

**Subsidy : National Roadmap for Large-scale Research Facilities**Project number : 184.032 306Project title : **CLARIAH - Common Lab Research Infrastructure for the Arts and Humanities**Main applicant : **Prof. dr. J.E.J.M. Odijk****Summary rebuttal**

Overall, we interpret the referee reports as encouraging and strongly supportive of our application. Three referees evaluate the proposal as "excellent", one as "very good to excellent". On the individual "cases"/aspects, we received 28 "excellent" scores, 13 "very good" scores, and, disappointingly, 1 "poor" score (to which we will return later on). One referee did not evaluate the business case, another skipped the technical case.

The "talent", "partnership", and "critical mass" cases are unanimously evaluated as "excellent". The reviewers do not seem to agree on the "social trends" case. In spite of their praise, they make a number of very useful remarks to which we would like to respond and they give us a number of suggestions for which we are grateful.

As both the proposal and the referee reports are structured according to the 11 cases, it makes sense to structure our rebuttal similarly.

**1. THE LIKELIHOOD OF SCIENTIFIC BREAKTHROUGHS (SCIENCE CASE)**

Ref. 1 states that it is important not to forget "the numerous researchers in Arts and Humanities (A&H) who are less familiar with digital data". Ref. 3 likewise warns not to "neglect [...] the very possible passive resistance on the part of older researchers in the humanities". We will not, as is stated under the Talent case (2).

Ref. 2 hopes that CLARIAH will consider something similar to the publication of workflows in biomedical research, allowing the replication of findings; and the referee envisages that even undergraduates will get hands-on experience with the infrastructure. Replication of findings will certainly be stimulated by CLARIAH. Workflow systems are only just emerging in the humanities domain, for instance in the context of the CLARIN project. We happily take the referee's advice to make publication of workflows an element of the CLARIAH infrastructure. The role of CLARIAH in academic curricula is discussed under the Talent case (2).

For Ref. 3 "It is not entirely clear how Dutch-oriented the facility will be". The facility will reflect the research interests of the humanities community in the Netherlands. Therefore it is our estimate that the majority of the material is Dutch, but many text, audio-visual and structured data resources in CLARIAH pertain to other languages, cultures, economies and societies. Since CLARIAH is part of the European infrastructures CLARIN and DARIAH, core metadata, software documentation and interfaces will be available in English or use language-independent coding.

The referee also states that "As regards audio-visual material, much work has been done over the last twenty years and the project does not seem to add anything spectacularly new to what is presently around". It seems fair to say that a major part of the work in the audio-visual realm of the past decades is coming from or has been initiated by projects involving CLARIAH partners (funded by the EU or otherwise). Notably UT, UvA, UU and VU have proven their scientific excellence in the development of human-computer interaction for multimedia, video search and visualization tools, retrieval of music, text and spoken word content, and multimedia semantics. Earlier this year, the public-private Dutch IT-programme COMMIT was launched (funding 50 M€) which builds on the heritage of its predecessor MultimediaN (including video search engine MediaMill and other e-culture access tools). Cultural heritage is again one of the core application areas, and Sound and Vision, one of Europe's most advanced audio-

visual archives, participates. CLARIAH will stimulate the teams involved to continue their role in advancing processing technology for audio-visual content.

CLARIAH will advance the state of the field for platforms supporting humanities researchers and other professional user groups that use audio-visual content as a primary source of scholarly research, such as oral historians, media professionals and art historians. The introduction of digital workflows for the audio-visual domain is lagging behind in comparison to what is available for the text, and the active involvement of scholarly user groups in the co-development of platforms supporting novel practices can bring a leap forwards in digital scholarship. This explains the CLARIAH focus for the audio-visual Kit on innovative functionalities such as visualisation of large corpora, recommendation technology for audio-visual, annotation platforms for scholars that integrate manual and (semi-)automatic labelling of audio-visual content, and the generation of links across collections and modalities, all in accordance with the emerging standards for open data.

This referee rightly notices that the “CLARIAH standards and procedures [...] are not spelt out” in the proposal text. However, we clearly state that these are based on existing standards and procedures adopted in CLARIN<sup>1</sup> and DARIAH<sup>2</sup>, which are available on their websites and will be integrated in the CLARIAH context.

