Common Language Resources and Technology Infrastructure

CLARIN Classification Guide for Deposition Licenses - First comprehensive summary about licensing problems

CLARIN-2010-4

EC FP7 project no. 212230

Milestone M7S-2.3
Version 2.1

Responsible: UHEL Krister Lindén, Ville Oksanen, Sari Bruun

Contributing Partners: BBAW, CSC, IDS, INL, ILSP, IMCS, MPI, UCPH, UHEL, UTU,
Contributing Members: IDS, Joensuu U, Jyväskylä U, KOTUS, Oulu U, Radboud U, Tampere U

www.clarin.eu

A guide for classification of Deposition License Agreements

The aim of this guide is to help with the classification of existing Deposition License Agreements. For a more complete picture of the set of legal contracts needed to set up the licensing and distribution of material through CLARIN, see Deliverable WP7S 2.1 “A guide for legal contracts for tools and resources in CLARIN”.

To distribute tools and resources to the research community through CLARIN, a Content Provider and a CLARIN Service Provider have to sign a Deposition Licensing Agreement. 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. Completely new tools and resources may also be directly deposited with a CLARIN Service Provider.

You can add or update tools or resources and categorize their distribution rights in the CLARIN LRT registry:

http://www.clarin.eu/view_resources
http://www.clarin.eu/view_tools

You need to login, select a resource and go to the edit tab of the resource where you can select the appropriate main distribution type. You will then be offered four additional distribution restrictions. Remember to save your choices before you leave the page.

At the end of this document is an initial list of frequently asked questions that will be updated as you provide feedback on problematic issues with regard to the classification. Please send your questions to Hanna.Westerlund@helsinki.fi.

This guide does not imply transfer of the ethical or legal responsibility for the implementation and distribution of a study from the researcher, service or content provider to the CLARIN infrastructure. All research as well as providing of service or content must comply with the legislation and other guidelines.

Content or Software Provider Agreements

Deposition Licensing Agreement (DELA)

There are three main categories for the shared tools and resources in CLARIN:

- Publicly Available
- Academic Use
- Restricted Use

PUB	ACA	RES

In conjunction with any main category, there can also be all or any of four additional requirements: non-commercial, usage report, redeposition or any other additional requirement, e.g. disclosure with separate permission only.

Main Categories

If the goal is to submit tools and resources to the **Publicly Available** (PUB) category, the following requirements have to be met:

- The license should allow distribution of the tools and resources from the CLARIN infrastructure.
- There are no limitations (based on status or geographical location etc.) on who can access and use the tools and resources.
- There are no limitations on the purpose the tools and resources are used for.

In other words, the license should follow as closely as possible the Protocol for Implementing Open Access Data.[1] For the new tools and resources, the preferable license is either the Creative Commons Zero (CC0)[2] or the Open Database License (ODbL). However, for the licensed tools and resources, relicensing is often not possible, and the submitting party should make a careful assessment of the terms of the existing licensing agreement. If necessary, use a CLARIN Upgrade Agreement to obtain the required rights.

For **Academic Use** (ACA) the license agreement includes an additional requirement that the use is somehow related to an academic institution. Here the problem may arise from the definition of academic use. To qualify under this category, the tools and resources:

[1] http://sciencecommons.org/projects/publishing/open-access-data-protocol/

[2] http://creativecommons.org/choose/zero

- should be available at least for anyone doing research or studying in an academic institution recognized by an Identity Federation (IdF) with which CLARIN has a service provider agreement;
- can be used for studying, research and teaching purposes.

If necessary, the CLARIN Academic Upgrade Agreement may be used to obtain the missing rights.

The last category, **Restricted Use** (RES) includes resources that do not fulfil the previous requirements but still could be offered to the End-Users if certain additional requirements are met. The most typical reasons for a resource to fall under the scope of RES are:

- a requirement to submit detailed information (e.g. an abstract) about the planned usage;
- specific ethical or data protection-related additional requirements. Content including Personal Data typically falls under the scope of RES.
- etc.

There may be **insufficient rights** to distribute the tools and resources for a CLARIN Service Provider, in which case the tools and resources may also be offered by an external Service Provider through a link from the CLARIN page. However, it is preferable to try to streamline the access process and to obtain the necessary additional rights by upgrading the main category to RES or ACA.

