EUROPEAN COMMISSION
Research Executive Agency (REA)
Inclusive, Innovative and Reflective Societies

HORIZON 2020

Simpatico

Project acronym:	SIMPATICO
Project full title:	SIMplifying the interaction with Public Administration Through Information technology for Citizens and cOmpanies
Call identifier:	EURO-6-2015
Type of action:	RIA
Start date:	1 March 2016
End date:	28 February 2019
Grant agreement no:	692819

D1.3 – Data management plan v.1

WP1	Project management
Due Date:	31/08/2016
Submission Date:	20/09/2016
Responsible Partner:	Fondazione Bruno Kessler (FBK)
Version:	1.0
Status:	Final
Author(s):	Matteo Gerosa (FBK), Serena Bressan (FBK)
Reviewer(s):	Ivan Pretel (DEUSTO), Zulf Choudhary (SPARTA)
Deliverable Type:	R: Report
Dissemination Level:	PU: Public

Version History

Version	Date	Author	Organisation	Description
v0.1	14/06/2016	M. Gerosa, S. Bressan	FBK	Table of contents
v0.2	21/07/2016	M. Gerosa, S. Bressan	FBK	First draft
v0.3	16/08/2016	M. Gerosa, S. Bressan	FBK	Second Draft. Implemented the suggestions of DEUSTO and SPARTA.
v0.5	16/09/2016	M. Gerosa, S. Bressan	FBK	Final revision.
v1.0	20/09/2016	M. Pistore	FBK	Final quality ckeck.

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Glossary

AEPD	Agencia Española de Protección de Datos
AES	Advanced Encryption Standard
API	Application Program Interface
CDV	Citizen Data Vault
DBMS	Database Management System
DMP	Data Management Plan
DPA	Data Protection Authority
DSM	Data Security Manager
EAB	Ethics Advisory Board
EC	European Commission
ES	Spain
EU	European Union
FIPS	Federal Information Processing Standard
GA	Grant Agreement
H2020	Horizon 2020
HTTPS	HyperText Transfer Protocol over Secure Socket
IAM	Identity and Access Management
ICO	Information Commissioner's Office
IPR	Intellectual Property Rights
ISO	International Organization for Standardization
IT	Italy
JSON	JavaScript Object Notation
MVCR	Minimum Viable Consent Record
N/A	Not Applicable
NLP	Natural Language Processing
ORD	Open Research Data
PA	Public Administration
PDF	Portable Document Format
REST	Representational State Transfer
RO	Research Objective

POPD	Protection of Personal Data
RDF	Resource Description Framework
SQL	Structured Query Language
SSL	Secure Sockets Layer
SUS	System Usability Testing
TDE	Transparent Data Encryption
TSL	Transfer Security Layer
UMA	User Managed Access
UREC	University Research Ethics Committee
UK	United Kingdom
WP	Workpackage
XACML	Extensible Access Control Markup Language
XML	Extensible Markup Language

Table of contents

Executive summary

This document is the deliverable **"D1.3 – Data management plan v.1"** of the European project "SIMPATICO - SIMplifying the interaction with Public Administration Through Information technology for Citizens and cOmpanies" (hereinafter also referred to as **"SIMPATICO"**, project reference: 692819).

The **Data Management Plan (DMP)** describes the types of data that will be generated and/or gathered during the project, the standards that will be used, the ways in which data will be exploited and shared (for verification or reuse), and in which way data will be preserved. This DMP has been prepared by taking into account the template of the **"Guidelines on Data Management in Horizon 2020"** [Version 2.1. of 15 February 2016]. The elaboration of the DMP will allow to SIMPATICO partners to address all issues related with data protection, including ethical concerns and security protection strategy. SIMPATICO takes part in the **Open Research Data Pilot (ORD pilot)**; this pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects, such as the Data generated by the SIMPATICO platform during its deployment and validation. Moreover, under Horizon 2020 each beneficiary must **ensure open access to all peer-reviewed scientific publications** relating to its results: these publications shall be made also available through the public section of the SIMPATICO website. All these aspects have been taken into account in the elaboration of the DMP.

Current deliverables "D1.3 – Data management plan v.1" is released at the beginning of the project (M6). An revised deliverable is expected at the end of the project ("D1.4 – Data management plan v.2", at M36). However, the DMP will be a **living document** throughout the project, and this initial version will evolve during the SIMPATICO lifespan according to the progress of project activities.

Starting from a brief illustration of the SIMPATICO project, and of the ethical concerns raised by the project activities, this report describes **the procedures of data collection, storing and processing**, with a final overview on **SIMPATICO security protection strategy**. This report does not cover the general concerns related to ethics and data protection, as they are the focus of dedicated deliverables already submitted – namely reports "D1.5 – Ethics compliance report", "D8.1 – H – Requirement no. 1", and "D8.2 – POPD – Requirement no. 2" (M3).

1. Introduction

The research activities undertaken in the SIMPATICO project have important data protection aspects, in particular due the foreseen involvement of public/private stakeholders and citizens and due to the necessity to collect, store and process personal data. This deliverable analyses the **data management implications** of the activities undertaken in the project, and describes the guidelines and procedures put in place in order to ensure compliance with data management requirements.

The rest of this section provides **background information on the SIMPATICO project** (Subsection 1.1) and identifies in brief the **ethical issues** raised by the project activities (Subsection 1.2). The project aims **to maximise access to and re-use of research data**, also ensuring **open access to all peer-reviewed scientific publications** relating to its results, in order pave the path for its data management plan according to the signed Grant Agreement - GA (Subsection 1.3.). Section 2 concerns the detailed **description of SIMPATICO datasets**, according to the requirements set out in Annex 1 - Data Management Plan template of the “Guidelines on Data Management in Horizon 2020” [1]: (a) the handling of research data during and after the project; (b) what data will be collected, processed or generated; (c) what methodology & standards will be applied; (d) whether data will be shared/made open access and how; (e) how data will be curated and preserved. Finally, Section 3 presents the **SIMPATICO security protection strategy**.

1.1. SIMPATICO in brief

SIMPATICO's goal is **to improve the experience of citizens and companies in their daily interactions with the public administration** by providing a **personalized delivery of e-services** based on advanced **cognitive system technologies** and by promoting an **active engagement of people** for the continuous improvement of the interaction with these services. The SIMPATICO approach is realized through a platform that can be deployed on top of an existing PA system and allows for **a personalized service delivery** without having to change or replace its internal systems: a process often too expensive for a public administration, especially considering the cuts in resources imposed by the current economic situation.

The goal of SIMPATICO is accomplished through a solution based on the **interplay of language processing, machine learning and the wisdom of the crowd** (represented by citizens, business organizations and civil servants) **to change for the better the way citizens interact with the PA. SIMPATICO will adapt the interaction process** to the characteristics of each user; **simplify** text and documents to make them understandable; **enable feedback for the users** on problems and difficulties in the interaction; **engage civil servants, citizens and professionals** so as to make use of their knowledge and integrate it in the system (Fig. 1).