Finally the referee criticizes our “frequent use of modal verbs (should, may, will, can) and expressions [...] which indicate, on occasion, a lack of certainty” (or, as stated in the summary overview “a desire for honesty”). Modal expressions are often used to indicate uncertainty, although there are also other uses. In the proposal text, all occurrences of the “uncertain” use of modal expressions are in the context of future events that are in most cases not (or not completely) under our control. The use of such modal expressions is in our view appropriate here, and it certainly does not indicate a lack of decidedness from our part.

Ref. 4 would have liked to see “more and clearer examples of individual research fields and inter-and trans-disciplinary collaborations”. Considerations of space have induced the applicants not to include specific ‘good examples’ in the main text, even though these are abundantly available; for telling cases of individual research and cross-disciplinary collaborations, see e.g. <http://www.clarin.nl/node/76>; <http://www.dans.knaw.nl/en/content/projects>; <http://alfalab.ehumanities.nl/>; <http://ehumanities.nl/projects/>; [http://www.nwo.nl/nwohome.nsf/pages/NWOP\\_6CCC3L\\_Eng](http://www.nwo.nl/nwohome.nsf/pages/NWOP_6CCC3L_Eng)

## **2. THE POTENTIAL FOR 'BRAIN GAIN' (TALENT CASE)**

The undergraduate and postgraduate curricula in the Netherlands will include the use of the CLARIAH infrastructure as a regular component in the training and use of advanced ICT methods in Arts and Humanities (as Ref. 1 urges; see also the proposal text, p. 15). An English interface is also provided (see above). Of course CLARIAH will offer training to established researchers as well (as Ref. 2 wishes). The proposal naturally often mentions ‘students’ in the context of education and training, which, in this context, is intended to encompass existing scholars and researchers that are elsewhere also explicitly mentioned (scholars, p. 16, and researchers p. 17).

## **3. SOCIAL AND COMMERCIAL RELEVANCE (INNOVATION CASE)**

Where Ref. 2 stresses the publishing industry, Ref. 3 mentions audio-visual and entertainment-based products, and Ref. 4 the communications industry as fertile areas for creative synergy and likely spin-offs of CLARIAH. We agree with Ref. 1 that “the Common Lab can be of utmost interest to a wide range of enterprises”. Tangent planes between the CLARIAH infrastructure and the Dutch “Top sector” of Creative Industry, ranging from the publishing industry to audio-visual and (serious) gaming industry, are manifold. In fact, the project will open up a vast area of potential economic and societal valorisation. This perspective of synergy between CLARIAH with ‘society at large’ plays a significant role in the national policy plans now under construction for boosting the Creative Industry sector.

<sup>1</sup> <http://www.clarin.eu/recommendations>

<sup>2</sup> [http://www.dariah.eu/index.php?option=com\\_docman&task=cat\\_view&gid=92&Itemid=200](http://www.dariah.eu/index.php?option=com_docman&task=cat_view&gid=92&Itemid=200)

#### **4. COLLABORATION AND COMPETITION (PARTNERSHIP CASE)**

The only slightly critical remark we could discover here is by Ref. 3: "some of the ventures proposed are qualified by the phrase 'it is expected that'." We already commented on this under section 1 (Science Case).

#### **5. FINANCIAL ASPECTS (BUSINESS CASE)**

All reviewers (except Ref. 1, who skips this case) seem to find the requested budget reasonable and justified (though with little detail) in the light of the ambitions. More detailed budgets, linked to the work plan, deliverables and partners are being worked on and can be discussed in a hearing session with the commission.

Ref. 2's question whether the costs for training users and for the "brain gain" activities are included under "Dissemination and outreach" can be answered affirmatively.

#### **6. TECHNICAL FEASIBILITY/TECHNICAL CHALLENGES (TECHNICAL CASE)**

Ref. 2 would have liked the proposal to be more specific about technical aims and methodology. Ref. 2 is "concerned about the possible expectation of development in search technology and NLP beyond the current state of the art (e.g., "tools for analysis of discrepancies between stories", p. 10) - it is possible that the proposers assume something relatively limited which is currently feasible, but it's unclear exactly what." CLARIAH offers a general facility to store data and tools, to find them there, and to apply tools to data in a seamless manner. Some of the tools that are candidates for inclusion in the CLARIAH infrastructure reflect cutting-edge research, and performance may currently be insufficient to make them available as generally usable tools; but the research into and the development of these technologies will be boosted by the CLARIAH infrastructure. In particular, we would like to stress that all tools mentioned as examples in the proposal are currently being used by groups of researchers. For instance, the "tools for analysis of discrepancies between stories", which Ref. 2 specifically refers to, have recently been developed on the basis of semi-automatically identifying different narrative structures and perspectives in stories. These tools are highly useful for many other researchers and students in both the humanities and the social sciences. It is precisely CLARIAH's goal to create an adequate infrastructure to facilitate this use. This referee's warning for "one size fits none" is certainly well taken, and we emphasize the importance of taking into account the user requirements from the start in several places in the proposal (e.g. on p. 20) as well as the need to develop the infrastructure in a stepwise manner and guided by early usage tests (p. 26).

