Technical Background Notes for Horizon 2020 Objective ICT-18-2016 Big Data PPP: privacy-preserving big data technologies

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-18 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 **Research and Innovation Actions**, 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.[1]

and **Coordination and Support Actions**, 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.[2]

In the first case research (and publication) activities are expected to be a substantial part of the activities proposed, in the second, they are specifically expected **not** to be the primary focus of proposals under this objective.

The topic addresses the following challenges:

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

Specific Challenge: In view of privacy considerations, businesses are often unsure about how to deal with the data collected through their operations. This data is of particularly high value to companies for offering personalised services or developing new business models. Data subjects (citizens, consumers) often feel that they have no control over the use of their personal data. This is aggravated by uncontrolled exploitation, aggregation and linking of personal data by large corporations and advertisers. The resulting lack of confidence undermines efficient and legitimate data sharing and value creation for agreed purposes. The challenge is to develop technologies that are inherently privacy-preserving and offer the basis for empowering the data subjects to understand and be informed of (and, where appropriate, control) the use of their personal data, and the entrepreneurs to develop and run their data driven business.

The motivation behind this topic is the need to develop the set of best practices, engineering principles and software components that will allow European enterprises and organisations to operate as efficiently as possible under the constraints of the upcoming European Data Protection regulation[3], i.e. to extract the maximum amount of societal and commercial value from data resources without violating any recognised rights or societal norms[4]. The recent European Court of Justice Judgement on the Safe Harbour Decision underscores the importance of these objectives[5].

The objective is articulated in two parts, one devoted to the study of how privacy rights recognised by European and national legislation interact with social norms[6] or expectations of privacy and data reuse and one devoted to laying the foundations and providing implementations of reliable software components in support of privacy preserving data management processes.

Data Protection Methods and Software

Research and Innovation Actions under this objective are expected to translate into best practices, reliable processes[7] and software components the privacy requirements deriving from legislation or societal expectations (which are the focus of a dedicated Coordination and Support Action, of which more below).

[3] http://ec.europa.eu/justice/data-protection/

[4] The obvious tension between the two objectives, when pursued one-sidedly, is explored at length in Zarski [2015] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2596822 here mentioned as an example of a sophisticated conceptual mapping of the solution space for this problem (i.e. not as an endorsement of the specific conclusions reached by the author).

[5] http://curia.europa.eu/jcms/upload/docs/application/pdf/2015-10/cp150117en.pdf

[6] One such norm or expectation might be a dislike of undisclosed price discrimination (charging customers different prices for the same product based on data available about the customers). For a recent review of the relationship between privacy and price discrimination see Borgesius [2015] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2652665

[7] For a recent review of the broader, process oriented, context in which the anonymization/de-anonymization arms race can be fruitfully understood and managed, see Rubinstein & Harzog [2015] http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2646185 and Narayanan et al. [2015] https://freedom-to-tinker.com/blog/randomwalker/what-should-we-do-about-re-identification-a-precautionary-approach-to-big-data-privacy/

As such, this part of the objective is inspired by the same broad goals that inform DARPA's Brandeis programme[8], i.e. " *break the tension between: (a) maintaining privacy and (b) being able to tap into the huge value of data*."

The work programme reads:

Research and Innovation actions will advance the state of the art in the definition of methods that will support protection of personal data for harvesting, sharing and querying data assets. The personal data protection methods shall be implemented in secure and robust software modules and be exposed to publicly administered penetration/hacking challenges, open to participants the world over.

The challenges foreseen to demonstrate the soundness of the methods and software components developed in the face of a determined adversary are meant to accomplish two objectives:

1. Determine if exploits are possible that were not foreseen by the developers of the privacy methods/components[9]
2. Measure exactly how determined an adversary would have to be in order to carry out the attack

This second point will provide empirical evidence relevant to a risk-based cost/benefit analysis of privacy schemes[10].

Naturally, given that they may turn out to be successful (i.e. breach privacy) such challenges cannot be carried out with real personal data: synthetic datasets will have to be used instead.

Cross-disciplinary consortia are required to conduct legally and methodologically sound field work and coordinate with the CSA to determine i) if the various formal notions of personal data protection implemented are consistent with EU legislation and with the ethical intuitions of the EU citizens such methods are designed to protect; ii) to what extent privacy protection measures can be personalised in a way that remains intelligible to the data subject while remaining consistent with EU legislation.