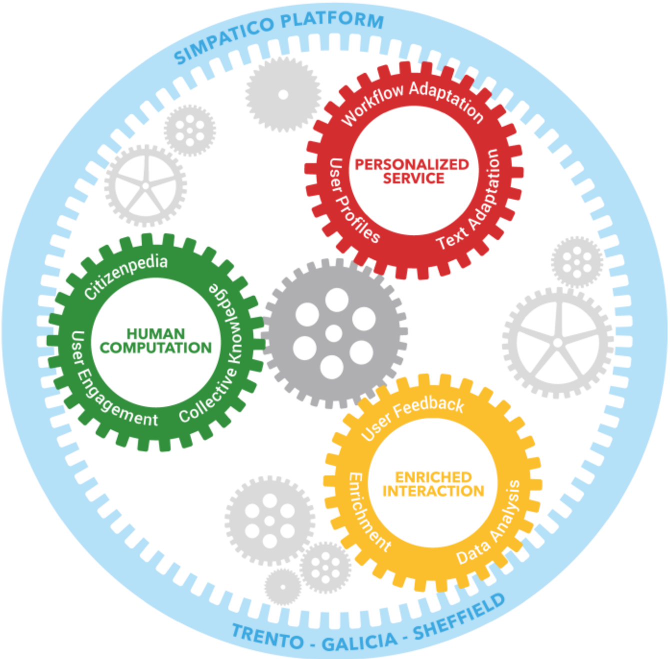

Figure 1: SIMPATICO concept as a glance

The project aims can be broken down into the following **smaller research objectives (ROs)**.

RO1. Adapt the interaction process with respect to the profile of each citizen and company (PA service consumer), in order **to make it clear, understandable and easy to follow**.

- A **text adaptation** framework**,** based on a **rich text information layer** and on machine learning algorithms capable of **inducing general text adaptation operations** from **few examples, and of customizing these adaptations to the user profiles.**
- **A workflow adaptation engine** that will take user characteristics and tailor the interaction according to the user's profile and needs.
- A feedback and annotation mechanism that **gives users the possibility to visualize, rate, comment, annotate, document the interaction process** (e.g., underlying the most difficult steps), so as to provide valuable feedback to the PA, further refine the adaptation process and enrich the interaction.

RO2. Exploit the wisdom of the crowd to enhance the entire e-service interaction process.

- An **advanced web-based social question answering engine (Citizenpedia)** where citizens, companies and civil servants will **discuss and suggest potential solutions and interpretation for the most problematic procedures and concepts.**
- A **collective knowledge** database on e-services that will be used to simplify these services and improve their understanding.
- An **award mechanism** that will **engage users and incentivize them to collaborate** by giving them **reputation** (a valuable asset for professionals and organizations) and **privileges** (for the government of Citizenpedia – a new public domain resource) according to their contributions.

RO3. Deliver the SIMPATICO Platform, an open software system that can interoperate with PA legacy systems.

- A platform that **combines consolidated e-government methodologies with innovative cognitive technologies** (language processing, machine learning) at different level of maturity, enabling their experimentation in more or less controlled operational settings.
- An interoperability platform that enables an **agile integration of SIMPATICO's solution with** PA legacy systems and that allows the exploitation of data and services from these systems with the SIMPATICO adaptation and personalization engines.

RO4. Evaluate and assess the impact of the SIMPATICO solution

- Customise, deploy, operate and evaluate the SIMPATICO solution on **three use-cases in two EU cities** – Trento (IT) and Sheffield (UK) – **and one EU region** – Galicia (ES).
- **Assess the impact** of the proposed solution in terms of **increase in competitiveness, efficiency of interaction and quality of experience.**

1.1.1. SIMPATICO technical framework and infrastructure

The SIMPATICO project will provide a **software platform** incorporating technical innovations to enhance **the efficiency, effectiveness and inclusiveness of public services**. To this aim, SIMPATICO collects, generates and utilizes both personal and other data in a complex way. For what concerns this deliverable (consumption, production and storage of data), the key SIMPATICO components are the following:

1. **Citizen Data Vault**: it represents the component that will take care of personal data exchange between a user and SIMPATICO components. It is a distributed repository of the citizen (or company) profile and related information. It is continuously updated through each interaction and is used to automatically pre-fill forms. In this way, the citizen will give to the PA the same information only once, as the information will be stored in the vault and used in all the following interactions;
2. **Human Computation (Citizenpedia):** SIMPATICO fosters citizens' involvement, by providing Citizenpedia, a hybrid of Wikipedia and a collaborative question answering engine, and sharing improvements on public resources in a semi-automatic basis. Citizens, companies and civil servants will discuss and suggest potential solutions and interpretation for the most problematic procedures and concepts. In addition, the user will be able to highlight portions of text that he/she considers unclear and ask for a simplified version. These interaction actions will further refine the user profile and will be stored in the citizen data vault to serve as the basis for the adaptation of future interactions. Public servants are able to moderate comments and suggestions of citizenships to prevent crowd's wisdom bias. The knowledge collected by a user on a specific e-service (e.g., a request of clarification or the explanation of a concept) can propagate and improve the understanding and interaction of potentially all users and e-services. An award mechanism that engages users and incentivize them to collaborate by giving them reputation (a valuable asset for professionals and organizations) and privileges is designed.
3. **SIMPATICO Adaptation Engine:** it is a cognitive system that will make use of innovative text processing and machine learning algorithms to adapt the text and the workflow of the interaction according to the user profile. The text adaptation engine will adapt the text of the forms and of the other documents to make it more understandable and to clarify complex elements, while the workflow adaptation engine will adapt the interaction process itself by presenting the citizen only the elements that are relevant for his/her profile (e.g., if the citizen is not a foreigner he/she will not be presented the section of a form reserved for

foreigners). The adaptation engine exploits data collected on the interactions of the users, exploiting both implicit and explicit techniques; these data are stored in the "User Profile" and "Log" components of SIMPATICO.

These components are highlighted in Figure 2, depicting SIMPATICO conceptual architecture.

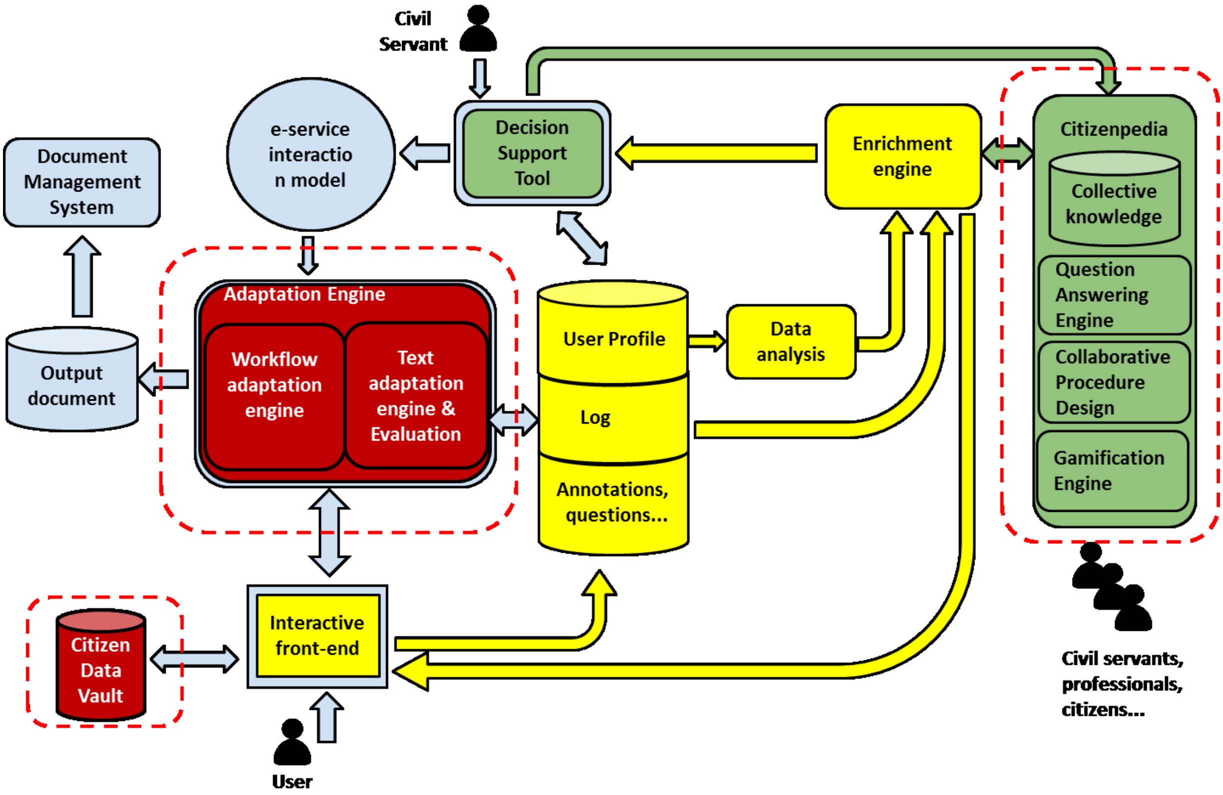
Civil Servant
Document Management System
e-service interactio n model
Decision Support Tool
Enrichment engine
Citizenpedia
Collective knowledge
Question Answering Engine
Collaborative Procedure Design
Gamification Engine
Output document
Adaptation Engine
Workflow adaptation engine
Text adaptation engine & Evaluation
User Profile
Log
Annotations, questions...
Data analysis
Citizen Data Vault
Interactive front-end
User
Civil servants, professionals, citizens...

