A report including Model Licensing Templates and Authorization and Authentication Scheme

CLARIN-2010-12

EC FP7 project no. 212230

Deliverable: D7S-2.1
Version 5, final

Responsible: UHEL

Contributing Partners: ATILF, BBAW, BAS, CSC, CU, DFKI, HASRIL, ILC, ILSP, INL, IMCS, MPI, RACAI, UCPH, UHEL, UPF, UTU, WROCLAW TU


www.clarin.eu
The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organisational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) and to secure sustainable support from the funding bodies in the (now 23) participating countries for the subsequent construction and exploitation phases beyond 2010.
A report including Model Licensing Templates and Authorization and Authentication Scheme

CLARIN-2010-12

EC FP7 project no. 212230
Responsible: Krister Lindén

Contributing Partners: ATILF, BBAW, BAS, CSC, CU, DFKI, HASRIL, ILC, ILSP, INL, IMCS, MPI, RACAI, UCPH, UHEL, UPF, UTU, WROCLAW TU
CLARIN D7S-2.1 A report including Model Licensing Templates and Authorization and Authentication Scheme

Final (version 5)
2011-03-03

Contents

CLARIN Internal References ............................................................................. 6
Other References ............................................................................................ 6
Introduction ......................................................................................................... 7
Terminology and abbreviations ........................................................................... 8

PART I CLARIN PARTIES ................................................................................. 10
1.1 CO, Content Owner ..................................................................................... 10
1.2 Content Provider CP .................................................................................... 11
1.3 Service Provider SP .................................................................................... 12
1.4 End-User, CU ............................................................................................. 13
1.5 CLARIN Identity Providers IdPs .................................................................... 14

PART II OVERVIEW OF THE AA SCHEME ............................................................ 15
2.1 End-User Authorization for Accessing CLARIN Materials ......................... 16
2.1.1 Identifying users and materials .............................................................. 16
2.1.2 Authorizing users ................................................................................. 16
2.1.3 Technical framework of user authorization and access .............................. 17

PART III CLARIN PROPOSAL FOR AN END-USER ACCESS MODEL WITH CONTENT
CLASSIFICATION PROCEDURE .............................................................................. 20
3.1 The Laundry Symbols introduced ................................................................ 21
3.1.1 PUB .................................................................................................... 22
3.1.2 ACA .................................................................................................... 23
3.1.3 RES .................................................................................................... 23
3.1.4 Additional requirements ........................................................................ 24
3.2 The Laundry Symbols assist the Copyright Owners ........................................ 26
3.3 The prerequisites for access rights ............................................................... 27
3.4 End-User Access Categories ....................................................................... 28
3.5 Extension of the CLARIN End-User Access Model ..................................... 28
3.6 CLARIN Prototype: AAI at CSC from the End-User perspective ................... 29
3.7 CLARIN prototype: AAI at CSC introducing the referee system ................... 30
PART IV  CASE STUDIES WITH REFERENCE TO THE CURRENT PRACTICES AND THE CLARIN PROTOTYPE SPECIFICATIONS FOR THE DISTRIBUTION TYPES .................................. 32

4.1 Case Study 1: A PUB resource deposited – the case of the FinnWordNet ........... 32
  4.1.1 Current (pre-CLARIN) practices concerning the PUB resources ............... 33
  4.1.2 CLARIN prototype specifications for PUB.................................................. 34

4.2 Case Study 2: An ACA resource deposited and access granted –
 the case of the JP data set ............................................................................. 35
  4.2.1 Current (pre-CLARIN) practices concerning the ACA resources ............ 36
  4.2.2 CLARIN prototype specifications for ACA ............................................... 38
  4.2.3 The End-User Raymond applies for a right to access
the CLARIN Material JP Dataset ................................................................. 39

4.3 Case Study 3: A RES resource deposited and access/user account
granted - the case of FDC .............................................................................. 39
  4.3.1 Current (pre-CLARIN) practices concerning the RES resources ............ 41
  4.3.2 CLARIN prototype specifications for RES................................................ 43
  4.3.3 Special issues concerning personal data; case study RES ....................... 44
  4.3.4 The referee procedure; case study RES ................................................ 47
  4.3.5 The End-User Raymond applies for a right to access the
CLARIN Material FDC ............................................................................. 47
  4.3.6 Overall Statistics from the survey based on the classification task
in the LRT Inventory .................................................................................. 48

PART V MODEL AGREEMENTS BETWEEN CLARIN PARTIES ................................. 50
  5.1.1 Agreement btw Service Providers: an agreement of uniform services...... 50
  5.1.2 Agreements btw Content Owners/Content Providers and Service Providers..
the End-User Licenses .................................................................................. 53

PART VI FUTURE COLLABORATION WITHIN CLARIN ........................................... 56

Appendices

(A) CLARIN Deposition & License Agreement....................................................... 58
(B) CLARIN Upgrade Agreement ........................................................................ 62
(C) CLARIN Terms of Service.............................................................................. 63
(D) CLARIN Privacy Policy.................................................................................. 66
(E) CLARIN Application for Use of Restricted Data from the CLARIN Database .... 70
(F) CLARIN Data User Agreement................................................................. 71
(G) CLARIN Data Processor Agreement between Content and Service Provider .... 73
(H) CLARIN Guidelines for Consent Forms for Collecting Personal Data............ 75
(I)  CLARIN Check list for CLARIN Agreement................................................ 76
CLARIN Internal References

[u1] Arppe, A., Bruun S. and Westerlund H. D7S-4.1: Set of Federation Agreements for CLARIN Centres
[u6] CLARIN Description of Work 2008 Annex I to the Grant agreement for Combination of Collaborative Project and Coordination and Support Actions
[u9] Lindén K., Oksanen, V. and Bruun, S. M7S-2.3 CLARIN Classification Guide for Deposition Licenses – First comprehensive summary of licensing problems
[u10] Lindén, K. 2010. CLARIN IPR and Business Model
[u13] Lindén, K. and Oksanen, V. 2010. EULAs in the context of an Eric, D-Spin Vienna 22.10.2010 (presentation)
[u18] Oksanen, V. 2010. CLARIN Deposition & License Agreement. Appendix (A)
[u19] Oksanen, V. 2010. CLARIN Upgrade Agreement Appendix (B)

Other References

[u16] EU Directive (95/46/EC)

We would like to thank all CLARIN partners and centres that took part in the surveys and questionnaires, provided us with information, licenses and also questions and feedback - especially CSC in Finland, IDS in Germany, ILSP in Greece, IMCS in Latvia, and MPI in the Netherlands.
Introduction

One of the objectives of WP7 as stated in the CLARIN Description of Work in 2008 [u6] was to investigate the Intellectual Property Rights (IPR) and Data Privacy Issues concerning the LRT linguistic materials in order to provide “a framework for licensing and authorization between CLARIN and external providers to enable incorporating new and existing resources and technology into CLARIN” [u6].

For this purpose, WP7 set out to study the state of the art in authorization and authentication procedures within the CLARIN centres, and based on the findings and needs of the centres, to draft model licensing templates to be of use for the CLARIN players.

By Authorization we mean procedures that are needed in order to grant access to resources or tools for authenticated users.

Authentication refers to the identification of users through user account, password of a trusted organization.

The current state of affairs in the CLARIN member institutions and centres were studied, a pattern of conditions regulating the distribution of CLARIN Materials were identified and model templates for various kinds of agreements were drafted.

For the AA Scheme the UHEL team prepared questionnaires and conducted studies on the current procedures and needs, as well as potential problems that might hinder the progress towards a true CLARIN LRT community with CLARIN Materials available for the whole research community, with a possible extension towards business models [u9]. Reliable distributed authentication schemes open the way to providing licenses for incorporating as many resources and tools as possible to be accessible for large number of users within CLARIN. The surveys would not have been possible without the active involvement of the CLARIN partners and centres.

LRT materials can only be made available for the research community subject to a variety of legal and ethical restrictions. The goal of WP7 was to work out such patterns during the prototype phase for preparing the construction phase, and as the end result, a network of agreements and licenses was drafted in order to achieve and maintain sufficient levels of trust conforming to the law. In addition to the AA Scheme, This deliverable presents the Model Licensing Templates as a result for the WP7 contribution. They are presented in a more concise form in M7S-2.4 [u11] and they should be read in conjunction with the CLARIN Agreement Check List [u12].

The prerequisite for successful cooperation towards a functioning AA Scheme is a federation of trusted organizations that can offer reliable identification of users and a reliable certification of the user reading the license and having signed it. Year 2009 saw the establishment of a federation of five CLARIN centres uniting their efforts towards achieving the CLARIN goals, and at the end of 2010 three new parties are in the process of joining the federation. Experience from a number of national identity federations within CLARIN forms a basis for the CLARIN Authorization and Authentication Scheme described in more detail in this deliverable.

The contents of this deliverable are organized as follows. The key concepts of the CLARIN parties involved in the Authentication and Authorization Scheme are first presented in this chapter, and the abbreviations that have been used are listed in table 1 below. The rest of the report presents findings
from surveys and questionnaires as well as technical specifications in six parts. Part I presents the CLARIN parties, and Part II introduces an overview of the AA Scheme and the CLARIN proposal for classification of data into various distribution types is described in Part III. The classification has a dual role in the AA Scheme: firstly, it provides a framework for organizing the CLARIN Materials for the depositor and the repository, and secondly, it determines the application process of the applicant wishing to obtain access to the resources or the tools offered by the CLARIN research community. The classification into distribution types is described in more detail in Part IV by presenting three case studies, each representing one of the distribution types and including issues that are distribution type specific, such as questions concerning personal data. The relevant information gathered from the CLARIN centres with surveys and questionnaires is also presented in conjunction with the case studies. Terms of Service and Privacy Policy presented in Part IV are aspects of the AA Scheme have a dual role: they concern the applicant and the repository, but at the same time they form an integral part of the licensing scheme as the most central conditions are included in these documents. Part V lists the model template agreements that the depositors and repositories can adapt to their needs and to the relevant national IPR legislation. Finally, some future prospects to be addressed in the construction phase are introduced briefly in Part VI.

**Terminology and abbreviations**

During the present CLARIN WP7 project, various terms for the key concepts have been used in the milestones and the presentations. It is only to be expected, since the key concepts themselves have evolved during the project period. The data on which the categorization and distribution of CLARIN Materials are based includes the agreements received from several CLARIN centres, multiple projects and Service Providers, and the terminological variation displays this fact perfectly. With the Upgrade Agreements it is possible that this variation will decrease and a more coherent conceptual space will be created within the CLARIN community in the future.

One of the key concepts within CLARIN, the linguistic resources, is a good example of this multiplicity of terms. In the agreements and licenses that were studied in order to identify the minimal set of core requirements and conditions of use in them, the linguistic resources are referred to as data, corpora, resource, (linguistic) material, digital data files, if at all - it is quite common to use the abbreviation of the corpus in the text instead of a more generic term. The following section outlines the key terminology used in this document and presents the abbreviations you may find in this document, mainly in the figures and the statistics but also in the documents we refer to.

**CLARIN Material** is the term chosen to cover all language resources offered for the End-User mostly for research but sometimes also for any other purpose. It includes the various types of data in any digital form and it also includes tools, dictionaries, WordNets, TreeBanks etc. In the CLARIN documentation, e.g. the Description of Work refers to Resources and Tools, as does the database LRT Inventory containing information about Resources and Tools within CLARIN. The term **CLARIN Material(s)** is used throughout this document to refer to all of these types of data and tools, including software programmes.

**CLARIN Content** and **CLARIN Services** are used in the model agreements and templates to cover all CLARIN Materials. Content is preferred in the deposition and upgrade agreements. Services refer both to the physical environment, i.e. the websites and the servers and to the materials, and it is used in the Terms of Service.
**Compiler** means any person or organisation collecting the sub-parts of the resource together.

**Depositor** means any person or organisation authorised to transfer and deposit CLARIN Material.

**Repository** refers to the CLARIN infrastructure described in more detail in this document.

**Abbreviations:**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA Scheme</td>
<td>Authentication and Authorization Scheme</td>
</tr>
<tr>
<td>AAI</td>
<td>Authentication, Authorization Infrastructure at CSC IT Center for Science for online applications to use the CLARIN Material</td>
</tr>
<tr>
<td>ACA</td>
<td>Academic Use</td>
</tr>
<tr>
<td>AGPL</td>
<td>Affero General Public License</td>
</tr>
<tr>
<td>AR</td>
<td>Authorization Record(s)</td>
</tr>
<tr>
<td>BSD</td>
<td>Berkley Software Distribution</td>
</tr>
<tr>
<td>CC</td>
<td>Creative Commons</td>
</tr>
<tr>
<td>CO</td>
<td>Content Owner, same as Copyright Owner or IPR owner</td>
</tr>
<tr>
<td>CP</td>
<td>Content Provider</td>
</tr>
<tr>
<td>CU</td>
<td>End-User</td>
</tr>
<tr>
<td>DELA</td>
<td>Deposition License Agreement</td>
</tr>
<tr>
<td>DRM</td>
<td>Digital Rights Management</td>
</tr>
<tr>
<td>DUA</td>
<td>Data User Agreement</td>
</tr>
<tr>
<td>EUL</td>
<td>End-User License</td>
</tr>
<tr>
<td>GNU GPL</td>
<td>GNU General Public License</td>
</tr>
<tr>
<td>Inf</td>
<td>Information required</td>
</tr>
<tr>
<td>LRT</td>
<td>Language Resources and Technology</td>
</tr>
<tr>
<td>LRT Inventory</td>
<td>CLARIN Inventory for language tools and resources <a href="http://www.clarin.eu">www.clarin.eu</a></td>
</tr>
<tr>
<td>NC</td>
<td>Non-Commercial Use</td>
</tr>
<tr>
<td>ODbL</td>
<td>Open Database License</td>
</tr>
<tr>
<td>PID</td>
<td>Persistent Identifier</td>
</tr>
<tr>
<td>PUB</td>
<td>Public Use</td>
</tr>
<tr>
<td>ReD</td>
<td>Redeposition required</td>
</tr>
<tr>
<td>RES</td>
<td>Restricted Use</td>
</tr>
<tr>
<td>SP</td>
<td>Service Provider</td>
</tr>
<tr>
<td>ToS</td>
<td>Terms of Service</td>
</tr>
</tbody>
</table>
PART I  CLARIN PARTIES

The CLARIN Infrastructure is a network of parties. These parties can have different tasks depending on the local organisational structure, and the relationships between the parties in each local organisational unit can vary. In this section, the parties and their relationships are introduced in a rather general level, and a more detailed description of the agreements between the parties is found in the latter parts of this document. [u7]

The general framework of CLARIN as seen from the Work Package 7 is roughly as in the following diagram where the dashed lines represent the flow of permissions, and solid lines stand for other types of communication.

![Figure 1. CLARIN parties in a local level](image)

The concepts described below are here understood as roles that an individual or an organization can assume within the CLARIN framework. The roles typically mean rights and obligations that are specified in agreements between the parties in question. Each individual or organization can assume multiple roles, which is also reflected by the versatility of the licensing agreements.

1.1 Content Owner, CO

The Content Owner possesses the copyright for the data.

Linguistic data (text or speech) is produced by an informant who ultimately might have partial copyright for that specific data. It is recorded or written down by a person for using it for a specific purpose. The person documenting the data typically obtains copyright for the material or dataset produced.
Example 1. A researcher collects data from informants and compiles a data set from the outcome of the informants. Each informant gives the researcher a permission to use the data for the purpose that is specified in beforehand. The informant is expected to transfer the copyright to the data to the researcher by signing an agreement. In the agreement the researcher specifies the purpose and the scope of distributing the data in the future. It is possible to delimit the rights of the researcher so that s/he cannot distribute the data at all, or the informant can grant the researcher with a right to distribute it in whichever form s/he feels appropriate in the future. Often the agreement also specifies the organization, such as the university department or research institute of the researcher as the holder of the data and the copyright.

Example 2. A journalist writes an article for a newspaper. The journalist and the publisher can both have the copyright for the article.

Example 3. A researcher creates a tool for processing the data. The copyright for the tool stays with the researcher but can also be co-owned by the organization employing the researcher.

The national legislation governs the scope and the details of the copyright. The copyrights can be chained.
Also software or tools can be examples of data.

1.2 Content Provider, CP

CLARIN Content Providers are typically organizations which acquire language materials and sufficient rights from the Content Owners, i.e. the authors and publishers or whoever possesses the rights. The rights needed by the Content Provider typically include the right to grant some End-Users the right to access and use the materials. The Content Owners may put some restrictions on who may use the materials and in which ways they may be used, e.g. only for research purposes or not to make copies other than customary citations.

The Content Provider is the party that compiles the data into a linguistic collection. This collection can be called a corpus. A researcher can function as a Content Provider and compile a corpus for conducting a study on it. An organization can function as a Content Provider when hosting data for research purposes.

Example 4. A researcher compiles a corpus of a specific genre in a specific language as part of the research project within a department. The data consists of a set of newspaper articles that the researcher has acquired from the publisher. In the agreement between the publisher and the researcher, the parties have agreed to grant the department a deposit right of the data and also a right to distribute the data for research purposes. The distribution is subject to citation and there is a given form for citation available. The researcher as well as the institution hosting the data has the role of a CP. The publisher represents the holder or the original copyrights of the newspaper articles, i.e. the journalists.
Common Language Resources and Technology Infrastructure

The Content Provider grants the applicants the permissions to access the data complying with the conditions specified in the agreement. The agreement between the Content/Copyright Owner and the Content Provider governs the scope of usage of the data in question.

