chist-era

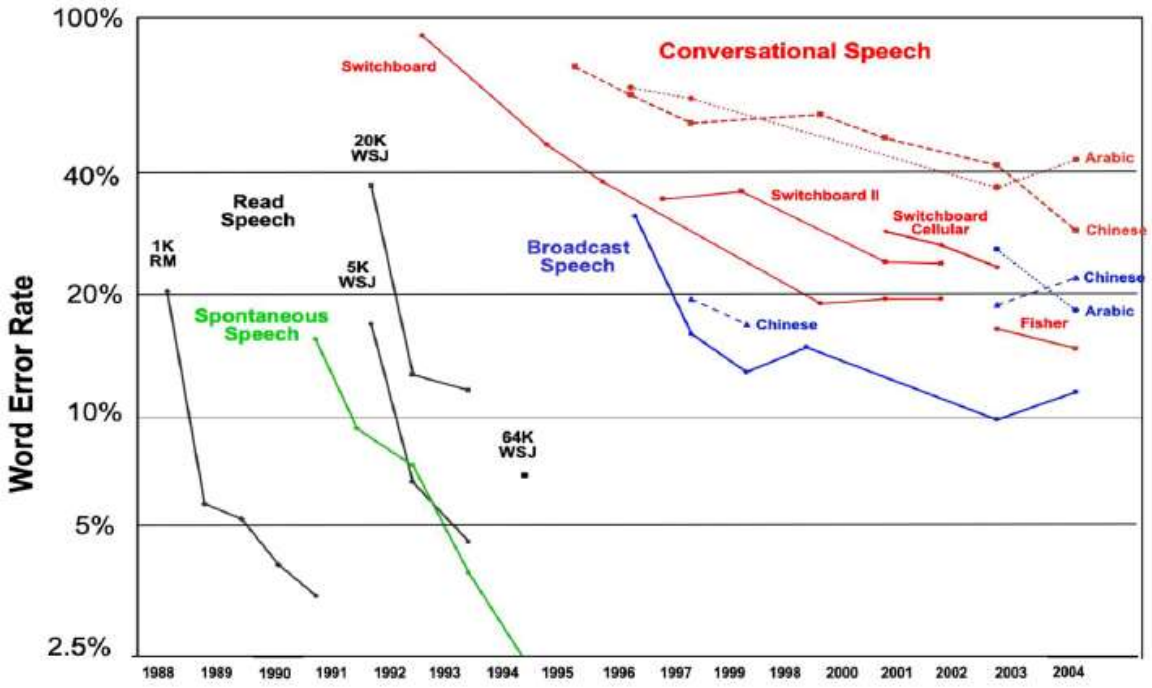
# CHIST-ERA Challenge Call

## Open Evaluation Methodology & Experiment Reproducibility in ICT Research

This document reminds the concept of a call following the Challenge format: What it is, why in CHIST-ERA and recommendations for such a call. It builds on the online workshop organised by CHIST-ERA and ICT-AGRI-FOOD ERA-NETs on November 23, 2020 and aims at guiding the topic selection for the Challenge as well as paving the way towards the launch of a Challenge in the fall of 2021.

In general, a Challenge is more focused than a standard call. Above all, it involves close coordination with the operation of evaluation campaigns to evaluate funded projects performance. Both specificities call for a deeper preparation methodology compared to the preparation of the yearly call of CHIST-ERA. This document represents a step in that direction.

At the strategic level, by listing connections between the Challenge concept and past CHIST-ERA activities and the added value of the Challenge approach with regard to Open Science, the document highlights the pathfinder role of CHIST-ERA in spreading the use of *open* evaluation methodologies and support to experiment *reproducibility* in ICT research in Europe.




*"The* Common Task *method, a research management technique developed in the 80s and applied increasingly widely since then to support long-term R&D, is arguably responsible for the success of all modern* AI *research"*

