Data Curation

CLARIN aims to develop a sustainable interoperable research infrastructure that supports humanities researchers while addressing their research questions. The infrastructure should make the resources (tools and data) that the researchers may want to use (better) visible and more easily accessible, for example via online methods and by integrating them in appealing new applications.

Whether the infrastructure will be successful in supporting the needs of the humanities research community does not just depend on the functionality it has to offer but also on the resources it gives access to. Thus there is a vast interest in preserving and sharing datasets that individual researchers and research groups have collected and annotated. CLARIN-NL actively supports their curation so that these resources can participate in the CLARIN infrastructure. It does so by funding curation projects, supplying the resource compilers/owners with the means to carry out the necessary actions, and by offering the services of the CLARIN-NL Data Curation Service (DCS). The DCS gives advice and may undertake the curation of a resource when the individual researchers or research groups lack the expertise and/or the means to take the necessary measures to ensure the future availability of language resources they have compiled.

The curation process involves the adaptation of a given resource such that it becomes visible,is suited for interoperability, is uniquely referable and can be accessed via the web. Visibility of the resources is mainly achieved by providing standardized metadata that can be harvested by service providers. In order to ensure the accessibility of the resource, it is assigned a unique persistent identifier (PID) and stored at computers that are accessible through the infrastructure in a format that adheres to best practices.

Data curation projects

Resource curation projects take existing data or tools as a basis and attempt to apply CLARIN-supported standards and best practices to make the data and tools CLARIN-compliant. Examples of such projects can be found elsewhere on the CLARIN-NL website (http://www.clarin.nl/showcases).

A proposal for a data curation project may be submitted in response to a Call for proposals. The project must involve a user, a data provider (DP), a technology provider (TP) and an infrastructure specialist (IS). The relations between the partners in a project must be agreed upon in a consortium agreement before the start of the project.

The project proposal should clearly describe the research data the DP has at his/her disposal that can be used to address specific research question(s), and how they can be used for this purpose. The research data must be existing digital language or language-related data. No new research data should be created in the project. The data provider must have the right to make the research data available on a CLARIN server running at a dedicated CLARIN centre.

The project proposal should contain a detailed description of the research data, its current state and format, the plans to convert it if needed, justification for using different formats if applicable, and a detailed plan for dealing with the data and metadata. The project proposal should describe all issues related to IPR and present a solution for them. Any restrictions on the use of the data as well as any ethical issues that apply or may arise must be properly documented in the proposal.

Resource curation involves a number of different aspects:

1. The resource should be brought into a format that adheres to widely accepted standards and best practices currently considered as likely candidates by CLARIN.
2. Proper metadata descriptions need to be created and made available. They must be compliant with the CLARIN component metadata infrastructure (CMDI) and it should be possible to harvest and access them.
3. Metadata descriptions should include persistent identifiers that can be resolved and the CLARIN requirements should hold for the PID system.
4. The linguistic encoding must be related to the data category registry, i.e. data categories used must be mapped to corresponding ISOCAT[1] data categories where they exist in a formal way (e.g. via an XML Schema) and new data categories must be added to ISOCAT if they do not exist there yet.
5. Provide proper documentation of the resource, at least in English. The plan for a curation project should describe in detail how these different aspects are going to be/have been addressed in the project.

The results of these aspects should be tested by the project participants. Setting up tests for this should be included in the project plan and the results of these tests will be included in the project's success criteria. Example tests are e.g. a metadata harvesting test and formal procedures such as testing against an XML Schema.

The resulting resource and its metadata must be made available on a server of a recognized CLARIN centre. The project proposal must specify which (candidate) CLARIN centre this will be and concrete arrangements must have been made with this centre.

The experiences gained in the process of curating a resource should be documented. More specifically it should be described what requirements and desiderata have emerged as regards the CLARIN infrastructure.