The Content Provider may seek to acquire more data for larger collections. It may also develop tools for the users for accessing the data. It may also be the case that the only permissible way to use the data is through such a tool, where the data itself is not accessed by the user who will have the results provided by the tool available.

Example 5. The Department of General Linguistics at UHEL holds a dataset for speech compiled within a national project. The researchers in the project can access the data only through a specific tool with which it is possible to scan the data fulfilling the pre-specified linguistic criteria. No researcher has access to the original data. In the publications the researchers are required to hide the identities of the informants.

The Content Provider monitors the usage of the data.

Example 6. The Research Institute for the Languages of Finland has compiled a corpus of Modern Finnish Lexicon, which is a freely accessible on-line data service under GNU LGPL. The user can access the lexicon via a web tool.

If the Content Provider maintains the technical environment required for the deposition of the data it also assumes the role of a Service Provider.

1.3 Service Provider, SP

The Service Provider provides the technical infrastructure required for hosting data and for user authentication.

A Content Provider institution deposits the material at a CLARIN Service Provider, usually a computing centre. The Content Provider and the Service Provider specify the conditions under which the Service Provider operates in the agreements between the Content Provider and the Service Provider. The Service Provider might be required to consult the Content Provider whenever a user applies for the right to access the data but the Content Provider may also transfer the right or obligation to make the decision when the pre-defined criteria are fulfilled. Among other things the
Service Provider agrees to protect the material against unauthorized access and to let End-Users only access the materials according to the explicit authorization by the Content Provider.

The agreements between the Content Provider and the Content Owner(s) permitting, the Content Provider can transfer the data to be hosted at a Service Provider which typically can offer computing power and better technical infrastructure.

CLARIN consists of several Content Provider organizations and some Service Provider centres. Each Content Provider is typically associated with one Service Provider centre. The various actors can sometimes be the one and the same organization, i.e. a unit can provide both the Content Provider and Service Provider functions: the Service Provider can also be active in acquiring data in which case it assumes the roles of a Content Provider and a Service Provider at the same time. Or a CLARIN Service Provider may simultaneously be a national organization that maintains an IdP federation (e.g. CSC IT Center in Finland).

Although the diagram just shows one Service Provider, CLARIN consists of several Service Providers linked together with agreements, which harmonizes their CLARIN services to the extent that the End-Users can identify themselves using their local IdP services and access the materials on any Service Provider centre according to their permissions in the Authorization Records. The CLARIN Service Provider Federation agreement states some obligations for each of the participating CLARIN centres, including the responsibility to enter into necessary agreements with the IdPs or their federations used within CLARIN.

Furthermore, there may be several Content Providers connected to each Service Provider.

---

Example 7. The Research Institute for the Languages of Finland has compiled a Swedish-Finnish Parallel Text Corpus from the 21st century. The data has been transferred to CSC (IT Center for Science) and it is part of the Finnish Language Bank. The application to access this corpus is sent to the Research Institute for the Languages of Finland for approval and authorization, after which CSC assumes responsibility for authentication and monitoring the use as well as technical assistance in accessing the data when required.

1.4 End-User, CU

The End-User follows a given procedure in applying for authorization to use the data. S/he might be required to e.g. prove his/her identity, to obtain a recommendation from his/her home organization or provide other relevant information for the Content Provider or the Service Provider before obtaining a permit to access the data. In the example 7 above, the user will enter into an agreement with the Service Provider and all relevant discussions will take place between the End-User and the Service Provider.
1.5 CLARIN Identity Providers IdPs

CLARIN Identity Providers are existing national or institutional identity services (or federations consisting of such) and they are used for identifying and authenticating large groups of people such as staffs of organizations or students in universities. CLARIN Service Providers and IdPs are connected to each other through agreements. The identity provided by IdPs is the basis for identifying End-Users. A Service Provider cooperates with all CLARIN IdPs.
PART II  OVERVIEW OF THE AA SCHEME

The AA Scheme is a description of a process whereby the potential user of CLARIN resources or tools is granted access to them in a way that satisfies all parties in question. The process looks different from the point of view of each party involved, and the role of each party depends on the resource, the applicant, and the organization.

This description of the AA Scheme proposed for the CLARIN construction phase seeks to

- offer an overview of the present processes that are common in the CLARIN LRT community, and
- present the flow of events in the prototype process through real case studies and exemplify in practice how the aims of CLARIN can be achieved, and
- discuss the IPR issues in each case and the problems that have been identified and the possible solutions to these problems.

The general overview of the process is first explained from the point of view of the material depositor and the Content Provider /Service Provider and it describes the overall structure of the system. The second part focuses on the End-User and the issues s/he needs to consider.

The Content Provider institutions control the authorization by maintaining the contents of CLARIN Authorization Records (ARs) which indicate what materials each End-User is allowed to use and how. As mentioned above, the Content Provider has acquired rights from the Content Owner to grant the End-Users licenses to use the materials. The ARs contains all relevant information in these licenses. The Service Providers are responsible for organizing the ARs. Through the ARs the Content Provider and the Service Provider agree that the Service Provider may offer users an access to CLARIN Materials only according to the licenses specified in the ARs. The ARs are based on the IdP identities and possible electronic signatures confirming the acceptance of relevant license terms. An electronic signature can be implemented e.g. by an End-User clicking “I accept” button on a web page which she is properly logged in with the timestamp, eduPersonPrincipalName (ePPN) value and hash value of the terms.\(^1\)

The Content Provider is also responsible for supplying in addition to the content, some metadata in a standard format, and exact information about the authorization scheme for the material. The authorization information has to indicate among other things: (i) who is authorized to grant the permissions for users, (ii) what qualifications the individual applicants must satisfy and (iii) what license agreement the applicants must sign (including the license text explaining the exact conditions of use). The Content Provider may also have to indicate the level of assurance needed for identifying and authenticating the End-Users when accepting identity information about them.

\(^1\) The possibility of this kind of signature is evident in the Finnish legislation, but in some other European country it maybe is not.
2.1 End-User Authorization for Accessing CLARIN Materials

This chapter describes the basic framework how an End-User can get permissions to use CLARIN Materials and how a Content Provider can grant such permissions. The presentation here is simplified by focusing our attention on activities at one Service Provider. [u7]

2.1.1 Identifying users and materials

Each CLARIN Material must be identified by unique Persistent Identifier (PID, cf. WP2) and accompanied with some metadata (cf. WP2 and WP5) that among other things indicates a way how an End-User can apply for or directly get a permission to use that material.

A person, who becomes an identified CLARIN End-User, must have a unique identifier within the whole CLARIN. A reliable identification is the EduPersonPrincipalName (ePPN) attribute provided by an Identity Provider (IdP). CLARIN Service Providers needs IdPs in order to enable single sign on services. For all CLARIN purposes, End-Users must be uniquely identified with ePPN.

The Identity Provider is the organization responsible for the authentication of the person who is affiliated to it. The ePPN is guaranteed to be unique within such an organization. N.B. Though the ePPN attribute contains a Scope segment designating the organization, it is properly uniquely identified with the SchacHomeOrganization attribute. IdPs and the institutions usually belong to some Identity Federation (IdF) that requires certain policies to be followed so that we can rely on the uniqueness and legal and technical quality of the identities.

2.1.2 Authorizing users

The Content Provider has agreed with the Copyright Owners of CLARIN Material that it can grant permissions for some users to use the CLARIN Material and, based on this agreement, the Content Provider can authorize individual End-Users to use the CLARIN Material in question. Usually the End-User must agree e.g. that the materials are used only for certain purposes, subject to data protection regulations and research ethics, or all copies of the material are kept within the same Service Provider site.

For the purpose of our discussion here, a license is an authorization by the Content Provider to use the licensed CLARIN Material by the End-User. We denote the license formally by L(CU,M). For us, a license is a document binding the two parties, the Content Provider and the End-User. When a user has to commit to obeying some conditions or assure the correctness of some relevant information, the license agreement must be signed by the user. The signature can be hand written on paper or an electronic signature the validity of which is based on the policies followed by the IdPs and IdFs. In principle, all restrictions and rights for the End-User using CLARIN Materials must be a part of this license.

---

2 The ePPN is unique at any point of time. A policy avoiding the recycling of user names within organizations might be necessary. The forthcoming eduGAIN may resolve this problem, but promoting such a development will require motivations from end-user communities such as CLARIN.
Three categories of materials can be handled. (They are described in more detail in Part III below):

1. **Free materials**, e.g. Materials (a) which are not subject to copyright or where the copyright has expired or (b) materials granted with different kinds open licenses (including but not limited to open source or open access licenses). No additional CLARIN licenses need to be signed for these kinds of material. However, the End-User will need to accept the Terms of Service to gain access to the materials (see Part V for more information). The End-User can access the materials without strict authentication and no actions by the Content Provider are needed to authorize the End-User. The End-User commits implicitly herself to obey the license terms without explicit assurances.

2. Materials where the Content Provider authorizes pre-defined groups of CLARIN End-Users (such as all members, students or personnel of universities) to use the materials provided that the user signs an appropriate license. The Content Provider grants the license automatically based on the user signature by which the user commits to obeying the terms of use. The copyright holder must have given sufficient rights for the Content Provider to do this. A reliable identification and authentication of the CLARIN End-User is necessary for the user to sign the license and access the materials. No additional actions by the Content Provider are needed for authorizing individual CLARIN End-Users to use the material.

3. Materials where the Content Provider considers applications individually before deciding whether to accept or reject the application. This class includes (but is not limited to) materials subject to data protection and materials with commercial interest such as copyrighted books still being sold. A reliable identification and authentication of the CLARIN End-User and the representative of the Content Provider is required, and some action by the Content Provider is needed for each CLARIN End-User.

The Content Provider can grant a user the right to access several materials through a single signature. Additionally or alternatively, the Content Provider can allow a certain group of users to get the authorization to use the materials if an individual belonging to such a group signs the license. Such groups could be e.g. personnel of any university within CLARIN or all full-time students. A liberal policy is likely to be possible mostly for materials which have neither copyright nor privacy restrictions. On the other hand, some collectors and copyright owners may be more willing to deposit their materials in CLARIN if the use of the materials is more closely controlled.

The essential documentation of this granting should be saved as the **Authorization Records** (ARs) at the Service Provider where the Content Provider has deposited the CLARIN Material. The Service Provider has agreed to grant technical access to the materials only according to these ARs.

### 2.1.3 Technical framework of user authorization and access

The Content Provider may coincide with the Service Provider in some cases, in others it may be a separate institution. Also the IdP may be run either by a separate organization, possibly in another country, or may sometimes be provided as a function of the same Service Provider.
The ARs contain all contracts including electronic signatures and other documents which are needed to permit the use. For controlling the technical access of CLARIN End-User to CLARIN Material we only need to have an Access Database (ADB) as a Policy Decision Point which tells whether the End-User (identified with her ePPN) has the permission to use material M (identified with its PID). The ADB contains information about the pairs of these unique identifiers. The register is technically maintained by the Service Provider where the CLARIN Material resides. When the properly authenticated and identified End-User tries to access a CLARIN Material, the CLARIN service at the Service Provider must give the CLARIN End-User the access if and only if the ARs permit.

In practice, a license often authorizes a CLARIN End-User to use all CLARIN Materials in some collection (each with a unique PID). Therefore the ARs and the ADB will most likely use such collections to simplify the processing. Thus, the End-User signs a license for all CLARIN Materials in a collection. The ADB then needs to contain only pairs of ePPNs and collection identifiers. The Single Sign On principle of CLARIN means that the user need not have an account on this particular Service Provider, but can authenticate herself at her (more) local IdP (which might be in a different country than this Service Provider).

In principle, one could maintain an ADB manually at the Service Provider by entering data from signed paper documents or similar sources. This might be laborious and make it somewhat unreliable in identifying the connection between the person's ePPN. IdPs belonging to an IdF which is trusted by CLARIN Service Providers can identify people more accurately. ADB can, of course, be created automatically from systematic ARs.

For materials in classes 2 and 3, some automated workflow for granting and recording a license may be used. In some more complicated cases, the applicant might need a recommendation and assurances for which the electronic signatures would be collected by the work flow. Furthermore,
such a work flow mechanism could be used to achieve a sufficient level of trust to the correctness and validity of additional information concerning the applicant such as the kind of affiliation (if an assurance by the applicant alone is not considered sufficient).
PART III CLARIN PROPOSAL FOR AN END-USER ACCESS MODEL WITH CONTENT CLASSIFICATION PROCEDURE

The following chapter will introduce the classification of distribution types of CLARIN Materials developed by WP7 within CLARIN in close collaboration with WP2 and WP5 as well as the partner institutions, especially IMCS in Latvia and ILSP in Greece.

Terms of Service & Privacy Policy

There will be plenty of materials and users in CLARIN. The number of Content Providers and distinct license formulas may also become problematic. In order to keep the system of licenses manageable, some rationalization is needed. License texts have been written in many languages, and some materials can be used more freely, e.g. with less restrictions and obligations than others. Licensing should be as simple as possible but it cannot be reduced into a single agreement template.

The WP7 proposes a small number of such combinations that serve as the recommended *standard licensing categories* that will be recognized and understood by the users and other parties. These access categories have received graphical representations PUB, ACA and RES that remind us about the laundry symbols found for example in garments and clothing items and that convey the information about how they should be cared for. These CLARIN Laundry Symbols are described in more detail below.

The End-User agreements, code of conducts, and licenses that have been used in the CLARIN centres are very different and take into account different copyright reflecting the national copyright laws. The rights a Content Provider has for *some old materials* may not be quite equivalent with the standard categories. The present agreements and licenses need to be modified only in so far as they do not comply with the classification system within CLARIN. In many cases, the Content Provider
can include even such materials, if the requirements for a CLARIN End-User of such a material (set by the copyright owner) are at most as strict and if the permissions that can be granted for the CLARIN End-User are at least as liberal as in the standard category. The model agreement template for an upgrade agreement (see Part V in this document) can be used to renew or clarify an existing agreement.

CLARIN recommends that all new materials would be deposited according to the common standard categories.

Some linguistic materials hosted by the organizations and institutions within CLARIN can be directly offered for End-Users without further agreements. The typical example of such a scenario is a resource freely available for anybody without any requirements neither from the Copyright Owner(s) nor the Service Provider. When no registration is needed, the questions to be considered are not legal but practical, e.g. how and where to publish the information and availability of such resources or tools. For other language resource materials the classification is not a mere labelling task but entails legal aspects either with respect to intellectual property rights or privacy issues. Information about language resources and tools has been gathered by MPI now hosting the database LRT Inventory where 848 language resources and 183 tools have been entered by December 2010.

The classification of the language resources and tools is a task fulfilling two needs:

The depositor wants to make his or her linguistic material or tool CLARIN compatible. She needs to know how to classify the resource or tool and what information is required so that the Service Provider is able to fulfil its task as the distributor of the resource or the tool in question.

The potential End-User is interested in a specific CLARIN Material and wishes to use it for academic use that is in the core function of the CLARIN community, but in some cases also for the development of LRT tools. He needs to find out the requirements and conditions under which the usage is allowed and what in practice is required from him to gain access to the linguistic material or tool.

In this section we introduce the classification types and discuss some questions that have come up during the project.

### 3.1 The Laundry Symbols introduced

The work behind the laundry symbols is based on several project team efforts. First of all the legal expertise within WP7 provided by CLARIN has been the key in analyzing the material where the categories and types emerged. The material was gathered with great help from the CLARIN centres and participating institutions and their web sites in the form of agreements, licenses, codes-of-conduct, web forms, and valuable comments received on the summaries distributed CLARIN wide as well as answers to questions concerning the various ways the potential user or the applicant is authorized and authenticated at present [u7] and [u15]. The other effort where the centres’ participation was invaluable was the pilot study for the classification of the resources in the LRT Inventory where the Content Owners or Content Providers were asked to classify at least one resource of all different types they hold [u9]. The statistics from the latter study are found below.
The questions from the participants in the classification pilot were to the point and assisted in locating the aspects that needed further thought and development.

Based on the findings from the surveys, these **three main content categories** and **three sub-categories** presented as laundry symbols will be attached to each CLARIN resource and tool available for the CLARIN End-Users.

The Content Owner or Content Provider needs to choose the main category first:

- **PUB** (licly) available resources, usually through web interfaces only
- Resources licensed to **ACA** (demic) use and researchers, the main CLARIN focus area
- **RES** (stricted) resources requiring individual authorization

![Figure 4. Access categories PUB, ACA and RES](image)

### 3.1.1 PUB

The resources for the **Public Use** (PUB) are usually available through web interfaces and not downloadable as texts. The role of CLARIN in providing the PUB resources is mostly in distributing information on the availability of CLARIN Materials. The main CLARIN effort is focused on providing and extending the ACA category as the resources in the category RES will in the future also require activities that are outside the scope of CLARIN, such as approval practices and other mutual tasks where the Content Owner or the Content Provider and the potential user are actively involved. [u13]

The licenses in the PUB category should follow as closely as possible the Protocol for Implementing Open Access Data. For the new CLARIN Materials, the preferable license is either the Creative Commons Zero (CC0) or the Open Database License (ODbL). Otherwise the licenses in this category should allow at least the Open Source definition 4-6 and 9-10, see [http://www.opensource.org/osd.html](http://www.opensource.org/osd.html):

---

4. [http://creativecommons.org/choose/zero](http://creativecommons.org/choose/zero)
Integrity of The Content  
No Discrimination Against Persons or Groups  
No Discrimination Against Fields of Endeavour  
License Must Not Restrict Other Content  
License Must Be Technology-Neutral

In addition the license should allow distribution of the tools and resources from the CLARIN infrastructure in all circumstances. However, there is no requirement to allow sub-licensing of the content for the user.