*Mark Liberman (CHIST-ERA Scientific Advisory Board member), Projects Seminar 2019*

chist-era

# Table of Content


Preparation Timeline ..... 3

Description ..... 4

Evaluation campaigns ..... 5

Why is it needed? ..... 6

Benefits and limitations ..... 7

Challenges and CHIST-ERA ..... 8

Recommendations ..... 10

Keep It Simple! ..... 10

Connect Early with Evaluation Campaigns Organiser: Squaring the Circle ..... 10

Conclusion ..... 12

chist-era

# Preparation Timeline

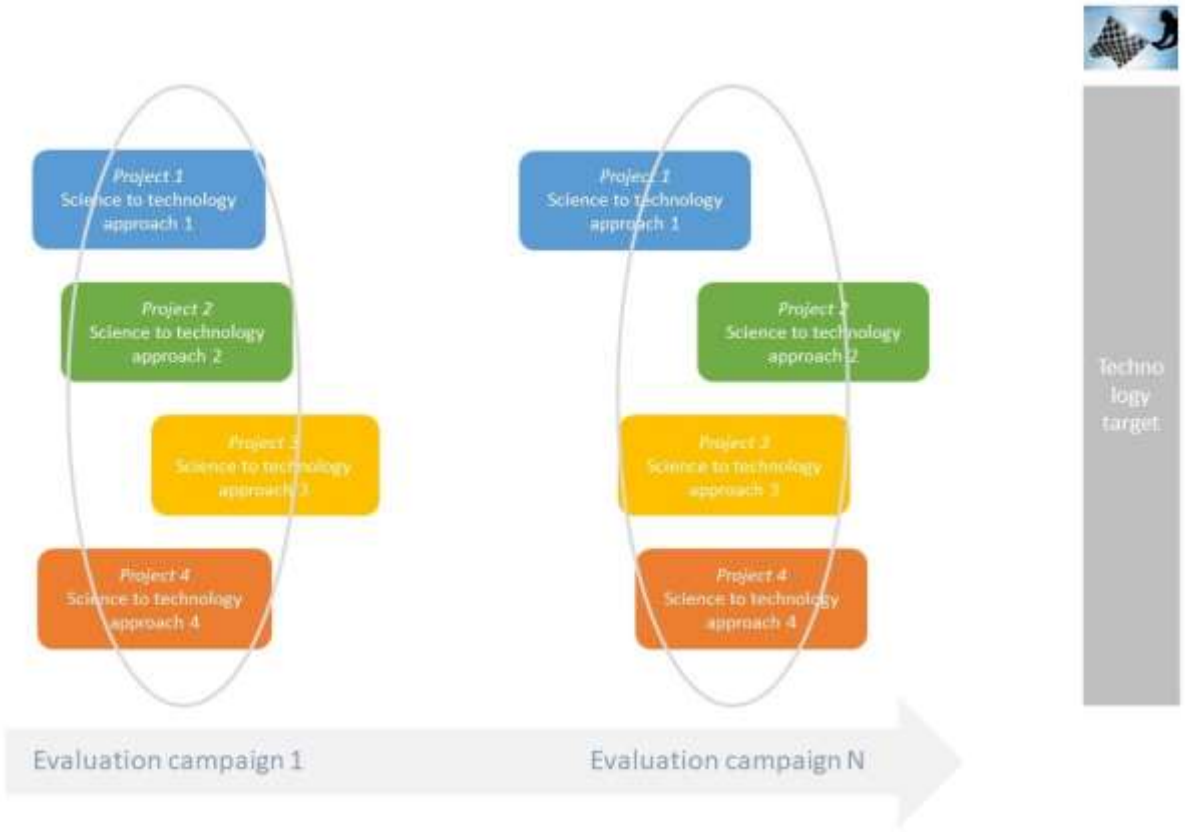
The table below lists the main steps of the Challenge preparation implemented so far as well as remaining steps towards the launch of the call end of 2021.

| Time | Action | Status |
|---|---|---|
| **1st semester 2020** | The opportunity of a Challenge call is presented by ANR | OK |
| **Mid-2020** | EC supports the idea<br>• The Challenge Call would be added to CHIST-ERA IV<br>• Its timeline would be aligned with the Call 2021<br>• Its organisation would rely on the CHIST-ERA call infrastructure (topics selection process, evaluation standards…) | OK |
| **Mid-2020** | Design of Challenge call topic suggestion form with the Scientific Advisory Board | OK |
| **November 2020** | Workshop organised with experts in Challenge organisation: US NIST, EU project METRICS, French LNE, CHIST-ERA project IMOTION, Challenges MALIN and ROSE | OK |
| **2nd semester 2020** | CHIST-ERA partners and the public propose topic candidates<br>• 4 topics proposed | OK |
| **Beginning of 2021** | Topic selection | Ongoing |
| **1st semester 2021** | Challenge call preparation:<br>• Identification of potential evaluation campaigns organisers<br>• Refinement of Challenge description<br>✓ Main and secondary objectives<br>✓ Evaluation scenarios, metrics, development, training and test data preparation | - |
| **1st semester 2021** | Call documents preparation | - |
| **May/June 2021** | Conference 2021<br>• Refinement of Challenge description with the scientific community | - |
| **October 2021** | Call launch | - |

chist-era

## Description

In a Challenge, a well-defined technology target is considered and a set of science-to-technology approaches to reach this target (proposed by the Challenge call applicants) is selected to join a competition made of regular evaluation campaigns, often of increasing difficulty. The objective of the campaigns is to compare these approaches.