Work expected under this objective is also particularly important both in domains that simultaneously represent significant economic opportunity but remain relatively unfamiliar in

[8] http://www.darpa.mil/program/brandeis
[9] See Chen et al. [2015] http://arxiv.org/abs/1508.07306 for an example of how subtle details may require the re-evaluation of the safety of certain methods
[10] See Wan et al. [2015] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120592 for a game theoretic analysis of re-identification risk.

their details to the broader public (e.g the emergent Internet of Things)[11] as well as in established domains where the application of privacy methods can create new opportunities[12].

Particularly welcome will be the development and implementation of methods that preserve both the utility of the data and the privacy of the individuals concerned[13].

Work to be carried out under this objective is also expected to offer a scientific assessment of the extent to which widespread ethical intuitions are, in fact, correct. Recent studies point at the possibility that, in equilibrium, the purposeful behaviour of those who act to protect their privacy may result in much less protection than commonly supposed because it leaks the very information that it is trying to shield[14]. Proposals to develop methods that support a range of privacy norms and expectations[15] will be positively evaluated.

Studies are also expected to provide a 'full-stack' assessment of methods, such as Homomorphic Encryption, that have recently been proposed as promising solutions to the privacy protection challenge, from the obstacles that need to be overcome to deploy them in the operations of an organization (e.g. hospitals[16]) to the algorithmic machinery needed to re-engineer known methods such as machine learning[17], clustering[18], outlier detection[19] to their use in cloud computing[20] or routing protocols[21] to their hardware implementation[22]. Homomorphic Encryption is offered here as an illustrative example only: other methods are also welcome when analysed and evaluated at comparable levels of scientific detail and rigour.

The implementation of privacy preserving methods with additional positive effects in other domains[23] will be considered an advantage, when convincingly documented.

Finally, it is crucial that the alternative settings of the software tools developed, however technically advanced their inner workings, should be easy to explain to individuals that are not scientist or programmers. Ideally, this should include the general public, but at the very least it should include professional figures such as privacy compliance officials in European

---

[11] For an example of privacy issues in an Internet of Things environment, see Perera et al.[2015] http://arxiv.org/abs/1506.08865

[12] For an example of privacy aware document summarization, see Marujo et al. [2015] http://arxiv.org/abs/1508.01420

[13] See Bindschaedler et al. [2015] http://arxiv.org/abs/1505.07499 for an example.

[14] For an example of this, see Cummings et al. [2015] http://arxiv.org/abs/1508.03080

[15] For examples, see Alaggan et al. [2015] http://arxiv.org/abs/1504.06998 , Peddinti et al. [2015] http://research.google.com/pubs/pub43426.html and Acquisti et al [2015] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2580411

[16] http://www.sciencedirect.com/science/article/pii/S1532046414000884

[17] http://arxiv.org/abs/1508.06574 or http://arxiv.org/abs/1505.06556

[18] http://arxiv.org/abs/1508.00192 or http://arxiv.org/abs/1504.05998

[19] http://arxiv.org/abs/1507.06763

[20] See Damiani et al. [2015] http://arxiv.org/abs/1503.07994

[21] http://arxiv.org/abs/1508.05411

[22] See Pöpplemann et al. [2015] http://research.microsoft.com/apps/pubs/default.aspx?id=256217

[23] See Dwork et al. [2015] http://www.sciencemag.org/content/349/6248/636.abstract for an application of privacy technology to statistically reliable data analysis in the scientific domain and Wang et al. [2015] http://arxiv.org/abs/1502.07645 for a connection between Bayesian learning and privacy.

corporations. Interactions with the public or such professional figures will be crucial to prove the usability of the tools developed.

The diversity (e.g. in terms of age, sex, gender, socio-economic class) of data subjects should be taken into account, as appropriate. The data experimentation and integration projects (ICT-14) are likely to provide real-world challenges and data to validate the privacy-preserving technologies.

The first point is related to the exploration of privacy norms and expectations that quite possibly might turn out to be very different across demographics or cultures. To this end, consortia will be well advised to plan for systematic interactions with the Coordination and Support action also planned under this objective (see below).

The second point refers to the fact that projects funded under objective ICT-14 of this work programme will be endowed with data assets and informed by business environments and objectives that are very likely to present their own set of privacy concerns. Once again, proposals under this objective are well advised to plan for interactions with such projects.

Finally, projects proposed under this part of the objective are expected to be quite compact, with a small number of partners with documented technical and software engineering skills. Collection of requirements or societal expectations is expected to happen through interviews with external communities and other Horizon 2020 projects, and the presence of a dedicated partner to manage this process might be advisable.

