way that makes fast searching and browsing possible. For the actual resources, federated search will issue search queries to local search engines for individual resources at the CLARIN-centres, where the local search will also take place on resources formatted and stored in a way that optimizes search.

In this way, visibility of the resources and their resource descriptions will be ensured.

5 Referability

There must be a simple way to refer to resources and resource descriptions. This is needed for humans (so that they know exactly which resource or resource description has been used in a particular research project), but also for machines. The search engines mentioned in the preceding section cannot work properly if they have no way of uniquely referring to resources and resource descriptions.

Natural Language One way of referring to a resource is by using a name or title for a resource in natural language (e.g. the title of a novel, article, etc.). This method is not suited for the purposes of CLARIN because it has all the disadvantages that natural language has as a means of communication. First, such names do not always refer to a unique resource (ambiguity). Names are often language-specific (e.g. *Corpus Gesproken Nederlands*), which leads to variants of the name in other languages (e.g. *Spoken Dutch Corpus*) (language-dependency). Furthermore, names and titles are typically long, which is inconvenient. But more importantly, names and titles are highly redundant. A little bit of redundancy is good for communication, but natural language has too much redundancy. This leads to shorter versions of the name (e.g. acronyms such as *CGN*), and to sloppiness with human users: typos (*Spken Dutch Cropus*) or changes in order (*Dutch Spoken Corpus*) are perhaps sometimes intelligible for humans but not (without special software) for computers.

URLs URLs are sometimes used to refer to resources and resource descriptions. URLs avoid most of the problems with natural language descriptions (though they tend to have too much redundancy) and have the additional advantage that they immediately specify where to find the resource. A big disadvantage of URLs, however, is that they are quite unstable and volatile (URLs are often changed or disappear completely).

PIDs What is needed is a means of referring that is not based on natural language, is a short as possible, has at most very little redundancy, and is stable. Persistent Identifiers (PIDs) have been proposed for this, accompanied by services to map from names/titles and/or URLs to PIDs and vice versa (resolution systems). PIDs are usually strings of digits and or letters. Familiar examples are ISBN numbers[21] for books and EAN numbers for products.[22]

A CLARIN-Centre must assign PIDs to the resources and resource descriptions it makes available. In CLARIN (and thus in CLARIN-NL) the preferred PID system is the Handle system[23], since it currently offers the most robust and best performing PID resolution system. Some centres, however, used the URN system[24] already before CLARIN started, and it is being investigated how this can be accommodated in the best way. Furthermore, there are also other PID systems[25] which may have to be accommodated.

The fact that CLARIN centres in the Netherlands assign a PID to each resource and resource description and offer the associated resolution services again take a way an obstacle for optimal and efficient sharing of resources.

6 Accessibility

The CLARIN infrastructure is a virtual web-based distributed infrastructure. The resources and resource descriptions are therefore accessible at virtually any time and from any place (with

[20] Where a computer is "slow" when the user has to wait an unacceptable amount of time for the computer's response. What is "unacceptable" may differ per application or circumstances.

[21] www.isbn-international.org/

[22] en.wikipedia.org/wiki/International_Article_Number_(EAN)

[23] /www.handle.net/

[24] www.w3.org/2001/tag/doc/URNsAndRegistries-50

[25] E.g. the DOI system: www.doi.org/

internet access[26]). Accessibility of the resources and resource descriptions for this aspect of access is therefore taken care of pretty well in CLARIN.

However, there are two other aspects of access: (2) intellectual property rights (IPR) and ethical issues, and (3) the attitude of researchers towards sharing resources.