Figure 2: SIMPATICO Platform conceptual architecture and main components

1.1.2. SIMPATICO pilots

The piloting of this platform in two European cities (**Trento and Sheffield**) and one region (**Galicia**) in Italy, Spain and the United Kingdom (UK), through a **two-phase use-case validation** is in line with the important objectives in the **EU eGovernment Action Plan 2011-2015**. The stakeholders engaged in the **three use-cases** were selected for their experience and interest in e-services, as well as for the different socio-cultural backgrounds of the three regions. In this way, the Consortium have the opportunity to validate the effectiveness of the project results in **contexts which differ on the number and heterogeneity of citizens and their social and cultural background**.

There are indeed important **differences in the technological ecosystems**, with Trento and Sheffield having just started the process of digitalization of their services to citizens and businesses (this process will actually happen in alignment and integration with the SIMPATICO activities), and Galicia having a mature and consolidated e-service delivery infrastructure (thus allowing to study the deployment of SIMPATICO on top of an already operating system). The contexts also **differ for the point of view of the number and heterogeneity of end-users and for the variety and maturity of e-services** (**see deliverable "D6.1 – Use-case planning & evaluation v1"**).

1.2. SIMPATICO ethical issues

The SIMPATICO consortium is committed to **perform a professional management of any ethical issue** that could emerge in the scope of the activities of the project, also through the support of its **Ethics Advisory Board** (see deliverable "D1.5 – Ethics compliance report"). For this reason, the consortium has identified relevant ethical concerns already during the preparation of the project proposal and, then, during the preparation of the Grant Agreement. During this phase, the European Commission has also carried out an ethics scrutiny of the proposal, with the objective of verifying the respect of ethical principles and legislation. With regard to SIMPATICO, the research entails specific ethical implications, involving human subjects and risks for the protection of personal data [2] [3]. In particular, the **SIMPATICO ethical issues (requirements)**, as reported in the European Commission ethics scrutiny report and acknowledged by the SIMPATICO project, are the following:

Humans - "D8.1 – H – Requirement no. 1"

1. *Details on the procedures and criteria that will be used to identify/recruit research participants must be provided.*
2. *Detailed information must be provided on the informed consent procedures that will be implemented.*

SIMPATICO involves **work with humans** ('research or study participants'): according to the EC, collection of personal data, interviews, observations, tracking or the secondary use of information provided for other purposes. End-users (i.e., citizens and businesses) will be **engaged in the project use-cases** to test the functionalities provided by the SIMPATICO solution for the usage of e-services. Specific **engagement campaigns** will be defined and executed for each use-case. The use-cases will involve **only voluntary participants aged 18 or older and capable to give consent**, who will be informed on the nature of their involvement and on the data collection/retention procedures through an **informed consent form** before the commencement of their participations. **Terms and conditions** will be transparently communicated to the end-users by means of an **information sheet** including descriptions of e.g., purpose of the research, adopted procedures, data protection and privacy policies. For further details, please see sections below and deliverables **"D1.5 – Ethics compliance report"** and **"D8.1 – H – Requirement no. 1"**.

Protection of personal data – "D8.2 – POPD – Requirement no. 2"

1. *Copies of ethical approvals for the collection of personal data by the competent University Data Protection Officer/National Data Protection authority must be submitted by the coordinator to REA before commencement of data gathering.*
2. *Clarification and if relevant justification must be given in case of collection and/or processing of personal sensitive data. Requirement needs to be met before commencement of relevant work.*
3. *The applicant must explicitly confirm that the existing data are publicly available.*
4. *In case of data not publicly available, relevant authorisations must be provided, requirements to be met before grant agreement signature.*

SIMPATICO involves **collecting and processing personal data** (i.e., any information which relates to an identified or identifiable natural person, such as name, address, email) and **sensitive data** (e.g., health, sexual life, ethnicity). The **Citizen Data Vault** represents the component that will take care of personal and sensitive data exchange between a user and SIMPATICO components. Personal and sensitive data will be made **publicly available** (e.g., for the data of **Citizenpedia**) only after an **informed consent** has been collected and suitable **aggregation and/or pseudonymization techniques** have been applied. Mechanisms for encryption, authentication, and authorization (e.g., TLS protocol, Single-Sign-On implementations, Policy Enforcement Point for XACML) will be exploited in the processes, so to ensure the satisfaction of core **security and data protection requirements**, namely confidentiality, integrity, and availability. For further details, please see sections below and deliverables **"D1.5 – Ethics compliance report"** and **"D8.2 – POPD – Requirement no. 2"**.

Vulnerable groups

In addition to the above-mentioned ethical requirements, in the context of this deliverable it is also important to specify that SIMPATICO pilots may involve certain **vulnerable groups**: e.g., **elderly people and immigrants** (see also deliverables **"D1.5 – Ethics compliance report"** and **"D8.1 – H – Requirement no. 1"**). Please note that all the research participants will have the **capacity to provide informed consent**: individuals who lack capacity to decide whether or not to participate in research will be appropriately excluded from research. Anyway taking into account the scope and objectives of the research, researchers should be **inclusive in selecting participants**. Researchers shall not exclude individuals from the opportunity to participate in research on the basis of attributes such as culture, language, religion, race, sexual orientation, ethnicity, linguistic proficiency, gender or age, unless there is a valid reason for the exclusion.

Vulnerable groups could be misapplied for stigmatisation, discrimination, harassment or intimidation. Concern for **the rights and wellbeing of research participants** lies at the root of ethical review. The perception of subjects as vulnerable is likely to be influenced by diverse cultural preconceptions and so regulated differentially by localised legislation. It is likely to be one of the areas where researchers **need extra vigilance to ensure compliance with laws and customs**. Some vulnerabilities may not even be obvious until research is actually being conducted.