3.1.2 ACA

For Academic Use (ACA), the license agreement includes a requirement that the use is somehow related to an academic institution. Here the problem may and will arise from the definition of academic use. The main points to check are:

- The content should be available at least for anyone working or/and studying in an academic institution recognized by an Identity Federation (IdF) with which CLARIN has a service provider agreement.
- The content can be used for general academic purposes. However, it is relatively hard to define exactly what this means, e.g. “teaching” is not normally listed as permitted but that may or may not mean that it is prohibited. In case of doubt, contact the copyright holder (or use the CLARIN Upgrade Agreement to obtain the missing right).

Due to legacy deposition license agreements, the Academic Use licenses may include non-standard provisions that the user is required to accept separately for a specific tool or resource.

3.1.3 RES

The category Restricted Use (RES) includes the resources that do not fulfil the previous requirements but still could be offered to the users if certain requirements are met. The first thing to check is that the license gives enough rights for CLARIN to distribute it in the first place. If the access is limited to a certain physical room or a certain library, there is not much point in adding it to CLARIN, and the use of an upgrade agreement is necessary.

However, the most typical reasons for a resource to fall under the scope of RES are less problematic leading to a requirement that the license be granted individually by the owner, i.e.

- a requirement to submit detailed information (e.g. an abstract) about the planned usage, or
- specific ethical or data protection-related requirements found in speech and video material, i.e. material including Personal Data. Even if RES is the typical license category for Personal Data, in some cases the category PUB may also be considered e.g. when there is an explicit consent (from the data subject/informant) for processing and disclosing Personal Data without any limitations, or when the legislation permits disclosing of Personal Data.
In addition to the aforementioned terms, the following terms are commonly part of the agreements and do not normally cause problems as such for CLARIN:

- No right to re-distribute the material in any manner, no right to sub-license the material
- Only insignificant parts of the materials may be extracted at a time; whole works or parts of works may not be extracted
- Materials will have to be destroyed after the termination of the contract
- End-User ensures that she will keep any password secure and not disclose it to third parties
- End-User ensures to preserve the confidentiality of, and not attempt to identify individuals, households or organizations in the data
- Acknowledging that the user’s personal data may be held for validation and statistical purposes according to the data protection regulations
- Any breach of the agreement will lead to immediate termination of the user’s access to the services and could result in legal action
- Applicable law
- Liability disclaimer

One question, which is seldom addressed in the deposition agreements, is a question of what happens to the content if the copyright holder loses its legal status, e.g. ceases to exist. While this does not affect the addition of the content to the service, it would be preferable to address the question in new agreements and in the update agreements e.g. by adding a term that puts the content in question to a public domain in such a situation. This is included in the CLARIN Deposition & License Agreement (Appendix A).

3.1.4 Additional requirements

The next step for the depositor is to consider whether there are additional requirements that are essential for the potential user to accept before s/he can have access rights to the resource or the tool. The sub-categories containing additional requirements are

- **NC** - A requirement for strictly non-commercial use. A term requiring non-commercial use of the content is commonly found in different licenses. It is problematic because there is no common definition of what non-commercial actually means in different jurisdictions. From the CLARIN point of view, this is not a problem as long as the user of the content is aware of the term.
- **Inf** - A requirement to inform the Content Owner or the Content Provider regarding the usage of the tools and/or the resources in published articles. This term is something that is almost part of good practice in writing academic articles. It seldom causes any problems as long as the user is aware of it.
- **ReD** - A requirement to re-deposit modified versions of the tools and resources with the Service Provider. In certain cases the right holder has an interest to collect the modified versions of the content.

In the future the aim in CLARIN is that these categories would cover the needs of the Content Providers and the Content Owners. However, in practice the present licenses and agreements entail other resource or tool specific requirements or conditions that the potential user needs to accept in
order to gain access to the resource or tool in question. These conditions are not standard, and they are formulated by the Content Provider or the Content Owner specifically for that particular resource or tool.

Examples of such resource or tool specific requirements would be a case where the Content Provider of the Content Owner requires a specific form in citations when the overall requirement to cite the resource is understood as common practice. The Content Provider or Content Owner might wish to follow the citations in the electronic articles and therefore will not accept any other form of citation. It is also possible that the End-User is required to compensate for the right to use the resource or tool for example by purchasing a CD-ROM format of the resource. Such requirements need normally not be handled by the Service Provider that in these cases has a role more like an intermediator of instructions for the applicant on how to proceed. A requirement to provide an abstract or a research plan will also be classified as a resource/tool specific requirement. [u13]

Most agreements and licenses include some requirements that do not derive from the agreement itself but reinforces the common practice and the ethical standards that are usually followed within the scientific community and that are considered almost self evident. The CLARIN prototype model suggest that these requirements are included in the Terms of Service (ToS) as it is the document that the CLARIN End-User will first read and sign when he enters the CLARIN Infrastructure, i.e. the website of the Service Provider or any CLARIN party who assumes the role of informing the CLARIN community about the CLARIN Materials available and of providing the CLARIN End-User with the AA system for access rights. A requirement of citation and the liability issues for the Content Provider are examples of such requirements.

Sometimes it is the data itself that defines the classification. Resources including Personal Data, such as speech corpora or interviews for example, typically fall in the category RES unless the Content Provider has agreed with the subjects that might be identified later or explicitly otherwise on beforehand. A more detailed discussion of personal data issues can be found in case study 2 in Part IV below.
3.2 The Laundry Symbols assist the Copyright Owners

When the researcher or the research project leader is aware of the potential of the CLARIN community to host the resources and/or tools created with the project, the main task is to familiarize oneself with the proposed classification, to consider the requirements and conditions that might apply, the potential user scenario for the data, and finally the agreements that need to be made between the Content Provider and/or the Content Owner(s). The next step is to implement the model agreement templates for the deposition of the materials and the guidelines necessary that are presented in the following section.

To distribute tools and resources to the research community through CLARIN, the research project leader as the Content Provider needs to localize the model Deposition Licensing Agreement and sign it with the Service Provider (see Part V below and appendix A for the model agreement).

There are several CLARIN Materials hosted by the Service Providers of Content Providers throughout the CLARIN community that can directly be classified with the Laundry Symbols as the deposition agreements do not need modification. But it is the task of the Content Provider to familiarize him/herself with the classification types and then check the agreements made between the original Content Owners whether these agreements require updating or not.

The aim is to reuse existing deposition license agreements for tools and resources. For practical purposes they will be classified into three main categories. However, the original licensing agreement may be close to one of the categories but not fully compatible, in which case it is desirable to upgrade the license agreement for CLARIN purposes. See Appendix B for the Upgrade Agreement model template.
The first two classification categories PUB and ACA are more demanding for the licensing agreements. In RES basically any kind of licensing language is accepted and it is up to the user to fulfil the requirement. However, the goal of CLARIN is to have as much of its content in the first two categories as possible, and therefore the use of the upgrade agreements is strongly advised.

Some of the resources may be protected because the identity of research participants is disclosed or there is a risk that the identity of research participants could be disclosed. Each Content Provider conducts a disclosure risk review of his resources in order to determine whether any data items could be used to identify individual respondents (Personal Data, see case 3 in Part IV below). The review is also used for assessing the option to offer the material for the CLARIN use, and for informing the Service Providers about resources including Personal Data and any restrictions on such data items. Personal Data is normally disclosed to an End-User only with the permission of the Content Provider.

The summary of the classification task for the Content Owner or the Content Provider is shown in Figure 5 above.

The CLARIN prototype specifications for PUB, ACA and RES are shown in the following picture. Examples of each process are found in conjunction with the case studies 1, 2 and 3 in Part IV below.

![Resource Categories](image)

Figure 6 Three main content categories PUB, ACA and RES and the three sub-categories indicating the additional requirements Inf, NC and ReD.

### 3.3 The prerequisites for access rights

- Authentication is a prerequisite but authorization can be automatic, e.g. the applicant belongs to a certain group and signs an agreement “I agree” to the conditions related to the resource in question.
Authorization cannot be automatic, e.g. access is subject to special conditions and cannot be granted to all applicants even if they belong to a certain group but is restricted. The data can be sensitive, or governed by IPR law.

Reliable certification of the user having read the license and having signed it.

### 3.4 End-User Access Categories

The Laundry Symbols for the access categories are graphical expressions for the End-User License conditions, and they convey the information that the potential End-User needs to know about the CLARIN Materials of interest. The technical requirements for the three access categories are as follows:

- **PUB** - authentication - individual authorization
- **ACA** + authentication - individual authorization
- **RES** + authentication + individual authorization

The authentication for ACA is provided by an Identity Service Provider providing access for academic researchers. The resource owners (Content Providers) grant a license to academic researchers provided that the researchers accept the general conditions for the Laundry Symbols listed in the Terms of Service Agreement. ACA licenses are group licenses made available by the resource provider to authenticated academic researchers. Some of the ACA licenses may contain special conditions requiring the user to electronically sign additional conditions or submit additional information for certain resource collections. These tasks do not require interaction from resource owner so the individual can have access immediately after the demands are met.

Resources in the RES category require individual authorization by the resource owner bringing an essential amount of legacy resources within reach to the CLARIN community. The RES category also provides the procedure for getting access to resources for an individual non-standard research purpose or some other special purpose that needs the attention of the resource owner, e.g. the payment of a fee. [u13]

### 3.5 Extension of the CLARIN End-User Access Model

CLARIN centres may accept logins from Identity Providers for non-academic groups, but in that case the option to automatically access resources for the academic community will not be offered. Resource requests by non-academic users will by default be handled through the procedure for granting access to RES resources. It is of course possible that non-academic entities, e.g. commercial companies or government agencies, negotiate a BUS(iness) category for BUS(iness-friendly) resources, e.g. against a membership fee. The collaboration is documented in detail in the WP 7 Deliverable D7S-3.1 Collaboration Plan.
3.6 CLARIN Prototype: AAI at CSC from the End-User perspective

At this point we have discussed authorization and authentication issues from the perspective of the Content Owner, Content Provider or depositor where the core comprises the mechanisms with which the resources and the tools become CLARIN compatible. These deposition schemes will be brought to a more concrete level in Part III below with three case studies of new resources to be deposited to CSC IT Centre of Science. CSC is one of the five CLARIN Centres in the Federation established in 2009 with the Federation Agreement presented in the Deliverable D7S-4.1 [u1]. In the examples CSC is used as a prototype Centre for practical reasons and as such it represents the prototype of the CLARIN Centre in the CLARIN building phase.

We now change the perspective from that of the depositor to the CLARIN End-User who first is the applicant who wishes to obtain a right to use the CLARIN resources or tools hosted by the Service Provider, i.e. the CLARIN Centre CSC in this case. CSC hosts a web site where all CLARIN Materials are displayed and the laundry symbols PUB, ACA and RES plus the sub-categories or additional requirements NC, Inf and ReD indicate the distribution tags for each resource and tool. The CLARIN End-User enters the CLARIN Centre web site. There are two licenses s/he needs to accept online, the Terms of Service (ToS) [u14] and the Privacy Policy [u5].

The ToS contains the general terms and conditions concerning the use of CSC’s services and CLARIN Materials. The Privacy Policy indicates how End-User data is collected, used, disclosed, transferred and stored. By agreeing to the terms of ToS and the Privacy Policy the CLARIN End-User gains automatic access to all PUB material. But the ToS also contains the specifications on the Laundry Symbols for the other main category and sub-category types and by accepting these terms the End-User also accepts the liability to comply with them concerning the usage of any CLARIN resources or tools. Each CLARIN Centre is advised to localize the specific conditions governing the service taking into account the national legislation as long as the distribution tags and also the privacy policy issues for the CLARIN Materials are included in the document. The model ToS can be found in Appendix C in this document and the model Privacy Policy in Appendix D.

The list of CLARIN Materials available indicates the level she needs to enter in order to obtain access rights to the CLARIN Materials. If she is interested in a resource or tool classified as PUB, there is no action needed. It can be that the PUB material can be accessed through web interface tools such as a concordancer, or if the licenses so permit, also larger amounts of data can be accessible or downloadable.

The End-User wishing to obtain a right to use the CLARIN Materials classified as ACA will now proceed to the next level in the interface. In CSC this interface is called the Scientist User’s Interface (SUI). To be able to log in SUI the End-User needs to be authenticated via a user account in one of the Identity Federations within CLARIN. In our example the authentication is provided by the Finnish HAKA Federation.

The ACA resources and tools as well as all the PUB resources and tools from the previous level can be accessed in the SUI. At this point the End-User has already accepted the terms for all other distribution types except for the resource or tool specific other requirements or special conditions described above.
The reason for not including the resource or tools specific requirement or special conditions in the ToS is quite evident. In theory it would be possible to modify the ToS so that it would include all possible restrictions and conditions that would apply in all CLARIN Materials. But the RES category requires that the Content Owner or Content Provider is contacted meaning that the right to access the CLARIN Materials cannot be granted automatically online. The process might take some time and require activities from the Service Provider e.g. in sending out instructions or distributing information about the application or the applicant.

In practice for applying for the right to access the resources or tools under category ACA with resource or tool specific requirements or special condition or under category RES, the applicant needs to proceed from the SUI onwards to the third level in the application procedure. For this purpose, CSC has created an interface called AAI that can be either accessed from the SUI or form the main web page in case the user does not wish to log in the SUI first. The AAI presents the list of CLARIN Materials in the distribution category RES or ACA with special conditions described above. The user selects the resource(s) and/or tool(s) s/he wishes to apply a right to access and the AAI system then guides the applicant onwards. For each resource, the AAI asks the applicant to accept the requirements there are for each resource or tool and once that has been done, the system proceeds to process the application along the lines determined by a) the resource or tool in question and b) the distribution condition or requirement applicable. For example it might be that the Content Provider has set such a special condition to the resource that it is only the Content Provider that can make the decision on granting the right to access the data, even if the applicant is affiliated to a research institution. It is recommended that in such cases the Content Owner or Content Provider would consider an exception through an Upgrade Agreement to always agree to grant the right to access the resource or tool in case the applicant uses it for research purposes in a research institute such as a university.

Perhaps the most usual case in this third layer where the AAI is used is the case when the Content Owner or Content Provider wants to know more about the research project in question in order to determine if the application is acceptable or not. In these cases the AAI will register an abstract or a research plan and send it per email to the Content Owner or Content Provider for consultation. The Content Owner or Content Provider will then need to answer the query either by granting the right or by denying the request.

### 3.7 The CLARIN prototype: AAI at CSC introducing the referee system

The AAI is currently tested both for its technical implementation and for its functionality. The functionality includes a referee system whereby the Content Owner or the Content Provider agrees with a referee that s/he is willing to comment on applications for access rights for a resource or a tool governed by the Content Owner or Content Provider. There can be one or several referees for the resource or tool in question, and the applicant may choose one of them presented for him/her by the AAI. It is also possible to apply for the right to access the resource or tool without selecting the referee. The motivation behind the referee system is to assist the Content Owner in the decision making process.

The referee does not have a legal role in the application procedure. If the Content Owner has approved or proposed a set of referees who have expertise in the resources or tool or groups of them in question and possibly the scientific community within the field of the resource or the tool. The purpose of the referee system is to lessen the workload of the Content Owner/Content Provider, i.e.
the real person making the decision of each application. With the recommendation from a referee it is easier for the Content Owner or the Content Provider to make the decision but it is important to note that the referee system does not take away or reduce the responsibility of the Content Owner/Content Provider. This is especially important in the future when the number of applications is expected to grow and one of the goals of CLARIN is to find ways to hinder extra workload for the CLARIN parties.
PART IV CASE STUDIES WITH REFERENCE TO THE CURRENT PRACTICES AND THE CLARIN PROTOTYPE SPECIFICATIONS FOR THE DISTRIBUTION TYPES

The best way to test the agreements available for the CLARIN community is to implement them in practice. Fortunately it was possible to test the agreements and the distribution types at the University of Helsinki Department of Modern Languages. This was due to the fact that several researchers at the department had compiled corpora recently and they were in practice ready to deposit the data to the repository. The CSC IT Center hosting the Language Bank of Finland and being one of the current parties in the CLARIN Service Provider Federation (see D7S-4.1 for further information on the Federation [u1]) is an obvious repository for the Finnish CLARIN Materials. There are eight deposition cases pending and three of them were also technically ready for deposition, whereas questions of format are still unresolved in the other cases. Out of the three cases available, one is an ongoing project within FINCLARIN and the second one was a dissertation project presenting an interesting case of a potential derivative, and the third one was a speech corpus case with agreements with the subjects having already signed an agreement were the data could be deposited directly to CSC.

In this chapter, the case studies are presented in the order PUB, ACA and finally RES, and it has been our aim to describe all questions and issues that have arisen in the discussions with the Content Owners.

4.1 Case Study 1:

A PUB resource deposited – the case of the FinnWordNet

The Content Owner presents the case for the Content Provider

FinnWordNet is a lexical database for Finnish. It is a part of the FIN-CLARIN project. FinnWordNet contains words (nouns, verbs, adjectives and adverbs) grouped by meaning into synonym groups representing concepts. These synonym groups are linked to each other with relations such as hyponymy and antonyms, creating a semantic network. FinnWordNet can be used in language technology research and applications. It can also be used interactively as an electronic thesaurus. The first version of FinnWordNet has been created by having the words of the original English (Princeton) WordNet (version 3.0) translated into Finnish by professional translators.