IPR CLARIN-NL promotes maximal open access of resources. It has issued a declaration on this matter[27] and had discussions about it at various occasions.[28] Important research organizations such as the *Royal Netherlands Academy of Arts and Sciences* (KNAW) and the Dutch foundation for Scientific Research (NWO) also stimulate or even require open access to results of scientific research, esp. data and tools.[29]

CLARIN-NL realizes that there are many legacy data with legacy IPR arrangements that also need to be accommodated in the CLARIN infrastructure. This may involve special licenses, in some cases even license fees, restrictions on the usage of resources, limited periods of inaccessibility of the resource, etc. In all resources, ethical issues (e.g. privacy concerns) may play a role as well, restricting the usage of certain (parts of) resources. Problems of this nature have actually been encountered in the data curation projects. In one case it has led to a clear separation of the resources (and resource descriptions) that are freely accessible on the hand, and the resources and resource descriptions for which additional licenses are required on the other. In a second case, the participants in the resource have been approached again to clear these matters (successfully). CLARIN-NL is working on plans to implement policies and functionality to properly handle IPR and ethical restrictions. For some centres, these could be extensions of existing systems (e.g. DANS has the EASY system[30] and soon its successor EASY II[31]).

Mindset A third aspect related to accessibility is the mindset of researchers. Many researchers in the humanities are hesitant or even unwilling to share their resources with others.[32] There is therefore a big task for CLARIN-NL to discuss these matters, listen carefully what arguments are adduced against sharing resources, counter these arguments where appropriate and promote maximal open access, e.g. by illustrating the great potential offered by sharing resources. In some cases, arguments against sharing must be accommodated (because they are reasonable objections), and CLARIN-NL has done so already in its declaration. CLARIN-NL also supports researchers (logistically, organizationally, financially and by means of training and education) to enable them to share their resources.

In short, CLARIN-NL has developed a range of policies and facilities to maximize *accessibility* of resources and resource descriptions for a range of aspects of this term, thus directly contributing to optimal sharing of resources.

7 Long term preservation

Resources should be shared not only with contemporary researchers, but also with future generation researchers. This makes it necessary to carry out *long term preservation* of resources. In CLARIN-NL, each CLARIN centre is required to provide a solution for the long term preservation of the resources they maintain. Usually the centres in the Netherlands do not carry out this long term preservation themselves but make use of centres dedicated to it. For example, MI outsources this to DANS, and the MPI outsources it to the organization within the Max Planck Gesellschaft dealing with long term preservation.

The requirement for long term preservation of resources imposed on the CLARIN centres thus makes it possible to preserve the resources and share them with future generation researchers.

8 Interoperability

Resources can be used by other researchers only if they are interoperable. Interoperability is thus a necessary condition for resource sharing to be useful.

[26] This might be an obstacle for certain researchers, e.g. descriptive linguists doing field research in remote locations with no internet access. Functionality that enables one to work off-line and replicate off-line data and tools with the on-line CLARIN infrastructure are therefore desirable, and some applications in CLARIN already have this functionality.

[27] www.clarin.nl/system/files/Call%20Open%20-Data%20English%20101018.pdf

[28] For example at the *Open and Persistent Access Panel Discussion* at SDH/NEERI 2011, Vienna, see ztwweb.trans.univie.ac.at/sdh2010/

[29] See www.knaw.nl/Pages/DEF/29/838.bGFuZz1F-Tkc.html for the KNAW and www.nwo.nl/nwohome.-nsf/pages/NWOP_89BBXM_Eng for NWO

[30] https://easy.dans.knaw.nl/dms

[31]

[32] Though I understand from representatives from other disciplines that the Humanities are not unique in this respect.

Interoperability of resources is the ability of resources to seamlessly work together. The need for interoperability of resources is more stringent in CLARIN than in other domains, since the targeted users, humanities researchers, usually do not have the technical skills to make ad-hoc conversions and adaptations to make resources work together. But of course, even for HLT researchers and developers, full interoperability will save a lot of (often duplicated) effort for ad-hoc readjustment of resources to make them interoperable.