The launch of a Challenge is motivated by a research management strategy and possibly by the expression of end-user need:

1. Research managers rely on this format because of its recognised capacity for specific scientific fields, in particular in knowledge processing fields, to accelerate research and structuration of the respective research community around high quality research standards. In this context, the technology target evolves to become more challenging as science progress.
2. The end-user observes the maturity of the scientific landscape for a new technology to emerge of direct interest for them. The Challenge aim is to accelerate the development of various technical options and to identify the most promising one(s) via the organisation of evaluation campaigns.

A Challenge may require a physical infrastructure or it can be organised online. It can be closed to the selected projects or open to any volunteering team if their participation generates a reasonable cost for the organiser. There are no strong *a priori* constraints on the technology readiness levels concerned. A Challenge can address low TRLs and pursue more applied/finalised research objectives at the end of the Challenge. This variety of configurations was highlighted during the November 2020 workshop by the NIST (US) presenter Mark Przybocki.

Actually, the distinctive feature of a Challenge is the built-in technology evaluation component, which expresses itself in the form of evaluation campaigns. Their role is essential and goes beyond standard project follow-up. They are entangled with the work of the competing projects as described in the following section.

## Evaluation campaigns

The funding of the selected projects goes hand in hand with the creation, development or reuse of a dedicated measuring instrument. This instrument takes the form of evaluation campaigns involving:

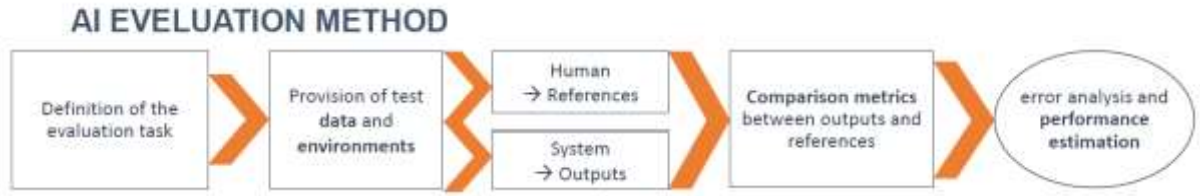
- A scenario, reflecting the typical situation of the Challenge[1];
- Metrics and protocols to compare the performance of the different projects against the scenario;
- Data sets allowing the projects to develop their system (their answer to the Challenge) and train it in advance of the competition and allowing the evaluation campaigns organisers to evaluate them during the competition.

The typical cycle of a Challenge structured around a yearly evaluation campaign is presented in the table below.

| Stage | Duration | Evaluation campaign organisers | Competing projects |
|---|---|---|---|
| Preparation | ~10 months | • Refinement of competition scenario, metrics, protocols and data sets<br>✓ In collaboration with projects<br>✓ Competition final rules known sufficiently in advance by the projects<br>• Logistics of upcoming evaluation campaign | • Development of system<br>• Interaction with organisers |
| Competition | ~1 week | • Projects scoring against evaluation scenario<br>• Organisation of project individual interviews | • Project system evaluated against campaign scenario<br>• Interview with organisers<br>• Interaction with other projects |
| Debriefing | ~2 months | • Analysis of evaluation campaign outputs<br>• Feedback to projects<br>• Campaign outputs will serve as input for designing of next campaign<br>✓ In collaboration with projects | • Feedback of organisers will serve to improve their device/system<br>• Interaction with organisers |

[1] E.g. in the Challenge MALIN, whose objective is to develop systems for indoor positioning, a path is defined that projects, one by one, must follow. This path is representative of the environment that a firefighter in a building in fire, for example, might encounter. The measured positions are then compared to the expected ones to score the projects.

Another illustration of this cycle is provided in the presentation of Guillaume Avrin at the November 2020 workshop in his presentation of LNE:




## Why is it needed?

Compared to the usual follow-up of funded projects, this may seem complex. Is it really worth? Why one may need such a measuring instrument? Does it concern all scientific fields?

