Technical background notes for Horizon 2020, Topic ICT-16 2015 "Big Data: Research"

DG CONNECT/G3
CNECT-G3@ec.europa.eu

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

http://cordis.europa.eu/fp7/ict/content-knowledge/home_en.html

This document is intended to provide background information and technical commentary on Topic ICT-16 2015 of the Horizon 2020 programme. The official text of the topic (including the timeline and procedures for applications) has been published as part of the 2014-15 Horizon 2020 work programme.

http://ec.europa.eu/programmes/horizon2020/en/h2020-section/information-and-communication-technologies,

http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-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 states that this topic addresses the following challenges:

"The activities supported within LEIT under this topic contribute to the Big Data challenge by addressing the fundamental research problems related to the scalability and responsiveness of analytics capabilities (such as privacy-aware machine learning, language understanding, data mining and visualization). Special focus is on industry-validated, user-defined challenges like predictions, and rigorous processes for monitoring and measurement..." (page 38)

The motivation behind this topic is to develop technologies that would increase the efficiency of all EU companies and organisations that need to manage vast and complex amounts of data and in particular the competitiveness of EU enterprises. The emphasis is on **rigorously measured** increases in performance in data processing at very large scale.

To quote again from the official text of the topic, if this is accomplished, the expected impact is:

- *Ability to track publicly and quantitatively progress in the performance and optimization of very large scale data analytics technologies in a European ecosystem consisting of hundreds of companies; the ability to track this progress is crucial for industrial planning and strategy development.*
- *Advanced real-time and predictive data analytics technologies thoroughly validated by means of rigorous experiments testing their scalability, accuracy and feasibility and ready to be turned over to thousands of innovators and large scale system developers.*
- *Demonstrated ability of developed technologies to keep abreast of growth in data volumes and variety by validation experiments.*
- *Demonstration of the technological and value-generation potential of the European Open Data documenting improvements in the market position and job creations of hundreds of European data intensive companies.*

This will require addressing at least all the following aspects:

- **usability**: the systems developed should be engineered so as to be usable by people whose primary occupation is not research or software development but rather running a business or an organisation. It is thus extremely important that usability be taken as a foremost concern at every step of software development and validation.
- **innovation**: the systems developed must be able to improve the processes of existing businesses or open new business opportunities. For this reason, it is extremely important that commercial companies with clearly defined use cases be at the core of consortia. Consortia should not be driven by researchers and technology developers adding commercial partners as use cases for technologies that are not originally motivated by actual working conditions. This requires also that the commercial partner(s) have in place both the commitment to open their data assets for experimentation to developers in the consortium and that they have in place a corporate process that guarantees confidentiality of commercially sensitive data and personal data when applicable.
- **robustness**: the systems developed must be designed under the assumption that the data handled may be missing, corrupted or inconsistent; furthermore, they must be engineered to be able to be deployed outside the lab in typical operating conditions (including memory, storage and network failures).
- **accountability**: the systems developed will need to specify the performance parameters they are trying to meet. They will need to give convincing evidence that such targets are meaningful when compared to documented industry trends (i.e. that the systems proposed, when finally delivered, are very unlikely to be outperformed by best-in-class industrial solutions produced elsewhere in the world). They will need to specify a credible plan for testing, ideally one in which the system proposed is being made to compete with alternative solutions developed elsewhere (outside the project).
- **privacy:** all systems designed to collect, manage and analyse personal data will need to provide a detailed explanation of how they support data protection regulations.

Notice that the topic is articulated in two distinct types of actions (Research and Innovation Actions on one side and Support Actions on the other) so clearly distinct from one another that each proposal is expected to address exactly one of them. In general, it is important to understand that proposal evaluation will be based strictly on excellence, impact and quality and efficiency of the implementation. Please see the appendix to this document for a list of points that must be addressed to avoid submitting mediocre proposals.

a) Research and Innovation Actions

Research and Innovation Actions are expected to address one or both of the following activities: research/innovation and benchmarks.

Research and Innovation actions are meant to cover:

Collaborative projects to develop novel data structures, algorithms, methodology, software architectures, optimisation methodologies and language understanding technologies for carrying out data analytics, data quality assessment and improvement, prediction and visualization tasks at extremely large scale and with diverse structured and unstructured data. Of specific interest is the real time cross-stream analysis of very large numbers of diverse, and, where appropriate, multilingual, multimodal data streams. The availability for testing and validation purposes of extremely large and realistically complex European data sets and/or streams is a strict requirement for participation as is the availability of appropriate populations of experimental subjects for human factors testing in the domain of usability and effectiveness of visualizations. Explicit experimental protocols and analyses of statistical power are required in the description of usability validation experiments for the systems proposed. Proposals are expected, where appropriate, to make best possible use of large volumes of diverse open data from the European Union Open Data portal and/or other European open data sources, including data coming from EU initiatives like Copernicus and Galileo.

