European Security in Health Data Exchange

Deliverable D1.5

Data Management Plan

Editor(s):	Xabier Larrucea, Jason Mansell, Alberto Berreteaga
Responsible Partner:	TECNALIA
Status-Version:	1.0
Date:	30/06/2017
Distribution level (CO, PU):	PU

Project Number:	GA 727301
Project Title:	SHIELD

Title of Deliverable:	Data Management Plan
Due Date of Delivery to the EC:	30/06/2017

Workpackage responsible for the Deliverable:	Tecnalia
Editor(s):	Tecnalia
Contributor(s):	Matthias (Stelar), Eleonora (FCSR)
Reviewer(s):	Antony (AIMES)
Approved by:	AIMES
Recommended/mandatory readers:	All WP's

Abstract:	This document specifies what the kind of data generated by the project is and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved.
Keyword List:	Data Management Plan, OpenAire, metadata
Disclaimer	This document reflects only the author's views and neither Agency nor the Commission are responsible for any use that may be made of the information contained therein

Document Description

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	18/05/2017	First draft version	Xabier Larrucea, Jason Mansell, Alberto Berreteaga (TECNALIA)
V0.2	19/06/2017	Contributions from FCSR	Eleonora Lavalle
V0.3	20/06/2017	Contributions from Stelar	Matthias Pocs
v0.4	18/05/2017	Compiled draft	Xabier Larrucea, Jason Mansell, Alberto Berreteaga (TECNALIA)
V0.5	23/06/2017	Osakidetza inputs	Eunate Arana
V1.0	27/06/2017	reviewed	Antony Shimmin

Table of Contents

List of Figures

List of Tables

[there is no tables]

Terms and abbreviations

DMP	Data Management Plan
EC	European Commission
GDPR	General Data Protection Regulation
KPI	Key Performance Indicators
KR	Key Results
PA	Public Administrations
WP	Work Package

Executive Summary

The objective of this deliverable is to present a data management plan for the SHIELD project. This document covers a wide set of activities such as data collection, generation, storage and preservation. In this action, we envision five different types of data: data related to the use cases, data coming from publications, public deliverables and open source software.
The document presents, following the EC template [1], how these different types of data will be collected, who the main beneficiaries are, and how SHIELD will store them, manage them, and make them accessible, findable and re-usable. The text continues with the foreseen resources needed for the openness and data to finalize with security and ethical aspects that will be taken into consideration in the context of SHIELD.

This plan is the first version of the data management plan, which will be updated in subsequent versions (M18 and M36) as part of the Technical Reports, having as input the work carried out in the use cases (WP6), the social and technical work packages (WP2 – WP5) and the dissemination activities (WP7).

1 Introduction

1.1 About this deliverable

This deliverable focuses on the management of the data in SHIELD. In this context there are two different types of data: those related to the publications generated as part of research activities, and those related to the data collected from citizens, users and non-users of digital public services, as well as from civil servants, that will be used as part of the implementation of the different key results established in the project.

1.2 Document structure

The document follows the established H2020 template for a Data Management Plan (DMP) [1]. Section 2 presents the data summary of what the purpose of the data collection and generation is. Section 3 explains how the data will be made fair, and thus findable, accessible, interoperable and reusable. Section 4 briefly explains how the financial resources for this openness are envisioned at this stage to be allocated. Section 5 and 6 outline security and ethical aspects respectively. And finally Section 7 presents the conclusions and future work.

2 Data Summary

2.1 Purpose of the data collection/generation and its relation to the objectives of the project

The following list of SHIELD's project objectives and related key results (KR) provides a description for each KR specifying the purpose of the data collection/generation (if any):

- **(O1) Systematic protection of health data against threats and cyber-attacks**.
    - **KR01: Knowledge base of generic security issues that may affect a system**. The purpose is to create a knowledge base which captures threats that should be managed by the architecture and regulatory data protection requirements (supporting objective O4). This knowledge base does not capture nor user's health data nor users, and it only manages threats and compliance issues in specific end-to-end applications. For the SHIELD use cases we will use fake data just to prove the benefits of the results.
    - **KR02: Tool that provides an automated analysis of data structures in order to identify sensitive elements that may be vulnerable to specific threats.** Data structure used to have flaws and weaknesses during the storage or exchange of data. The purpose is to analyse/collect the schema of these structures. SHiELD pilots will be used to identify sensitive data, and it will be traced during the pilots to ensure its privacy aspects and that access rights requirements are kept.
    - **KR03: Security requirements identification tool**: this tool will allow models of end-to-end applications to be created, and security threats and compliance issues affecting that application to be automatically identified. We will just list security threats and compliance issues according to 'security by design' principles.
- **(O2) Definition of a common architecture for secure exchange of health data across European borders.**