Full interoperability is only possible if the resources meet the requirements (1) of formal or *syntactic interoperability* and (2) of meaning or *semantic interoperability*. Projects in CLARIN-NL must attempt to meet these requirements, and report when problems for achieving this arise. In this way we learn about the limitations of various proposed standards and can make proposals to deal with them and make suggestions for improved standards and best practices. We will discuss syntactic and semantic interoperability in more detail in the next subsections,

8.1 Syntactic Interoperability

Syntactic interoperability in CLARIN is the requirement that the formats of data are selected from a limited set of (de facto) standards or best practices supported by CLARIN, and that software tools and applications take input and yield output in these formats. A list of the formats currently supported is provided by CLARIN. [33] Though currently this list is in a fixed document, it is evident that experience is teaching us that the list is incomplete and needs constant refinement and updating.

Applying the recommended standards and best practices is not easy. In many projects we have found that many standards are not fully applicable to existing data and need adaptations. For example, the DUELME database of Dutch multiword expressions (Grégoire 2010a) which was represented in an idiosyncratic format was converted to an XML format in accordance with the Lexical Markup Framework (LMF). [34] But the new representation requires properties that are not covered by LMF and should be considered as candidate extensions to LMF (Grégoire 2010b). Many resources are stored in relational databases or Excel files. No format supported by CLARIN can accommodate such data. The CSV format is mentioned but not explicitly recommended. An XML format implementing (a set of) CSV files using XML markup may have to be developed here. Such a format will also be able to provide facilities for semantic interoperability of such resources not offered by the CSV format.

Nevertheless, the only way to make any progress towards syntactic interoperability is by trying out the supported formats with existing data, learning about their opportunities and limitations, making concrete proposals to deal with these limitations and constructive proposals for extensions and/or adaptations of the standardized format. And this is exactly what CLARIN-NL is doing in a wide variety of projects and for a wide variety of data, including lexical databases, text corpora with various levels of annotations, audio and video data with their annotations, typological and other linguistic databases, and for a variety of tools and applications, *inter alia* data-specific search engines, part-of-speech taggers, lemmatizers, parsers, speech technology tools for recognition, alignment and diarisation, and many others.

Resource descriptions play a crucial role in ensuring syntactic interoperability. The resource description of a data resource should specify, in quite some detail, the format of the resource, and the resource description of a software resource should specify, in quite some detail, which format(s) it accepts as input and which one(s) it yields as output. Such specifications will prevent a non-technical user from applying software to data it is not suited for or warn the users for the limited validity of the results (e.g. textual resources with the wrong character encoding; a desktop speech recognizer applied to telephone speech, etc.)

By actively trying out the recommended standards and best practices for syntactic interoperability CLARIN-NL contributes directly to enabling sharing of resources and it makes the problems that arise with this explicit so that evidence-based recommendations can be made for extensions and adaptations.

8.2 Semantic Interoperability

Semantic interoperability of resources requires explicit semantics of elements in their contents (in the case of data) or interface (in the case of software). In CLARIN, the semantics of elements of resources is limited to the semantics of data categories (DCs). The basic idea is that the semantics of DCs is captured as follows: a privi-

[33] www.clarin.eu/system/files/Standards%20for%20LRT-v6.pdf
[34] www.lexicalmarkupframework.org/

leged data category registry (DCR) is set up containing (inter alia) DCs, unique persistent identifiers for DCs (PIDs), their semantics, a definition, examples and lexicalizations in various languages. The semantics of each data category (DC) used in a specific resource must be specified by mapping this resource-specific DC to a DC from the privileged DCR. This enables every researcher to use resource-specific DCs but at the same time guarantees that different DCs from different resources can be interpreted in the same way, via the DC of the privileged DCR, which acts as a pivot.

In CLARIN, ISOCAT[35] is used as one of the privileged DCRs.[36] In each CLARIN-NL project, all resource-specific data categories must be mapped to ISOCAT DCs, or, when no DC with the right interpretation exists, a new DC must be added to ISOCAT. ISOCAT can incorporate results of independent initiatives for defining DCs, and it actually incorporates a subset of the GOLD ontology[37] for linguistic description.