The text of the call stresses two requirements that are in keeping with the Horizon 2020 new orientation towards industrial competitiveness and innovation.

The first requirement is the availability of *extremely large and realistically complex European data sets*. What this means is that a strong proposal would typically include at least one **European** industrial/commercial partner that:

1. produces or harvests as part of its core business extremely large data sets;
2. can provide a precise definition of one or more data-related business problem(s) or opportunity(ies) that cannot be addressed with existing technology;
3. can provide a credible projection of the performance parameters required to address that problem/opportunity in the medium term (i.e. around 2020, when the project to be funded will have come to its conclusion);

4. has the ability to bring its data assets in an environment in which the other members of the consortium can freely experiment. This could for example be its own internal computing environment (to which consortium researchers are given efficient access, i.e. the ability to perform rapid cycle experiments and validation) or a shared cloud computing environment

5. has, where applicable, a well-defined corporate process that protects the commercial confidentiality of its data assets as they are made available for researchers to experiment. It similarly has a well-defined process to protect personal data of the individuals to which they refer.

Note the emphasis on the fact that the partner(s) providing the realistically-sized data assets should be **European**. This is a direct consequence of the fact that the Horizon 2020 programme is designed to improve the competitiveness of the European economy, an objective confirmed by the orientation of the incoming Commission. In this context, "European" means "owned by a legal entity established in the European Research Area[1]".

In the past (in the late Framework Programme 7 calls, for example), it was not uncommon for proposals to outline a research plan and explain that the results obtained would be tested against data obtained through the APIs of, for example, non-European large social networks or search engines.

This is **not** be encouraged in H2020 topic ICT-16, and is likely to lead to low scoring.

What, on the contrary, **will** be encouraged is the use of non-European data assets to complement European ones. For example, a proposal could develop and test a framework using the data assets form a European partner and then, additionally, cross-check its results against data assets from non-European sources. Or a proposal could build a data analytics system that **provably** delivers value when used on the data assets of a European partner, with the option of delivering additional value when the European data assets are cross-analysed with non-European ones.

The point of this requirement is to support the general innovation and knowledge transfer goals of the Horizon 2020 programme by requiring that the solutions proposed by the projects selected for funding be developed within the concrete context of European business or societal challenges.

In this context, it is important to note ongoing work towards a European Public Private Partnership (PPP) devoted to the development of data technologies. The roadmap[2] produced as part of this work provides a very comprehensive (and prioritised) overview of what kind of challenges and opportunities the European industry recognises in this domain.

[1] http://ec.europa.eu/research/era/index_en.htm

[2] http://bigdatavalue.eu/index.php/downloads/viewdownload/3-big-data-value/14-big-data-value-strategic-research-and-innovation-agenda

In the preparation of their proposals for topic ICT-16, consortia will benefit from studying the roadmap in question by getting inspiration for challenges and opportunities identified by European industry.

Another point emphasised in the text of the call is that the data assets available for consortia to experiment on should be **extremely large and realistically complex**. This often prompts questions as to what exactly is meant with "extremely large". To these questions, one can answer in three stages:

1. ICT-16, like all topics in Horizon 2020, is a **competitive** call for proposals. In the context of scalability of data processing, what this means is that if two proposals are equally good on all other respects, the evaluation process will favour the one that comes with the larger **European** data assets. In this context, "extremely large" means "larger than what is made available by the next proposal that is otherwise as good as yours".

2. ICT-16 is a call for Research and Innovation activities. This means that consortia must demonstrate their innovative work on **European** data assets of a volume, velocity and/or variety that **cannot** be handled with current technologies. In this context, "extremely large" means "so large that today no amount of money could buy a system capable of handling it".