The parties and the agreements involved are

- The resource: FinnWordNet
- CO+CP: UHEL/Department of Modern Languages
- SP: CSC IT Centre
- DELA: New Deposition License Agreement
- Problems: no
The discussion between the CO+CP and the SP includes the following issues:

- Copyright issues concerning the translations of the entries: it was noted that the translations are not running text but individual words and therefore no copyright can be issued for the translators.
- A log for saving information on IP-addresses and what kinds of searches have been made will be automatically incorporated in FinnWordNet
- The Material will be available also as a downloadable resource later
- There are no additional requirements. The PUB Materials are seldom problematic as the requirements that the End-User needs to comply with are usually those included in the ToS. See Appendix (C).

Finally, the metadata was briefly discussed with the CO and it was agreed that the CO will provide the metadata required for the data set for CLARIN use.

4.1.1 Current (pre-CLARIN) practices concerning the PUB resources

In this section the central findings about the PUB resources from the data gathered for the CLARIN project during 2009 and 2010 are presented. The goal of the survey was to find out how the CLARIN centres have classified their resources so far and also to test how the laundry symbols work in practice. [u7] [u9]

Materials typically classified as PUB

Typical PUB materials are dictionaries, treebanks, parsebanks, wordnets and tools such as parsers. Some of the resources are available only via a web interface and the researcher can only access the results and has no access to the actual data, e.g. national corpora such as National Corpus of Polish or Hungarian National Corpus are available as PUB only through a web interface.

Typical additional requirements

There are examples of all additional requirements attached to PUB resources and tools, but the requirement ReD is very rare. The most common additional requirement is NC.

PUB Statistics [13]

Out of the LRT Inventory distribution tags approximately 59% have been categorised as PUB, and almost half of them with an additional requirement (NC). Nearly all centres have resources that are freely available in the web, usually as free web applications for retrieving results of queries but not the original resources, or as download packages, and signing licenses or any kind of registration is not required. The CU are not registered by the centre. This kind of guest user is the most common user type as the resources in most centres are mostly freely accessible.
Administrative points on the PUB materials hosted by the Content Providers or Service Providers [u9]

In most cases there are no restrictions or any special requirements as the resource is free for anyone to use (often via a web tool for retrieving results or hits, e.g. concordances or word lists or non-specified free access via web interface [ATILF/free resources, HASRIL, ILSP/HNC with limited results and without the full set of functionalities, Meertens, MPI, OTA, Sheffield, Språkbanken, Tartu, Tübingen/some resources]. When the resource is freely available in the web, obviously no authorization is needed [ATILF, DANS, INL, MPI, OTA etc.]. In most cases the resources are not downloadable or accessible but the End-User can make queries and searches with various tools in the web, e.g. concordancers, search engines etc. [IDS, ILSP, HASRIL, Sheffield, Tübingen]

Problems identified and discussed: no questions but a general remark [u11]

It is often enough for the prospective End-User to send in personal information usually via a web form. No other means for authorization or authentication are needed. When the prospective End-User sends in a web form, the identity of the applicant is trusted by the Content Provider.

4.1.2 CLARIN prototype specifications for PUB

PUB resources are publicly available on the internet and they need to be made available mostly for convenience. For new PUB resources, the preferable license is Creative Commons Zero (CC0) or the Open Database License (ODbL). However, the licenses should allow at least Open Source definition 4-6 and 9-10, http://www.opensource.org/osd.html: Integrity of The Author’s Work

- No Discrimination Against Persons or Groups, i.e. there are no limitations (based on status or geographical location etc.) on who can access and use the tools and resources
- No Discrimination Against Fields of Endeavour, i.e. there are no limitations on the purpose the tools and resources are used for.
- License Must Not Restrict Other Content
- License Must Be Technology-Neutral
- Closely follow Protocol for Implementing Open Access Data.
- The license should allow distribution via CLARIN under all circumstances. [u13]
However, there is no requirement to allow sub-licensing of the content. [u8]

In the survey, all the resources that were classified to the category PUB but which also had additional requirements NC, Inf or ReD, will not be problematic cases as the additional requirements and also the general obligation to cite the Material in all publications are included in the ToS and agreed by and thus binding the End-Users. In the statistics the only other requirement that would need consideration happened to be the obligation to cite so there were no resources classified as PUB that would not serve as CLARIN PUB Materials.

In theory, other than the requirements specified in the ToS would preferably be renegotiated with the Content Owner in order to remove the requirement in question, or the resource or tool would maybe need to be moved from the category PUB to the category ACA or even RES; or in the worst scenario the resource or tool would be need to be removed from the CLARIN database of Materials.

4.2 Case Study 2: An ACA resource deposited and access granted – the case of the JP data set

The Content Owner presents the case for the Content Provider

The resource is a Finnish lexical dataset created in the course of the PhD research of the CO and consists of two parts. The first part of the dataset is derived from the Helsingin Sanomat 1995-1997 material that is now available in the Language Bank of Finland, using an early version of Connexor's (then Conexor) fi-fdg parser for language analysis. This dataset contains a distributional representation for all Finnish nouns that occurred in the corpus more than one hundred times. Such distributional representations consist of word lemmas that occurred as heads or dependents of the represented word, together with the dependency relation identified by the parser, and the cooccurrence counts, or distilled dependency concordances.

The second part of the dataset is derived from the first part using software that the CO developed himself. This dataset consists of distributional similarity lists for all the frequent nouns, each list giving the one hundred nouns that were most similar to the head of the list.

The corpus and the parser were available to the CO in the Department of General Linguistics. Can they be treated now as if they were used in the Language Bank?

The first dataset contains lots of word cooccurrences identified by the parser, including linguistic labels introduced by the parser. The parser also provides the low level language analysis (tokens and lemmas). In the second dataset, only the frequent noun lemmas are visible. The corpus is to be submitted to CSC in order to make it accessible for other researchers interested in similar language research questions.

The parties and the agreements involved are

- The resource: dataset created by Jussi Piitulainen
- The type: dataset
The discussion between the CO+CP and the SP includes the following issues:

The dataset includes word lists from the original corpora, so now copyrights derive for the Content Owners of the corpus Helsingin Sanomat, but the compilers of the parser may have copyright.

The researcher did not use the parser at CSC but it was used at the Department of General Linguistics. Are there derived copyright issues that have to do with this?

So we assume that the researcher Jussi Piitulainen is the sole owner of the copyright.

Does the Content Owner wish to add additional requirements? We discussed NC but that is not a good idea since the dataset might prove useful for the open source community.

Citation requirement in ToS is sufficient but Jussi Piitulainen will provide a predefined form for the metadata that is not obligatory but recommended.

Redeposition is not required since it might hinder the use of this dataset as part of a free software development, e.g. under GPL.

Finally, the metadata was briefly discussed with the Content Owner and it was agreed upon that the Content Owner will provide the metadata required for the data set for CLARIN use.

4.2.1 Current (pre-CLARIN) practices concerning the ACA resources

In this section the central findings about the ACA resources from the data gathered for the CLARIN project during 2009 and 2010 are presented. The goal of the survey was to find out how the CLARIN centres have classified their resources so far and also to test how the laundry symbols work in practice. [u7] [u9]

Materials typically classified as ACA

ACA materials contain text corpora of various kinds. There is no evident reason behind the choice of PUB, ACA and RES other than maybe the wish of the CO to be able to grant the user accounts individually. Historical corpora can be in any category, whereas parallel and multilingual corpora tend to be either PUB or ACA.

Typical additional requirements

There are examples of all additional requirements attached to PUB resources and tools, but the requirement ReD is very rare. The most common additional requirement is NC.
ACA Statistics [13]

Out of the LRT Inventory distribution tags approximately 30% have been categorised as ACA, and in practice all of them with an additional requirement s (NC, Inf or other restriction or a combination of these).

The most usual prospective CU is researcher/student/other personnel in a research institution [CSC, IDS, Vienna U, ILSP, ILC, MPI, Tartu]. The resources can also be offered for in house use also [Vienna U] or solely [IDS, some resources].

ATILF/Frantext and Wroclaw U invite applications from institutions only. It is also possible to define a predefined group of users who can access the resources without any form of registration or application procedure [Språkbanken, DANS/not for linguistic data but archaeology]

<table>
<thead>
<tr>
<th>ACA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ACA</td>
</tr>
<tr>
<td>32</td>
<td>ACA+NC</td>
</tr>
<tr>
<td>8</td>
<td>ACA+NC+Inf</td>
</tr>
<tr>
<td>2</td>
<td>ACA+NC+Other restriction</td>
</tr>
<tr>
<td>1</td>
<td>ACA+Inf</td>
</tr>
<tr>
<td>0</td>
<td>ACA+Other restriction</td>
</tr>
<tr>
<td><strong>44</strong></td>
<td>sum</td>
</tr>
</tbody>
</table>

Administrative points on the ACA materials hosted by the Content Providers or Service Providers [u9] [13]

To be granted access to the resources, the applicant may also need:

- green light from the CO [CSC, MPI, OTA, probably more...], this is often done by the SP or the CP
- privacy act statement [IDS]
- to agree to an existing license GNU GPL/ Creative Commons for non-commercial use [IDS/word lists etc. Språkbanken, Tübingen, most tools in general]
- to pronounce the obligations and rights of both parties [ILSP]

Even though a minority with respect to the resources hosted by the centres, the prospective CU quite often needs to enter into an agreement with the CP/CO. The prospective CU usually fills in a form (User Agreement/Licensing Agreement/End-User agreement), prints it out, and sends it to the CP per regular mail [IDS, ILSP, INL, MPI, OTA Tübingen, Språkbanken, Wroclaw U], or alternatively scans it in and sends it as an attachment to an email. CSC requires the web form both to be filled in, sent to both CSC and CO (automatically distributed by the web form), and also printed out, signed and sent in per regular mail to CSC.
Problems identified and discussed: the FAQ [u11]

Q: The CC license is not tailored for software. Should we use ODbL?
A: For software, popular open source licenses like the GNU GPL, the BSD or the EU Public License would be better.

Q: Probably it might be helpful, if you could give some examples for categorizing e.g. common resource-licenses like GPL or CC.
A: CC and GPL licenses fall into the category PUB. Also LGPL belong to this category. For CC, we have similar additional restrictions like CC-by corresponding to our PUB+Inf, CC-sa corresponding to our PUB+ReD, and CC-nc corresponding to our PUB+NC. Tools and resources that do not allow derivative works like the CC-nd license can note this restriction in the field for other additional requirements.

4.2.2 CLARIN prototype specifications for ACA

ACA resources are resources for which an academic group license has been prepared and which license can be electronically signed by an identifiable academic researcher without further scrutiny by the resource owner. Academic Use requires that the use is somehow related to an academic institution, and the resources are thus available at least for anyone working and/or studying in an academic institution.

The content can be used for general academic purposes, e.g. “teaching” is not normally listed as permitted but that may or may not mean that it is prohibited. [u13]

For Academic Use (ACA) the license agreement includes an additional requirement that the use is somehow related to an academic institution/use. Here the problem may and will arise from the definition of academic use. The main points to check are:

- The content should be available at least for anyone working or/and studying in an academic institution recognized by an Identity Federation (IdF) with which CLARIN has a service provider agreement.
- The content can be used for general academic purposes. However, it is relatively hard to define exactly what this means e.g. “teaching” is not normally listed as permitted act but that may or may not mean that it is also prohibited. In case of doubt, contact the right holder and ask (or the CLARIN Academic Upgrade Agreement may be used to obtain the missing rights). [u8]

The Academic licenses may also include a provision, which requires the user to accept the licensing terms separately for each CLARIN Material.

As a prototype of a deposition case study 2, the following IPR issues need to be considered

- Is the corpus or data set a derivative of the original data?
- Who has the copyright to the new corpus or data set?
What kinds of agreements are needed in order to safeguard the copyright of the original Content Owners?

Within the CLARIN context, the question of derivatives is highly relevant, as the data and corpora available are of such magnitude that in the future it is probable that the borderline of derivatives / not derivatives will become an issue in language research projects and individuals working with corpora or language materials offered by Content Providers.

4.2.3 The End-User Raymond applies for a right to access the CLARIN Material JP Dataset

1. Raymond Researcher from the MPI in Nijmegen wants to use language resource dataset JP, stored at CSC in Helsinki/Espoo
2. Raymond goes to CLARIN resource listing at www.clarin.eu as a new End-User
3. Raymond reads the ToS and the Privacy Policy and accepts the terms by clicking “I agree”
4. Raymond selects resource dataset JP – with unique PID(G) – from a list
   Service informs Raymond that he has to agree to CLARIN general End-User License terms concerning research use
5. Raymond clicks link
   ”Apply for access to resource dataset JP”
6. Raymond is redirected to the AAI service at CSC https://aai.csc.fi via logging in through his Dutch national IdF service SurfFederatie (specifically his local IdP: MPI/Nijmegen)
   Raymond is shown the general CLARIN terms of use for research purposes for the specific resource dataset JP
7. Raymond ticks the box ”I have read and understood these terms of use for research and agree to abide by them” and presses the ”Agree” button
   Raymond's Identity Attributes raymondr@mpi.nl (eduPersonPrincipalName@Domain) as provided by his IdP (MPI/Nijmegen) are now linked with the resource identifier PID(G) in Authorization Records (AR) at CSC
8. Raymond proceeds to get access to resource dataset JP

4.3 Case Study 3: A RES resource deposited and access/user account granted - the case of FDC

The Content Owner presents the case for the Content Provider

The Finnish Dialogue Corpus (FDC) consists of 10 informal Finnish dialogues by 20 native adult speakers from the capital city area in Finland. Each dialogue is represented by two sound files of equal duration stored in AIFF format. The total duration of the sound files is approximately 8 hours. The sound files have been annotated with the Praat program, and the annotations are included in the corpus as Praat TextGrid files.
The parties and the agreements involved are

- The resource: data set created by Mietta Lennes
- type: speech corpora
- CO chain: see below
- Agreements: between the subjects and the compiler; CSC is granted permission to distribute the data to academic purposes
- CP: IBS (Institute of Behavioural Sciences)
- SP: CSC
- DELA: New Deposition License Agreement, ok except for paragraph 9
- Motivation: sensitive data

The discussion between the CO+CP and the SP includes the following issues:

In the compilation phase the compiler of the corpus and the subjects signed an agreement in which the subjects explicitly grant the transfer of the data set to CSC as well as access for the researchers in the Finnish universities for research purposes. The chain of IPR rights looks like the following:

CO Mietta Lennes/compiler – CO(2) CP Institute of Behavioural Sciences at the University of Helsinki (IBS) – SP CSC

The compiler of the data will now enter into an agreement with CSC for the distribution of the data set for the researchers in the Finnish universities under the category RES + NC. This is because of the data type, i.e. speech corpora, is sensitive data because of the personal data that might be possible to incur. When the data set will in the future be upgraded to category ACA + NC, either new consent form will be signed between the subjects and the CO or, if possible, the data will be checked to find out if there are personal data that can lead to identification of the subjects or in principle the data might be anonymised in order to remove all personal data from the files. But unfortunately the nature of FDC is such that it is not possible to anonymise because it is important for research purposes to be able to retrieve the whole discussion file from the data. With speech corpora it needs to be noted that anonymisation eventually affects the data as it is no more possible to study the anonymised items e.g. the pronunciation of proper names.

It is also important to preserve and hand out some personal data of the subjects. Publishing databases of persons might not be possible due to personal data issues.

Classification: RES + NC+ license from the CO

In this context the following special conditions were discussed:
Inf – the compiler of the data set does not wish to activate this condition because in the future it might potentially mean lots of emails informing her about how the data set has been used and in which publications. On the other hand, it would be optimal for the collector of the data set that this kind of information would be stored at the SP and if the compiler would be able to get reports from SP when required.
ReD – the compiler of the data set did not wish to add this condition because in speech corpora it is difficult to draw the line as in when the data would require redeposition. For example, if the user would change some details in the annotation of the layers or the boundaries of some phonemes, the resulting data set might or might not represent a change worth redeposition. It would also require a sophisticated version control system from the SP. Redeposition would also require thorough documentation from the depositor and this documentation should be available in the metadata.

The CO(2) also wishes to add the referee system to her data set. In practice she would be willing to name 1-3 referees who would be able to comment on the applicability of the potential user applying for a right to access the data set in the future.

Citation – the general provision in the ToS cover the citation requirement sufficiently. There is no need to explicitly state the citation model at this stage, although it might be an option to include a model citation paragraph in the metadata. On the other hand if the citation would be in a predefined form, SP could follow how the resource has been cited by a search engine or similar.

Finally, the metadata was briefly discussed with the CO and it was agreed upon that the CO will provide the metadata required for the data set for CLARIN use. Now the corpus is ready to be deposited to the Language Bank of Finland.

**Summary of the procedure:**

1. Authorization: CO, referees have a consulting role
2. Authentication: SP via AAI
3. New Deposition License Agreement Appendix A.
4. Towards upgrading resources from RES to ACA: possible if requested by the CO

### 4.3.1 Current (pre-CLARIN) practices concerning the RES resources

In this section the central findings about the RES resources from the data gathered for the CLARIN project during 2009 and 2010 are presented. The goal of the survey was to find out how the CLARIN centres have classified their resources so far and also to test how the laundry symbols work in practice. [u7] [u9]

**Materials typically classified as RES**

RES materials are typically speech corpora of various kinds. Also resources that are freely available (PUB) via a web interface can be accessible as local versions under the category RES.
Typical additional requirements

There additional requirements attached to RES resources and tools are mostly a requirement to obtain a license from the CO but also Inf and NC are common as additional requirements. The most common additional requirement is NC.