To reduce the risk of enhancing the vulnerability/stigmatisation of the above-mentioned individuals, the SIMPATICO **Ethics Advisory Board** (see **"D1.5 – Ethics compliance report"**) will provide **specific assessment on vulnerable groups** that may be involved, prior of the commencement of the pilots' activities. Such an assessment will be included in the expected versions of the reports "**Project progress report"** (M12), **"D1.2 – Intermediate activity report"** (M22), as well as in the reports concerning the use-case planning. In particular, it will include further details on the:

- type of vulnerability;
- recruitment/inclusion criteria and informed consent procedures;
- appropriate efforts to ensure fully informed understanding of the implications of participation.

Language, educational, administrative and technical barriers affecting certain societal collectives at risk of exclusion will be captured, recognized, analysed and tackled by the dynamic adaptation of services towards **maximizing user experience**. Diversity will be tackled by matchmaking citizen and organizations profiles, securely preserved within the SIMPATICO platform, with the services and range of adaptations readily available by SIMPATICO. The process of e-service delivery will be transformed by allowing service providers and consumers to **cooperate through the SIMPATICO Platform**, mutually enriching themselves and contributing to the co-creation of the knowledge base of **Citizenpedia**. **Political, legal and cultural obstacles** and factors affecting the acceptability and effectiveness of this transformation will be analysed and addressed throughout the project. For the above-mentioned reasons, it remains vital on ethical grounds that all the participants (and also vulnerable groups) should be able **to freely decide for themselves**, with advocacy support if needed. If a research study like SIMPATICO enhances the provision of services to the community, then **study participants may gain both directly and indirectly**. Research which reflects the needs and perspectives of service users may even be more likely to produce **successful policy and practice recommendations**, also with regard to the vulnerable groups of populations involved.

SIMPATICO Ethics Advisory Board

All the above-mentioned deliverables will be assessed and validated during the first meeting of the **SIMPATICO Ethics Advisory Board (EAB)** (see “D1.5 – Ethics compliance report”). It is **competent to provide the necessary authorizations** when the collection and processing of personal (or sensitive) data is part of the planned research, with the validation of national and/or local Data Protection Authorities if needed. This board is led by an **ethics adviser** external to the project and to the host institution, totally independent and free from any conflict of interest. In addition to the external ethics adviser, the EAB is composed of **one expert representative from all members of the SIMPATICO Consortium** [4]. Members of the Ethics Board are listed in “D1.5 – Ethics compliance report” with the name and contact information for persons appointed, the terms of reference for their involvement, and their declarations of no conflict of interest. The **reference national and/or local Data Protection Authorities** competent to provide the above-mentioned SIMPATICO EAB with the necessary **instructions/authorizations/notifications** for each pilot are the following [5] [6] [7]:

Trento pilot (Italy): the Italian Data Protection Authority (DPA - http://www.garanteprivacy.it/).
According to the “Italian Data Protection Code” (Legislative Decree no. 196/2003), an authorisation by the Italian DPA is required to enable private (and public) bodies to process specific typologies of personal and sensitive data (see Section 26 of the Italian Data Protection Code). More precisely, the DPA needs to be notified (also thorough an electronic form) whenever a public or private body undertakes a personal data collection, or personal data processing activity, as data controller. A data controller is required under the law to only notify the processing operations that concern e.g., data suitable for disclosing health and sex life, data processed with the help of electronic means aimed at profiling the data subject and/or his/her personality, analysing consumption patterns and/or choices. In such context, the DPA is also responsible for evaluating and expressing opinions on specific arguments concerning data protection (see “Simplification of Notification Requirements and Forms. Decision of the DPA dated 22 October 2008, as published in Italy's Official Journal no. 287 of 9 December 2008”).
In the case of Trento pilot, we consider this public authority appropriate for providing the SIMPATICO EAB with the necessary instructions/authorizations/notifications.

Sheffield pilot (United Kingdom): the University Research Ethics Committee (UREC) of the University of Sheffield (https://www.sheffield.ac.uk/ris/other/committees/ethicscommittee).
The University Research Ethics Committee (UREC) of the University of Sheffield is an independent, unbiased and interdisciplinary university-wide body that scrutinizes any potential issues related to research ethics for staff and students of the University of Sheffield, including collaborative research deriving from external funding. The key tasks this committee is in charge of are:

- Advise on any ethical matters in research that are referred to it from within the University;
- Keep abreast of the external research ethics environment and ensure that the University responds to all external requirements.

In the case of the Sheffield pilot, we consider this committee appropriate for providing the SIMPATICO EAB with the necessary instructions/authorizations/notifications. We remark that, in the case of Sheffield Council, Sheffield University and Sparta Technologies Ltd, all the involved entities comply with the UK data protection regulations and intend though the committee to ensure that act is enforced. Only if necessary, the AEB will engage the UK Information Commissioner's Office (ICO - https://ico.org.uk/).

Galicia pilot (Spain): the Research Ethics Committee of the University of Deusto (http://research.deusto.es/cs/Satellite/deustoresearch/en/home/research-ethics-comittee).
This committee is an independent, unbiased and interdisciplinary body that is both consultative and advisory in nature, and reports to the Vice-Rector's Office for Research. Among other responsibilities, this committee is in charge of:

- Conducting the ethical assessment of research projects and drawing up the ethical suitability reports requested by institutions and researchers.
- Ensuring compliance with best research and experimentation practices with regard to individuals' fundamental rights and the concerns related to environmental defense and protection.
- Supervising assessment processes or ethical requirements in research carried out by institutions and public bodies.
- Preparing reports for the University's governing bodies on the ethical problems that may arise from R+D+I activities.
- Ensuring compliance with the Policy on Scientific Integrity and Best Research Practices of the University of Deusto.
- Providing guidance on laws, regulations and reports on research ethics.
- Reviewing procedures that have already been assessed, or proposing the suspension of any experimentation already started if there are objective reasons to do so.

In the case of the Galicia pilot, we consider this committee appropriate for providing the SIMPATICO EAB with the necessary instructions/authorizations/notifications. Only if necessary, the AEB will engage the Spanish Data Protection Authority, i.e., Agencia Española de Protección de Datos (AEPD - http://www.agpd.es/).

1.3. Open access and data management

The Consortium adheres to **the pilot for open access to research data (ORD pilot)** adopting an open access policy of all projects results, guidelines and reports, providing on-line access to scientific information that is free of charge to the reader [8]. Open access typically refers to two main categories: **scientific publication** (e.g., peer-reviewed scientific research articles, primarily published in academic journals) (Subsection 1.3.1) and **research data** (Subsection 1.3.2).

1.3.1. Open access to scientific publications

According to the European Commission, "under Horizon 2020, each beneficiary must ensure open access to all peer-reviewed scientific publications relating to its results" (see also Article 29.2 of the GA). The SIMPATICO Consortium adheres to the EU open access to publications policy, choosing as most appropriate route towards open access **self-archiving** (hereinafter also referred to as **'green' open access**), namely "a published article or the final peer-reviewed manuscript is archived (deposited) in an online repository before, alongside or after its publication. Repository software usually allows authors to delay access to the article ('embargo period')". The Consortium will ensure open access to the publication within a maximum of six months.
The dissemination of SIMPATICO results will occur by mean of the activities identified in the implementation plan, such as international publications, participation in international events (exhibitions, conferences, seminars, courses, etc.). In compliance with the Consortium Agreement, **free-online access will be privileged for scientific publication**, following the above-mentioned rules

of ‘green’ open access. All relevant information and the platform textual material (papers, deliverables, etc.) will be **also freely available on the project website**. In order to guarantee that also people who are visually impaired have access to all textual materials we will provide also **accessible PDF files**. In specific cases and according to the rules on open access set out below, the dissemination of research results will be managed by **adopting precautionary IPR protection tools**, in order not to obstacle with preventive disclosures the possibility of protecting the achieved foreground.