3. Projects selected for funding under ICT-16 will start in early 2016 and typically come to an end in late 2018. A defensible way to estimate what volumes of data **European** companies will need to handle in 2018 could be to observe what the volumes were in 2010 and work out what the rate of growth has been until 2014 (the previous four-year period). And even this could be an under-estimate, given that parallel developments in areas such as smart cities, smart grids, Internet of Things, Earth observations, transportation systems, in-car sensors will greatly add to what had been sources of data in the 2010-2014 period. Another useful rule of thumb is to observe the current level of data processing performance at those companies that are today global leaders in big data technologies. One could conjecture that those levels of performance will be industry-wide expectations by 2018[3]. In this context, "extremely large" means "as

[3] A few examples. As of September 2014, MapR can ingest 100M data points per second on a four node Hadoop cluster: http://www.datanami.com/2014/09/09/mapr-reports-accelerated-opentsdb-performance .As of September 2014, Google's Sybil machine learning framework can handle training sets of 200B examples with up to 100 features : http://www.kdnuggets.com/2014/08/sibyl-google-system-large-scale-machine-learning.html . As of June 2014, DataTorrent can process 1.5B events per second on its 34 node Hadoop cluster: http://www.datanami.com/2014/06/03/datatorrent-rts-clocks-1-5b-events-per-second/ . As of June 2014, Google's web index is 100PB. Its BigTable system serves more than 2 Exabytes at 600M queries per second: http://lintool.github.io/my-data-is-bigger-than-your-data/ . As of April 2014, Facebook stores 300PB of data, with 600TB added daily: https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/ . As of September 2014, certain forms of feature selection in machine learning can be performed in matters of minutes on sparse datasets with 1-billion dimensions and 1-billion nonzero features:

large as you would expect the largest 10% of data assets to be in 2018, based on current growth trends", however, keeping in mind that complexity/variety is also an element to be taken into account, not sheer volume.

Proposal evaluators will be specifically instructed to look for (and comment explicitly on) the availability of extremely large datasets (or lack thereof).

The European Commission intends to contribute in making available "realistically complex" datasets for consortia to use in their experiments. Consortia are in particular encouraged to reuse datasets made available by EU institutions. Examples include data from the EU's Open Data portal[4] and the Copernicus earth observation programme[5]. This is done with the expectation that large numbers of experts systematically (re)using such data resources will extract from them insights of value for the European Union and set in motion a positive feedback mechanism that will allow for the quality and coverage of those sources to improve constantly over time.

In addition to measuring improvements on technical performance parameters such as speed, scalability and robustness, proposals will also need to demonstrate how the software frameworks or components they intend to develop would work in the hands of their intended users. These will typically be subject matter experts in various industrial domains or in organisations. For example, if a software stack is being proposed to improve the effectiveness of workers on a given task, it is important that the performance of workers in those tasks be appropriately measured.

This means in particular that:

1. Proposals should explain their plans to recruit as testers an appropriate number of typical users from the relevant activity domains. To exemplify, a system built by software engineers to make traffic management more efficient should be tested by traffic managers (whose skills and professional experience are concentrated on the management of traffic) and not by software engineers (whose skills are to build and operate complex software systems). Also, a system built to assist high level decision makers in companies from a certain sector should be tested by high level decision makers of companies in that sector and not by entry level staff or summer interns who cannot be presumed to follow the same decision patterns.

2. Proposals should clearly specify the experimental protocol (definition of variables, experimental design, number of subjects, …) they intend to follow in order to test their hypothesis that the system(s) they intend to build will have a positive impact in one or

http://arxiv.org/abs/1409.7794 . As of October 2014, triangular meshes can be used to efficiently classify in geographic regions billions of data points: http://arxiv.org/abs/1410.0709 .

[4] https://open-data.europa.eu/en/data/

[5] http://www.copernicus.eu/pages-principales/infrastructure/data-access/

more given domain of activity. Instead of generically stating that they will test the system with real users, proposals will need to specify the dependent and independent variables of the experiments to be conducted. In addition, given the design of the experiment(s) envisaged, proposals will need to present a statistical power analysis[6] in order to determine the number of experimental subjects required by the experiment and a proof that the consortium will have access to the required number of subjects on a schedule consistent with the work plan presented.

Finally, **benchmarks** are expected to cover:

Collaborative projects to define relevant benchmarks in domains of industrial relevance, assemble the data resources and infrastructure necessary for administering and validating the benchmarks and organise evaluation campaigns with a commitment to producing public reports on the performance of participants against the defined benchmarks. Since the goal is to create big data analysis and prediction benchmarking environments of sufficient general usefulness to be able to become self-sustaining after the end of funding, proposals will have to provide detailed and convincing exit strategies.

There are two important points to stress here.

The first one 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, and involve them (not necessarily as members of a consortium) in the definition of the benchmarks and performance goals.

The second point is that collaborative projects funded to define technology benchmarks and administer tests against the same are 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.

b) Support Actions

[6] http://en.wikipedia.org/wiki/Statistical_power

The text of the call states:

Support actions to define challenges and prize schemes for verifiable performance in tasks requiring extremely large scale prediction and deep analysis. Compact consortia are required to organise and run well-publicised fast turn-around prediction competitions based on European datasets of a significant size. Proposals in this category are expected to be short in duration and are not required to provide sustainability strategies past the end of the project.