RES Statistics

Out of the LRT Inventory distribution tags approximately 20% have been categorised as RES, and in practice they must always have an additional restriction, typically that the CO needs to approve the application individually.

<table>
<thead>
<tr>
<th>RES</th>
<th>4 RES</th>
</tr>
</thead>
<tbody>
<tr>
<td>RES+NC</td>
<td>3</td>
</tr>
<tr>
<td>RES+NC+Inf</td>
<td>3</td>
</tr>
<tr>
<td>RES+NC+Inf+Other restriction</td>
<td>1</td>
</tr>
<tr>
<td>RES+NC+ReD</td>
<td>1</td>
</tr>
<tr>
<td>RES+NC+Inf+ReD</td>
<td>4</td>
</tr>
<tr>
<td>RES+Inf</td>
<td>13</td>
</tr>
<tr>
<td>RES+Other restriction</td>
<td>30</td>
</tr>
</tbody>
</table>

Administrative points on the ACA materials hosted by the Content Providers or Service Providers [u9]

If the resource is available on special request only, the prospective CU usually sends an email to the contact person indicated on the web site [IDS, ILC, HASRIL, DANS, INL] and receives information by email about how to proceed.
The user can also be required to contact the CO directly [Tartu].
The SP will in some cases need the CO to actively approve the application before they can create an account or provide access to the resources [CSC, OTA].

Problems identified and discussed: the FAQ [u11]

Q: Most spoken resources will probably fall into the RES category – how do we deal with that? Can we use the CLARIN infrastructure for trustworthy limited redistribution rights?

A: The Trust Federation makes it possible to grant licenses based on individual applications under RES, which will make more resources containing, e.g., personal data accessible as they will always require individual licenses unless they are anonymised.
In addition, a CLARIN centre can offer the End-User an option to use a resource on a server without the need to download the resource to his computer. However, if the license strictly forbids redistribution, there is nothing CLARIN can do but to bring this to the End-User’s attention. CLARIN can still display the metadata of the resource, to show that it exists and what it contains and leave it to the individual researcher to approach the owner in order to negotiate a licence.
Q: What protocol should be applied to the resources that cannot be made available under the Protocol for Implementing Open Access Data, Creative Commons, ODbL? Can supplements to the agreement be developed to include the benefits granted for the CO such as terms of payment?

A: In the RES category, the license can be almost anything but even there it would make sense to take an existing agreement and amend it to take into consideration the requirements of the CO. However, the questions regarding payments are political and may have to be negotiated separately. One solution is to include metadata about the resource in the LRT Inventory with clear indications of the payment and access restrictions while providing a link to an external web site with more detailed information about the payment procedure.

Q: Are resources with excessive and firm conditions under the category RES welcome in CLARIN?

A: Metadata about the resource in the LRT Inventory with clear indications of the restrictions are always welcome.

Q: Should authentic written texts (such as letters, essays) be classified as RES?

A: If they contain, e.g., personal data, RES is a probable category, but with the consent of the owner other categories are also applicable.

Q: Limited distribution rights do not convince the right holders (COs) unless these limits are controlled technically and the liabilities are agreed upon in the license agreements.

A: Building a full DRM system inside CLARIN is an extremely hard task. It would make the use of the material much more cumbersome. Furthermore, the experiences from the music and book business seem to indicate that dropping the DRM systems does not significantly add piracy.

In practice, limited distribution means that CLARIN can grant researchers the permission to use the data, but not the right to distribute the data further. If we wish to be able to process samples of real data in new ways and not only read small samples of words in context, it is not really possible to control the data samples.

However, if the CO really is concerned only with the fact that the original data could be reproduced, sold and read as an entire work, many application developers and research projects will happily accept the sentences of the work in a scrambled order and with some random sentences left out so that the data no longer resembles the original work and the original work cannot be reconstructed. This of course only applies to regular books and newspapers, but they constitute the current majority of works with linguistic content eligible for DRM control.

4.3.2 CLARIN prototype specifications for RES

**Restricted Use** (RES) includes the resources that do not fulfil the requirements for ACA but still could be offered to the users if certain additional requirements are met.
Special issues due to the sensitivity of the data require that the data set is deposited as a RES category with a special condition that the application must be approved by the CO as well as the condition that the data set is available for strictly non-commercial use only. [u11]

The first thing to check is that the license gives enough rights so that CLARIN may share it in the first place. If the access is limited to a certain room in a certain library, there is no much point of adding it to CLARIN and the use of upgrade agreement is required. However, the most typical reasons for a resource to fall under the scope of RES are less problematic:

- a requirement to submit detailed information (e.g. abstract) about the planned usage, or
- specific data protection-related additional requirements typically applied for speech and video material, i.e. material including Personal Data

**Terms of Service (ToS) and End-User License (EUL)**

The ToS includes the requirements and conditions that are generally applicable to all resources under PUB, ACA and RES. The additional requirements and also the CLARIN Material specific requirements and conditions are included in the End-User License the applicant is expected to agree to separately. See Part VI below for more information on ToS and EUL.

The End-User applying for the right to access any RES resource or tool will first sign the ToS agreement and the Privacy Policy by accepting the terms on the web page. He/she will now be able to move forward to the part of the CLARIN Service and the level where the special conditions will be displayed for the End-User to consider and accept. As the regular conditions and requirements already agreed upon are embedded in the RES –category and therefore understood by the End-User. Therefore, the End-User has already agreed upon part of the licensing terms governing the CLARIN Material in question and s/he will be accepting the other set of distribution tags presented by the AA Infrastructure at the RES level in the CLARIN Services system. [u13]

### 4.3.3 Special issues concerning personal data; case study RES

As the RES resources typically contain sensitive data or data that could be used for disclosing the identity of data subject, it is important to consider issues concerning personal data. This chapter introduces different aspects that need to be considered in addition with the personal data.

**Personal Data**

The legal framework around the use of personal data in research is a complicated. Within EU, each country has implemented the Directives according to its own national interpretation (including the directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Directive 95/46/EC; [u16]). Hence there are many differences in some aspects of the legislations in EU, i.e. the guidelines on personal data can only give general advice and should be adapted to each country.

The Directive 95/46/EC defines personal data as:

> Any information relating to an identified or identifiable natural person (‘data subject’); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.
To determine whether a person is identifiable, all likely and reasonable means for identifying a person should be considered.

**Personal Data in Research**

For new personal data, the best approach is to procure sufficient consent for research and secondary use from the research subjects. (see the guidelines for the Consent Form)

If personal data have been collected with insufficient rights for distribution or secondary use, there may still be some options, e.g., anonymisation for distribution or certain exemptions for scientific, historical or statistical research purposes. These will be briefly outlined below and in the section on End-User agreements.

**Anonymisation**

Anonymisation can sometimes be a solution to avoid complicated issues involving Personal Data, as the principles of data protection do not apply to data rendered anonymous in such a way that the data subject is no longer identifiable.

When necessary, the protection of privacy should be ensured through anonymisation measures. In case of anonymisation, please notice:

- It is not always clear who can remove identifiers from stored data and under what circumstances. The act of anonymisation can itself be seen as a form of data processing, in which case the anonymisation requires consent from the data subject or should at least be done by the person who originally collected the data. Also the adequate level of anonymisation can cause problems.

**Categorising the Resources Including Personal Data**

Content that has Personal Data typically falls into the categories RES with an additional mention that it requires personal access permission.

However in some cases the category PUB may also be considered with Personal Data:

1) There is an explicit consent (from the data subject) for processing and disclosing Personal Data without any limitations.

2) Legislation permits disclosing of Personal Data, e.g. exemptions for journalistic purposes such as naming a politician or someone having a public office or a leading corporate position in a piece of news for the public good.

**Secondary Use of Personal Data**

If the resources contain information that could be used for disclosing the identity of human subjects (Personal Data), the End-Users of such resources can be required to fill in an Application for use of restricted data from the CLARIN database [u3] and to accept an additional Data User Agreement (DUA). [u4]

**Special Issues in Secondary Use of Personal Data**

The protection of the subjects' privacy must be ensured by setting strict conditions on the secondary use of Personal Data. In general, Personal Data, which have been obtained for a particular purpose, cannot be used for another purpose without the consent of the data subject(s). End-Users of the data can be requested to sign an agreement on the conditions for secondary research (DUA) and if needed also a pledge of confidentiality.

Personal Data may only be processed and disclosed for secondary use when there is a clear legal purpose for doing so. Legal purpose can mean:
The data subject has given his consent for processing and disclosing.

The data subject will be informed about disclosing beforehand or afterwards and this is accepted by national laws or regulations.

Special legislation at the national level allows processing and disclosing of Personal Data.

Typically there are certain exemptions, which allow Personal Data to be used for scientific, historical or statistical research purposes (even if the data were originally gathered for other purposes). For example, the protection of privacy may not apply to materials that are in the public domain or to published data, which can concern individuals and their activities in the fields of politics, business, official activities and culture. [14]

Particular attention needs to be paid to the following:

- Data subjects should be fully informed on how and for what purposes their Personal Data will be used. Personal Data should not be disclosed for secondary purposes without the consent of the subject or without the authority of law, e.g. exceptions for research use. Research data including Personal Data typically may not be used or handed over for other uses besides research. It is particularly unacceptable to reveal information on research data or hand over data in such a way that it could influence the evaluation, treatment or position of individual subjects.

- Usually the data subject's consent shall mean any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to Personal Data relating to him being processed.

- Particular care should be taken when the processing involves sensitive Personal Data (racial or ethnic origin, political opinions, religious beliefs, trade union membership, health, sex life, criminal convictions). Many times disclosing the sensitive Personal Data is prohibited or it can only be disclosed if you have the explicit consent of the data subject.

- Collected Personal Data must be adequate, relevant and not excessive in relation to the purposes for which they are processed. A basic principle regarding the collection and storage of Personal Data is the need for Personal Data in research. Personal Data may not be collected and stored unnecessarily. If research data can be reasonably analysed without direct identifiers and there are no research grounds for storing identifiers, only data from which identifiers have been removed may be produced for research purposes and stored for secondary research. Data with identifiers can be collected and used when this is appropriate from the viewpoint of that particular research. With the consent of the subject, data might also be stored for secondary research with identifiers.

- In some cases, data subjects have a right to object to the processing of their Personal Data.

- Find out if you have an obligation to notify the data protection officers of collecting and disclosing the Personal Data.

- Directive 95/46/EC imposes a prohibition on the transfer of Personal Data to countries outside Europe unless the country of destination provides an adequate level of protection.
4.3.4 The referee procedure; case study RES

As the application procedure is handled with AAI it is possible if desired to add a component to the system for a special referee round. It is the sole responsibility in the referee round of the Content Owner to initiate the round in the first place, and also to choose the referee candidates who can then be asked to register himself or herself as a referee for this particular CLARIN Material. In our case study the flow of events could be the following:

Mietta Lennes considers the need for referees as part of the application process for her speech corpora. In the future Mietta wishes to upgrade her resource from RES to ACA but for the time being, her resource will be classified as RES for the personal data issues. She believes that choosing some referees would assist her in her decision making process, and so she decides to ask Stefan Werner from the University of Eastern Finland and Martti Vainio from the University of Helsinki to function as referees. Both accept the invitation and register as referees for Mietta’s resource. Both referees work with speech corpora.

The End-User Raymond who applies for a right to use Mietta’s resource will find Stefan and Martti as referees in the AAI referee page and s/he will choose Stefan as a referee. It is also possible for her or him not to choose a referee. Raymond writes a short abstract and describes how he will use the resource in his research and when he has sent the application forward after accepting the license term, the AAI sends an email to Stefan presenting the data in the application form. Stefan reads the application and decides to recommend that Mietta accepts Raymond’s application. Mietta then receives the original application but also Stefan’s statement of recommendation. Mietta approves of the application.

4.3.5 The End-User Raymond applies for a right to access the CLARIN Material FDC

1. Raymond Researcher from the MPI in Nijmegen wants to use language resource FDC, stored at CSC in Helsinki/Espoo
2. Raymond goes to CLARIN resource listing at www.clarin.eu as a new End-User
3. Raymond reads the ToS and the Privacy Policy and accepts the terms by clicking “I agree”
4. Raymond selects resource FDC – with unique PID(G) – from a list
   Service informs Raymond that he has to agree to CLARIN general End-User License terms concerning research use
5. Raymond clicks link
   ”Apply for access to resource FDC”
6. Raymond is redirected to the AAI service at CSC https://aai.csc.fi via logging in through his Dutch national IdF service SurfFederatie (specifically his local IdP: MPI/Nijmegen)
   Raymond is shown the general CLARIN terms of use for research purposes for the specific resource FDC
7. Raymond ticks the box "I have read and understood these terms of use for research and agree to abide by them" and presses the "Agree" button

Raymond's Identity Attributes raymondr@mpi.nl (eduPersonPrincipalName@Domain) as provided by his IdP (MPI/Nijmegen) are now linked with the resource identifier PID(G) in Authorization Records (AR) at CSC but subject to authorization from the CO.

8. Raymond is redirected to a page where there is a list of referees for the specific resource FDC. Raymond chooses Stefan Werner as a referee from the list.

9. Stefan Werner at the University of Eastern Finland receives email from CSC and proceeds as instructed to send in his recommendation for the CO.

10. The CO Mietta Lennes receives Raymond’s application as well as Stefan’s recommendation. Mietta approves the application.

8. Raymond gets his user account information and proceeds to have access to resource FDC.

   [NB! This process is in a test phase at CSC and subject to change]

4.3.6 Overall Statistics from the survey based on the classification task in the LRT Inventory

The statistics presented here are from the Survey conducted during 2010 by WP7. In the survey we asked approximately 400 Content Owners or Content Providers in the LRT Inventory for Resources and Tools (www.clarin.eu/inventory) to participate in an effort to classify their materials according to the Guidelines [u9]. We hoped each Content Provider to try to classify at least one resource or tool that would fall under each category offered and that would require a certain set of Laundry Symbols. Some of the comments and classification decisions were made in conjunction of discussions over the classification effort per phone and/or email within the FIN-CLARIN consortium in Finland. By the end of September 2010, a total of 153 resources or tools had been classified.

The particular findings on what kinds of CLARIN Materials tend to fall in these categories shown in the table below can be found in the section describing the distribution type PUB, ACA or RES above.

The feedback received and the questions asked during the survey now form the FAQ questions and answers that are presented in this document in conjunction with the case studies above and also in the last part of this document on some issues concerning the future of CLARIN.
RESOURCES AND TOOLS; data both from the LRT Inventory and other sources (emails)

<table>
<thead>
<tr>
<th></th>
<th>PUB</th>
<th></th>
<th>ACA</th>
<th></th>
<th>RES</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>42 PUB</td>
<td>23 tools, 19 resources</td>
<td></td>
<td>1 ACA</td>
<td>1 resource</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26 PUB+NC</td>
<td>6 tools, 20 resources</td>
<td></td>
<td>32 ACA+NC</td>
<td>6 tools, 26 resources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 PUB+NC+Inf</td>
<td>5 tools, 1 resource</td>
<td></td>
<td>8 ACA+NC+Inf</td>
<td>1 tool, 7 resources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 PUB+NC+ReD</td>
<td>1 tool</td>
<td></td>
<td>2 ACA+NC+Other restriction</td>
<td>2 resources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 PUB+ReD</td>
<td>1 tool, 1 resource</td>
<td></td>
<td>1 ACA+Inf</td>
<td>1 resource</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 PUB+ReD+Other restriction</td>
<td>1 resource</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>78 sum</strong></td>
<td><strong>36 tools, 42 resources</strong></td>
<td></td>
<td><strong>44 sum</strong></td>
<td><strong>7 tools, 37 resources</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 RES</td>
<td>4 resources</td>
<td></td>
<td>3 RES+NC</td>
<td>3 resources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 RES+NC</td>
<td>3 resources</td>
<td></td>
<td>3 RES+NC+Inf</td>
<td>3 resources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 RES+NC+Inf+Other restriction</td>
<td>1 resource</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 RES+NC+ReD</td>
<td>1 resource</td>
<td></td>
<td>4 RES+NC+Inf+ReD</td>
<td>4 resources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 RES+Inf</td>
<td>1 resource</td>
<td></td>
<td>1 RES+Inf</td>
<td>1 resource</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13 RES+Other restriction</td>
<td>13 resources (11 license from CO)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>30 sum</strong></td>
<td><strong>30 resources</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**152 Total Oct 15, 2010**

Table 1. Statistics on the distribution types registered in the LRT Inventory per October 15, 2010
PART V MODEL AGREEMENTS BETWEEN CLARIN PARTIES

Part V introduces the various agreements that the CLARIN parties enter into in order to safeguard the mutual trust relations and in order to offer the potential End-Users access to the CLARIN Materials they wish to apply for. These agreements are shortly outlined in this section. Some questions presented by the CLARIN centres and Content Providers in the surveys 2009-2010 [u15], [u9] and the Frequently Asked Questions and answers to them are listed under each agreement.

NB. The Agreements in the Appendices are model agreements and the national legislation and binding regulations need to be taken into account.