The rationale of the Challenge format is detailed in the publication An Economic View on Human Language Technology Evaluation (Edouard Geoffrois, 2008, LREC 2008 Proceedings). In summary: Evaluation campaigns and their organisation by an independent third party provide a relevant instrument for measuring the performance of knowledge processing related research projects in a reproducible and unbiased way.

**Insight on Challenge Format Rationale**

In most other scientific domains, the availability of precalibrated measuring instruments relying on benchmarks and units defined by physical phenomena provides direct means to measure and compare the performance of systems, objectively and in a reproducible way. However, in knowledge processing fields, which form an important subset of ICT, similar experimental set up do not exist in this form due to the following characteristics (let us consider translation):

I. The translation system involves learning: The test can be used to improve the system capability to succeed to the same test. Therefore, to avoid any bias, the test set of texts to translate should not be known in advance.
II. The judge is human: Since the technology is about reproducing human capabilities, testing it requires manual work to judge it. The evaluation tool cannot be fully automatised.
III. The set of texts to translate is infinite: But for practical reason, the translation test can concern a small sample of texts only. Therefore, for the measure to be meaningful (reproducible), the test data set must be published as well.

Combining characteristics I and II imply that an independent third party should be responsible for the evaluation organisation. In addition, combining I and III leads to the requirement to test the competing translation systems at the same time.

An analogy can be drawn with students taking an exam to illustrate with a well-known daily configuration what an evaluation campaign is. In this analogy, the student correspond to the intelligent system, the exam to a representation of the teaching target (the Challenge) and the teacher acts as trusted third party.

Another reason why CHIST-ERA would consider the Challenge format is that it works. The talk of Mark Liberman at the Open Science workshop organised in 2019 reminds as an example the history of the progress made in Human Language Technology and how the 'Common Task' method (another wording for Challenge) has unlocked years of stagnation.

Finally, it happens that in spite of the scientific interest of the method, its constraints (cost to develop and maintain data sets, need to involve an independent third party), make it difficult to integrate the

Challenge format as part of the daily research without specific support. This calls for a dedicated support from the research funders to fill this gap.

## Benefits and limitations

By promoting objectively evaluable research in knowledge processing fields, the main benefit of the Challenge format is to promote progress with an unprecedented efficiency:

- Science and technology targets are made explicit and shared by the community.
- It facilitates risk taking at the projects selection stage: The Challenge will allow evaluate their innovative and risky technology approach anyway.
- It eases comparison between competing technology approaches and technology transfer.
- Challenges also foster exchanges between the researchers: The competition involves a lot of emulation and cross-project sharing of know-how

Moreover, Challenges have broader scientific impacts beyond their duration at the level of the research community, by promoting:

- Reproducibility
- Open Access to methodologies, reference and possibly large data sets.

A limitation of the format may lie in the focused objective of the Challenge. If not relevant, it may waste resources of several research teams. However, there are ways to avoid that pitfall. When the Challenge is motivated by science, its target should be defined consistently with the state of the art. When the Challenge is motivated by end-user need, they often invest much time to assess the opportunity of launching the Challenge.

chist-era

# Challenges and CHIST-ERA

The proposal of launching a Challenge call became more concrete in 2018 at the occasion of the CHIST-ERA IV ERA-NET proposal writing. However, and this is not surprising given the ICT scope of CHIST-ERA, the concept behind is floating in CHIST-ERA since a longer time: From the perspective of quantitative ICT evaluation or in Open Science related discussions.

Already in the Call 2012 (topic Intelligent User Interfaces), applicants were informed that:

- "*In all cases, projects should address the question of measuring progress toward the foreseen applications and the proposals should provide a detailed description of how ideas and systems will be experimentally tested (evaluation data, metrics and protocols). Projects are encouraged to include the means for objective, significant and reproducible experiments when these are not already available elsewhere.*"
- In addition, they are expected to *"make the outcomes of the research, including data sets and test protocols, available to other researchers and to industry […]."*

In other words, the projects were invited to take on board as far as possible the principles of technology evaluation associated to the Challenge idea. The objective was to promote via the calls of CHIST-ERA an evaluation culture in ICT research:

- In parallel to setting the project objectives, applicants should aim at establishing a reliable methodology to evaluate whether they are approached.
- In addition, they should aim at opening access to their methodology.