The Commission considers that proposals requesting a contribution from the EU of between EUR 2 and 4 million would allow this area to be addressed appropriately. Nonetheless, this does not preclude submission and selection of proposals requesting other amounts.

Societal and Ethical Implications

In order to study societal expectations with respect to data privacy exactly one Coordination and Support Action (CSA) will be funded. The expected funding requested for this activity is €1M. This is an expectation and not a requirement: proposals can request alternative amounts, if they can appropriately justify them, up to a maximum of €1M.

Coordination and Support Actions will complement the research by exploring the societal and ethical implications and provide a broad basis and wider context to validate privacy-preserving technologies. The CSA is expected to liaise with a broad and multidisciplinary community of stakeholders (including public administrations, research community, companies, civil society, citizens) to advise the research and innovation in privacy-preserving (Big) Data technologies, promoting an integrated societally and ethically valid approach.

This should be understood as a strict requirement to consult with communities whose primary purpose is **not** that of producing academic publications. The conclusions and reports produced by the selected project, while based on high quality conceptual analysis drawing on disciplines as diverse as legal theory, psychology, sociology, economics, etc… should, nonetheless also be intelligible to the broader European public whose interests they are meant

to foster. In other words, a CSA that only produced highly technical legal/philosophical papers unintelligible to the broader public would **not** be acceptable under this objective.

Another task is to observe, map and report on ethical and Responsible Research and Innovation (RRI) issues in the field of Big Data, including technology, research, markets and education. The action is expected to organise networking, awareness-raising and consultation among its communities, connect with the technical RIAs to inform their thinking and issue reports, analyses and recommendations.

This second task refers to the need for the selected proposal to interact formally with the Responsible Research and Innovation activities of existing or future Horizon 2020 projects (see, for example, objective ICT-28-2017 Robotics, of the present work programme) for the specific purpose of eliciting, recording, analysing and harmonising their perspective concerning issues of privacy. As in the previous case, emphasis is placed on the ability of these interactive communities to identify a common set of problems and to explain these problems in a language intelligible to the policy maker and the wider public (as opposed to only producing highly technical publications unintelligible to outsiders).

Given that the output of the consultations carried out by the CSA is expected to inform technical work done in privacy research and software engineering, it is highly recommended that proposal for the CSA include a dissemination schedule that does not defer all analyses results to the very end of the project.

The impact expected for Coordination and Support Action proposals under this objective is:

- *Appropriate consideration and attention towards an ethically sound approach to big data processing, and effective involvement of the relevant actors and stakeholders;*
- *Improving the dialogue between data subjects and Big Data communities (industry, research, policy makers, regulators), thereby improving the confidence of citizens towards Big Data technologies and data markets.*

The last point concerning improved citizen confidence underscores why consultations with the widest possible range of relevant communities is essential and why technical academic publications cannot be the only outcome of the activities proposed.

Proposal evaluators will be **specifically instructed** to look for evidence of commitment and realistic deployment of resources towards these objectives.

Appendix: a list of questions that proposals must answer in order to be in scope of objective ICT-18 2016 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)

Data Protection Methods and Software

1. What is your specific plan to ensure that the privacy preserving data management tools you will be developing will be understood by the intended users, in terms of their purpose, limitations and interactions with other tools?
2. What kind of challenges will you organise to prove the security of the tools you will be developing? How will you attract world-wide participation to said challenges? What (synthetic) datasets will you use to carry out the challenge?
3. If the techniques or tools you develop reveal the existence of trade-offs, what methodology will you put in place to allow their intended users to identify the choice that creates most value for them?
4. How do you plan to identify and manage cases in which strong popular intuitions are at variance with an analytic treatment of a privacy problem?[24]
5. If the software tools you will be developing are proprietary, what is your business plan to sell them or deployed in production by your organization?
6. If you are developing such tools as open source components, how will you create a community that will continue to maintain the tools after the end of the project?

Societal and Ethical Implications

1. What specific communities will you interact with in order to collect as broad and as representative a spectrum of opinion on ethical or societal expectations of privacy?
2. What methodology will you use to verify that such expectations are widespread (i.e. don't change across European member states, demographics and social groups) and stable (they don't change under a wide range of conditions, including incentives of various nature, from simple convenience to monetary incentives)?
3. What methodology will you use to synthesise the expectations of various groups and communities into a coherent set of recommendations (as opposed to simply providing a list of who believes what)?
4. How do you plan to manage cases in which different groups of individuals display strong and strictly incompatible privacy preferences?

[24] See https://en.wikipedia.org/wiki/Conjunction_fallacy for the existence of such situations.