# CLARIN-NL First Open Call

CLARIN-NL launches an open call for project proposals. This call is open from Tuesday June 2, 2009. The total budget for this call is limited to a maximum of € 600,000. This call is specifically open for proposals targeting resource curation projects or demonstrator projects. Only those proposals that specifically target this priority will be eligible. The maximum budget per project is 60,000€.

## 1 Introduction

The first phase of the CLARIN-NL project focuses on the specification and design of the infrastructure. Obviously, the CLARIN infrastructure should be designed in such a way that it can incorporate the data and tools currently used by humanities researchers to address their research questions. On the other hand, the CLARIN infrastructure can only be successful if these data and tools meet certain requirements with regard to standards and interoperability, not only with regard to the data and tools themselves but also with regard to their visibility and accessibility. Many of the data and tools currently in use do not meet these requirements, or meet them only partially. One of the aims of the first call is to make existing data and tools compliant with the requirements on standards and interoperability currently advocated in CLARIN.

However, the list of standards, best practices and interoperability requirements currently advocated by CLARIN has not been completely fixed yet. There are good reasons for this: (1) there may be crucial data or tools for which none of the currently advocated standards or best practices is suited; (2) the currently advocated standards and best practices may be incomplete, insufficiently specific, inconvenient or even incompatible with crucial data and tools. Therefore, a second aim of the first call is to test whether and to which extent the currently existing data and tools can be made compatible with the standards and best practices currently advocated in CLARIN, and to get a good overview of any incompatibilities as well as suggestions for adaptations of these standards.

These matters can best be investigated by actually attempting the necessary conversions and adaptations. When applied to tools and other software this will result in CLARIN-compliant web applications that can also serve as showcases of the kind of functionality CLARIN aims to offer. A third aim of the first call is therefore to make sure that the applications developed become available in demonstrators which can help promote the particular applications but also the CLARIN infrastructure as a whole.

Finally, by investigating these matters using specific data and tools that are currently in use a lot can be learnt about requirements the CLARIN infrastructure should meet or desiderata that it should offer. The fourth aim of this call is therefore obtaining a detailed list of such requirements and desiderata.

# 2 Project Types

In the first call, CLARIN-NL therefore solicits (1) projects that carry out resource (data or tools) curation, and (2) demonstrator projects.

**Curation projects**

CLARIN-NL aims to support the curation of digital language resources (data and tools) so that these resources can participate in the CLARIN infrastructure, more easily be accessed by interested researchers via online methods and become part of appealing new applications. Many language resources are neither visible nor accessible. Visibility is mainly achieved by standardized metadata that are being harvested by service providers. Accessibility has many different aspects:

1. The resource needs to be stored at computers that are accessible.
2. The resource needs to be identified in a persistent manner.
3. The resource needs to be interpretable, which requires a format that adheres to best practices and it requires references to registries were the used concepts are defined.

**Demonstrator projects**

CLARIN-NL aims to support projects that create appealing showcases of functionality that the CLARIN infrastructure should support. Such projects should make available web applications that can be used as demonstrators of functionality that supports addressing research questions of the CLARIN-NL intended user group. The development of these demonstrators will also be used to inventory a list of requirements the CLARIN infrastructure should meet and desiderata it preferably should offer.

# 3 Goals

The goal of a **curation project** is:

- Adapting specific resources so that they are visible, uniquely referable and accessible via the web, and properly documented.

The goal of a **demonstrator project** is:

- Creating a documented web application starting from an existing tool or application that can be used as a demonstrator and function as a showcase of the type of functionality CLARIN will incorporate and support. Within the web application there must be a clear separation between the web-based user interface and the core component. For the latter an API must be defined and documented.

Important goals **common** to both types of projects are:

- Applying standards and best practices and makes use of the suggested CLARIN architecture and agreements to understand their limitations and the requirements for extensions.
- Establishing requirements and desiderata for the CLARIN infrastructure

In both project types the use of CLARIN-supported standards and best practices is essential.

The selection of CLARIN-supported standards and best practices is currently ongoing and it is the intention that the projects supported in this call contribute to this. A preliminary set of candidates for CLARIN-supported standards and best practices is available,[1] and though it is by no means final we will refer to this list in this document by the term "CLARIN-standards".