The problem of finding (enough) suitable personnel was discussed extensively within the proposal preparation team. We think this point should also be seen in relation to point 2 (brain gain): a part of the personnel will almost certainly be hired from beyond the Netherlands in order to meet the volume of high-quality executive force needed for this project.

Ref. 3 finds that "the technical feasibility of the project, while not 100%, is reasonably convincing". We agree that much of the technical work requires more detailed description and planning, which is already incorporated in the outline of our work plan.

#### **7. POSSIBLE FOCUS FOR THE NETHERLANDS**

All referees agree that The Netherlands already has a strong position in digital humanities research, and that we are perfectly placed for hosting the proposed infrastructure, which will also strengthen the European counterparts.

Ref. 4: "If the collaborative and partnership dimensions involve regular seminar based discussion of emerging questions it is highly likely that the Netherlands will be at the centre of questions of intersection between humanities disciplines and public policy problems. This dimension was not brought out in the application but it seems to me an important 'lurking' possibility." We thank this referee for pointing out this potential of CLARIAH. Technology stimulation in the humanities, when used in the context of the "Topsector" Creative Industry, will indeed put the Netherlands in the centre of the solutions for public policy enabled by language technology.

**8. CRITICAL MASS**

Although we are proud indeed that we managed to unite virtually all humanities research organizations in The Netherlands in CLARIAH, our claim is not as “grandiose”, as Ref. 3 states, that we think that “the entire humanities community in the Netherlands [...] is involved”. Of course, even though our ambitions are high and the scale of the infrastructure substantial, the deliverables of CLARIAH will need to be focused in order to be useful, and hence will inevitably have a more direct impact on some communities than on others.

**9. EMBEDDING**

All referees find CLARIAH well embedded; ref. 3 remarks that “the foreign connections are less clear but undoubtedly exist”, which does not prevent him from marking our embedding as “excellent”. The main applicant and other participants of the project are leading partners in related international projects described in section 9 (p. 27-28).

**10. PROVEN WILLINGNESS TO COLLABORATE**

The referees are happy with this item; no further remarks needed.

**11. REFLECTION OF SOCIAL TRENDS**

The referees have different views on this aspect of our proposal.

On the one hand, Ref. 2 and 3 find our claims “unconvincing” or even “pure speculation”, especially where we attempted to describe the relevance of techniques like data mining, information filtering and content analysis in a social policy area such as safety. Of course, we did not want to imply that CLARIAH in itself could prevent incidents such as 9/11 or the murder of Theo Van Gogh. However, we do maintain that such techniques, when applied to resources of political and social interest, make it possible to analyse (textual, structured and audio-visual) data from these domains more effectively.

Ref. 3 finds that our “claim that problems such as the negative effect of parental divorce or the question of below-replacement fertility can be better understood through longitudinal analysis of large quantities of data, is certainly of great interest but to be totally convincing we will need some concrete proof.” Below-replacement fertility has often been associated with the ‘second demographic transition’, the supposedly irreversible change in values and life styles that has begun in the 1970s and which is wreaking havoc with population forecasting. However, recent research has shown that already in the 1920s and 1930s fertility fell drastically, a situation that was reversed during the ‘baby boom’. We need to understand these dramatic fluctuations of demographic behaviour in order to improve our grasp on current developments.

On the other hand, ref. 4 is disappointed by our “failure of imagination” in this dimension and scores us a disappointing “poor”, although this referee recognizes the value of “monitoring the attitudes and positions of Dutch politicians with regard to important national and international problems”, which we described elsewhere. We chose to restrict ourselves to two social issues that are central in the Dutch government’s social policy, but of course we can report other national social developments and trends. We mention once more two of the core partner projects that will benefit from and contribute to the CLARIAH infrastructure: Clio-Infra, addressing the topic of global inequality, the increasing divergence between rich and poor countries, which is one of most pressing concerns of our time; and the Historical Sample of the Netherlands, providing insight in the long-term changes in Dutch demography and society. We are of course happy to provide additional examples, for instance in a hearing with the Roadmap committee.