1.3.2. Open access to research data (Open Research Data Pilot)

According to the European Commission, “research data is information (particularly facts or numbers) collected to be examined and considered, and to serve as a basis for reasoning, discussion, or calculation”. Open access to research data is **the right to access and reuse digital research data** under the terms and conditions set out in the Grant Agreement.

Regarding the digital research data generated in the action, according to the Article 29.3 of the GA, the SIMPATICO Consortium will:

Deposit in a research data repository *and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:*

- *(i) the data, including associated metadata, needed to validate the results presented in scientific publications;*
- *(ii) other data, including associated metadata, as specified and within the deadlines laid down in this data management plan;*
- *(i) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results.*

In order to discuss the **public availability of data**, there are three different types of datasets within the SIMPATICO project: 1. not publicly available personal and sensitive data; 2. data treated according to open access policy of all project results; 3. data connected to Citizenpedia. These datasets will be discussed in detail in Section 2 below. Please note that a portion of the relevant data for SIMPATICO comes from **existing data sets of the PAs** (e.g., service usage data, citizens’ data), while **new data sources** will be defined by this deliverable and will be identified as a result of requirements analysis in SIMPATICO (see Section 2 below on SIMPATICO datasets). **Whenever possible**, these additional data sources will also be made available **as open data or through open services**. However, some of the collected data, in particular that concerning **user profiles and personal data**, is highly sensitive and will not be made available. All relevant information and the platform textual material (papers, deliverables, etc.) will be also **freely available on the project website**. In order to guarantee that also people who are visually impaired have access to all textual materials we will provide also **accessible PDF files.**

2. SIMPATICO datasets

This **Data Management Plan (DMP)** has been prepared by taking into account the current template of the "Guidelines on Data Management in Horizon 2020" [1]. The elaboration of the DMP will allow to SIMPATICO partners to address all issues related with data. An updated version will be created and submitted on Month 36 ("D1.4 – Data management plan v.2"). However, the DMP will be a **live document** throughout the project. This initial version will evolve during the project according to the progress of research activities. The Consortium will comply with the requirements of **Directive 95/46/EC** of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, as well as to the **national legislation** of SIMPATICO pilots (i.e., Italy, Spain, and the United Kingdom) in the field[1] [9] (see also "D1.5 – Ethics compliance report").

Special attention is given to the **security of data sources**, **services and interfaces** as well as to the **data protection and privacy of persons**, which are important aspects for stakeholders and components in SIMPATICO solution. In particular, privacy is managed by the **Citizen Data Vault (CDV)**. It represents the component that takes care of the storage of personal and sensitive data of the users, and of the exchange of these data with the users, the legacy systems of the PA, and components of the SIMPATICO Platform. Mechanisms for **encryption, authentication, and authorization** are exploited in the storage and exchange of data, so to ensure the satisfaction of core **security and data protection requirements**, namely confidentiality, integrity, and availability (for further details, please see Section 3).

In order to discuss the **public availability of data**, as outlined above, it is convenient to distinguish three different types of datasets within the SIMPATICO project:

1. **Not publicly available personal and sensitive data will be collected and processed as part of the execution of the SIMPATICO use-cases**, more specifically for the execution of the e-services. Specifically, the use-cases will involve only voluntary participants aged 18 or older and capable to give consent, who will be informed on the nature of their involvement and on the data collection/retention procedures through an informed consent form before the commencement of their participations. Informed consent will follow procedures and mechanisms compliant with European and national regulations in the field on ethics, data protection and privacy (see also deliverables "D1.5 – Ethics compliance report", "D8.1 – H – Requirement no. 1", and "D8.2 – POPD – Requirement no. 2").
2. **SIMPATICO adheres to the open access policy of all project results.** Specifically, we are committed to make available, whenever possible, the data collected during the execution of SIMPATICO, in particular data collected during the use-cases, also to researchers and other relevant stakeholders outside the project Consortium. Whenever possible, these additional data sources will also be made available as open data or through open services. In this context, any personal data will only be published after suitable aggregation and/or pseudonymization techniques have been applied, and after an informed consent that explicitly authorize this usage has been collected.

[1] Please note that on 4 May 2016, the official texts of the new Regulation and the Directive on EU data protection have been published in the EU Official Journal in all the official languages. While the Regulation entered into force on 24 May 2016, it shall apply from 25 May 2018. The Directive entered into force on 5 May 2016 and EU Member States will have to transpose it into their national law by 6 May 2018.

3. **SIMPATICO intends to build an open knowledge base on public services and processes through Citizenpedia**, released as a new public domain resource co-created and co-operated by the community (i.e., citizens, professionals and civil servants). The initial content of Citizenpedia will be based on datasets and other digital goods that are publicly available. In the case of datasets and other digital goods owned by the PAs and not already publicly available, the Consortium will pursue to obtain an authorization for public release, as open content, before inclusion in the Citizenpedia. For what concerns the data contributed to Citizenpedia by the community, SIMPATICO will require that they are made available as open content (e.g., with licenses such as Creative Commons).

This Data Management Plan and its updated versions describe **datasets characteristics** and **define principles and rules for the distribution of data** within SIMPATICO. In particular, in this first version of the DMP we present in details the procedures of creating **'primary data'** (i.e., data not available from any other sources) and of their management. As such, only the datasets named "SIMPATICO 01 - Citizenpedia Dataset" (Section 2.1), "SIMPATICO 02 - Logging/Feedback Dataset" (Section 2.2), and "SIMPATICO 03 – Citizen Data Vault (CDV) Dataset" (Section 2.3) are described in detail in the following sections, as any other datasets already exist and their creation is not foreseen in the GA. Specifically, the following subsections illustrate all SIMPATICO datasets taking into account:

- the need to balance openness and protection of scientific information;
- commercialisation and IPR;
- privacy concerns;
- security;
- data management and preservation questions.

2.1. SIMPATICO 01 - Citizenpedia Dataset

2.1.1. Dataset reference and name

The dataset reference and name is **“SIMPATICO 01 - Citizenpedia Dataset”**.

2.1.2. Dataset description

Citizenpedia is the **human computation framework** inside the SIMPATICO platform. Its aim is to be a place where citizens can find useful information regarding e-services and public administration. Thus, **most of the content will be created and consumed by humans**. It will be mainly stored in JSON format.

Citizenpedia is composed of **two main interactive parts** for the users, a Question Answering Engine and a Collaborative Procedure Designer. Thus, the typology of data is twofold:

1. **Question Answering Engine:** questions, answers, comments and terms/definitions, generated in the Question Answering Engine. All of them will be created, stored and retrieved in JSON format.
2. **Collaborative Procedure Designer:** diagrams representing procedures, and comments to these diagrams. The diagrams will be stored and encoded, in a computer processable manner, and not as a bitmap. Comments will be stored in JSON format.

Both types of data will be stored in the **same database**, within Citizenpedia.

Citizenpedia, along with the SIMPATICO platform, is intended to be deployed in **three different cities/regions of different countries** (i.e., Italy, Spain, and the United Kingdom). Each country speaks its own language (i.e., Italian, Spanish and English), and the human-generated data in each Citizenpedia will be in **different languages**. For that reason, we are using a different database in each pilot.