  - **KR04: SHiELD open architecture and open secure interoperability API:** the purpose is to create a SHiELD architecture which is composed by the results of epSOS project but also with tools brought by SHIELD partners such as the anonymisation mechanisms. Furthermore the health data interchanged is fake, and we do not use real users data. SHIELD pilots will invent users for each scenario. Basically the approach is to allow citizens and healthcare providers the possibility for accessing their health data from other countries
  - **KR05: SHiELD (Sec)DevOps tool:** the purpose is twofold. During development time, a set of architectural patterns (mainly in Java) are stored in order to check data protection security mechanisms. During run time a set of tools provide monitoring facilities alerting the operator of the system that a threat is likely to occur.
- **(O3) Assurance of the protection and privacy of the health data exchange.** This objective is addressed mostly in WP5, led by IBM based on their expertise in novel data security mechanisms for securing the exchanged data among the different Member States. This data is protected before, during and after it is exchanged.
  - **KR06: Data protection mechanisms**: the purpose is to collect a suite of security mechanisms to address data protection threats and regulatory compliance issues in end-to-end heterogeneous systems. This includes (but not limited to) tamper detection for mobile devices, data protection mechanisms, and consent-based access control mechanisms.
  - **KR07: Privacy protection mechanisms:** these privacy mechanisms address different aspects of privacy protection and regulation of data. These include methods for sensitive information identification. The purpose is to use and develop methods to mask private sensitive information dynamically on the fly as well as methods able to anonymize data while enabling analysis on the data.
- **(O4) To understand the legal/regulatory requirements in each member state, which are only partly aligned by previous EU directives and regulations and provide recommendations to regulators for the development of new/improved regulations**.
  - **KR08: Legal recommendations report.** For this KY we are not going to use private data. The purpose is to create a common regulatory framework where the legal requirements regarding security among the state members are aligned.
- **(O5) Validation of SHiELD in different pilots across three Member States**
  - **KR09: Pilots:** the purpose is to test implementations which are deployed in three Member States, supporting validation scenarios defined. The collected data will be used to prove that scenarios are working.
  - **KR10: Best practices:** the purpose of the data used is to describe lessons learned and best practices for protecting health data
- **(O6) Dissemination of SHiELD results**
  - **KR11: Publications:** the purpose is to collect the scientific papers, white papers, popular press articles, media and social media we are producing**.**
  - **KR12: Take up opportunities: its purpose is to identify the main** users, standards bodies and regulators.

2.2 Types and formats

- During these first 6 months we are just considering the format suggested in [2] we are considering a Patient Summary as an identifiable "dataset of essential and understandable health information" that is made available "at the point of care to deliver safe patient care during unscheduled care [and planned care] with its maximal impact in unscheduled care"; it can also be defined at a high level as: "the minimum set of information needed to assure healthcare coordination and the continuity of care" [2]. From a technical point of view, we will use as readable formats such as CSV, XML or JSON. Examples of the XML format are described in [3] which is the official Metada Registry. SHIELD project manages structured and unstructured data collected during current and past patient hospitalizations. For example:**Structured data** refers to kinds of data with a high level of organization, such as information in a relational database. For example:
    - SDO (discharge form) that contains 5 .txt files where each field is separated by ";"
    - ED (Emergency Department) dataset.
    - Medical images
    - Treatment forms.
    - Constant collection forms.
- **Unstructured data** refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. Examples of are:
    - Reports of complementary tests (radiology, pathological anatomy, endoscopy, etc.)
    - Discharge letter.
    - Monitoring of evolutions in external consultations.

Both documents are typically uploaded in PDF format.

2.3 Re-use of existing data

We will reuse the existing and available data provided in epSOS (http://www.epsos.eu/home/epsos-results-outlook.html) just to check the feasibility of the solutions provided in SHIELD.