These are key ingredients of the Challenge full concept. Missing elements were the competition and the thorough cross-project exchanges and emulation it creates, and the involvement of a trusted third party for managing the evaluation with the required expertise and able to guarantee the evaluation independence.

With regard to the latter, the Open Science session of the Projects Seminar 2019 in its Open Science workshop aimed at introducing to the funded projects the option to partnership with institutions specialised in technology evaluation (see here for instance presentation of LNE at the November 2020 workshop). More generally, the objective of the workshop was to highlight the instrumental role such institutions have to support open evaluation methodology and experiment reproducibility in ICT research and their need of public support.

The impact of the action of CHIST-ERA towards opening evaluation methodology and fostering experiment reproducibility is visible in the work programme of the funded projects. At the Projects Seminar 2019 in its Open Session, three of them were put forward:

- ALLIES (Call 2016, Lifelong Learning for Intelligent Systems)
  One of the tasks of ALLIES was to *"develop, evaluate and disseminate […] metrics and protocols. They will be available to European actors via an open evaluation platform dedicated to reproducible research. An evaluation campaign and a workshop will be organised to engage the community on this path"*.
  Their evaluation campaign is presented here.
  LNE is a partner of the project.
  They have also proposed a topic for the CHIST-ERA Challenge: *People identification in videos with on-the-job learning*.

- CORSMAL (Call 2017, Object Recognition and Manipulation by Robots: Data Sharing and Experiment Reproducibility)
  One of the goals of CORSMAL was to *"evaluate the robustness of the proposed framework with prototype implementations in different environments. Importantly, during the project we will organise two community challenges to favour data sharing and support experiment reproducibility in additional sites".*
  The first Challenge organised by CORSMAL is presented here.
  The coordinator of the project is a member of the Open Science Advisory Board of CHIST-ERA.
- Cocoon (Call 2015, User-Centric Security, Privacy and Trust in the Internet of Things)
  The coordinator of the project Etienne Roesch presented the UK Reproducibility Network in which he is involved and whose mission is to *"understand the factors that contribute to poor research reproducibility and replicability, and develop approaches to counter these, in order to improve the trustworthiness and quality of research"*.

To end this review of past CHIST-ERA activities related to the idea of a Challenge, one should mention the following two call topics:

- Call 2017 - Object Recognition and Manipulation by Robots: Data Sharing and Experiment Reproducibility
- Call 2014 - Human Language Understanding: Grounding Language Learning

They belong to fields accustomed to the Challenge format, robotics and human language technology, and the emphasis of the call on the requirement to measure progress made adequately and openly is even strengthened (up to a mention in the topic title for the robotics topic). In accordance with this spirit of developing and opening evaluation methodologies (incl. access to the necessary data sets), the projects of the Call 2014 topic spontaneously organised joint workshops. In addition, the projects of the Call 2017 topic planned to do so as well (to be confirmed due to the complexity induced by the COVID-19 context).

In conclusion, with regard to CHIST-ERA, a Challenge would represent:

- A step further in the support of CHIST-ERA to open evaluation methodology component of Open Science and experiment reproducibility;
- A continuity with the influence of CHIST-ERA in that direction since almost a decade;
- A rare funding opportunity for the research community given the still limited level of public support in spite of the scientific relevance and the researchers' demands.

A Challenge would strengthen CHIST-ERA in its role as pathfinder; with this time not only a future or emerging technology in view, but also an emerging research method of significant efficiency for many ICT domains. The emerging nature is even more pronounced from the funders' perspective, who have invested limited resources so far in comparison with the demands of the researchers.

# Recommendations

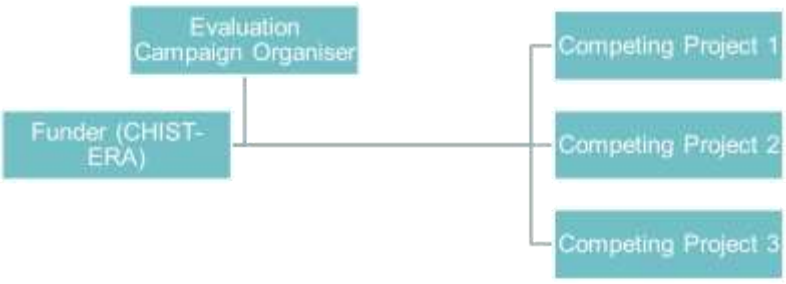
## Keep It Simple!