5.1.1 Agreement btw Service Providers: an agreement of uniform services

In the Service Provider Federation Agreement, five CLARIN Centres agreed on principles of cooperation within CLARIN. See D7S-4.1 Set of Service Provider Federation Agreements for CLARIN Centres for more information [u1]. These agreements are shown in Figure 7 above as arrows between the Service Providers.
The agreements between the Identity Federations and the Identity Providers and the Service Providers are also presented in figure 7 above and incorporated in the Service Provider Federation Agreement and for that reason they are only mentioned here.

5.1.2 Agreements btw Content Owners/Content Providers and Service Providers

CLARIN Deposition & License Agreement

With the Deposition Agreement the depositor agrees to make new CLARIN Material available through the CLARIN Infrastructure. For this purpose the depositor provides the Service Provider a metadata file including the general information on the material in question but also the distribution categories applicable, i.e. the laundry symbols and the eventual other restrictions that the depositor sees fit to require. In the agreement, this requirement to make the metadata public is expressed as follows:

The Repository shall publish the metadata and make them freely available, on the basis of the documentation that the Depositor provides with Content. The term metadata refers to the information that describes the digital files. Other documentation that relates to the dataset and is provided by the Depositor shall be published and made freely available, unless the Depositor has specified that certain documents must not be made freely available. Documents that contain personal data will not be made freely available.

The model agreement template: Appendix A

The summary about licensing problems within the survey carried out by WP7 during 2010 is presented in the form of frequently asked questions. Some of the questions concerned a situation where a new resource or tool is deposited to a repository, or an agreement over a resource or tool is to be upgraded to become CLARIN compatible. The questions and answers are listed below.

FAQ:

Q: What happens if after signing the DELA the CO wants to amend the agreement so that it no longer fits into the CLARIN categories?
A: It is possible to change the classification to a more restricted category. If no more restricted category is suitable, the material is no longer distributed through CLARIN. If there is a paragraph in the original license on termination of rights once the agreement is terminated, it is basically the task of the CO/CP to inform the CU about it and ask for removal of data. If the rights have been provided permanently, the CU can use the resource in the same way than before. Within CLARIN, it is important that the CP is not made responsible for the removal of the resource from the CU computer or whatever means the data is stored (can also be CDROM etc). The metadata for the resource can still be listed in the LRT Inventory with a link to the external resource provider.

Q: How far can the License differ from suggested examples and guidelines? Is it possible to participate in CLARIN even if the SP or the CP concludes a different License agreement?
A: As long as RES is used, there are not that many requirements for the license. Basically the content should be somehow available for the users EU-wide. As a general principle, the license can be broader than the category but not stricter. Each category has a set of minimal requirements that all resources of that category need to comply with.
Q: Commercial usage is a tricky question – we would not like to exclude it totally but we would also not like to promote commercial product development based on our tools or resources. Can there be a restriction allowing commercial use under a separate agreement between the content owner and the end-user?

A: As it is by default always possible to negotiate additional rights with the content owner, only the more restricted conditions need to be mentioned. However, two entries can be provided for the same resource with the first one e.g. as ACA+NC and the second one e.g. as RES with commercial usage possible under special conditions.

Upgrade Agreement

The upgrade agreement is used when the Content Owner wishes to transfer the CLARIN Material from the category RES to either category ACA or category PUB. In the upgrade agreement model template there is also a provision available with which the Content Owner and/or the Content Provider can authorize the Service Provider to automatically grant access to the CLARIN Material in question under specific conditions, e.g. if the applicant is affiliated in a research institute.

Model agreement template: Appendix B

FAQ:

Q: Does the upgrade agreement effectively replace the original agreement?
A: No – it changes certain relevant parts, but the original agreement is still applicable.

Q: Upgrading agreements can be costly – who will cover the costs?
A: It is indeed true that negotiations and lawyers add costs – anyway it is up to CLARIN to decide if it wants to subsidize this process. The model agreements are made in order to save some costs.
If the CO requires additional compensation to grant additional rights, it is to be decided on a national level or by the local CP/SP whether they will pay for these additional rights.

Q: Who is responsible for upgrading agreements made during already expired projects?
A: Unfortunately no-one is. If the original right holder no longer exists, the material is in effect in the public domain (true orphan work), i.e. when there is no CO, there is nobody to take the case to court. In case the resource is compiled from material owned by existing COs, e.g. an article database with authors alive, all COs need to be contacted for negotiating a new license agreement. This is taken into account in the model agreements.

Q: Upgrading the present agreements might be structurally difficult using the model upgrade agreement. Does it matter how the original agreement is formulated? Is it possible that the upgrade agreement replaces the original agreement?
A: If the original agreement cannot be amended for some reason, there are two options: use the model DELA agreement or adapt it to write your own.

Q: Can the concepts be so different between the original agreement and the upgrade agreement that the original agreement cannot be updated but needs to be replaced?
A: That is unlikely but still possible.
Data Processor Agreement btw Content Provider and Service Providers

If any of the Service Providers are processing Personal Data inherent in the resources on behalf of the Content Provider, the terms of data processing should be agreed in writing. In the consideration whether the Service Provider is acting in the role of Processor, the essential element is that the Processor only acts "on behalf of the Controller" and thus only subject to his instructions.

According to the Directive 95/46/EC [16]:
- Data Controller shall mean the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data
- Data Processor shall mean a natural or legal person, public authority, agency or any other body which processes personal data on behalf of the Data Controller.

The purpose of the Data Processor Agreement is to ensure that the processing of Personal Data is carried out in accordance with the articles of directive 95/46/EC as set out in the legislation of the member state in which Data Controller is established, subject to that legislation being in accordance with the directive. These Articles require the Data Processor to take appropriate technical and organizational measures to protect the Personal Data.

Sometimes the Processor may choose not to process the data himself, but may have recourse to a subcontractor who processes the data on his behalf (e.g. Identity Federations). Use of subcontractors has to be agreed in the Data Processor Agreement.

A model agreement is available in the appendix but depending on the content of the processing services and the national legislation of the Content Provider, the model agreement may need some modification by the parties. Please note that the model agreement is intended for use only within the EU.
Model agreement template: Appendix G

5.1.3 Agreements btw End-User and Content Provider/Service Provider: the End-User Licenses

End-User License (included in the ToS)

The purpose of the End-User License (EUL) is to give the user right to use the resource or tool through presenting the minimal set of requirements that the Content Owner or the Content Provider needs to specify in order to be able to offer this right. The applicant agrees on the terms of the resources or tools either by accepting the terms stated in the ToS or resource or tool specific terms in the categories RES and the ACA + special requirements.

All hosted tools and resources in the CLARIN Infrastructure should include the licensing information based on the above categories and their graphical representation, the Laundry Symbols. The terms of the End-User license are to offer CLARIN Materials and tools, i.e. Content in the following categories:
- Public Content (PUB)
- Academic Content (ACA)
- Restricted Content (RES)
User understands and acknowledges that depending on the content the access may be limited based on this categorization and the User Identity. Additional licensing or usage terms for Academic and Restricted Content may be required, and the offered content may thus also belong to a certain subcategory when there are requirements that the AA System can handle automatically:

- Non-Commercial Use (NC)
- Usage Information (INF)
- Redeposition (RED)

In addition to the aforementioned categories, some content may have its own more permissive license terms specifying how the content can be, e.g., distributed to third parties with a GNU General Public License or Creative Commons, which may give the user rights in addition to the minimal set of terms defined by the categories and subcategories. [u14], Model agreement template: Appendix C

**ToS and Privacy Policy**

CLARIN End-User agreement templates are divided in two parts: ToS (Terms of Service Agreement) and Privacy Policy. Terms and conditions of these user agreements apply to CLARIN End-Users. In an ideal system, there would be only one common ToS and Privacy Policy. However, since this may not be the case, each CLARIN centres user agreements should follow when possible the model agreements presented as CLARIN ToS and Privacy Policy. In the ToS the End-User's attention is drawn to the provisions limiting access to specific classes of resources. By using the CLARIN service End-User agrees to follow the aforementioned categorization of CLARIN content as described above in this document. In addition to the categories most of the content has its own license, which may set additional terms how content can be e.g. used or distributed to third parties, which may go further (e.g. GPL v.3 and DRM) than the category definitions. In such a case the End-User agrees to follow terms of these licenses in addition to categorization.

In addition to the licensing terms, the ToS contains e.g.:

- A term describing how and when the agreement comes into force
- Definitions for distribution types PUB, ACA and RES with subtypes, see above
  - User Identity
  - Clarin Material
- Explanation of the content categories and a requirement for End-Users to adhere to them
- A research ethics clause
- A liability / availability disclaimer
- A choice for the court of Law
- A reference to the Privacy Notice

By accessing the ToS, the End-Users agree to comply with the Privacy Policy. The Privacy Policy describes how the personal data of the End-User will be processed. The Privacy Policy describes e.g.:

- What kind of personal information is gathered about the End-Users?
- How is the personal information used?
- Who has access to the personal information?
- How to make a request for access to personal data (application form template).
What security measures are there to safeguard the personal information?

How are cookies used?

Model agreement template: Appendices C and D

Data User Agreement

CLARIN Application for Use of Restricted Data from the CLARIN Database

If the resources contain information that could be used for disclosing the identity of human subjects (Personal Data), the End-Users of such resources can be required to fill in an Application for use of restricted data from the CLARIN database and to accept an additional Data User Agreement (DUA) to ensure that they will use resources only for research, report the Personal Data only in aggregates expunging names and other personal and individual identifiers, not try to identify the persons based on the data, delete the data when it is no longer needed, etc. The Application has to be approved by the Content Provider prior to the disclosure of the restricted data to ensure that data intended for research purposes is used only for such purposes and that researchers are aware of the limitations on these data. Guidelines for the Application and the Data User Agreement have been drafted.

Model agreement template: Data User Agreement, Appendix F
Model application template: Application for Use of Restricted Data from the CLARIN Database, Appendix E
PART VI FUTURE COLLABORATION WITHIN CLARIN

The implementation phase in making the CLARIN Materials, i.e. the LRT resources and tools available for the CLARIN community will require efforts in classifying the resources and tools hosted by the CLARIN centres using the laundry symbols presented in this document. There will no doubt be problems to which this report does not offer answers, as it is only by applying the guidelines in practice that the problems become visible and can be tackled together in the future as the CLARIN community takes shape. In every sense this is only the beginning, and it is up to the CLARIN Centres to update the agreements between the Content Owners or Content Providers and the Service Providers in order to make more materials available for everybody or for the research community. The goal of CLARIN is to upgrade as many RES materials as possible to the category ACA to make this progress possible and we in WP7 hope that this report will assist in the practical tasks required to reach this common goal.

Some of the questions we received as feedback for the summary on licensing problems concerned the CLARIN Infrastructure and its future prospects. These questions and answers to them are displayed below.

FAQ

Q: CLARIN needs to be extremely trustworthy with respect to distribution rights within CLARIN. What will be the legal form for CLARIN?
A: In the future, the CLARIN ERIC will be a competent legal entity. However, the CLARIN centres will actually have a more crucial position with respect to data security issues, such as the quality of their authentication and authorization procedures. Quality management audits will be integral to the operation of the centres.

Q: Who takes care of the authorization for web service access?
A: Each CLARIN centre takes care of the authorization of the access to the resources at its facilities. There may be authentication and authorization software for automating the procedure.

Q: Who or what entity signs the deposition license agreements for CLARIN?
A: The CLARIN ERIC is one option as the ERIC is intended as a forum for cooperation and coordination between various national activities. However, each CLARIN centre can also sign their resource distribution agreements separately, and even legally it is not a problem to leave the name of the licensee open for PUB category resources as the identity of the licensee is important only for resources in the ACA and RES category, where there is no automatic granting of rights. If the CLARIN Service Provider Federation is one of the parties in the agreement, the resource can be located at any CLARIN centre. However, the CLARIN Service Provider Federation is not a competent legal entity, which means that all SP Federation members may need to sign the agreements separately unless the CLARIN ERIC or some specific service provider is granted the right to sign on behalf of the others.

Q: Could we move the code to the data if the data is not allowed to move, i.e. could we migrate from web services to cloud services?
A: This is an interesting challenge and open source tools and licensing policies are being developed to cope with this scenario, e.g. the AGPL license.
Q: Should the abstract legal concepts such as academic and non-commercial be defined? No exact definitions are good for CO-CP agreements but the CP and the CUs may wish for legal clarity.

A: It makes the system rather inflexible and there is no precise definition for NC anyway. It is always a fuzzy term, see *Community created content* by Hietanen, Oksanen, and Välimäki in Law, Business and Policy, 2007: http://www.turre.com/wp-content/uploads/webkirja_koko_optimoitu2.pdf.

Q: Who will provide general guidelines on the national level?
A: It will be the task of the national CLARIN committees.

Q: Will there be a registry of license terms like IsoCAT with a check-list for license agreements and declarations of consent?
A: We are considering the option to include the legal “laundry symbols” in the data category registry defining widely accepted linguistic concepts, http://www.isocat.org/
D7S-2.1 Appendix A

CLARIN Deposition & License Agreement

1. Parties

1.1 The organisation or person authorised to transfer and deposit the digital dataset(s), hereafter referred to as the Depositor:

1.2 Clarin infrastructure (the Repository)

1. Licence

a. The Depositor grants the Repository a non-exclusive licence for digital data files, hereafter referred to as Content.

b. The Repository is authorised to include the Content in its data archive. The Repository shall transfer Content to an available carrier, through any method and in any form.

c. The Repository is authorised to make Content (or substantial parts thereof) available to third parties by means of on-line transmission. In addition, the Repository has the right, on the instruction of third parties or otherwise, to make a copy of the dataset or to grant third parties permission to download a copy.

2. The Depositor

a. The Depositor declares that he is a holder of rights to Content, or the only holder of rights to the Content, under the relevant legislation or otherwise, and/or is entitled to act in the present matter with the permission of other parties that hold rights.

b. The Depositor indemnifies the Repository against all claims made by other parties against the Repository with regard to Content.

3. The Repository

a. The Repository shall ensure, to the best of its ability and resources, that the deposited Content is archived in a sustainable manner and remains legible and accessible.

b. The Repository shall, as far as possible, preserve Content unchanged in its original digital format, taking account of current technology and the costs of implementation. The Repository has the right to modify the format and/or functionality of Content if this is necessary in order to facilitate the digital sustainability, distribution or re-use of Content.

c. If the access categories "Restricted Access" or "Academic Access", as specified at the end of this Agreement, are selected, the Repository shall, to the best of its ability and resources, ensure that effective technical and other measures are in place to prevent unauthorised third parties from gaining access to and/or consulting the Content or substantial parts thereof.

4. The Content

a. Content to which the licence relates is specified in the appendix to this Agreement. The appendix forms an integral part of this Agreement.

b. The Depositor declares that the Content corresponds to the specification provided.

c. The Depositor declares that Content contains no data or other elements that are contrary to the law or public regulations.
d. The Depositor indemnifies the Repository against all claims by third parties relating to Content.
e. The Depositor will supply Content by means of a method and medium deemed acceptable by the Repository.

5. Removal of Content / changes to access conditions
a. If sufficient indispensable grounds exist, the Depositor has the right to request the Repository not to make Content available for a temporary period or permanently. In such cases, the Repository shall retain Content in the data archive, but shall no longer allow third parties to access the Content or substantial parts thereof.
b. If sufficient indispensable grounds exist, the Repository has the right to remove Content from the archive wholly or in part, or to restrict or prevent access to Content on a temporary or permanent basis. The Repository shall inform the Depositor in such cases.

6. Availability to third parties:
a. The Repository shall make the Content available to third parties in accordance with the access conditions agreed with the Depositor: "Public", "Academic Access" or “Restricted Access”.
b. The Repository shall make Content available only to third parties who have agreed to comply with the conditions of use. Unless agreed otherwise with the Depositor, the use of Content is subject to the General Terms of Use laid down by the Repository.
c. If the access category "Restricted Access" has been agreed, the Repository shall make the Content available only to the persons and/or organisations specified by the Depositor.
d. The Repository can make Content (or substantial parts thereof) available to third parties:
   - if the Repository is required to do so by legislation or regulations, a court decision, or by a regulatory or other institution
   - if this is necessary for the preservation of Content and/or the data archive
   - (to a similar institution) if the Repository ceases to exist and/or its activities in the field of data-archiving are terminated

e. The Repository shall publish the metadata and make them freely available, on the basis of the documentation that the Depositor provides with Content. The term metadata refers to the information that describes the digital files. Other documentation that relates to the dataset and is provided by the Depositor shall be published and made freely available, unless the Depositor has specified that certain documents must not be made freely available. Documents that contain personal data will not be made freely available.
f. The general information about the research and the metadata relating to Content shall be included in the Repository’s databases and publications that are freely accessible to all persons.

7. Provisions relating to use by third parties
a. The Repository shall require third parties to whom the Content (or substantial parts thereof) is made available to include in the research results a clear reference to the Content from which data have been used. The reference must comply with the Clarin Infrastructure Terms of Use.

8. Death of the Depositor
Following the death of the Depositor, or in the event that the Depositor's organization ceases to exist, Content in the ‘Restricted Access’ category shall automatically be transferred to the ‘Public’ category. This is not applicable if Content contains personal data or such material, which copyright moves to the heirs of the original copyright owner.

9. Liability
a. The Repository accepts no liability in the event that all or part of Content is lost.
b. The Repository accepts no liability for any damage or losses resulting from acts or omissions by third parties to whom the Repository has made Content available.