# 4 Roles

Four roles of persons involved in the projects can be distinguished: the *user*, the *data provider* (DP), the *technology provider (TP)*, and the *infrastructure specialist (IS)*. The *user* is a researcher from a linguistics or humanities institute who aims to investigate one or more specific research questions. The DP has a certain set of digital language-related data at his/her disposal (the 'research data') that can be used for addressing the research questions of the user. The TP has a certain technology (e.g. language or speech technology) at its disposal and a thorough understanding of this technology (e.g. because it was developed by the TP). This technology offers functionality that makes it possible to (better) address the user's research question by applying this functionality to the research data. The *infrastructure specialist* (IS) is a specialist who has a deep understanding of the CLARIN service-oriented architecture and its requirements, and/or a specialist in data and tool format standards supported in CLARIN. In many cases the different roles of user, DP, TP and IS will be played by persons from different organizations, but they may originate from a single organization, and occasionally even be played by a single person. CLARIN-NL can offer assistance in bringing the right experts together, if desired. CLARIN-NL will assign IS specialist to awarded projects.

## 4.1 User

The *user* is a researcher from a linguistics or humanities institute who aims to investigate one or more specific research questions. The project proposal should clearly describe the research question(s) of the user, and the research question(s) must be in the domain of research in the humanities in general and the study of language in particular. The project must be led by the user.

## 4.2 Data Provider

The DP has a certain set of digital language-related data at his/her disposal (the 'research data') that can be used for addressing the research questions of the user. The project

[1] http://www.clarin.eu/recommendations

proposal should clearly describe the research data the DP has at his/her disposal that can be used to address the research question(s), and how they can be used for this purpose. The research data must be existing digital language or language-related data. No new research data should be created in the project. The DP must have the right to make the research data available on a CLARIN server running at a dedicated centre. If the data are in a format that is not currently on the list of CLARIN standards, a resource curation project is in order. Otherwise, the data can be used in a demonstrator project. The project proposal should contain a detailed description of the research data, its current state and format, the plans to convert it if needed, justification for using different formats if applicable, and a detailed plan for dealing with the data and its metadata (see below). Any restrictions on the use of the data as well as any ethical issues that apply or may arise must be properly documented in the proposal.

## 4.3 The Technology Provider

The TP has a certain technology at his/her disposal that can be used as a basis for the development of a web-based application (possibly web-services based) and concomitant demonstrator, or that can be used for resource curation. Since the research data are language data, the technology will in most case be language or speech technology.

The proposal should contain a detailed description of the available technology and its current status. It should make clear that the TP has a thorough understanding of this technology and describe how the TP obtained this understanding (e.g. because the TP developed the technology).

The intended use of the technology in the project should be described, as well as any extensions or modifications that have to be made to the technology in the project, and a plan to achieve this.

The TP must have the right to use this technology and indicate how it will be used in the project.

## 4.4 Infrastructure Specialist

The infrastructure specialist (IS) is a specialist who has a deep understanding of the CLARIN service-oriented architecture and its requirements, and/or a specialist in data, metadata and tool format standards and best practices supported in CLARIN. The IS will assist the user and the TP to turn the research data and/or the technology into CLARIN-supported formats and architectures. The IS will advise and assist the project partners and may also contribute by providing reusable CLARIN components to the project partners. The IS will assist selected projects to turn its web application into a web service, and will assist the project with testing the project results.

A research data resource often consists of information of various kinds contained in multiple folders and multiple files of varying types. The information contained in such a resource can include documentation, source data, annotations of the source data, aggregate statistics tables on the source data and/or annotations, etc. A web service should find out in a fully

automated manner whether the research data selected by the user are appropriate input for the web-service, and, if so, that it is applied to the right information (e.g. to the source data but not to the documentation). To achieve this it will integrate wrappers that read and write metadata and provenance information provided by the CLARIN infrastructure. Any requirements or desiderata that follow from this for metadata and data contents and formats should be properly documented in the documents with the requirements and desiderata for the CLARIN infrastructure.

# 5 Project Types

## 5.1 Resource Curation

Resource curation involves a number of different aspects:

1. The resource should be brought into a format that adheres to widely accepted standards and best practices currently considered as likely candidates by CLARIN.
2. Proper metadata descriptions need to be created and made available. They must be compliant with the CLARIN component metadata infrastructure (CMDI) and it should be possible to harvest and access them.
3. Metadata descriptions should include persistent identifiers that can be resolved and the CLARIN requirements should hold for the PID system.
4. As far as possible, the linguistic encoding must be related to the data category registry, i.e. an auxiliary resource should be created and made available that includes formally represented mappings between the categories specific to the research data and those that are registered in ISOcat[2] or with new entries registered in the user space of ISOcat that you create if existing categories do not match.
5. Provide proper documentation of the resource.