CLARIN-NL Help desk

The CLARIN-NL Helpdesk can be approached with questions regarding all kinds of technical questions pertaining to for example metadata, data categories, standards, and web services. The helpdesk can be contacted by e-mail: helpdesk@clarin.nl

[1] http::://www.isocat.org/

CLARIN Centres

All resources must be made accessible through a CLARIN Centre. A list of the current (candidate) CLARIN centres and their contact persons can be found on the CLARIN-NL website:
http://www.clarin.nl/node/130

Data Curation Service

The Data Curation Service (DCS) was established in October 2011 and has been operational since January 2012. It aims to contribute the research infrastructure that CLARIN is implementing by salvaging resources and advising on best practices and the use of standards. Set up as a service, the DCS maintains close contact with the research communities as a mediator between these and the CLARIN Data Centres. The DCS prepares resources for archiving at the Data centres, but does not archive any resources itself. Accordingly, the tasks of the DCS are defined as follows:

1. the 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 form the core of the work to be undertaken by the DCS. The DCS is run from the Centre for Language and Speech Technology (CLST) at Radboud University Nijmegen and can be contacted at http://www.clarin.nl/page/about/147 or http://www.ru.nl/letteren/datacuratieservice/

Language data that can be curated include text corpora, lexicons, audio-visual resources containing language and their associated transcripts and annotations. While considering whether the DCS will actually undertake the curation of a particular resource, a number of factors are taken into consideration. These include the resources' relevance to the research community, its uniqueness, and urgency. Especially resources that are at risk of being lost, either because their owners/caretakers are no longer around, or because of the limited lifetime of the carriers that are used to store them or the software and devices needed to retrieve them are no longer available.

Intellectual Property Rights (IPR)

Most resources up for curation involve (huge quantities of) text and text-tools (=software) that were originally written by one or more authors and is owned by these authors or by a legal entity that has bought the rights. In other words: there is always a legal owner of the text and the text cannot be used without solving these legal issues.

Now that a lot of text can be found on the Internet, people often believe that any use of these texts is free just because it is on the Internet. However, this is not the case! Without any official announcement of the owners of the content (text and tools), this content is not free of IPR.

Text and tools used in CLARIN projects will be made publically available on CLARIN servers and therefore all IPR issues need to be solved before a CLARIN project can be granted or the DCS can undertake the curation.

Ownership of original data and software remains with the original owners. If the owners of the original data and software are not identical to the project applicants, the applicants are required and therefore must have the rights to make the research data, the application, its core component, and any run time auxiliary data or software available on a CLARIN server for use by researchers having access to the CLARIN infrastructure. An agreement must be in place between the owners of the original data/software and the project participants on the IPR of the adapted data/software before the submission date of a proposal.

Where the ownership of the created adaptations and extensions is with project applicants, the project applicants have the obligation - and therefore must have the rights - to make the research data, the application, its core component(s), and any run-time auxiliary data or software available on a CLARIN server. This is a sine qua non. Any proposal not satisfying this requirement or being insufficiently clear about this matter will be considered to be formally non-compliant and will be rejected on these grounds.

FAQs

Q: What are the general rules that apply in relation to IPR and Ethical Issues for projects and project proposals submitted in response to a CLARIN-NL Call for proposals ?

A: These are described in the Call Text. See also the CLARIN-NL policy with regard to Open Data and the new NWO Rules and Open Access Policy

Q: The IPR of my resource is arranged in a way as required by CLARIN-NL but not all subjects whose speech or audio occur in it have given explicit permission. Can I put such data on a server of a CLARIN centre?

A: In principle yes, though it may depend a little bit on the policy of the specific CLARIN centre where the data reside, and provided you have a procedure in place so that subjects who object to this can make this known to your organization and measures can be taken to accommodate the subject's objections (e.g. by restricting access to these data or in the extreme case by removal of the relevant data from a CLARIN server). You also will have to describe this procedure in your project deliverables and list its functionality as one of the functionalities that should be offered by the CLARIN infrastructure.

Q: What happens to derivatives created by CLARIN partners that have made use of data for which a subject has requested removal?

A: If such a derivative contains or presents the relevant data in a recognizable way, they are subject to the same measures as applied to the original data (see the answer to the previous question). However, if such a derivative does not contain the relevant data in a recognizable way, CLARIN-NL will support all parties involved in obtaining an agreement that allows the derivative to stay available and maximally accessible in the CLARIN infrastructure.

