Technical Background Notes for Horizon 2020 Objective ICT-17-2016-2017 Big Data PPP: Support, Industrial Skills, Benchmarking and Evaluation

DG CONNECT/G3

CNECT-G3@ec.europa.eu

http://ec.europa.eu/digital-agenda/en/content-and-media/data

http://ec.europa.eu/digital-agenda/en/language-technologies-and-big-data

This document is intended to provide background information and technical commentary on Topic ICT-17 2016 published as part of the 2016-17 Horizon 2020 work programme:

http://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-leit-ict_en.pdf

The official work programme text is the only legally binding source of information on the topic. Should any inconsistency between the present explanatory document and the official text be detected it is always to be resolved in favour of the work programme text.

Motivation of the Topic and Scope of this Document

The official text of the work programme specifies that proposals under this objective should be for part a) **Coordination and Support Actions (CSA)**, defined as:

An action consisting primarily of accompanying measures such as standardisation, dissemination, awareness raising and communication, networking, coordination or support services, policy dialogues and mutual learning exercises and studies, including design studies for new infrastructure and may also include complementary activities of networking and coordination between programmes in different countries.[1]

and for part b) **Research and Innovation Actions (RIA)**, defined as:

Action primarily consisting of activities aiming to establish new knowledge and/or to explore the feasibility of a new or improved technology, product, process, service or solution. For this purpose they may include basic and applied research, technology development and integration, testing and validation on a small-scale prototype in a laboratory or simulated environment. Projects may contain closely connected but limited demonstration or pilot activities aiming to show technical feasibility in a near to operational environment.[2]

[1] http://ec.europa.eu/research/participants/portal/desktop/en/support/reference_terms.html
[2] Same as previous footnote.

This means that research (and publication) activities are appropriate but should **not** be the **only** focus of proposals under this objective.

The topic addresses the following challenges:

Specific Challenge: The newly created Big Data Value contractual public-private partnership (cPPP) needs strong operational support for community outreach, coordination and consolidation, as well as widely recognised benchmarks and performance evaluation schemes to avoid fragmentation or overlaps, and to allow measuring progress in (Big) Data challenges by solid methodology, especially in emerging areas where the significance of Big Data is rapidly increasing. Also, there is an urgent need to improve the education, professional training and career dynamics (including addressing the existing gender gaps in ICT) so that the profiles of data professionals better respond to the rapidly evolving needs of data intensive industry sectors.

The objective is articulated into two specific types of activities: **coordination, support and skills** and **Big Data technologies benchmarking**.

a) Coordination, Support and Skills (CSA)

Under this part of the objective, proposals are solicited to fund a **single** Coordination and Support Action that is charged with two tasks:

- *support the community building, the administration and governance of the cPPP, in close collaboration with the cPPP governance bodies; facilitate discussion on relevant topics such as the framework conditions of the data economy; organise events and contribute to synergies and coordination between the actors and stakeholders of the cPPP and beyond;*

- *liaise with and build on related actions and support the establishment of national centres of excellence in all Member states, and exchange knowledge on the universities' data scientist programmes across all Member States; to align curricula and training programmes to industry needs; to stimulate and promote (among the organisations participating in the Big Data PPP actions) exchanges of students, confirmed data professionals and domain experts that would acquire data skills and let them work on a specific Big Data challenge/project in a company or a research centre/university in another Member State.*

The first task must be understood with reference to the agreement of October 2014[3] with which the European Commission and the Big Data Value Association[4] agree to enter in a Public Private Partnership (PPP) with the objective of promoting the use of Big Data technologies in Europe so as to make European organisations more efficient and European companies more competitive.

[3] http://europa.eu/rapid/press-release_IP-14-1129_en.htm
[4] http://www.bdva.eu/

What is expected from proposals in addition to plans for supporting the administration and governance of the PPP, are plans for raising the interest and obtaining the support of European companies in those sectors that can most be expected to benefit from the introduction and systematic use of Big Data technology. The credibility of such plans will depend on a systematic analysis of the global competitive landscape in which European companies operate and their specific strengths or opportunities.

The second task requires building on existing initiatives such as

1. European Data Science Academy: http://edsa-project.eu
2. European Institute of Technology Digital: http://eit.europa.eu/eit-community/eit-digital
3. Marie Sklodowska-Curie Actions: http://ec.europa.eu/programmes/horizon2020/en/h2020-section/marie-sklodowska-curie-actions

in order to make sure that European Big Data jobs will be filled by professionals with the desired set of skills. A first aspect of this concerns the development and harmonization of academic curricula, a second, equally important aspect, will be the movement of people, for example:

1. Data professionals interested in 'going back to school' for short courses on emerging ideas, techniques and technologies
2. Students of data technologies find short term internships in companies where they can see the practical applications of what they are studying, in the context of the actual operations of a commercial entities

Plans for the practical professional development of data scientists and engineers will have to promote the circulation of such individuals across Europe and will be credible to the extent that they can credibly obtain the support of European companies or incubators.