10. Term and termination of the Agreement
a. This Agreement shall come into effect on the date on which the Repository receives the Content (hereafter the deposit date) and shall remain valid for an indefinite period. Cancellation of this Agreement is subject to a period of notice of six months, and notice shall be given in writing. It is possible to change the agreed access category at any time during the term of the Agreement.
b. Notwithstanding point (a), this Agreement shall end when Content is removed from the data archive in accordance with Article 5 of this Agreement.
c. If the Repository ceases to exist or terminates its data-archiving activities, the Repository shall attempt to transfer the data files to a similar organisation that will continue the Agreement with the Depositor under similar conditions if possible.

11. Applicable law
[WHAT] law is applicable to this agreement.

Access categories for Content
The Repository is permitted to distribute Content and make it available by means of the method mentioned below and, if indicated below making use of the additional option non-commercial

---

You have chosen:

[Public: unrestricted access]

The Repository is permitted to make content available to all persons, legal entities and organisations for any purposes.

[Academic Access only]

The Repository is permitted to make Content available to all persons and organisations registered with the Repository for the purpose of Academic research, study or teaching.

[Restricted Access: access with the permission of the Repository]
The Repository is permitted to make the Content available to persons, legal entities and organisations registered with the Depositor only after receiving express permission from the Depositor.

You have additionally chosen:

[Restriction: Non-Commercial]

Content is not be used in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation.

[Restriction: Grant back]

If a Derivate version of Content is created, it has to be licensed with the same license as the original Content.

[Restriction: Notification]

If the Content is used in a published article or other work, a notification to Depositor about the publication is required.

The Depositor hereby agrees to the above provisions and the general code(s) of conduct referred to in this document.
D7S-2.1 Appendix B

CLARIN Upgrade Agreement

1. Parties
Licensor:
Licensee:

2. Definitions
2.1. Licensed Content
[A brief description of material in question]
2.2. CLARIN Infrastructure
[Definition]

3. Purpose of the Agreement
The purpose of this agreement is to clarify the content of the existing licensing agreement between the Parties to allow the use the Licensed Content also as a part of the CLARIN Infrastructure. Thus agreement does not supersede the existing agreement(s) but only adds the needed Rights to CLARIN Infrastructure.

4. License Grant and Rights
Licensor agrees that the CLARIN Infrastructure has the same rights and duties pertaining Content, including but not limited to the right to distribute and use Licensed Content, as Licensee has based on the existing licensing agreement.

For this purpose, Licensor grants CLARIN Infrastructure worldwide, royalty-free, non-exclusive, terminable license to use Licensed Content for the duration of any applicable copyright and Database Rights Academic and Teaching purposes. To the extent possible in the relevant jurisdiction, these rights may be exercised in all media and formats whether now known or created in the future. [However, the rights granted shall not be used in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation.]

The granted rights are: Extraction and re-utilisation of the whole or a substantial part of Licensed Content; Creation of derivative content; Creation of collective content; Creation of temporary or permanent reproductions by any means and in any form, in whole or in part of Licensed Content; and distribution, communication, display, making available, or performance to the public by any means and in any form, in whole or in part of the Licensed Content.

CLARIN Infrastructure may not transfer these rights to any 3rd party without permission from the Licensor.

5. Additional Information
The Licensor requires that CLARIN Infrastructure [DOES WHAT] before the access is given automatically to the User to the Material.

6. Termination
This Agreement and the rights granted will terminate automatically upon any substantial breach by Licensor or the CLARIN Infrastructure of the terms of this Agreement or the original licensing agreement. Individuals or entities who have received Adaptations or Collections from CLARIN Infrastructure under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses.

Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Licensed Content). Notwithstanding the above, Licensor reserves the right to release the Licensed Content under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.
D7S-2.1 Appendix C

**CLARIN Terms of Service**

**Preamble**

Welcome to the CLARIN Centre website. The mission of CLARIN is to provide as wide access as possible to language research tools and materials around Europe [hosted in different CLARIN centres]. To archive this goal, we set out some ground rules in this document [so that the CLARIN centres can legally distribute the tools and materials]. By accessing or using CLARIN Materials, and in consideration of the Services provided to you, you agree to abide by the Terms Below.

**Terms of Use**

1. Governing Terms
1.1 The access to and use of CLARIN services are governed by this Agreement and by its terms and conditions, which are hereby incorporated into this Agreement. The User's attention is drawn to the provisions limiting access to certain specific classes of resources.
1.2 Any use of the website by means of the User Identity granted by the local identity provider also signifies User's acceptance of the terms and User's agreement to be bound thereby.

2. Definitions
2.1 All terms used in this Agreement, unless specifically defined herein, will have the meanings attributed to them in the Agreement. The following terms shall have the following meaning:
   “CLARIN” means CLARIN ERIC and national CLARIN Service Providers or centres.
   "Authorised Use" means use by a User who is accessing the website via the identity provided by national identity organizations.
   “Academic Use” means use by a User who fulfils the criteria set by the national identity organizations for Academic users.
   “Non-Commercial Use” refers to any use that does not generate income or is not used for promoting the generation of income.
   “User Identity” refers to the identity granted by a local identity provider [or CLARIN].
   “CLARIN Services” refers to all tools and materials hosted in different CLARIN centres.

3. Access to CLARIN
3.1 The CLARIN centre hereby grants the User a limited non-exclusive non-transferable licence to access and use the CLARIN resources it hosts under the terms of this Agreement and via the chosen User Identity.
3.2 Access to certain Features and databases may be restricted by CLARIN.
3.3 User will be responsible for all access to CLARIN by means of the chosen User Identity whether or not the User has knowledge of such access and use. The CLARIN centre reserves the right to cancel any User Identity without notice. If a User suspects that somebody may misuse an identity, CLARIN should be notified immediately.
3.4 User understands and acknowledges that the User Identity is for his or her use only and that the User is prohibited from permitting any third party from accessing CLARIN Services by means of the User Identity.
3.5 Content Categories
CLARIN offers content in three different categories:
   - Public content (PUB)
   - Academic Content (ACA)
• Restricted Content (RES)

User understands and acknowledges that depending on the content the access may be limited based on this categorization and the User Identity. Also, CLARIN may require additional licensing or usage terms for Academic and Restricted Content and there may be steps like submitting more info for this purpose. User agrees to follow the categorization.

3.6 Sub-categories
The offered content may also belong to a certain sub-category:
- A requirement for strictly Non-Commercial use (NC)
- A requirement to inform the copyright holder regarding the usage of the tools and/or the resources in published articles (INF)
- A requirement to redeposit modified versions of the tools and resources with the Service Provider (ReD)

User agrees to follow the requirements for sub-categories.

3.7 Content Specific Licenses
In addition to the aforementioned categories, most of the content has its own license, which may set additional terms how content can be e.g. used or distributed to third parties, which may go further (e.g. GPL v.3 and DRM) than the category-definitions. In such a case User agrees to follow the terms.

4. Research Ethics
The User agrees to follow the best practices regarding research ethics. This includes treating colleagues, stakeholders, customers, suppliers and the public respectfully and professionally, taking into account confidentiality issues when appropriate as well as protecting the privacy of the informants, respecting cultural differences and having an open and explicit relationship with government, the public, the private sector and other funders. In addition, The User agrees to follow best academic practices while adding citations regarding the Content.

5. Access to Service and Liability
The CLARIN centre disclaims all responsibility and liability for the availability, timeliness, security or reliability of the services, software or other content provided through the site. Similarly, the original owner or source of the content disclaims all responsibility and liability for the availability, timeliness, security or reliability of the Content. The CLARIN centre reserves the right to modify, suspend, or discontinue the services or access to the services without any notice at any time and without any liability to you.

6. Governing Law and Entire Agreement
These Terms are governed by the laws of the [ERIC’s / Centre’s home venue] without regard to the rules of conflict of law that may cause the laws of another jurisdiction to apply. User agrees to the sole and exclusive jurisdiction and venue of the [CLARIN ERIC/ Centre’s home venue] in the event of any dispute of any kind arising from or relating to the Site, or User’s use or review of it. However, the CLARIN centre has the right to use the laws of another jurisdiction to get injunctive measures or relieves against the misuse of the service.
The Terms constitute the entire agreement between the parties with respect to the subject matter hereof and supersedes and replaces all prior or contemporaneous understandings or agreements, written or oral, regarding such subject matter. If for any reason a court of competent jurisdiction finds any provision or portion of these Terms to be unenforceable, the remainder of the Terms will continue in full force and effect.

7. Data Protection and Privacy
User agrees with and agrees to follow the Privacy Notice set out in the separate agreement [link]. [check box]

8. Usage Statistics and Automated Querying
CLARIN strives to maintain the integrity of usage statistics as a measure of readership and other use of the CLARIN sites by authors and researchers. It is a violation of CLARIN policy for a party to directly or indirectly use CLARIN Services with a view to affecting downloads and other usage statistics, or to encourage others to do so. As part of its general right to refuse or terminate service and remove or edit site content, the CLARIN centre reserves the right in its sole discretion to limit access, remove content, and adjust usage statistics to respond to any activity that appears likely to have such an effect.

9. Termination
If User violates the letter or spirit of this agreement, or otherwise creates risk or possible legal exposure for CLARIN services, the CLARIN centre can stop providing all or part of the service. The CLARIN centre will notify User the next time he or she attempts to access the service.
A model agreement for the Privacy Policy has been drafted below. Depending on the content of the processing services and the national legislation, the model agreement may need some modifications.

CLARIN Privacy Policy

This Privacy Policy applies to (ADD ADDRESS) site and services provided therein (both together “CLARIN service”), operated by (ADD THE NAME; ADDRESS) as a CLARIN service provider (later on “CLARIN Service Provider”).

This Privacy Policy describes the personal information that we collect from you concerning the CLARIN service and how we may use that information.

By using the CLARIN service, you agree to accept the terms of this CLARIN Privacy Policy. This Privacy Policy does not necessarily apply to the practices of the other CLARIN Service Providers. Such CLARIN Service Providers may have their own privacy policies, which may be viewed on their CLARIN sites (links). The other CLARIN Service Providers may require that you accept their privacy policy when using their services.

However, by joining CLARIN, each CLARIN Service Provider has committed to the obligations imposed by the currently valid legislation on personal data protection, concerning without limitation processing and disclosure of personal data, obligations of non-disclosure and secrecy, as well as providing a sufficient data security level for the services.

Collecting Your Personal Information

This Privacy Policy applies to personal information collected about you by the CLARIN Service Provider.

Personal information means any information that may be used to identify an individual, including, but not limited to, a name, home address and email addresses.

Registration Information

Registration may be required in order to use the CLARIN service. In such case, users who request services may be required to provide certain personal information, such as name, email address, user name and password, when registering.

Information Collected During the Use of the Clarin Service

When you visit the CLARIN services, the following kind of information may be collected automatically:

- the name of the domain from which you entered the internet,
- the IP address of the user,
- the date and time you accessed our site,
- the pages you visited within the CLARIN infrastructure.
This information is used for purposes such as assessing what information is of most interest to users and system enhancements.

In addition we store information that you provide to us during your use of the CLARIN service. Such information may include e.g. personal correspondence, such as e-mails or letters.

**Cookies or Other Tracking Technologies**

The CLARIN service uses cookies to automate the process of logging on to certain areas of the site. Your Web browser can be set to allow you to control whether you will accept cookies, reject cookies, or to notify you each time a cookie is sent to you.

**Use of the Personal Information**

We use personal information to provide you with the CLARIN services you have requested and for administration purposes.

Personal information is used by the CLARIN Service Provider in at least all of the following ways:

- to identify you in order to provide the CLARIN material you have requested,
- to respond to your questions and otherwise communicate with you,
- for the auditing and research analysis in order to maintain, protect and improve the CLARIN service and to protect rights of property of the CLARIN material and service, and
- to ensure the technical operation of the CLARIN service and network.

**Disclosure of Personal Information**

Expect as authorized by law or this Privacy Policy document, personal information will not be disclosed to other organisations or persons.

We do not sell or rent your personal information to third parties for marketing purposes. We may need to share your information with our service providers for the purposes of data processing and storage on our behalf. We require that these parties agree to process your information based on our instructions and in compliance with this Privacy Policy and valid legislation on personal data protection.

In addition, we may need to share your information with the other CLARIN Service Providers in order to provide you with access to the CLARIN material of the other CLARIN Service Providers. We only share the following limited information with other CLARIN Service Providers:

- the necessary identification information, e.g. your name, user id and nationality, and
- the limits of your access rights to the CLARIN service.

*(ADD BULLET POINTS IF THERE IS SOME OTHER INFORMATION THAT NEEDS TO BE SHARED!)*
Data Retention

We will not retain your personal information longer than is necessary to fulfil the purposes for which it was collected or as required by applicable laws or regulations.

Accessing and Correcting the Personal Information

You have the right to ask for a copy of the information we hold about you and to have any inaccuracies in your information corrected.

Updating Your Information

If your information changes please email us at: (ADD ADDRESS)

Security

We will take reasonable and appropriate steps to protect the personal information that you share with us from unauthorized access or disclosure. These measures include e.g. the following:

- only authorized employees have access to the data collected for processing and maintenance of the database that stores list information,
- internal reviews of our data collection, storage and processing practices.

The personal information is stored by (ADD)

Links to Other Web Sites

The CLARIN service may contain links to other web sites. Please note that when you click on one of these links, you may be entering another web site for which the CLARIN Service Provider has no responsibility or control.

The CLARIN Service Provider does not share personal information with those sites unless otherwise agreed.

Changes to This Privacy Policy

We reserve the right to revise our Privacy Policy, and such a revised policy will be posted on our Web site in order to keep you informed of how we collect, use, manage, disclose and protect the Personal Data. Such changes will be posted on our Web site at this location and the effective date of the policy will be updated accordingly.

You are responsible for regularly reviewing this Privacy Policy. Continued use of the site shall constitute your consent to such changes.
Contact Information

Additional questions about this Privacy Policy should be directed to:
(ADD)

Governing Law

This Privacy Policy and the privacy practices of the CLARIN Service Provider will be subject exclusively to the laws of (ADD). The CLARIN Service Provider makes no representation that this Privacy Policy and such practices comply with the laws of any other country. Visitors who use the CLARIN service outside Europe do so on their own initiative and are responsible for compliance with local laws, if and to the extent local laws are applicable. If you reside outside of Europe, by using the CLARIN service, you consent to the transfer and use of your information outside your country.

Privacy Policy Effective Date:
D7S-2.1 Appendix E

Application for Use of Restricted Data from the CLARIN Database

This application has to be approved by the Content Provider prior to the disclosure of data from the CLARIN database to ensure that the disclosure will comply with the requirements of the legislation.

Personal information of principal researcher ("User"): (Name, position)

Organization:

Complete mailing address:

Telephone number:

Name(s), position, organization of additional researchers who will be given access to the CLARIN data: (Please note that these persons have to be directly supervised by the User, or to be an employee of the institution or entity that employs the User.)

Name of the Research project:

Summary of the Research: (Provide a brief description of the study, timelines for the study, its objectives and how the CLARIN data will be analyzed.)

By filling and accepting this application for use of restricted data from the CLARIN database, the User affirms that CLARIN data will be used for research and analysis purposes only, and that these data will be used solely for the purposes specified above.

The User affirms that the User has read and accepted the associated agreement for use of restricted data (Data User Agreement) and shall comply fully with the terms of that agreement.

All researchers as named above working with the CLARIN data must comply with the provisions of this Application and the Data User Agreement concerning the use of CLARIN data and the protection of personal privacy. The User (principal researcher) bears the ultimate responsibility for ensuring this compliance.

Please note that a separate application form must be submitted for each study or project, even if you plan to use data you previously obtained from the CLARIN database.

Please return your completed application to:
D7S-2.1 Appendix F

Data User Agreement

Content Provider, operating through the CLARIN network has collected certain data (“CLARIN data”). Content Provider shall make these CLARIN data available, for the research user (“User”) subject to the terms of this Data User Agreement (“DUA”).

By accepting this DUA, the User agrees to the conditions set out in this DUA and affirms that CLARIN data will be used for research and analysis purposes only, and that these data will be used solely for the purposes specified in the Application for use of restricted data from the CLARIN database (“Application”).

By accepting the Application, the Content Provider authorizes the User and possible other researchers named in the Application, subject to the terms and conditions of this DUA a non-assignable limited term to use the CLARIN data for the purposes set forth in this DUA and the Application.

Licensing and other terms between User and Content Provider and/or CLARIN will be agreed separately.

1. Research Use. The User agrees to use the CLARIN data only for the research purposes specified in this DUA and the attached Application. Any use not expressly described in the DUA or the Application is prohibited. Prohibited use includes, but is not limited to, any commercial or business use.

Use of the CLARIN data for a research project other than the one described in the Application will not be undertaken until after a separate Application for that project has been submitted to, and approved by, the Content Provider.

2. Confidentiality. The User agrees to maintain the confidentiality of the CLARIN data, including not releasing, and prohibiting others from releasing, any information that identifies persons, directly or indirectly.

User will ensure that the CLARIN data with identifiers are kept in a physically/electronically secured environment and access will be given only to authorized researchers as named in the Application. User shall not try to identify persons based on the CLARIN data obtained via this agreement, including through such means as linkage with other databases.

3. Non-Assignability and other researchers. Neither this agreement nor the use of the CLARIN data may be assigned by the Users to other persons without the prior written consent of the Content Provider. User agrees not to distribute, sell, or permit others to use the CLARIN data. The User shall not allow access to these CLARIN data in full or in part to any person without the separate written permission, unless these researchers are named in the Application and accepted by the Content Provider.