CLARIN formats and standards

Data

The resource should be brought into a format that adheres to widely accepted standards and best practices currently considered as likely candidates by CLARIN. A list of standards, best practices and

interoperability requirements currently advocated by CLARIN can be found at
http://www.clarin.nl/system/files/Standards%20for%20LRT-v6.pdf

Please note that the list has not yet been completely fixed. There are good reasons for this: (1) there may be crucial data or tools for which none of the currently advocated standards or best practices is suited; (2) the currently advocated standards and best practices may be incomplete, insufficiently specific, inconvenient or even incompatible with crucial data and tools. Where CLARIN currently does not provide a standard that is suitable for sustaining the resource to be curated, you are urged to contact the CLARIN Helpdesk (helpdesk@clarin.nl).

For further information see

URL:	*http://www.clarin.eu/files/standards-text-CLARIN-ShortGuide.pdf*
Title*:*	*Standards for Text Encoding*
Date/Version*:*	*2009-05*
Content:	In this A4 shortguide, some introductory information is provided on Standards for Text Encoding following the generic shortguide layout.

Metadata

Descriptive metadata is used to characterize data resources (and tools) to facilitate discovery and management in large (virtual) infrastructures and repositories, i.e. they make resources visible to everyone. Metadata descriptions must be made for each resource. They should be in accordance with the CLARIN metadata standard (CMDI).[2] CMDI provides a flexible component-based framework for dealing with metadata. Thus you can combine several metadata components (sets of metadata elements) into a self-defined scheme that suits your particular needs. Wherever possible it is advised that you share your profile with others. If sharing the full profile is not an option, you still can use common components, e.g. a component to describe a sound recording. In case that still does not address your needs, it is even possible to create components yourself.

More information about creating components, profiles and using profiles created by others can be found at http://www.clarin.eu/cmdi

Of course data that are being curated may require the development of new CMDI components or the adaptation of existing components and thus can contribute to the further development of the CMDI framework.[3] Any required or desirable extension or modifications of the CMDI framework must be properly documented and be included in the CLARIN Requirements and Desiderata document.

For further information see

URL:	http://www.clarin.eu/files/metadata-CLARIN-ShortGuide.pdf
Title:	Component Metadata

[2] CMDI (Component MetaData Infrastructure) is the CLARIN Component Metadata Framework. See also http://www.clarin.eu/cmdi Login is required for editing.
[3] Existing components can be found in the Metadata Component Registry: http://catalog.clarin.eu/ds/ComponentRegistry/#

Contents: In this A4 shortguide, some introductory information is provided on Component Metadata following the generic shortguide layout of "*What is it?*", "*What is it for?*", "*Who can use it?*", "*When can it be used?*" and "*How does it work?*".

URL: http://www.clarin.nl/system/files/BestPracticeGuide-V4.docx
Title: Best Practice Guide for using CLARIN metadata components
Date/version: 2010
Contents: The Dutch CLARIN project "Creating and using CLARIN metadata components" was the first to actually test the use of components and to try to create metadata descriptions for resources available in two Dutch language resource centers: the Institute for Dutch Lexicology (INL) and the Meertens Institute. This "Best Practice Guide" is the result of this project. It will however in the future be extended with new experiences gained by new projects that will make use of the CMDI.

Arbil

For creating metadata a tool is available: Arbil. It can be found at http://www.lat-mpi.eu/tools/arbil.

An introduction to Arbil can be found here

URL: http://www.mpi.nl/corpus/a4guides/a4-guide-arbil.pdf
Title: ARBIL
Date/Version: 2009-11
Content: This A4 guide features practical information on: "*1. Starting ARBIL*", "*2. Getting your metadata*"", "*3. Changing your metadata*"and "*4. Saving and exporting your metadata*".

Validation: e.g. testing against an XML Schema

Metadata harvesting

Metadata harvesting is a term used for gathering metadata descriptions from several locations and storing it in a central database. You can find the results of such a harvesting process at http://catalog.clarin.eu/ (click on OLAC data providers). While metadata harvesting is the responsibility of the CLARIN Centres, a metadata harvesting test ….