The support actions expected under point b) of ICT-16 can be thought of as the mirror image of the benchmark initiatives under point a).

Under point a) what is fixed is a technology of interest to Europe (because Europe wants to increase the competitiveness of its technology vendors) and what is sought is a list of benchmarks that are credible to the target industries. Under point b) what is fixed is a data analytics problem of European interest and what is open is the technology used to solve it.

For example, a logistics company may wish to improve its planning based on past data from its operations. A challenge can then be defined to develop the best planning mechanism (where "best" must itself be precisely defined as a desired combination of features of interest, plausibly including effectiveness, speed, etc…). Given such a characterisation of the task, anyone is then free to bring their own planning mechanism to the task to a challenge administered transparently by a challenge committee. Prizes could be awarded as an inducement to participate in the challenge.

As for the part a) of ICT-16, here in part b) of ICT-16 too the requirement is that such challenges be based on European data assets of realistic size and complexity. As before, datasets from EU institutions should be considered and their use (when properly motivated in the general structure of the work proposed) will be consider a positive feature of a proposal.

Given that the goal of the challenge exercise is to demonstrate advances in data analytics capabilities of European interest by whatever technical means, consortia will need to detail their plans to plan, define, and administer the sectoral inducement prize challenge, and advertise it as widely as possible across communities endorsing very different approaches to the same underlying issue, so that the merits of their respective approaches could be evaluated according to a methodology that is fair to all.

Consortia will be responsible for administering all relevant aspects of the challenge. This is likely to include the infrastructure to store and/or distribute the relevant datasets to the participants of the challenge. It may also include the provision of (cloud based) computational environments designed to host software submissions from the challenge entrants and monitor the actual use of computational resources in the solution of the challenge.

As the task of the Support action is to plan, prepare and administer the sectoral inducement price scheme, it should not use its own budget for the actual prize money. The prize money for inducement prices is planned to be made available through the prize instrument in the upcoming work programmes, and the Support action will be tasked to manage the administration of such inducement prizes. The timing of inducement prizes will be subject to decision on the future work programme(s), and therefore, the Support action is not expected to present fixed time planning of the competitions.

Consortia will be expected to provide comprehensive risk management plans to include risks ranging from lack of participation in the challenge, to failure of infrastructural components needed to run the challenge, to competition from well-known alternative venues on which similar challenges are being run[7].

[7] See for example http://www.kaggle.com

Appendix: a list of questions that proposals must answer in order to fulfil topic ICT-16 2015

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 for the various outcomes it places itself at a serious disadvantage in a very competitive selection process (because the evaluators of the proposals will be specifically instructed to look for the answers to these and other questions).

a) Research and Innovation Actions

1. What European datasets will define the proposal's problem to be solved and form the basis for testing?
2. What is the current size of those datasets and what is their current and expected rate of growth?
3. Who is the owner of these datasets and what right of access and reuse will the remaining members of the consortium enjoy for the duration of the project, if funded? Who will be owner of the datasets after the project end?
4. On what schedule will these datasets become available for the consortium to work on? Is this consistent with the technical development objectives and the need for systematic testing?
5. What are the specific data analytics performance parameters on which the consortium expects to improve?
6. For each such parameter, what is the expected performance improvement and what is this expectation based on?
7. For each such parameter, what is the best currently reported performance? What performance improvements can be expected based on historical trends?
8. Is the consortium reusing any datasets made available by European institutions? What is the consortium's plan to provide to the relevant EU institutions feedback on the quality of their datasets or requests for additional datasets to be published?
9. What is the precise experimental protocol that the consortium will use to prove that the system proposed improves data analytics capabilities?
10. Given the protocol and the associated statistical power analysis, how many experimental subjects will be required to determine if the data technologies developed are as effective as hoped?
11. What is the consortium's plan to recruit the required number of subjects of the required type on a schedule consistent with the need for experimental testing?

b) Support Actions

1. What is the precise definition of the data challenge to which contestants will be invited?
2. What evidence is there that this challenge is industrially or societally meaningful for Europe?
3. What European dataset(s) will be needed to administer the challenge? By what means these dataset(s) are or will be made available to the consortium?

4. What methodology will the consortium follow to make sure that the challenge will be fairly administered and results from participants are publicly reported?
5. What storage and/or computational infrastructure will be needed to administer the challenge and how will it be made available to the project?
6. What is the consortium's plan to ensure the widest possible participation in the challenge?
7. What is the consortium's plan to ensure that the definition of the challenge is not unfairly favouring any of the possible alternative technical approaches?