5. Ethics and security

5.1 Ethics

Due to the nature of the proposed research infrastructure, **SoBigData** needs access to digital records of personal activities that potentially contain sensitive information: user mobility data, online social network and other web and internet data, handling of personal data in web pages, transaction records, etc. In this context, **SoBigData** has a commitment to promoting and adopting legally and ethically grounded collection, management and analysis of data, as demonstrated by effort of the project in considering and investigating ethical and privacy issues in the definition of big data analytics and social mining technologies (see NA1 work package). This aspect is shown also by the adoption of privacy enhancing tools. Also gender issues are considered proactively into the project work plan.

5.1.1 Proactive approach

The **SoBigData** consortium does not consider ethics as simply an external constraint on its research, but proactively addresses ethical issues as part of the research activities. In order to address ethical issues in an adequate manner, a number of measures (see work package NA1) will be taken: An ethics board will be installed, an ethical and legal framework to be used within the project will be designed, and issues related to privacy and data ownership in the context of **SoBigData** will be proactively explored.

The ethics board serves as an internal review process and provides guidance to researchers as to how to address issues concerning research ethics and data protection. All experiments, including the analysis of existing datasets that include personal information, will be reviewed by the ethics board and it will be ensured that adequate safeguards are in place. Where applicable, all experiments will be furthermore submitted for review at the corresponding ethics board of the research institutions conducting the experiment.

The development of an ethical and legal framework as part of work package NA1, but also the application of the value-sensitive design methodology to the **SoBigData** objectives, add further proactive elements to these safeguards. Issues that are known to be problematic with regard to data protection and the application of relevant legal provisions will be actively explored as part of the research objectives. These will in turn inform the research activities that are executed as part of other work packages (e.g. JRA2).

Besides, we consider Value Sensitive Design (VSD) as a normative base for **SoBigData** and for the education and research of data scientists, reflecting the idea that legal norms and moral values – such as privacy and justice – should be construed as non-functional requirements of any product of software engineering, architecture building, modelling, simulations and serious gaming and should be clearly and precisely represented in software and system development. VSD should allow for a systematic, sustained and seamless collaboration between all involved Partners on the articulation of moral principles of a just information society and their effective implementation. The formulation of a robust notion of VSD requires:

- A detailed account of how values and moral considerations can be given their due place in software engineering (methodology)
- A detailed account of how they can be used as requirements in the design of ICT tools, applications and services.

Instruments to increase societal ethical awareness and participation: **SoBigData** will develop and implement a research infrastructure that requires ethical awareness, which should be coupled to the ability of scientists to act responsibly and to consider possible consequences. We will therefore

adopt appropriate training for allowing individual researchers to reflect, compare and deliberate upon their values and upon their resources and competences to deal with moral issues.

5.1.2 Data protection

Applicable data protection regulations will be investigated and measures to enforce privacy will be implemented as an integral part of the systems that will be developed in the project. However, scientists within the consortium and outside it (within TA) will manage potentially re-identifiable data for the purpose of conducting the planned research. While doing so, all institutions and individual researchers in the **SoBigData** consortium shall comply with the applicable legislation on EU and national level, in particular Directive 95/46/EC (Data Protection Directive). Furthermore, relevant guidelines such as the "code of conduct applying to processing of personal data for statistical and scientific purposes" set forth by the Italian National Data Protection Commission will be considered.

Whenever possible, data subjects will be informed about the data processing and asked for their consent. In cases where existing datasets are analysed (e.g. for the crawling of web-data and the import of data sets), the guidelines set forth by the Article 29 Working Party on *purpose limitation* will be taken into account[2]. This means that data processing for the goal of statistical processing takes place separated from the original purposes for which this data was collected. The results of such processing are not linked back to the data subjects and all data is anonymized as early as possible.

Researchers will be trained in applying the necessary procedural safeguards and the project will take the necessary steps to ensure that applicable legislation concerning data protection is respected, and the steps taken during the project lifetime to ensure conformance to established guidelines will be documented. To this end, a specific *SoBigData Ethics Board* will be established (see NA1 work package).

It should be noted, besides, that a number of privacy issues that could arise when handling the above data, do not arise within **SoBigData**; specifically:

1. It is not planned to process, to combine or to re-link the data to be analyzed with personal data such as name, address, social security ID etc., and other information that can be used for directly identifying persons. If directly identifiable data is collected (e.g. as part of user provided datasets), such data will be removed or anonymized as early as possible. All data that is used for analysis can therefore at most be used to indirectly identify individuals.
2. The data sets and the potentially sensitive information contained in then will be handled within the data management provisions of the Big Data Ecosystem (see JRA1 work package), securing the different sharing/diffusion constraints of each data source. All data sets will be stored with appropriate security restrictions in place. This comprises access restrictions to IT systems and storage such as password protection, encryption, etc., complying with all national and EU legislation.
3. **SoBigData** investigates the data sets solely for the purpose of research. Publications need not and will not disclose specific privacy-relevant information.