While CHIST-ERA is not new to the Challenge idea, organising such a competition represents an additional workload and requires further attention to the methodology to define the Challenge scope. In that respect, the recommendation of the Scientific Advisory Board in its July 2020 meeting was to select a topic for which the community is already well structured and is quite used to the Challenge functioning: Minimum readiness of metrics and evaluation protocols, of data sets.

For this purpose, it was envisaged that the selected topic would build on past CHIST-ERA call topics.

## Connect Early with Evaluation Campaigns Organiser: Squaring the Circle

In a Challenge, in addition to the selected projects and the funders, there is a third stakeholder, the evaluation campaigns organiser.




Here lies the main challenge of the Challenge from the point of view of the funders. The goal is first to succeed in assembling two diverging axis:

- The involvement of an independent third party is essential for the proper implementation of a Challenge.
- Funding this third party is not, in general, perceived as essential by the funders, because this party does not perform standard research and development. Rather they provide an infrastructure and the expertise to operate it in order to support high quality research.

Secondly, the goal is to connect early with the evaluation campaigns organiser because the definition of the evaluation campaigns goes along with the definition of the call to select the projects.

In France the required expertise is at LNE. Other European organisations are quoted in the presentation of LNE at the November 2020 workshop (see list below). However the dedicated public support that LNE represent is quite unique in Europe. While in the United States, the NIST is a well established institution cooperating with DARPA for the organisation of many challenges, similar public support in Europe is less common and technolgy evaluation in ICT research is shared mainly across the research performers on a bottom-up basis rather than the subject of a specific public support. This landscape is slightly evolving, and with the Horizon 2020 project METRICS, LNE is now coordinating a large consortium reflecting that the question is gaining momentum.

This situation raises two comments for CHIST-ERA:

1. With at minimum the availability of LNE, the organisation of the evaluation campaigns of the CHIST-ERA Challenge is possible.
2. The support of CHIST-ERA to these technology evaluation activities corresponds to a rare opportuniy proposed to the researchers.

Depending on the selected topic the lead organiser may need to associate with other partner(s) to combine both expertise in campaigns orgnaisation and scientific expertise related to the topic. Similarly to the selected projects which will involve at least 3 partners from 3 countries, the organiser could form a partnership. This would contribute to spreading in Europe skills to organise evaluation campaigns.

CHIST-ERA funders willing to join the call should assess to which extent they wish to contribute to the call: Via the fundind of the selected projects only or by funding the organiser as well.

## A CHALLENGE NEEDS TRUSTWORTHY ORGANIZERS

Organizers:

- Competent in organizing such challenges
- Expert in AI evaluation

| Country coordinating | Challenge |
|---|---|
| Estonia | Robotex |
| France | Repere, Quaero, Rose, Metrics, Allies, Etiseo, Argos |
| Germany | Elrob |
| Italy | EuRoC, Promise (CLEF) |
| Netherlands | MediaEval |
| Portugal | RoCKIn |
| Spain | Albayzín evaluation |
| United-Kingdom | euRathlon |

And many more, in Europe and abroad (cf. **NIST** presentation)

Recommendations to CHIST-ERA:

- Each funder can seize the opportunity to strengthen the support to technology evaluation in ICT research by joining ANR to fund the organiser. Volunteers should identify potential organisers in their country.
- Once the topic is selected, start the identification of the organiser (single partner or consortium) and discuss the Challenge definition.

# Conclusion

Open ICT evaluation methodology and experiment reproducibility in knowledge processing domains is at stake behind the Challenge concept. Its organisation shaped as evaluation campaigns is theoretically grounded and it has demonstrated its efficiency experimentally. While this research management method is gaining momentum across the ICT research communities, spreading from experienced fields to new ones, dedicated public support is still nascent in Europe.

Since 2012, CHIST-ERA promotes this evaluation culture to support high quality research via its calls and Open Science activities, up to some projects that propose Challenges to their respective research communities (e.g. ALLIES and CORSMAL). A Challenge call of CHIST-ERA would represent an important step forward aimed at closing the loop. The existence of LNE guarantees the feasibility of a Challenge. Moreover, with the current limited level of public support, it would provide CHIST-ERA with a valuable contribution to answer the researchers' demands, fully in line with its ICT scope and its ambition to act as pathfinder.

To succeed, CHIST-ERA needs to define properly the Challenge in coordination with the evaluation campaigns organiser(s) as early as possible, which involves identifying potential candidates and clarifying what funding structure will be proposed to them.