Component Registry

The Component Registry can be found at http://www.clarin.eu/cmdi. For an overview of existing components see http://catalog.clarin.eu/ds/ComponentRegistry/#

ISOCat

ISOcat is a web-based implementation to store and make accessible concepts (a concept registry), more specifically data categories, that are relevant for the CLARIN infrastructure and for encoding linguistic phenomena. Basically it provides a persistent identifier for each data category, and a variety of properties of the data category. It allows one to uniquely refer to a data category (using a PID, e.g. http://www.isocat.org/datcat/DC-1333) under abstraction from language-specific (e.g. English 'noun' v. French 'nom') and arbitrary differences in notations for data categories (e.g. 'noun' v. 'n'). This will make it possible for all kinds of resources and tools to 'interoperate', not only on the format level,

but also on the level of content. A large list of data categories, mainly originating from the ISO TC 37/SC 4 project (which itself based its selection on earlier projects for best practices and standards such as EAGLES and ISLE) has been created. These data categories are currently considered as candidates for official inclusion in ISOcat and some have already been accepted (but all are already accessible for inspection, comments etc.). Of course, in some cases the same expression is used for two or more different concepts, sometimes dependent on a specific theoretical view on the matter. But ISOcat is open, one can add one's own concepts, and even organize a whole group of related concepts in a so-called 'profile'. Currently, ISOcat is basically a flat list of data categories, each with its properties. Data category specifications can be associated with a variety of data element names and with language-specific versions of definitions, names, value domains and other attributes. It is the intention to add, in a next stage, relations between concepts. This will allow one to specify many types of relations between concepts, e.g. that one concept is a hyponym of another one; that two concepts are not completely identical but very close; using such relations one can specify multiple hierarchical ontology's on these concepts, etc. etc.

For each concept that occurs in your resource or in the metadata of your resource, you should check whether a corresponding concept already exists in ISOcat. If this is not the case, you will have to add the concept in ISOcat. You will also have to make a formally represented mapping between the notations for concepts that occur in your resource or its metadata, and the PIDs of the corresponding ISOcat data categories. For example, if you use "zn" as the notation for the concept of 'noun', this mapping will have to include: zn is-a http://www.isocat.org/datcat/DC-1333. The way to express this relationship between a concept and an ISOcat data category depends on the resource format. In XML documents the DC Reference vocabulary (see http://www.isocat.org/12620/schemas/DCR.html) can be used.Thus the references can be directly inserted in the instance document, but it is also possible and in many cases preferable to include them in the schema of the resource, e.g., in an XML Schema or Relax NG document. Also resource specific languages, e.g. ODD or TBX, have their own specific ways to declare the relationship between a data element and a data category.

Further information about ISOcat can be found at http://www.isocat.org

See also

URL:	http://www.isocat.org/files/manual.html
Title:	Manual
Date/Version:	This page is updated regularly
Content:	This page contains links to guides on how to use ISOcat and the Data Category Registry in general, as they gradually come available. It is a very good place to find information about ISOcat.

Persistent Identifiers (PIs)

In order to ensure its long term persistence and accessibility a resource should be assigned a persistent identifier. A persistent identifier (PID) is a stable (persistent) and unique reference (identifier) to identify a resource, in the case of CLARIN a digital language resource. A well-known example of PIDs outside of CLARIN is formed by ISBN numbers, which are persistent identifiers for books. PIDs for resources are surely needed for tools, applications and services running on the CLARIN infrastructure to provide unique identifiers for resources but they can be useful for humans as well.

For further information see

URL: http://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf
Title: Persistent Identifier Service
Content: In this A4 shortguide, some introductory information is provided on the Persistent Identifier Service following the generic shortguide layout of "*What is it?*", "*What is it for?*", "*Who can use it?*", "*When can it be used?*" and "*How does it work?*".

FAQ

Q: What do I have to do to obtain a PID for my resource?
A: Make a request using the Persistence Identifier Service provided by CLARIN-NL. You will be asked to provide some minimal information about your resource such as a small subset of the metadata which you have to provide anyway in the context of your project. The exact nature of this minimal set of resource metadata will be made known ultimately at the start of your project.

Documentation

With the curation of any kind of resource it is essential to provide appropriate documentation. Two types of documentation can be distinguished:

- documentation of the resource itself; this documentation includes all relevant information about its design, the content, the format(s), the metadata, IPR etc. etc.
- a report on the experiences at all levels of the curation (such as the conversion of one format to another , the adaptation of an existing metadata profile).