The plan for a curation project should describe in detail how these different aspects are going to be addressed in the project.

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

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

Depending on the type of resource (text corpora, annotated corpora, lexica, audio, etc.), different expertise from the IS will be needed. If the project is awarded, CLARIN NL will select and assign one or more IS specialists to the project.

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

## 5.2 Demonstrator

In a demonstrator project a demonstrator is developed using a documented web-based application based on a technology that the TP currently has at his/her disposal. The development is carried out in close cooperation with the user and the IS.

The project proposal must contain a detailed description of the targeted functionality, including input and output specifications, and how it can contribute to addressing the user's research questions. More generic functionality, i.e. functionality that can serve multiple different research questions from linguistics and humanities research will be preferred over less generic or completely idiosyncratic functionality. See below for additional criteria related to the functionality that will be used to rank proposals. Since a demonstrator project is short in duration and is relatively small, this functionality must already be available to the TP, though perhaps not in the form of a web-based application, and it perhaps only operates on data formats other than the ones listed in the CLARIN standards. The project proposal should contain a detailed description of the functionality in its current state, the targeted web-based application and its components, and a plan to achieve this. The application includes a web-based user interface that takes care of user interactions and method invocations to the core component. An Application Programming Interface (API) to the core component must be provided and documented. The TP must have the rights to make the targeted core component as well as the web-application available on a CLARIN server running at a dedicated centre.

The core component of the web-application must at least be able to operate on the research data and yield output in the formats agreed upon between user and TP. It is a pre if it can apply to other formats from the CLARIN standard list and yield additional output formats. The web-application and its core component will be used to obtain requirements and specifications of the architectural framework that is being worked out in CLARIN and may be used to test it.

The demonstrator consists, as a minimum, of a web application, the research data, and a demonstration scenario. A demonstration scenario is a detailed description of example (sequences of) actions a user can take to have the application applied to the research data and the corresponding system responses in order to get a representative picture of the functionality offered. A movie or sequence of screen captures to illustrate the functionality is nice to have. The application will have to be installed on a CLARIN server, and the project proposal must contain a plan for doing this. It is the intention to have the demonstrator applications available for the lifetime of the CLARIN-NL project (2009-2014), so occasional support may be needed from the original developers even after the demonstrator project has finished.

The application must be tested with at least one of the common web browsers on the client side (MS IE, Firefox). Agreements about additional technical details (operating system, programming language, workspace requirements, etc) need to be made with the dedicated centre where the services should be executed.

Any vendor, platform or operating system dependent aspects of the application must be made explicit in the proposal and properly documented in the project.

The web-application and its core component should be properly documented, for users (user documentation), for application developers who want to use the core component (documentation of the API), and for technology developers who want to modify or extend the basic functionality of the application.

**Auxiliary Resources**

The web-application may require data and other software (auxiliary resources) while running.

It must be documented which auxiliary data (e.g. a lexicon) and software (e.g. a library, converters) are needed during runtime for the application. The TP must have the right to make these auxiliary data and software available on a CLARIN server. Any restrictions on their usage (including costs) should be properly documented in the project proposal and in the documentation of the resulting application.

The application and the core component must be able to run on a dedicated CLARIN-server. An application or core component that can run only on a specific (non CLARIN) server (e.g. because it contains auxiliary resources that cannot be made available otherwise) is not acceptable.

# 6 Metadata

For the web application and its core component, the research data and all runtime auxiliary data used in the application, metadata descriptions must be made in accordance with the CLARIN metadata standard (CMDI). Since the CMDI is currently under development and we hope that the projects of this call can contribute to this development, metadata should be created using the IMDI metadata format.[3] Metadata in IMDI format will be harvestable and can be integrated in the future CMDI-compatible metadata repositories. If the IMDI format is not applicable or suitable, this should be thoroughly justified and documented and discussed with an IS with excellent IMDI expertise. If unavoidable, deviations from this standard are allowed. Any such deviation must be properly documented and be included in the CLARIN Requirements and Desiderata document.

# 7 Requirements and Desiderata for CLARIN infrastructure

One important result of both demonstrator and curation projects is a document or series of documents describing requirements and desiderata for the CLARIN infrastructure resulting from the experiences gained with the curation of the research data and/or tools, and with the development of the application, its core component and web-services derived from it.