The impacts expected from the activities of the Coordination and Support Action are:

- *At least 10 major sectors and major domains supported by Big Data technologies and applications developed in the PPP;*
- *50% annual increase in the number of organisations that participate actively in the PPP;*
- *Significant involvement of SMEs and web entrepreneurs to the PPP;*
- *Constant increase in the number of data professionals in different sectors, domains and various operational functions within businesses;*

- *Networking of national centres of excellence and the industry, contributing to industrially valid training programs.*

Some of these objectives can be easily tracked from the internal bookkeeping of the PPP (e.g. membership increase), others must be monitored using data from the wider European economy. Proposals are strongly encouraged to describe what indicators will be appropriate in order to track this second kind of objectives.

b) Big Data Technologies Benchmarking (RIA)

For this part of the objective, the work programme specifies that:

The benchmarking action will identify specific data management and analytics technologies of European significance, define benchmarks and organise evaluations that allow following their certifiable progress on performance parameters (including energy efficiency) of industrial significance. The benchmarking and evaluation schemes will liaise closely with data experimentation/integration (ICT-14) and Large Scale Pilot (ICT-15) projects to reach out to key industrial communities, to ensure that benchmarking responds to their real needs and problems, and to provide a basis for measuring the success of the PPP. The "European significance" of industry/technology sectors should be determined and documented by objective criteria such as turnover, world-wide market share and growth rates of the European companies who provide or use such technologies. When real datasets cannot be made available for benchmarking, synthetic datasets will be acceptable, provided that they are produced by models that certifiably produce data distributions approximating real datasets in all respects that are industrially relevant. The action shall address areas of activity that do not yet have a benchmarking/evaluation scheme.

The first important point to note is that the benchmarks proposed must be of industrial relevance, i.e. relevant for European developers and providers of data technologies. The main goal of the benchmarking activities is to give European developers the means to continuously improve their performance (and thus their competitiveness) as measured against benchmarks that their customers acknowledge as unbiased and representative of realistic business conditions. This implies that, instead of starting from a technology of interest (initially to scientists and software developers) and then defining benchmark tasks that could in the abstract be presumed to be of industrial interest, consortia are rather advised to identify from the very beginning industrial actors that have expressed interest in the technology for very specific business reasons (to be appropriately documented in the proposal in terms of business plans, market projections, etc…), and involve them (not necessarily as members of a consortium) in the definition of the benchmarks and performance goals.

When the consortium opts for administering benchmarks against synthetic datasets, it is important that they should specify a methodology that guarantees that the descriptive statistics of the generated dataset will approximate, in the relevant business/operations respects, those of real datasets. For examples of existing efforts in this direction, see

1. http://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70+&imm_mid=0d16f4&cmp=em-data-na-na-newsltr_20150506

2. http://www.paralleldatageneration.org

3. http://ldbcouncil.org/developer/snb

The second point to note is that benchmarks are invited for those technologies for which they don't already exist. So, to exemplify, a proposal to develop a benchmark for graph databases would **not** be welcome, given such a benchmark has already been developed in the context of previous EU funding activities, notably http://ldbcouncil.org .

Finally, in order to be effective, technology benchmark initiatives must give reasonable guarantees that they will continue to exist throughout the entire life-cycle of the relevant technology. Credible proposals are thus expected to define and follow a process that would lead to structures that will continue to refine, extend and administer the benchmarks past the end of the Horizon 2020 grant agreement. In order to create such sustainable organisations, once again, it is imperative that consortia develop and follow in earnest credible plans to secure industrial involvement. Since the resulting benchmarks are expected to help European technology providers to become globally competitive, consortia should have a global plan to involve potential clients in the definition of the benchmarks.

The impacts expected from this part of the objective are:

- *Availability of solid, relevant, consistent and comparable metrics for measuring progress in Big Data processing and analytics performance;*
- *Availability of metrics for measuring the quality, diversity and value of data assets;*
- *Sustainable and globally supported and recognized Big Data benchmarks of industrial significance.*

€2M are available in total for benchmarking proposals. This means that the selection process may decide to recommend for funding a single proposal worth the entire €2M or, in alternative, several distinct proposals collectively asking for €2M in funding. The selection decision will be based on the merits of the proposals and considerations of strategic coverage and complementarity, with no pre-conceived preference for small vs large proposals.

Appendix: a list of questions that proposals must answer in order to be in scope of objective ICT-17 2016-17 of Horizon 2020

This appendix contains a list of simple questions that a consortium should ask about the proposal to be submitted. If the proposal as submitted does not contain a clear answer to the majority of the relevant questions it places itself at a serious disadvantage in a very competitive selection process (because the evaluators will be specifically instructed to look for the answers to these and other questions).

a) Coordination, Support and Skills (CSA)

1. What is your plan for making the Big Data Value Public Private Partnership (PPP) attractive for European companies that are not already part of it?
    a. How well do you understand the competitive position, business plans, research and development investment plans of those companies?
    b. What arguments for joining the PPP will you develop that are attractive to a European company's CEO, as well as to its researchers?
    c. What evidence do you need to make such a case and how will you gather it?
2. Do you have a **strategic** plan for the expansion of the PPP?
    a. Are any ten new European companies equally likely to join the PPP or are there companies that can act as catalysts (in the sense that their presence in the PPP is more likely to attract new members than the presence of other companies)?
    b. How will you identify such catalysts and prioritise your outreach events and networking activities based on that premise?
3. Skills: how do you liaise with existing initiatives in skills building (e.g. those mentioned in the text above)?
4. Skills: what innovative exchange schemes are you suggesting, which existing skills gaps do they address, and how do you ensure that they will become popular among students, companies and research centres?

b) Big Data Technologies Benchmarking (RIA)

1. Can you provide a precise definition of the Big Data technology for which you intend to develop a benchmark?
2. Why does this specific technology deserve to have a benchmark developed?
    a. Which industry sector or what type of operations depends on this technology?
    b. Do you have evidence that European companies in the relevant sectors are interested in the development of such benchmarks?
3. What is your plan to ensure the sustainability of the benchmarking activities past the end of Horizon 2020 funding?