Additional Requirements

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

- A requirement for strictly non-commercial use (NC)
- A requirement to report the published articles, which use the tools and resources (Inf)
- A requirement to redeposit modified versions of the tools and resources (ReD)
- Any other requirement, e.g. a separate permission from the Content Provider. (Other requirements are typically applicable mainly to resources in the main category RES.)

NC	Inf	ReD

Frequently Asked Questions:

List of abbreviations:

ACA	Academic Use
AGPL	Affero General Public License
BSD	Berkley Software Distribution
CC	Creative Commons
CO	Content Owner, same as copyright owner or IPR owner
CP	Content Provider
CU	CLARIN User or End-User
DELA	Deposition License Agreement
DRM	Digital Rights Management
GNU GPL	GNU General Public License
Inf	Information required
LRT Inventory	CLARIN Inventory for language tools and resources www.clarin.eu/inventory
NC	Non-Commercial Use
ODbL	Open Database License
PID	Persistent Identifier
PUB	Public Use
ReD	Redeposition required
RES	Restricted Use
SP	Service Provider

TOPIC: RESOURCES, CATEGORIES, AVAILABILITY

1a). Can types of usage be added as further categories under RES? For example "query is allowed without restrictions", whereas "downloading is not allowed". Can we include approaches with several different options for transferability of rights that are resource type specific in RES?

If you provide clearly distinct end-user licences for different types of usages, the current LRT inventory is best suited for having each usage type as a separate entry.

Sometimes a tool has been built only for one particular resource, e.g. to serve as a lookup interface to an electronic dictionary. The end-user license covers both the interface and the underlying resource. You only need to record the usage types that you are allowed to offer for such a bundle.

Some examples:

- A dictionary query service on the web has a name of its own, so it can be classified e.g. as a PUB tool if it is publicly accessible.
- The underlying resource, i.e. the proprietary dictionary, can be mentioned without distribution rights, i.e. no main distribution category. The resource then shows up as a separate entry serving merely as metadata for someone who is interested in contacting the owner to procure the rights to the resource.
- Quotations are covered by what is normally allowed by the law, so no license is needed and no entry. Only if there is a specific restriction that quotations are not allowed should such a

peculiarity be recorded in the field for other restrictions of the usage category.

These questions also touch on the limits of how restricted a restricted resource can be. We cannot list all the conditions that are possible for a given resource in advance, but we are interested in finding out about possible restrictions, so for this purpose we have provided an open field for other restrictions.

1b). Concerning the specification of the licensed resources: on the one hand, data (texts, audios, videos), and on the other hand, services and software, it would be important to differentiate between these varied resources. For example, Our COSMAS II service could probably get an ACA-NC tag, the corpus texts it operates on, however, are RES and the license restrictions for the output of the service might be a third issue.

We recommend that you provide separate entries in the LRT Inventory for the different access modes. However, currently you can classify according to the most important access mode for the End-User and simply add a comment on the other options in the field for other restrictions.

2. 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?

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 anonymized.

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.

3. Can we find points in the RES space that serve as attractor points?

From a chaos theoretic point of view there probably are attractors, i.e. bundles of restrictions that tend to go together, around which most resources will flock. After the study is completed we are better equipped to answer this question.

4. 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?

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.

5. Are resources with excessive and firm conditions under the category RES welcome in CLARIN?

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

6. Should authentic written texts (such as letters, essays) be classified as RES?

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

7. Can resources that are currently part of a project be categorized as RES?

It is up to the project that collected the resources to determine how widely and when they wish to distribute their resources, but metadata mentioning their existence in the LRT Inventory is still recommendable.

8. If a resource includes several subcategories, e.g. newspaper articles from different newspapers, and they require different conditions. Can we keep it as one resource or should it be split into minor resources corresponding to the different conditions?

Both alternatives are possible. The resource can consist of parts with varying requirements as long as the strictest condition is applied to all of them. It is also possible to list the parts as separate resources.

9. Should the agreement mention the version of the resource or tool that is covered by the license?

It is good practice to identify the versions of a tool or resource to which the license applies. Note that each version will typically have its own PID, so the license may apply to a range of PIDs.

TOPIC: COMMERCIAL USAGE

10. 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?

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.

11. 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?

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.

TOPIC: AGREEMENTS/LICENSES

12. What happens if after signing the DELA the CO wants to amend the agreement so that it no longer fits into the CLARIN categories?

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.