2.1.3. Standards and metadata

DEUSTO team, after an initial investigation, did not find a standard format for the storage/management of the generated data/metadata. Initially, we will follow the **structure used in already deployed human-computation platforms**, such as the ones reviewed in the deliverable “D4.1 Citizenpedia framework specification and architecture”. Throughout the course of the project, we will come to the **best representation of the gathered information**.

a) Data capture methods

As regards data capture methods, there are **two ways of creating content** in Citizenpedia:

1. **Using the web interface:** citizens/civil servants will use the platform and write the information using their browsers. Then, data will be stored in the Citizenpedia database.
2. **Programmatically, via a REST interface:** Citizenpedia will expose a REST API for other SIMPATICO components/third party applications to query/insert data in the system.

At the point of writing this document, there is **no fixed structure** defined for all the data that Citizenpedia will contain. In a more probable manner, the data will be stored in a **relational SQL database**, which means that the data will be stored structured in tables.

b) Metadata

We consider **two types of metadata** to be generated in Citizenpedia:

1. **Usage statistics:** this information will be created under demand. E.g. as an answer to the query "number of registered questions related to the Law XYZ/2015". Throughout the development of the project, we will decide at a later stage whether these statistics will be persistently stored or not.
2. **Indexing engine metadata:** an indexing engine included inside Citizenpedia, such as Apache SolR or ElasticSearch, will create this data. This metadata will be consumed by the indexing engine to provide better searching capabilities over plain-text data. Initially, we do not attempt to control or modify the creation and use of this metadata.

2.1.4. Data sharing

a) Method for data sharing

Method for data sharing is twofold:

1. **Human generated data:** first, human generated data (such as questions/answers/comments) will be shared publicly. It could be checked using the web interface or programmatically through a REST API. Given that this data is created by users, they will be warned in their first time using Citizenpedia that any content they make will be publicly available.
2. **Metadata:** second, as regards the metadata generated from the usage of data (such as statistics), some of the statistics (aggregated data) will be publicly available through the REST API, e.g., the number of questions related to a certain topic. The entire metadata will only be used for research purposes: in the case of the releasing some scientific publication from the usage data of Citizenpedia, information will be completely aggregated and/or pseudonymized.

b) Restrictions on sharing

What is described **in the previous section** applies here.

c) Data repository

All the information related to Citizenpedia (i.e., both user-generated data and metadata) will be storedin the **Citizenpedia internal database**. DEUSTO, as responsible of this WP within the SIMPATICO Consortium, will ensure to handle security and privacy issues, ensuring access to the internal database only via **secure connections** and using **access control systems**.

2.1.5. Archiving and preservation (including storage and backup)

The SIMPATICO Consortium and, in particular, DEUSTO (as responsible of this WP) consider **to retain generated data** during the length of the project. Statistical data can be retained longer, after the end of the project lifespan, for research purposes. If so, DEUSTO estimates no additional cost for this.

a) Preservation plan

If collected metadata and statistical data will be retained after the length of the project, DEUSTO has the infrastructure **to retain data safely**.

b) Resourcing

No need for additional resources is envisaged beyond the duration of the project to handle data.

2.2. SIMPATICO 02 - Logging/Feedback Dataset

2.2.1. Dataset reference and name

The dataset reference and name is **"SIMPATICO 02 - Logging/Feedback Dataset".**

2.2.2. Dataset description

The SIMPATICO project provides a series of interactive front-end components as depicted in the yellow blocks in the diagram below, i.e., the **user interaction and feedback analysis layer** of the SIMPATICO Platform (see Figure 2).
During the project's inception, it was proposed that valuable information would be generated by the users during this interaction. This was argued to occur by two different mechanisms:

- **Explicit information gathering**, e.g., asking users directly to assess their interaction after it has happened. This is widely done in the industry and can be performed by a number of different mechanisms.
- **Implicit information collection**, e.g., analysing metrics of interest in the interaction without requiring the users to be providing any extra information. As an example, upon the execution of a e-service request information about the time spent for each step may be collected and then analysed to find insights such as bottlenecks.

Both of these data generation sets were conceptualized in the platform's architecture (see Figure 3 above), as the blocks depicted in the red box with a dotted line. This includes **two data storage modules** for explicit and implicit data plus a data analysis step to generate new insights (e.g., statistics) from gathered data elements.

2.2.3. Standards and metadata

As of the writing of this document, the SIMPATICO team has not found relevant standards about the representation of these metadata. One of the tasks during the project set up phase (M1-M6) is to come up with a **concrete data model for the representation of such data elements**. Inspiration for this will be taken from common representation data models of usability evaluation such as the **System Usability Testing (SUS)** [10], and standards such as the **ISO 9241**[2] for desktop application ergonomics.

a) Data capture methods

The data will be collected in different ways according to the mechanism of information gathering (i.e., implicit/explicit), as explained above:

Explicit information gathering

- Questionnaires to the users with predefined ('canned') responses such as emoticons or "Likert scale" values.
- Open ended questions and free form responses, which can then be analyzed with the projects NLP tools:
  - Sentiment analysis to capture the general sentiment generated by the system.
  - Topic clustering to detect potential pain points or concerns of the users.

[2] ISO 9241-1:1997 "Ergonomic requirements for office work with visual display terminals (VDTs)".

Implicit information gathering

- Capture metrics such as click areas, time spent in different steps of the process.

b) Metadata

The metadata will be generated from the data in the section above using the following analysis steps:

Explicit information analysis

- Statistics about general feelings or ratings for particular areas or topics identified in the first stage of analysis.
- Summaries of multiple answers of open ended questions as proposed.

Implicit information analysis

- Statistical analysis of the data captured: average time spent by users, segmentation by age groups or target groups, etc.

2.2.4. Data sharing

a) Method for data sharing

The data and metadata generated by the module expected to be useful beyond SIMPATICO mainly to researchers due to the particularities of the application. **The sharing of corpora of the data** is envisaged at the end of the project as open data, after a suitable aggregation and/or pseudonymization is performed. Specific scientific data will also be used for **publications** of the Consortium members.

b) Restrictions on sharing

No sharing is envisaged at this stage of work.

c) Data repository

No sharing is envisaged at this stage of work.

2.2.5. Archiving and preservation (including storage and backup)

All storing and preservation procedures will be carried out internally to the project. At the end of the project data will shared as open data after a suitable aggregation and/or pseudonymization is performed.

a) Preservation plan

In principle, the results are expected to be very use-case specific and **no long-term storage is envisaged** beyond the needs for the SIMPATICO project execution, except the publication as open data discussed in the previous items.

b) Resourcing

No long term storing is envisaged beyond the duration of the project.

2.3. SIMPATICO 03 - Citizen Data Vault (CDV) Dataset

2.3.1. Dataset reference and name

The dataset reference and name is **“SIMPATICO 03 – Citizen Data Vault (CDV) Dataset”**.

2.3.2. Dataset description

The **Citizen Data Vault (CDV)** is a **repository of the citizen personal data, profile and information**. It is continuously updated through **each citizen interaction** and is used mainly to automatically fill e-service forms. In this way, citizens will give to the P.A. each information only once, as the information will be stored in the vault and used in all the following interactions. As regards the CDV, for personal data we will use the **definition provided by the World Economic Forum (June 2010)**, namely [11]:

"Personal data is defined as data (and metadata) created by and about people”, encompassing:

- **Volunteered data** – created and explicitly shared by individuals, e.g., social network profiles.
- **Observed data** – captured by recording the actions of individuals, e.g., location data when using cell phones.
- **Inferred data** – data about individuals based on analysis of volunteered or observed information, e.g., credit scores."