An example may illustrate how this could be useful. A search engine searching for occurrences of strings that are annotated for the ISOCAT DC *Part of Speech*[38] with as value the ISOCAT DC *noun*[39] will also find occurrences of data with resource-specific DCs *Substantiv*, *Nom* or *ZN*, if these resource-specific DCs have been mapped onto the ISOCAT DC *Noun.*

Achieving semantic interoperability is not easy, and even with the ISOCAT data category registry many problems arise once one really starts doing this. It would require a separate paper to discuss this in more detail, but such problems have been noted, have been discussed in workshops,[40] and for most problems solutions have been proposed in these workshops, including the set-up of a different registry to register relations between DCs, called RELCAT (Windhouwer 2011), and the proposed solutions are currently being tested.

However, one can only encounter such problems, and make progress in solving them, when one actually systematically attempts to achieve semantic interoperability for real resources. That is exactly what is being done in CLARIN-NL, and by doing so, CLARIN-NL contributes to optimizing the use of shared resources.

9 Conclusions

In this paper we have described how the CLARIN-NL project addresses crucial issues for maximizing the sharing of resources. We have described how CLARIN-NL addresses *documentation*, *visibility*, *referability*, *accessibility*, and *long term preservation* of the resources, as well as syntactic and semantic *interoperability*. None of adopted solutions is without problems, but it is only by systematically working on them that any progress can be made on these topics. And that is exactly what is being done in CLARIN-NL. We submit that the approach taken in CLARIN-NL is an exemplary approach that deserves adoption by other research communities, possibly slightly adapted to their own needs and requirements.

Acknowledgments

This work was funded by the NWO CLARIN-NL project. (www.clarin.nl).

References

Nicole Grégoire. 2010a. DuELME: A Dutch Electronic Lexicon of Multiword Expressions. *Journal of Language Resources and Evaluation* 44(1/2), 23-40. DOI 10.1007/s10579-009-9094-z

Nicole Grégoire. 2010b. En Garde Project. The redesign of a Dutch Electronic Lexicon of Multiword Expressions. Presentation held at the workshop *Lexicon Tools en Standards*, August 4, 2010, Max Planck Institute, Nijmegen. [**pdf**]

Jan Odijk. 2010. The CLARIN-NL Project. *Proceedings of LREC 2010*: 48-53. [pdf]

Dirk Roorda *et al*. 2009. CLARIN Centres. CLARIN Document. [pdf].

Stelios Piperidis. 2010. META-SHARE. Presentation held at the LREC 2010 Workshop on Language Technology issues for International Cooperation, Malta, 22 May 2010. [pdf]

Menzo Windhouwer. ISOcat. Presentation at the *Standards Workshop* (NEERI 09), Helsinki, Finland, September 30, 2009. [pdf]

[35] www.isocat.org/

[36] CLARIN supports multiple preferred DCRs if they are complementary. For example, CLARIN supports the use of ISO639 language codes contained in a different DCR (www.sil.org/iso639-3/codes.asp). In CLARIN-NL a project (CLAVAS) has started up to create a common interface to multiple DCRs.

[37] http://linguistics-ontology.org/

[38] More precisely, the ISOCAT DC with PID www.isocat.org/datcat/DC-396

[39] More precisely, the ISOCAT DC with PID www.isocat.org/datcat/DC-1333

[40] For example in the CLARIN Relation Registry Workshop, 8 Jan 2010 (www.isocat.org/2010-RR/) and in the CLARIN-NL ISOCAT Workshop, 21 Sep 2010 (www.isocat.org/2010-ISOcat-status/), both at MPI, Nijmegen.

Menzo Windhouwer. 2011. RELCAT and Friends. Presentation held at the CLARIN-NL ISOCAT Workshop, Utrecht, 5 May 2011. [pdf]