2.4 Origin of the data

The data is based on the scenarios provided in SHIELD, and more precisely on the three member states requirements (UK, Italy, Spain (Basque country)). The data used in the scenarios are not real, and do not belong nor describe any individual. Additionally we follow the principle of informed consent (see section 6 "Ethical aspects"). The use of these data can help us to demonstrate the technology developed.

2.5 Expected size of the data

At this stage of the project it still hard to define precisely the data size and ingestion rate. However, it can be useful to go into details regarding the dimension of the most important data involved in the use cases:

- **Medical images**: include all the bio-images such as ultrasound scan, MRI (magnetic resonance imaging) or CT (computer tomography) scan. Considering that the **Computerized Tomography** uses 3D x-rays to make detailed pictures of structures

inside of the body, it takes **pictures in slices**, like a loaf of bread. This means that each slice it's a picture, the number of pictures can be from 30 for simple examinations to 1000+ for sophisticated examinations. This scan can be repeated several times (2-6) to reduce noise and to ensure high quality of the exam. In conclusion we will have from 30 to 1000 images each one with the size of 5 MB times 2-6 series; we can say that a single CT examination for a patient will have the size of 300 MB – 30 GB depending on the kind of investigation;

- **SDO and ED dataset**: is around 1 Kb per patient since any images are included and since the information is codified (.txt format).

2.6 To whom might it be useful ('data utility')?

The results of SHIELD might be useful for healthcare providers, governments, and patients.

3 FAIR DATA

This data management plan follows the FAIR (Findable Accessible Interoperable Reusable) principles.

3.1 Data findable

There are different types of data:

- Data related to the use cases
- Data coming from publications
- Data coming from public deliverables
- Open source software

3.1.1 Data related to the use cases

During the lifetime of the project and especially during the trials execution, SHiELD partners expect several types of data to be generated, mainly health data, location data, personal data ("fake" names, addresses, contact information, IP addresses, etc.), pseudonymised data (user names, device identifiers, etc.), traffic data, as well as others.

The first step in development of the use case studies will be to produce a high level outline of the scenario to be used in the project. Starting from epSOS data exchange gateway, a set up for subsequent validation experiments will be deployed. Since these experiments will involve some novel security mechanisms whose value is not yet proven, current patient data will not be used directly in the use cases. Instead, an equivalent test system will be implemented by using synthetic patient data to verify that security is effective without compromising the data exchange interoperability requirements and that SHiELD solutions are compliant to European General Data Protection Regulation 679/2016.

The second step of the project will see the creation of synthetic data sets which may be sampled or combined randomly and associated with fictitious Patients.

This synthetic set of medical information will include the minimum patient summary dataset for electronic exchange developed in the epSOS project [4] defined according to the clinical point of view keeping in mind the medical perspective of the final users (medical doctors and patients).

SHIELD WP6 deliverable 6.1 describes a set of scenarios, and all digitalized data included in Electronic Health Records (EHR) includes as example:

- Patient's personal data
- Medical histories & Progress notes
- Diagnoses
- Acute and chronic medications
- Allergies
- Vacinations
- Radiology images
- Lab and test results
- Clinical parameters (blood pressure, heart rate, capillary glucose,…)

For each scenario it is going to be necessary to stablish which are going to be the minimum clinical data in order to try to manage the patient in the most efficient way. On the one hand, it will be necessary to establish the sensitivity and security of the data, but on the other hand it is essential to provide the health professionals with the minimum imprescriptible data in order to perform an efficient management and also provide security in the management of the patient. One of the scopes of SHiELD, is to stablish the minimum data necessary for each scenario just to improve the clinical management of foreign patients while traveling along Europe.

In this way we need to:

- Denominate the fields to include, its format and range of values that can adopt.
- The classification of the field as part of the minimum set or if its inclusion is recommended, corresponding to each Health Service the final decision to include it or not.
- Inclusion of the field and its value as part of the attributes of the document as a "tag" to identify the essential elements of its content without having to open (decrypt) the document.

To codify different fields of the minimum dataset that will be exchange we have:

- **SNOMED CT** or **SNOMED Clinical Terms:** is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. The primary purpose of SNOMED CT is to encode the meanings that are used in health information and to support the effective clinical recording of data with the aim of improving patient care. SNOMED CT provides the core general terminology for electronic health records. SNOMED CT comprehensive coverage includes: clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimens. between different Health Systems we have:
- **ICD-10** is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems, a medical classification list by the World Health Organization. It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. The code set allows more than 14,400 different codes and permits the tracking of many new diagnoses. The codes can be expanded to over 16,000 codes by using optional sub-classifications.