According to this definition, through the CDV **citizens have a practical mean to manage their personal data** with the ability to grant and withdraw **consent** to third parties for access to data about themselves (see “D1.5 – Ethics compliance report” – Annex I “Informed consent form”).

In summary, data collected by the means of CDV is referring on the context of personal data. In a first stage, we have identified a **first categorization of such personal data**, referring to:

1. Government Records
2. Profile
3. Education
4. Relationship
5. Banking and Finance
6. Health
7. Communication & Media
8. Energy
9. Mobility
10. Activities

For each category **several data fields** are going to be defined. This is a first version of the “**Personal Data Category”**, and it will be refined, reduced or modified according to the actual personal data that each citizen could manage by the means of CDV against the **three use cases** identified by the three SIMPATICO pilots (i.e., Trento, Sheffield, and Galicia).

The personal data collected or linked by the CDV will **never be shared at any time**. **Each citizen** has the control and the ability of **removing all data from CDV**.

2.3.3. Standards and metadata

CDV will collect personal data with a reference to a specific element of **Personal Data Taxonomy**. In order to assure semantic interoperability several options and tools are going to be considered, in particular **RDF and Linked Data**[12], **XML and JSON**.

a) Data capture methods

Personal data will be collected in two ways:

1. **Data will be inserted by citizens by the means of the CDV dashboard.** The user will be able to insert, collect and modify personal data fields by the means of interactive web forms provided by the CDV.
2. **Data will be collected during the interactions of the user with the e-service forms provided by the PAs.** The e-services and the related types of data will be the ones identified by the three pilots. During each interaction, users decide if the data inserted in the e-service forms can be stored in the CDV.

At any time, users can **view (through the dashboard)**, and possibly **remove the data collected**. No version of data collected is provided. Thanks to the approach used to collect data, the stored information can be **retrieved by using Personal Data Taxonomy**.

b) Metadata

As reported above, each information will be related to a **specific personal data category** and **data field**. Personal Data Taxonomy can be accessed **in JSON and RDF format**.

2.3.4. Data sharing

The personal data collected or linked by the CDV will **never be shared at any time**. Therefore, this subsection will only provide information on **data repository**.

a) Method for data sharing

N/A

b) Restrictions on sharing

N/A

c) Data repository

The CDV will provide an **ad-hoc repository to collect personal data**, adopting a **multiple key based data encryption**. According to the specific deployment strategy and the e-services that will be adopted in each use case, the CDV could refer to **multiple data stores** (i.e., legacy systems just provided by the PAs).

2.3.5. Archiving and preservation (including storage and backup)

SIMPATICO Project considers **to retain the collected personal data only for the lifetime of the grant**.

a) Preservation plan

In principle, the results are expected to be very use-case specific and **no long-term storage is envisaged** beyond the needs for the SIMPATICO project execution.

b) Resourcing

In principle, the results are expected to be very use-case specific and **no long-term storage is envisaged** beyond the needs for the SIMPATICO project execution.

3. SIMPATICO security protection strategy

This final section is dedicated to the **SIMPATICO security protection strategy** and will develop as the project progresses. It reflects the current status within the Consortium about the security of data that will be collected and produced. In the SIMPATICO project **we do not perform activities, neither produce results, raising any large scale security issues**. The project does not have the potential for military applications, and also does not involve the use of elements that may cause any harm to humans, animals, plants or environment. However, the process of collecting, processing, storing data might hide some pitfalls. To reduce the **risk of potential malevolent, criminal and/or terrorist abuse**, which might be perpetrated also by malicious people authorized to access the information, the SIMPATICO Consortium is examining the deployment of a **twofold security protection strategy**:

1. by ensuring that the employed **security layers and privacy-preserving measures** will work properly, keeping access logs and following best practices for system administration;
2. by employing techniques to prevent information leakage “on-the-fly”, i.e., through the adoption of the **aggregation and pseudonymization approach** of personal and sensitive information at collection, communication, and storage time (e.g. via an encryption scheme, hash functions, and/or tokenization). Such an approach will neutralise eavesdropping and/or similarly dangerous hack attempts in the unlikely event of successful retrieval, since it will secure data, making them completely meaningless to the possible attacker.

3.1. Authentication, authorization, and encryption

State-of-the-art mechanisms for **authentication, authorization, and encryption** will be exploited in the implemented processes (concerning data collection, storage, protection, retention and destruction), so to ensure the satisfaction of core security and data protection requirements, namely **confidentiality, integrity, and availability**. In context of SIMPATICO, the crucial legal challenges are primarily the security measures concerning authentication and authorization issues: pursue to the above-mentioned **Directive 95/46/EC** the implementation of both computerized authentication and procedures for managing authorization credentials is required. To assure the security of and the trust in the system, it is fundamental to provide technical solutions aimed at allowing the **circulation of digital identities** and the **access to the e-services**. For identity management and data protection mechanisms, SIMPATICO will follow the standard practice in the security research community.

Identity management deals with identifying individuals (**authentication**) and controlling access (**authorization**) to resources in a system. All the Privacy Enhancing Technologies associated with identity management aim at identity verification with minimum identity disclosure, and protection against identity theft. Due to internetworked services and in general to Cloud technology, the need of a secure identities management has grown increasingly. Identity and access management (IAM) is the security and business discipline that “enables the right individuals to access the right resources at the right times and for the right reasons”. It addresses the need to ensure appropriate access to resources across increasingly heterogeneous technology environments and to meet increasingly rigorous compliance requirements. Technologies, services and terms related to identity management will be exploited including Directory services, Digital Cards, Service Providers, Identity Providers, Digital Password Managers, Single Sign-on, JSON Web Token and JSON Web Key from OpenID Connect’s model, OpenID Connect , OAuth and XACML. In particular SIMPATICO’s solutions fo IAM are, and will be, influenced by many existing and upcoming standards: OAuth 2.0, User Managed

Access (UMA) and OpenID Connect as well as the upcoming Minimum Viable Consent Record (MVCR) specification from Kantara Initiative.

According to the “**Privacy by default and by design**” principles, the SIMPATICO platform will adopt an integrated and multilevel approach to protect the user information from the fraudulent access and consumption. In order to ensure the data consumption only by authorized applications and users. The open standard for authorization Oauth protocol ensures the exchange of personal data in a trusted context. In particular, SIMPATICO adopts two levels of security for personal data: data transport level and data storing level. In order to allow the secure transmission of personal data, the SIMPATICO APIs support the HTTPS communication protocol. The input and output data are transmitted as “plain text” over HTTPS and encrypted by the TSL (Transfer Security Layer), or by the SSL (Security Socket Layer). HTTPS is based over certificates and ensure the client and server mutual authentication.