[2] See Article 29 Working Party (2013). Opinion 03/2013 on purpose limitation. Available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

5.1.3 Human participants

Some of the planned activities include the direct involvement of human participants in the collection of data. This concerns, for example, the possibility to upload social network datasets (work package NA5) or participatory sensing methods (work package JRA1). To the extent that those activities constitute human subjects research, relevant principles, such as those set forth in the Declaration of Helsinki, will be applied. The experiments will furthermore be reviewed by the **SoBigData** ethics board and by review committees of the corresponding research institutions.

When describing issues relating to informed consent, it will be necessary to illustrate an appropriate level of ethical sensitivity. All participating users will have given signed consent, following the EU requirements using a standardized form, which will be printed in the appropriate language.

Prior to the start of the process, the partners will agree the method of approach to the individual, the handling of the completion of forms, the confidentiality of the process, the ability of the user to decline, and the handling, access, and storage of data will all follow a predetermined design at all sites. All users will also be informed of the process of withdrawal from the project. Each user will be given a user identity number and only this will appear on all other documentation during the project. This will assist in the anonymity of all users.

5.1.4 Integrating the gender dimension in SoBigData

Following the European policy of equal opportunities between women and men, the Commission has adopted a gender mainstreaming strategy by which each policy area, including that of research, must contribute to promoting gender equality. The consortium is highly concerned with this issue and all the partners are committed to promote women's participation in SoBigData.

We will specifically address gender aspects in recruitment. The **SoBigData** consortium is also engaged to take actions to increase the involvement of women scientists in order to reach a 40% minimum target. The main actions will be articulated around these ideas:

- Women participation in research must be encouraged both as scientists/technologists and within the evaluation, consultation, and implementation processes,
- There should be a gender balance between men and women in laboratories, in senior management, which reflects their roles in society.

We note that, specifically, **SoBigData** has the potential and the necessary instruments to promote excellence through mainstreaming gender equality since it has:

- a female global coordinator (Fosca Giannotti – CNR)
- a female deputy-coordinator (Kalina L. Bontcheva – USFD) who is also work package leader of TA1 – Transnational Access
- leading female researchers, such as Natalia Andrienko (FRH) and Donatella Castelli (CNR).

The European Machine Learning and Knowledge Discovery community, chaired by F. Giannotti, has already adopted a proactive approach in promoting gender equality, while taking into account the specific circumstances and interests of the different women scientists active in the community, and will serve as a channel for recruiting female scientists both in the consortium and as **SoBigData** users.

5.1.5 Ethical Review

SoBigData involves the groups at TUDelft and LUH as internal IT ethics and IT law experts for consultation and is open to ethical review by the Commission. The consortium shall ensure that the

research carried out under the project fully complies with the national legal and ethical requirements of the countries where the tasks raising potential ethical issues are to be carried out. To this purpose, the consortium shall submit an ethics self-assessment, where we shall explain in detail how to address the ethical issues, if any, raised by the research objectives, methodology and results.

5.1.6 Ethics requirements summary

SoBigData has a dedicated WPs NA1 (Legal and Ethical framework) that addresses the ethical requirements. In particular, **SoBigData** will establish a dedicated ethical and legal board with experts in Ethics of IT and IT Law, who will:

- Define and implement the legal and ethical framework of the **SoBigData** research infrastructure, in accordance with the European and national legislations (including data protection and intellectual property rights) as they develop
- Monitor the compliance of experiments and research protocols with the framework,
- Act as a continuous consultant for the JRAs on the development of big data analytics and social mining tools with Value-Sensitive Design and privacy-by-design methodologies, and
- Investigate, design and promote novel architectures, protocols and procedures for the safe and fair use of big data for research purposes, in order to boost excellence and international competitiveness of Europe's big data research.

The following table summarize how the Ethics requirements will be tackled.

Category HUMANS	
1. *Details on the procedures and criteria that will be used to identify/recruit research participants must be provided*	See section 5.1.3 (Human participants) The **SoBigData** Ethics Board will review the research procedures to monitor their compliance with Declaration of Helsinki.
2. *Detailed information must be provided on the informed consent procedures that will be implemented.*	See section 5.1.3 (Human participants) All participating users will have given signed consent, following EU requirements and using a standardized form in the appropriate language.
*3. The applicant has to further clarify how passive actors are involved for the purpose of data collection, and what **informed consent** procedures will be implemented for that case.*	These issues will be considered by the **legal and ethical framework** proposed by the Ethics Board, it will be revised after year 2 of the project, in order to cover necessary modifications in the development of the project or changed legislation. As stated in Task 2.2 a key legal and ethical issue will be the question under which provisions and codes of conduct collected datasets can be (re)used for secondary research purposes. Guidelines will also take into account existing national codes; for example, the Italian code of conduct stipulates that in certain cases information to users can be transmitted to through newspaper or other media instead of personal communication.

