|  |  |  |
| --- | --- | --- |
| **Title** | CLARIN Resource Families Workplan 2021 | A picture containing lamp, drawing  Description automatically generated |
| **Version** | 1.1 |
| **Author(s)** | DF, FdJ |
| **Date** | 28 September 2020 |
| **Status** | Draft |
| **Distribution** | BoD |
| **ID** | CE-2020-1724 |

# Background and motivation

Since the launch of the CLARIN Resource Families (CRF) in April 2018, the initiative currently comprises overviews of

* 10 corpus families (parliamentary, computer-mediated communication, parallel, newspaper, L2-learner, historical, spoken, manually curated, literary, and academic corpora)
* 5 families of lexical resources (lexica, dictionaries, conceptual resources, glossaries, and wordlists), and
* 3 families of language tools (tools for normalization, tools for named entity recognition, and part-of-speech taggers and lemmatizers),

which together amount to 877 manually curated tools and resources.

When the overviews were prepared, a total 355 remaining issues pertaining to missing metadata on annotation, size, licence, or input/output formats, while only 84 issues have been solved since June 2018. In addition to the analysis and reporting of the issues, we have also launched a ticketing system for the issues on GitHub, where for each issue identified we indicated the CLARIN consortium that is responsible or most closely associated with the issue, thereby further incentivizing individual CLARIN consortia to help with the curation process. In this memo we present a work plan for the coordination and support the improving the curation of the CLARIN Resource Families.

# Proposed activities

In this section we list the activities which we believe to be achievable as well as sustainable. They would have very direct, very tangible and long-term improvements of the usability of our infrastructure.

## CRF Taskforce

In order to increase the usability of the resources and tools, we suggest to form a special taskforce that will help us:

* Curate the existing resources and tools in the CLARIN infrastructure
* Develop and implement preventive measures which will minimize the number of metadata issues for any newly deposited resources and tools in the future
* Develop a user-friendly best practice guide for depositing resources and tools
* Perform a gap analysis for the CRF initiative
* Evaluate the resource and tool families from a more qualitative perspective, taking not only into account the availability or absence of metadata, but also in which way metadata are reported for the observed categories. For instance, it is often the case that resource size is reported in broadly different ways even for resources in the same resource family. For example, where some corpora in the same family report size in terms of sentences, others report it in terms of tokens, words, hours or file size, which extent hinders the cross-comparability of the resources. In addition to different measurement units for size, certain resources specify their annotation layers very imprecisely, using vague descriptors such as “multitagged” or “tagged with lexical relations” without further qualifying them, which is far from informative and user-friendly. In the case of licence information, some tools and resource list unhelpful values such as an “other” licence.
* Promote a greater inclusion of key publications describing the tools and resources. Listing key publications is not only important from the perspective of ensuring and facilitating author attribution, but the publications themselves generally provide the most detailed descriptions of the resources and tools, thereby crucially complementing the metadata presented in CLARIN repositories with documentation that enable appropriate reuse of resources and tools as well as interpretation of research results.

Such a concentrated and increased effort in curation is of crucial importance because existing issues hinder the reuse of any resources and tools in the CLARIN infrastructure but especially those featured in CLARIN Resource Families, which has proven to be a highly visible initiative appreciated by a broad spectrum CLARIN users and therefore warrants continued and careful upkeep.

The proposed composition of the taskforce is as follows:

* Technical Director
* Director of User Involvement
* a SCCTC representative
* a representative of the Interoperability Committee
* a representative of the Standards Committee
* a KSIC representative
* a representative of the UI Committee
* CRF Officer (Jakob)

## Dissemination and acquisition of additional impactful existing resources

* Develop and implement a dissemination plan for the existing resources, e.g.,
  + follow the conversations on various channels, such as CorporaList and Twitter, and
  + provide input whenever resources and tools are sought.
* Develop a strategy and proactively encourage dissemination of existing impactful resources with CLARIN, such as the Distant Reading multilingual corpus (we list the ones we have identified that are not yet in CLARIN but are well known in the community in each overview of the Resource Families Overview already).

# Proposed workplan

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Activity | Who | When | Budget | Comment |
| 0. Coordination of the initiative and the taskforce | DF | on-going | 2.5 PM (additional budget) |  |
| 1. Curation of existing resources and tools | | | | |
| 1.1 management and administrative support of the curation of existing resources and tools in the CLARIN infrastructure | JL | on-going | 3 PM (SSHOC) |  |
| 2. Streamlining of new deposits | | | | |
| 2.1 Development and implementation of preventive measures which will minimize the number of metadata issues for any newly deposited resources and tools | taskforce | March 2021 | 2 F2F meetings in 2021 | the rest monthly virtual meetings |
| 2.2 Evaluation of the resource and tool families | taskforce,  JL | Sept 2021 | 1 PM (additional budget) |  |
| 2.3 Promotion of a greater inclusion of key publications describing the tools and resources | taskforce | on-going |  |  |
| 2.4 Best practice guide on depositing resources and tools | taskforce | June 2021 |  |  |
| 3. Dissemination of the existing resources | | | | |
| 3.1 Improve dissemination of the existing resources | EG | dissemination plan Feb 2021, dissemination on-going | no additional budget |  |
| 3.2 Follow the conversations on various channels, such as CorporaList and Twitter, and provide input whenever resources and tools are sought | JL | on-going | 0.5 PM (additional budget) |  |
| 3.3 Proactively encourage depositing of existing impactful resources in CLARIN | strategy development (taskforce),  implementation at ERIC level (JL),  implementation at national level (UIC) | strategy Sept 2021,  implementation Dec 2021 | 1 PM (additional budget) | considering a workshop on the topic to devise a strategy and best practice guide with input from other RIs, also important to consider the extra technical burden with the new deposits |
| 4. Enrichment of CRF | | | | |
| 4.1 Gap analysis and development of workplan for 2022 | taskforce | Dec 2021 |  |  |
| 4.2 Call for interest in small project proposals from the CLARIN network on any topic related to the CRF | BoD | on-going | 3-6 PMs per project, up to 3 in total |  |
|  |  | Total additional budget | 5 PMs + travel budget for 2 meetings of the taskforce (if possible) + 12 PMs for small projects |  |