For the CDV component of SIMATICO, which is the main storage of personal data in the project, particular attention is dedicate to reduce the **server side vulnerabilities,** applying all the best **security practices and policies** about the configuration of the user privileges, remote access and connections. In order to get the database unreadable by unauthorized users/applications, the CDV architecture includes a module named Data Security Manager (DSM) that, implementing the Transparent Data Encryption (TDE) approach, enables the encryption/decryption of the CDV data in transparent way from users and application point of view. In order to distribute the security knowledge about the encryption keys and increase the data security, the CDV keys and encrypted data will be periodically backuped and stored in different places. According to the best practice and the architectural solution adopted by the most important DBMS, see Oracle[3] and Microsoft SQL Server[4], the CDV TDE implementation is based on the following concepts:

- **Master Key**: a key adopted to encrypt the Keys Table. It will be stored into a read only file in the filesystem and access restricted exclusivelly to each single user registered in the CDV.
- **User Key**: a key associated to a single CDV user.
- **Keys Table**: a table to store the User Keys. It will be located in a different server than the Master Key and the Personal Data Table one.
- **Encryption Key**: a key generated using the Master Key and the User Key.
- **AES Cipher Algorithm**: the CDV adopts the Advanced Encryption Standard (AES) at 128 bit, defined in the Federal Information Processing (FIPS) standard no. 197[5].
- **Personal Data Table**: it will contain the personal data encrypted/decrypted applying the AES and the Encryption Key.

[3] Transparent Data Encryption (TDE) adopted by Oracle: http://www.oracle.com/technetwork/database/security/twp-transparent-data-encryption-bes-130696.pdf

[4] Transparent Data Encryption (TDE) adopted by Microsoft: https://msdn.microsoft.com/en-us/library/bb934049.aspx

[5] National Institute of Standards and Technology (NIST) , Federal Information Processing (FIPS) standard no. 197, http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

3.2. Focus on data aggregation and pseudonymization techniques

Personal and sensitive data will be made publicly available only after an **informed consent** has been collected and **suitable aggregation and/or pseudonymization techniques** have been applied. Before starting the project activities that require user involvement, a careful investigation on privacy and security issues has been and will be undertaken, covering in particular **Italian, Spanish and UK privacy laws**, according to the procedures stated in deliverable "D1.5 – Ethics compliance report". In this Data Management Plan, data pseudonymization and aggregation techniques will be identified and applied to personal/sensitive data before their public release.

As regards aggregation techniques, data confidentiality, integrity and privacy will be assured **when collecting and processing data**. The information for each person contained in the release cannot be distinguished from a given number of other individuals whose information also appear in the release. Moreover, the pseudonymization of data is another method of ensuring confidentiality, according to **the Article 29 Working Party Opinion on Anonymization Techniques** and in relation to the upcoming EU General Data Protection Regulation [13]. Where data are particularly sensitive (e.g. data using detailed personal narratives) then risks to confidentiality increase. In this case, participants will be carefully informed of the nature of the possible risks. This does not preclude the responsibility of the applicant to ensure that maximal pseudonymization procedures are implemented. A detailed description of the measures that will be implemented to prevent improper use, improper data disclosure scenarios and 'mission creep' (i.e., unforeseen usage of data by any third party), within the above-mentioned security protection strategy, will be provided before the commencement of validation activities as update of this deliverable.

The optimal solution will be decided by using a combination of different techniques, while taking into account the practical recommendations developed in the above-mentioned **Article 29 Working Party Opinion on Anonymization Techniques**. Pseudonymization approaches reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure. These techniques have to adhere certain requirements to comply with data protection and privacy-related legislation in the EU [14]. The following set of requirements (among others) has been extracted from the Directive 95/46/EC and the Article 29 Working Party Opinion on Anonymization Techniques and will be the guidelines for security protection strategy drafting [13] [15]:

- **User authentication:** the system has to provide adequate mechanisms for user authentication.
- **Limited access:** the system must ensure that data is only provided to authenticated and authorized persons.
- **Protection against unauthorized and authorized access:** the records of an individual have to be protected against unauthorized access.
- **Notice about use of data:** the users should be informed about any access to their records.
- **Access and copy users' own data:** the system has to provide mechanisms to access and copy the users' own data.
- **Fall-back mechanism:** the system should provide mechanisms to back up and restore the security token used for pseudonymization.
- **Unobservability:** pseudonymized data should not be observable and linkable to a specific individual in the system.
- **Secondary use:** the system should provide a mechanism to export pseudonymized data for secondary use and a possibility to notify the owner of the exported data.
- **Modification of the database:** if an attacker breaks into the system, the system must detect modifications and inform the system administrator about this attack.

The above-mentioned potential “unforeseen usage” implications of this project will be examined by the SIMPATICO Ethics Advisory Board (see “D1.5 – Ethics compliance report”).

4. References

[1] European Commission, “Guidelines on Data Management in Horizon 2020 [Version 2.1. of 15 February 2016],” European Commission, Brussels, 2016.

[2] European Group on Ethics in Science and New technologies (EGE), “Opinion No. 26 of 22 February 2012 on “Ethics of information and communication technologies”,” European Group on Ethics in Science and New technologies (EGE), 2012. [Online]. Available: http://ftp.infoeuropa.eurocid.pt/database/000049001-000050000/000049238.pdf. [Accessed 5 May 2016].

[3] Council of Europe, Handbook on European data protection law, Luxembourg: Publications Office of the European Union, 2014.

[4] European Commission, “H2020 - How to complete your ethics Self-Assessment,” European Commission, 2014. [Online]. Available: http://ec.europa.eu/research/participants/portal/doc/call/h2020/h2020-msca-itn-2015/1620147-h2020_-_guidance_ethics_self_assess_en.pd. [Accessed 20 April 2016].

[5] Garante per la protezione dei dati personali, “Italian Legislation – Data Protection Code,” 2016. [Online]. Available: http://www.garanteprivacy.it/web/guest/home_en/italian-legislation#1. [Accessed 9 May 2016].

[6] ICO - Information’s Commissioner Office, “Protecting personal data in online services: learning from the mistakes of others,” 2014. [Online]. Available: https://ico.org.uk/media/for-organisations/documents/1042221/protecting-personal-data-in-online-services-learning-from-the-mistakes-of-others.pdf.. [Accessed 2 May 2016].

[7] Agencia Española de Protección de Datos, “Spanish Legislation – Data Protection Act,” 2016. [Online]. Available: http://www.agpd.es/portalwebAGPD/canaldocumentacion/legislacion/index-iden-idphp.ph. [Accessed 6 May 2016].

[8] European Commission, “Open access & Data management,” European Commission, 2016. [Online]. Available: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm. [Accessed 15 June 2016].

[9] European Commission, “Reform of EU data protection rules,” 4 July 2016. [Online]. Available: http://ec.europa.eu/justice/data-protection/reform/index_en.htm.

[10] J. Brooke, “SUS: a "quick and dirty" usability scale,” in *Usability Evaluation in Industry*, London, Taylor and Francis, 1996.

[11] World Economic Forum, “Rethinking Personal Data: Strengthening Trust,” 2012. [Online].

[Accessed 14 July 2016].

[12] W3C, “Linked Data,” 2016. [Online]. Available: https://www.w3.org/standards/semanticweb/data. [Accessed 14 July 2016].

[13] Article 29 Working Party, “Opinion no. 5 of 10 April 2014 on Anonymization Techniques,” 2014. [Online]. Available: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf. [Accessed 3 May 2016].

[14] Article 29 Data Protection Working Party, “Opinion No. 3 of 2 April 2013 on “Purpose limitation”,” [Online]. Available: http://ec.europa.eu/justice/dataprotection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.. [Accessed 4 July 2016].

[15] N. &. Kolb, “A Legal Evaluation of Pseudonymization Approaches,” *International Journal on Advances in Security,* vol. 2, no. 3, pp. 190-202, 2009.