13. Does the upgrade agreement effectively replace the original agreement?

No – it changes certain relevant parts, but the original agreement is still applicable.

14. Upgrading agreements can be costly – who will cover the costs?

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.

15. Who is responsible for upgrading agreements made during already expired projects?

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.

16. 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?

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.

17. 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?

If the original agreement cannot be amended for some reason, there are two options: use the model DELA agreement or write your own.

18. 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?

That is unlikely but still possible.

19. The CC license is not tailored for software. Should we use ODbL?

For software, popular open source licenses like the GNU GPL, the BSD or the EU Public License would be better.

20. Probably it might be helpful, if you could give some examples for categorizing e.g. common resource-licenses like GPL or CC.

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.

21. Deanonymisation of personal data may be possible for text but it is unlikely for audio or video, as it requires a massive effort. Can it be done by replacing names if de facto anonymisation is not possible?

The anonymisation task afterwards can be a massive and difficult project. There is also a question of who has the right to carry out the anonymisation in the first place. If the procedure is planned in advance, the best solution when compiling the data is to obtain a sufficient consent from the informants to use the data for research and redistribution through CLARIN.

The effectiveness of the anonymisation strategy has to be considered carefully as identification might be achieved by matching data from various sources, e.g. data from the original registers. Sometimes replacing names is not enough.

22. Will a model declaration of consent for already recorded speech be drafted? This should include a limitation to use the data for research only, include deletion date for personal data. The consent needs to be revocable.

General guidelines for a consent form are provided in the Milestone 7S2.4 version 3 "**CLARIN Model Contracts for Tools and Resources**", from which specific templates can be edited.

TOPIC: TECHNICAL AND RELATED

23. How can we convince the CO that the proposed technical solution for storing and distributing the data are better than the ones the CO uses at the moment if the CO does not want to make changes to the license and they are afraid of losing control of the data?

There is no simple answer for this question. One can at least argue that CLARIN as a Trust Federation has rather good control over who is logged in and to whom the files have been distributed or who have accessed the data. In addition, the CLARIN centres will have technical infrastructure according to the CLARIN specifications for hosting and distributing resources and the centres will be monitored and evaluated.

24. The CP is allowed to publish the resource on a specific website and anyone can access it through a search interface. If the data cannot be moved, is the RES category the correct one?

The search interface is on the web and as such a publicly accessible tool, but the underlying data is a clear default RES. For further details, see question 1.

25. 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.

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.

26. Technical measures to control the adherence to the license terms, and a specification are needed. This is essential in order to convince CO even to pass agreement drafts to their legal departments. CO-CP agreements can refer to these.

See answer to previous question.

27. 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?

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.

TOPIC: CLARIN

28. CLARIN needs to be extremely trustworthy with respect to distribution rights within CLARIN. What will be the legal form for CLARIN?

In 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 auditions will be integral to the operation of the centers.

29. What is the structure of CLARIN?

The members of the CLARIN ERIC will be the EC member states and not the research institutions. The CLARIN ERIC has a general assembly. The general assembly appoints the Board of Directors. The Board of Directors appoints and dissolves thematic committees. However, most of the activities of CLARIN will take place on a national level with nationally funded CLARIN committees responsible for the content development. In many countries there will also be CLARIN centers handling the infrastructure. Both of these national bodies will have standing committees in the CLARIN ERIC where they meet and discuss best practices and opportunities for cooperation.

30. Who takes care of the authorization for web service access?

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.

31. Who or what entity signs the deposition license agreements for CLARIN?

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 center 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 SP Federation is one of the parties in the agreement, the resource can be located at any CLARIN center. However, the CLARIN SP 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.

TOPIC: TEXT OF THE GUIDELINES

32. The paragraph ”Granted rights” could be more elaborate, i.e. defining what the user is allowed to do with the tools or resources could be documented in more detail.

The exact details are in the separate license agreements.

33. A section devoted to problems with tools and resources not under Creative Commons or ODbL where the rights of the CO would be secured would be welcome in the guidelines.

This is something we need more empirical data for, i.e. the current cases differ too much.

34. The links in the documents can change – maybe more detailed references can be provided in the future?

A separate list of references will be provided.

35. 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.

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.

36. Who will provide general guidelines on the national level?

It will be the task of the national CLARIN committees.

37. Will there be a registry of license terms like IsoCAT with a check-list for license agreements and declarations of consent?

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/