[3] http://www.mpi.nl/IMDI/

These requirements and desiderata can concern many aspects. The following is a non-exhaustive list of aspects that should be considered:

- Requirements for data formats and encoding standards
- Web-service wrappers
- Metadata elements and formats
- Processing requirements
- Memory requirements
- Network Bandwidth requirements
- User workspace requirements
- API requirements (e.g. Calling conventions)
- IPR / restricted use / ethical issues requirements
- Documentation requirements
- Repository Requirements
- Requirements for registering and resolving PIDs
- Requirements related to semantic interoperability

# 8 Evaluation Criteria

Proposals for projects will be evaluated and ranked according to criteria listed in the CLARIN-NL Long Term Working Plan. For convenience, these are also listed here:

- **Quality**
  - Clarity and originality of the project proposal, in particular of the problem and the proposed approach
  - Suitability of the method and plan for the problem at hand
  - Feasibility of the project targets: can they be realized within the specified amount of time and with the instruments proposed?
  - Adequate balance between requested instruments and funds and proposed targets
  - Clearly specified and realistic work plan
  - Conformance to established standards and protocols as supported within CLARIN, or contribute to the development such standards and protocols.
- **Project Participants**
  - Competence of the participating partners (including their past performance);
  - Balanced cooperation and task assignments within the project. Justification of the composition of the team.
  - Availability of the infrastructure required for the project to be successful
  - Embedding of the work in other research programmes or projects, and/or additional funding from other funding sources is an advantage
- **User-orientation of the project**
  - Does the project address needs of the targeted infrastructure users (linguists and humanities researchers)?
  - Is there cooperation with or support from the targeted (future) infrastructure users?
  - Is the resulting tool/ service user-friendly, i.e. will non-technical linguistic and humanities researchers be able to use it?

  - Is dissemination of the results to the targeted users and (where appropriate) training of them planned?
- **Contribution to CLARIN-NL as a whole**
  - Conformance to the goals of CLARIN-NL in particular and CLARIN in general and the priorities set within them
  - Contribution to knowledge transfer and network creation. In particular, cooperation between the intended users (linguists and humanities researchers) and technology and service providers (researchers in language and speech technology, computer science, etc.) is an advantage.
- **Intellectual Property Rights and Synergy**
  - Each proposal must contain clear statements about the situation of the IPR of the data and tools/technologies used, and a detailed plan to resolve any open issues.
  - The project participants have the obligation and must therefore have the rights to incorporate the core data and tools used in a project into the CLARIN infrastructure (this is a sine qua non). There has to be a clear specification and justification of the use of any data or tools needed in the project that cannot be incorporated into the CLARIN infrastructure.
  - Each proposal must show that the submitters have adequate and up-to-date knowledge of data, tools and services that are already available, so that any duplication of effort can be avoided.
- **Formal compliance**
  - A proposal must meet the formal requirements imposed by the CLARIN-NL organization for proposals, such as
    - conformance to the prescribed format and proposal template
    - submission before the set deadline, using the means prescribed
    - conformance to the prescribed language of the proposal

In addition, more generic applications and data, i.e. applications and data that can serve multiple different research questions will be preferred over less generic applications and data. In this call, a project that does not meet the IPR-requirements stated or is insufficiently clear about it will be considered formally noncompliant.

# 9 Duration

The duration of the project must be justified. The default maximum duration is 6 months. Any duration longer than 6 months requires thorough justification.

# 10 Budget

The project budget must be in accordance with the tasks to be carried out, and this must be justified in the project proposal. The maximum budget is 60k€ (the approximate costs of 2 FTEs for half a year).

# 11 Intellectual Property Rights (IPR)

Ownership of all original data and software remains with the original owners.

An agreement must be in place between the owners of the original data and software and the project participants on the IPR of the adapted data and software before the submission date of a proposal if the owners of the original data and software are not identical to the project participants. If applicable, a copy of this agreement must be uploaded together with the project proposal. Otherwise ownership of the created adaptations and extensions will be with the creator(s).

The project participants have the obligation and therefore must have the rights to make the research data, the application, its core component, and any runtime auxiliary data or software available on a CLARIN server for use by researchers having access to the CLARIN infrastructure. This is a *sine qua non*. Any proposal not satisfying this requirement or being insufficiently clear about this matter will be considered to be formally noncompliant and will be rejected on these grounds.

The project proposal should describe all issues related to IPR and present solution for them. The relations between the partners in a project must be agreed upon in a consortium agreement before the start of the project.

**Practical details**
**On the submission and evaluation procedure regarding the**
**CLARIN-NL First Open call**

**Call for Proposals**

The CLARIN-NL First Open Call is open from Tuesday June 2, 2009. The total budget for this call is limited to a maximum of € 600,000. This call is specifically open for proposals targeting resource curation projects or demonstrator projects. Only those proposals that specifically target this priority will be eligible.