The User agrees to inform these researchers who use the CLARIN data and are named in the Application about the policies and procedures related to this DUA. All researchers named in the Application working with the CLARIN data must comply with the provisions of this DUA. The User bears the ultimate responsibility for ensuring this compliance.
4. **Reporting of Research.** Reporting of research will be in aggregate, with all names and other personal individual identifiers expunged. User will not present or publish these CLARIN data in a manner in which any individual can be identified.

5. **No Warranties.** The CLARIN data is released “as is” without warranty of any kind. The entire risk of the quality and performance of the CLARIN data is assumed by the User. The User agrees that he or she will not hold, or attempt to hold, Content Provider or CLARIN responsible for any damages, including any incidental or consequential damages arising from the use of, or inability to access the requested CLARIN data files to the extent allowed by national law.

6. **Termination.** User shall delete the CLARIN data that is no longer needed. However The CLARIN data shall always be destroyed one year after the research project in question has been completed.

   The term of this agreement and hence, the authorization to use the CLARIN data, is the shorter of two years from the date of accepting this agreement, or upon written notice of termination by the parties. Content Provider and/or CLARIN may terminate this agreement at any time for any reason by providing written notice by email to the User or to any one of researchers named in the Application.

   In addition, this agreement and the rights granted will terminate automatically upon any substantial breach by User or other researchers named in the Application of the terms of this agreement. At the time of termination of this agreement, the User agrees to destroy immediately the CLARIN data files. The provisions of the Application and the DUA which protect personal information survive the termination of this DUA as well as the termination or completion of the research project and the research thereunder.
D7S-2.1 Appendix G

Data Processor Agreement between Content and Service Provider

1. Parties
   Data Controller:
   Data Processor:

2. Definitions
   2.1 CLARIN services means the services in which the service providers including Data Processor as part of CLARIN provide online access to CLARIN material, e.g. linguistic resources or technologies, to the end users.
   2.2 Data means the data described in Annex 1.
      [Or a brief description here of the data including Personal Data in question.]
   2.3 End-User means a person, who can have access to the Data.
   2.4 Parties mean the Data Controller and the Data Processor.
   2.5 Processing means any operation or set of operations which are performed on Data, whether or not by automatic means, and shall have the meaning as described to it as set out in the data protection legislation, including the EU Directive 95/46/EC ("On the protection of individuals with regard to the processing of personal data, and on the free movement of such data").

3. Purpose of the Agreement. The purpose of this agreement is to state the terms and conditions under which Data Processor will provide data facilities and services as defined below to the Data Controller.

   The main purpose of this agreement is to ensure that the Processing of Data including Personal Data is carried out in accordance with the European Union Directive 95/46/EC, as set out in the legislation of the Member State in which Controller is established, subject to that legislation being in accordance with the Directive.

4. Application of the Agreement. This Agreement shall apply to all Data accessed by the Data Controller to the Data Processor for processing it in relation to the CLARIN services.

5. Purpose of the Processing. The Data Processor shall process the Data it receives from the Data Controller solely for the purpose necessary for disclosing the Data for the End-User of the CLARIN services as agreed between the Parties and strictly in accordance with the terms of this agreement and written guidance from the Content Provider.

   The Data Processor shall promptly carry out any request from the Data Controller requiring the Data Processor to amend, transfer or delete all or any part of the Data.

6. Security and Confidentiality. The Data Processor shall use appropriate technical and organisational measures to safeguard the Data from unauthorised or unlawful processing or accidental loss, destruction or damage.

   The Data Processor shall ensure that each of its employees are made aware of its obligations with regard to the security and protection of the Data and shall require that they enter into binding obligations with the Data Processor in order to maintain the levels of security and protection provided for in this agreement.
Security aspects are described in more details in Annex 2.

7. Discloser of the Personal Data. The Data Processor will only disclose Data in accordance with instructions from the Data Controller as described in Annex 3. Data Processor will take appropriate security measures to ensure that no unauthorised disclosure occurs. Data Processor shall not divulge the Data whether directly or indirectly to any person, firm or company without the express consent of the Data Controller except to those of its employees, agents and subcontractors who are engaged in the processing of the Data and are subject to the binding obligations referred to in Annex 2. The Data Processor shall ensure by written contract that any subcontractor employed by the Data Processor to process Data to which this agreement relates also provides the Data Processor with details of the technical and organisational means it has adopted to prevent unauthorised or unlawful processing or accidental loss or destruction of the Data and confirms to the Data Processor the implementation of those means.

8. Obligations of the Data Controller. The Data Controller shall ensure that any disclosure of Data made by it to the Data Processor is made with the data subject’s consent or is otherwise lawful. The Data Controller agrees that it will comply with the relevant provisions of the data protection, privacy and other legislation.

9. Liability. These provisions set out the entire liability of the Data Processor to the Data Controller with respect to data protection and confidentiality obligations.

10. Ownership. All Data stored and processed under the terms of this agreement by Data Processor on behalf of Data Controller are and shall remain exclusively the property of Data Controller, unless otherwise agreed in writing.

11. Termination. This agreement shall terminate automatically upon termination or expiry of the Data Processor's obligations in relation to the CLARIN Services, and on termination of this agreement the Data Processor shall forthwith deliver to the Data Controller or if allowed by the Data Controller destroy all the Data Controller's Data in its possession or under its control. Either party may terminate this agreement immediately by giving written notice to the other in the event that the other Party:
   1) commits a material breach of its obligations under this agreement and, in the case of a breach which is capable of remedy, fails to remedy it after being given [14] days written notice specifying the breach and requiring it to be remedied; or
   2) becomes the subject of an insolvency event. Either party may terminate this agreement upon giving one month's prior written notice to the other. Upon receipt of written notice from the Data Controller or upon giving written notice of termination to the Data Controller, the Data Processor shall either return or if allowed by the Data Controller destroy any Data received from the Data Controller to the Data Controller forthwith.

12. Assignment. This agreement shall not be transferred or assigned by either party except with the prior written consent of the other.

13. Applicable Law. This agreement shall be construed according to and governed by the laws of [add].
D7S-2.1 Appendix H

Guidelines for Consent Forms for Collecting Personal Data

Data Subjects Consent Form
The European Union Directive (95/46/EC) "on the protection of individuals with regard to the processing of personal data and on the free movement of such data" describes consent of a data subject as “any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him to be processed”.

These guidelines indicate the kind of information that should be included in the consent form of the data subject. The consent form is divided in two separate parts: information sheet and consent sheet. You may use this as a template for developing necessary forms for your research needs. A general template is not provided because consent forms are study specific.

How the data is collected influences the chosen method of how to inform the data subject and obtain consent. However, use of written forms whenever possible is recommended. The data subject should be given a copy of the consent form and the researcher should retain the signed original.

Information Sheet
Data subjects must be adequately informed about the research to ensure that they have understood what is involved in the research and how the personal data will be used and processed. To promote ease of understanding, the information should be presented in plain language.

At least the following information should be considered for the Information Sheet:
- title of the research or study
- contact information (including the name of the organization delivering the study and the name of the researcher collecting the personal data and information on how to contact the researcher)
- purpose of the research
- why the personal data is collected
- present and possible future use of the personal data
- sharing of the personal data (who will have access to the personal data and under which circumstances)
- data security aspects (including information how the confidentiality of the data is maintained, who will have access to the personal data and how the data will be stored)
- how long the personal data will be kept
- any secondary use of the data (including option to use the data in future research projects through CLARIN)

Consent Sheet
For the consent to be valid, it must be fully informed and freely given.

Please notice that a study involving young people (typically under 18) also needs the consent of a parent or a person with parental responsibility.

The Consent Sheet could include at least the following information:
- identification of the person to whom the consent is given
- declarations that the data subject:
  - has read and understood the Information Sheet or has been told the main aspects of the study
  - agrees to the planned forms of data collection
  - agrees to the use of personal data in research, publications, sharing etc. as described
o agrees to the secondary use of personal data through CLARIN (including the storage of the data) and the terms of the secondary use (possible involving a new separate consent, signing of a DUA, etc.)
o agrees to the planned forms of data processing, including distribution of personal data
o understands the right to withdrawal of the consent
o acknowledges the right to the erasure, correction, updating or completion of personal data
- special restrictions on the use of the personal data (e.g. that the data subject must not be named in any publication)
- signatures and date for both the data subject and the researcher
A check list for CLARIN Agreements

Deposition Licensing Agreements (DELA) ................................................................. 77
  Publicly Available ................................................................................................. 77
  Academic Use ........................................................................................................ 78
  Restricted Use ....................................................................................................... 78
  Additional Requirements ....................................................................................... 79
  Secondary Use of Personal Data ........................................................................... 79
Privacy Policy ............................................................................................................ 80
Terms of Service (ToS) and End-User License Agreement (EULA) .................... 81

Deposition Licensing Agreements (DELA)

As explained in other documents in more detail, there are three main categories for the shared tools and resources in CLARIN:

- Publicly Available
- Academic Use
- Restricted Use

The first two categories are more demanding for the licensing agreements. In RES basically any kind of licensing language is accepted and it is up to the user to fulfil the requirements. However, the goal of CLARIN is to have as much as possible of its content in the first two categories and therefore the use of Upgrade Agreements is strongly advised.

Publicly Available

The licenses in this category should follow as closely as possible the Protocol for Implementing Open Access Data.\(^5\) For the new tools and resources, the preferable license is either the Creative Commons Zero (CC0)\(^6\) or the Open Database License (ODbL). Otherwise the licenses in this category should allow at least Open Source definition 4-6 and 9-10, see http://www.opensource.org/osd.html:

- Integrity of The Content
- No Discrimination Against Persons or Groups
- No Discrimination Against Fields of Endeavour

\(^5\) http://sciencecommons.org/projects/publishing/open-access-data-protocol/
\(^6\) http://creativecommons.org/choose/zero
- License Must Not Restrict Other Content
- License Must Be Technology-Neutral

In addition the license should allow distribution of the tools and resources from the CLARIN infrastructure in all circumstances. However, there is no requirement to allow sub-licensing of the content for the user.

**Academic Use**

For **Academic Use** (ACA), the license agreement includes a requirement that the use is somehow related to an academic institution. Here the problem may and will arise from the definition of academic use. The main points to check are:

- The content should be available at least for anyone working or/and studying in an academic institution recognized by an Identity Federation (IdF) with which CLARIN has a service provider agreement.
- The content can be used for general academic purposes. However, it is relatively hard to define exactly what this means, e.g. “teaching” is not normally listed as permitted but that may or may not mean that it is prohibited. In case of doubt, contact the right holder (or use the CLARIN Academic Upgrade Agreement to obtain the missing right).

Due to legacy deposition license agreements, the Academic Use licenses may include non-standard provisions which the user is required to accept separately for a specific tool or resource.

**Restricted Use**

The category **Restricted Use** (RES) includes the resources that do not fulfil the previous requirements but still could be offered to the users if certain requirements are met. The first thing to check is that the license gives enough rights for CLARIN to distribute it in the first place. If the access is limited to a certain room or a certain library, there is not much point in adding it to CLARIN, and the use of an upgrade agreement is necessary.

However, the most typical reasons for a resource to fall under the scope of RES are less problematic leading to a requirement that the license be granted individually by the owner, i.e.

- a requirement to submit detailed information (e.g. an abstract) about the planned usage, or
- specific ethical or data protection-related requirements found in speech and video material, i.e. material including Personal Data.

In addition to the aforementioned terms, the following terms are commonly part of the tool or resource agreements and do **not** normally cause problems for CLARIN:

- No right to re-distribute the material in any manner, no right to sub-license the material.

---

7 Even if RES is the typical license category for Personal data, in some cases the category PUB may also be considered:

1) there is an explicit consent (from the data subject) for processing and disclosing Personal Data without any limitations, or
2) the legislation permits disclosing of Personal Data.
- Only insignificant parts of the materials may be extracted at a time; whole works or parts of works may not be extracted.
- Materials will have to be destroyed after the termination of the contract
- The end-user ensures that he will keep any password secure and not disclose it a third party
- The end-user promises to preserve the confidentiality of, and not attempt to identify, individuals, households or organizations in the data.
- Acknowledging that personal data of the end-user may be held for verification and statistical purposes according to the data protection regulations.
- Any breach of the agreement will lead to immediate termination of the access of the user to the services and could result in legal action.
- Applicable law
- Liability disclaimer

One question, which is seldom addressed in the resource and tool deposition agreements, is what happens to the content if the right holder loses its legal status. While this does not affect adding the content to the service, it would be preferable to address the question in an update agreement, e.g. by adding a term that puts the content in question in the public domain in such a situation.

**Additional Requirements**

In conjunction with a main licence category, there can also be all or any of three additional requirements:

- **A requirement for strictly non-commercial use (NC):** A term requiring non-commercial use of the content is commonly found in different licenses. It is problematic because there is no common definition of what non-commercial actually means in different jurisdictions. However, this does not affect CLARIN as such, it just makes the life of the end-users more difficult.

- **A requirement to inform the tool or resource owner of published articles, which use the tools and resources (INF):** This term is something that is almost part of good practice in writing academic articles. It seldom causes any problems as long as the user is aware of it.

- **Redeposition (RED) of modified versions of the tools and resources with CLARIN:** In certain cases the right holder has an interest to collect the modified versions of the content. From CLARINs point of view, this is not a problem as long as the user of the content is aware of the term.

**Secondary Use of Personal Data**

Information that could be used for disclosing the identity of human subjects is Personal Data. In general, Personal Data, which has been obtained for a particular purpose, must not be used for any other purpose without the consent of the data subject(s) to whom they apply.

Personal data may only be processed and disclosed when there is a clear legal purpose for doing so. Legal purpose can mean:

1) The data subject has given his consent for processing and disclosing.
2) The data subject will be informed about disclosing before or afterwards and this is accepted in the national laws or regulations. (Directive 95/46/EC)
3) The special legislation at the national level allows processing and disclosing personal data. (Directive 95/46/EC)

4) Typically there are certain exemptions, which allow Personal Data to be used for research purposes (even if the data were originally gathered for other purposes). For example the protection of privacy may not apply to materials that are in the public domain or to published data which can concern individuals and their activities in the fields of politics, business, official activities and culture.

Users of resources containing Personal Data can be required to fill in an application for use of restricted data from the CLARIN database and accept an additional Data User Agreement (DUA) to ensure that they will use the resources only for research, report the Personal Data only in aggregates expunging names and other personal and individual identifiers, not try to identify the persons based on the data and delete the data when it is no longer needed. The application has to be approved by the Content Provider prior to the disclosure of the restricted data to ensure that the data intended for research purposes is used only for those purposes and that researchers are aware of the limitations of these data.

For additional information, see the proposal for ethical review, ethical principles of research in the humanities and social and behavioural sciences by the National Advisory Board on Research Ethics, Helsinki 2009. However, note the following:

- It is particularly unacceptable to reveal information on research data or hand over data in such a way that it could influence the evaluation, treatment or position of individual subjects.
- Particular care should be taken when the processing involves sensitive Personal Data (racial or ethnic origin, political opinions, religious beliefs, trade union membership, health, sex life, criminal convictions). Many times disclosing sensitive Personal Data is prohibited or it can only be disclosed if you have the explicit consent of the Data Subject.
- Collected Personal Data must be adequate, relevant and not excessive in relation to the purposes for which they are processed. If research data can be reasonably analysed without direct identifiers and there are no research grounds for storing identifiers, only data from which identifiers have been removed may be produced for research purposes and stored for secondary research. Data with identifiers can be collected and used when it is appropriate from the viewpoint of that particular research. With the consent of the subject, data can also be stored for secondary research with identifiers. Research in the humanities and social and behavioural sciences may require the processing and storage of identifiers. This may be based on the need to analyse data, later contacts with subjects or the historical and cultural significance of data.
- The Eighth Principle of the EC Data Protection Directive imposes a prohibition on the transfer of Personal Data to countries outside Europe unless the country of destination provides an adequate level of protection.

**Privacy Policy**

In addition to these agreements, there is a separate privacy policy document, which describes how the CLARIN Service Providers collect and use Personal Data. The Privacy Policy describes:

- What kind of personal information is gathered about the end-users?
- How is the personal information used?
- Who has access to the personal information?
- How to make a request for access to personal data (application form template).
- What security measures are there to safeguard the personal information?
- How are cookies used?

The principles on collecting and using Personal Data apply, so the End-User of CLARIN needs to accept the Privacy Policy of the CLARIN Service Provider concerning the personal data collected about End-Users.

**Terms of Service (ToS) and End-User License Agreement (EULA)**

In an ideal system, there is only one common Terms of Service Agreement for CLARIN also containing the essential parts of a general EULA. However, since this may not be the case, a separate local ToS Agreement should follow as closely as possible the model agreement presented as the CLARIN ToS. In practice, this means that the ToS should contain at least:

- A term describing how and when the agreement comes into force
- Definitions for
  - Restricted Use (RES)
  - Academic Use (ACA)
  - Non-Commercial Use (NC)
  - Usage Information (INF)
  - Redeposition (RED)
  - User Identity
  - CLARIN Material
- Explanation of the content categories and a requirement for end-users to adhere to them
- A research ethics clause
- A liability / availability disclaimer
- A choice for the court of Law
- A reference to the Privacy Notice

It is preferable that the language used in the ToS agreement be as clear as possible since it should be understandable for “laymen”.