	Basic principles are however already presented in section 5.1.2 (Data Protection) and are inspired to Article 29 Working Party on *purpose limitation*[3] and the Data Protection Directive "*either by a valid informed consent or a legal exception, such as "processing for scientific purposes"*.
Category PROTECTION OF PERSONAL DATA	
1. *The applicants must provide a policy for dealing with re-identifiable data. Moreover, the applicants must provide a policy for dealing with findings which are unrelated to the proposed work.*	One major duty of the Ethics Board is to design and update the **ethical and legal framework** to be adopted in SoBigData. Basic principles of the policy for dealing with re-identifiable data are established in section 5.1.2 (Data Protection).
2. Copies of ethical approvals for the collection of personal data by the competent University Data Protection Officer / National Data Protection authority must be submitted.	It is part of the planned procedures that will be implemented by the ethics board as established in Task 2.1.
3. Justification must be given in case of collection and/or processing of personal sensitive data.	The collection of personal sensitive data is not an objective in SoBigData. If needed, SoBigData will provide the appropriate justification and will put in place the appropriate security procedures.
4. *Detailed information must be provided on the procedures that will be implemented for data collection, storage, protection, retention and destruction and confirmation that they comply with national and EU legislation.*	Details will be provided as part of Task 8.1 (Data Management and Integration Plan) following the guidelines established by the Ethics Board in the Ethics and Legal framework (T2.2) so that they will be guaranteed to be compliant with the Data Protection Directive.
5. Detailed information must be provided on the informed consent procedures that will be implemented.	See point 3 of Category "Humans".
6. The applicant must explicitly confirm that the existing data are publicly available.	Dataset that are publicly available under different access provisions will be curated and managed as established in task8.1 (Data Management and Integration Plan).
7. In case of data not publicly available, relevant authorisations must be provided.	Dataset not publicly available will have appropriate authorization and appropriate access procedures to be used by researchers on site as established in task8.1 (Data Management and Integration Plan).
8. In the course of the project, it is possible that the consortium encounters new ethics issues. Any such new issues and how they have been addressed are to be documented to contribute to the understanding of ethics in big data research.	The Ethics Board installed and managed in work package NA1 proactively addresses ethical issues as part of the research activities in order to tackle current and new ethical issues. The Ethics Board will provide guidance to researchers as to how to address issues concerning research ethics and data.
Category NON-EU COUNTRIES	
1. The applicants must provide confirmation whether the proposed work requires local/national ethics clearance in those countries where the research will take place.	We do not foresee such cases, but if it happens the Ethics Board will take care of these cases.
2. The applicant must confirm that the ethical standards and guidelines of Horizon2020 will be	Yes, we confirm.

[3] See Article 29 Working Party (2013). Opinion 03/2013 on purpose limitation. Available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

rigorously applied, regardless of the country in which the research is carried out.	
3. The applicant must provide details on the material which will be imported to/exported from EU and provide the adequate authorisations.	These details concerning data exchange with Swiss and non-EU countries will be inserted in the (Data Management and Integration Plan).

5.2 Security

We do not identify issues of security related to the current proposal, other than the above mentioned provisions for secure data management of potentially sensitive or otherwise restricted access to data.

1.4. Ethics Requirements

Ethics Issue Category	Ethics Requirement Description
HUMANS	- Details on the procedures and criteria that will be used to identify/ recruit research participants must be provided
HUMANS	- Detailed information must be provided on the informed consent procedures that will be implemented.
HUMANS	- The applicant has to further clarify how passive actors are involved for the purpose of data collection, and what informed consent procedures will be implemented for that case. The applicants must provide a policy for dealing with re-identifiable data. Moreover, the applicants must provide a policy for dealing with findings which are unrelated to the proposed work.
PROTECTION OF PERSONAL DATA	- Copies of ethical approvals for the collection of personal data by the competent University Data Protection Officer / National Data Protection authority must be submitted
PROTECTION OF PERSONAL DATA	- Justification must be given in case of collection and/or processing of personal sensitive data
PROTECTION OF PERSONAL DATA	- Detailed information must be provided on the procedures that will be implemented for data collection, storage, protection, retention and destruction and confirmation that they comply with national and EU legislation
PROTECTION OF PERSONAL DATA	- Detailed information must be provided on the informed consent procedures that will be implemented
PROTECTION OF PERSONAL DATA	- The applicant must explicitly confirm that the existing data are publicly available
PROTECTION OF PERSONAL DATA	- In case of data not publicly available, relevant authorisations must be provided
PROTECTION OF PERSONAL DATA	- In the course of the project, it is possible that the consortium encounters new ethics issues. Any such new issues and how they have been addressed are to be documented to contribute to the understanding of ethics in big data research. The applicants must provide confirmation whether the proposed work requires local/ national ethics clearance in those countries where the research will take place.
NON-EU COUNTRIES	- The applicant must confirm that the ethical standards and guidelines of Horizon2020 will be rigorously applied, regardless of the country in which the research is carried out
NON-EU COUNTRIES	- The applicant must provide details on the material which will be imported to/exported from EU and provide the adequate authorisations