Full proposals must be submitted in English and in PDF format to the CLARIN-NL electronic proposal submission system using the prescribed template (which can be found on the CLARIN-NL website). The deadline for submitting full proposals in this call has been set for Monday August 17, 2009 13:00 hours CET.

The CLARIN-NL electronic proposal submission system can be accessed as of Monday June 22, 2009 via the CLARIN-NL website.

**Who can apply?**

Applications can be submitted only by researchers affiliated to CLARIN-NL participants that have signed the CLARIN-NL consortium agreement. The list of CLARIN-NL participants is available on the CLARIN-NL website. The main applicant and coordinator of the project must be a researcher from linguistics or humanities more broadly.

CLARIN-NL is in principle open to new participants. If your institute is not currently a CLARIN-NL participant, contact the CLARIN-NL office if you want your organization to become a participant in CLARIN-NL well in advance of the submission deadline.

Applicants who are planning to submit a proposal are strongly advised to contact the CLARIN-NL Office for an eligibility check of their plans.

**Eligible costs**

- Personnel costs directly related to the project up to a maximum of 60,000€, in accordance with the *Akkoord NWO-VSNU 2008* (and any additions to it).[4]
- A fee of maximally 3.000 € per FTE per year (or a pro rata part for less than 1 FTE per year) for covering travel and subsistence costs

**Evaluation procedure full proposals**

[4] http://www.nwo.nl/nwohome.nsf/pages/NWOP_67QK4E, more specifically http://www.nwo.nl/files.nsf/pages/NWOA_7LYGWY/$file/Ondertekende_Overeenkomst_NWO_VSNU.pdf

All eligible full proposals submitted in this call will be presented to a panel of international experts in the humanities, language and speech technology and infrastructures (International Advisory Panel, IAP). The composition of the IAP will be published on the CLARIN-NL website as soon as it is available. If the proposals require this, the CLARIN-NL Executive Board can decide to involve additional experts in the evaluation.

This international panel will assess and rank the eligible applications based on the assessment criteria relevant to this call and may formulate a set of recommendations for improving the individual proposals. The panel's assessment and recommendations will be presented to the National Advisory Panel (NAP). The members of the NAP who are not directly involved in the submitted proposals will assess the applications and the IAP's assessment and recommendations and determine the order of priority of the eligible proposals. On the basis of the IAP's advice and the NAP's advice, the CLARIN-NL Board will finally determine which projects will be funded.

Projects should start within two months after the applicant has received the formal notification of funding.

**CLARIN-NL *consortium agreement***

More information as to which legal rules apply for this specific CLARIN-NL granting scheme are laid down in the *CLARIN-NL consortium agreement*, which can be found on the CLARIN-NL website.

**Timetable**

| **Activity** | **Date** |
|---|---|
| CLARIN-NL Call Open | Tuesday June 2, 2009 |
| CLARIN-NL Kickoff Meeting | Wednesday May 27, 2009 |
| Deadline Proposal Submission | Monday August 17, 2009 13:00hrs |
| Assessment IAP | Wednesday September 16, 2009 |
| Assessment NAP | Friday October 16, 2009 |
| Decision on Funding by Board | Monday November 2, 2009 |

**CLARIN-NL Organization**

The CLARIN-NL project is funded by NWO.

The CLARIN-NL project is coordinated by the Programme Director, prof.dr. J.E.J.M. Odijk, who is a member of the CLARIN-NL Executive Board.

The International Advisory Panel (IAP) is a group of international experts in the areas of humanities, in particular linguistics, language and speech technology, and infrastructures for scientific research.

The National Advisory Panel is a group of national researchers representative for the fields of linguistics and humanities, language and speech technologies and infrastructures for scientific research.

The CLARIN-NL Board consists of national senior researchers with great expertise in governance and/or relevant technical expertise

The composition of these CLARIN-NL governance bodies can be found on the CLARIN-NL website.

**CLARIN-NL Office**

Contact the CLARIN-NL office for any questions related to this call, e.g.

- Additional clarification
- Advice on eligibility of your plans
- Assistance with finding experts, data or technology required
- Etc.

Jan Odijk
Tel: +31 30 253 6006
Fax: +31 30 253 6000
Postal Address: Janskerkhof 13, 3511 BL Utrecht
Visiting Address: Achter de Dom 24, 3512 JP Utrecht
E-mail: clarinnl@uu.nl (or j.odijk@uu.nl )
CLARIN-NL Website: http://www.hum.uu.nl/clarin-nl/.