This is just a brief list of medical data; indeed, it represents only a subset of the whole set of medical information that could be involved in SHiELD project. For documents including

sensitive information we will remove and hide all data. For example, Figure 1 represents a dismissal letter in which sensitive information can be found. It can be seen that in the discharge letter the sensitive information are **name and surname** of the patient and a **clinical record number** that is bound to the patient. All these data will be removed.

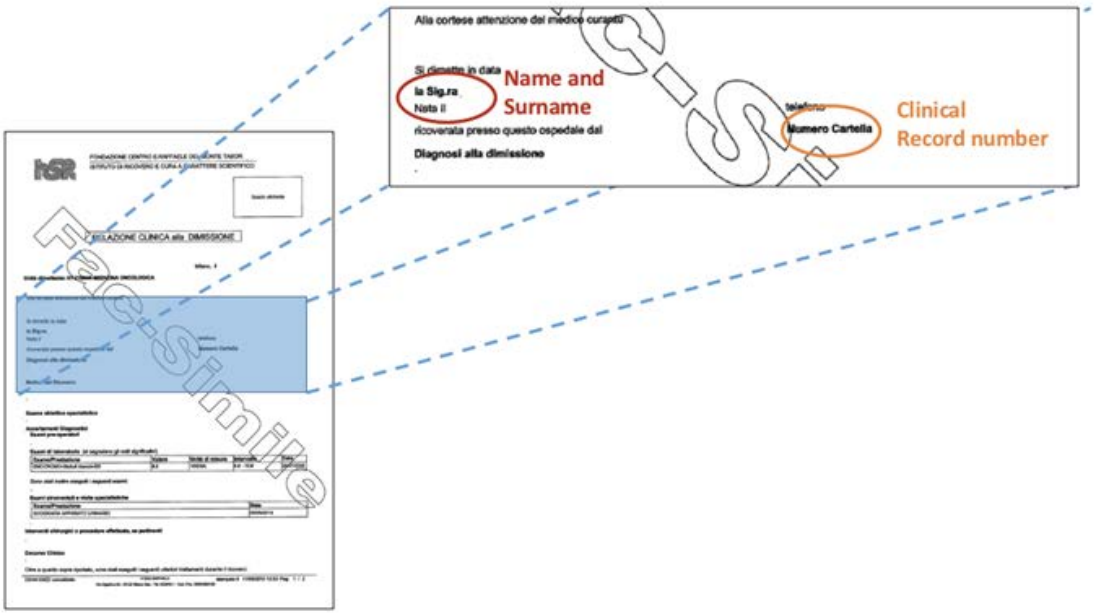


Figure 1: Document containing clinical record number and name.

Regarding medical images, Figure 2represents a slice of a simulated patient. Within this figure some sensitive information are circled in blue:

- **FANTOCCIO** is the space dedicated to the patient name and surname;
- **PID** is the internal patient ID, it means that the code identifies the patient within hospital internal system;
- **Acc.num** is a progressive number in the hospital internal system;

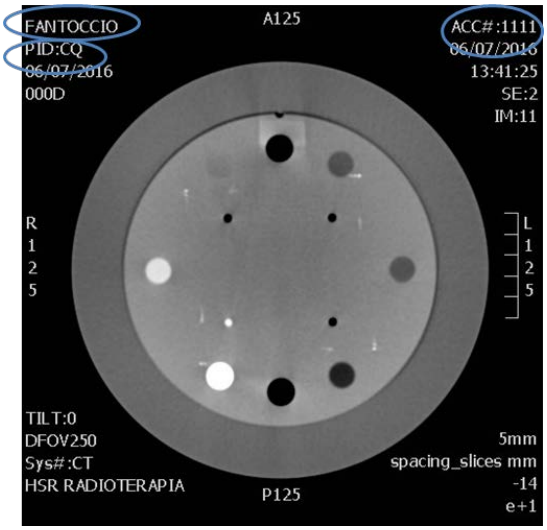


Figure 2: Slice of a simulated patient

Additional to synthetic data regarding to patient past hospitalization, SHiELD project can include mobile data that can be useful for diagnostic purposes. Data could come from both mobile and wearable devices; some examples of datasets are provided:

- GPS tracks (*e.g.* localization);
- Posts (*e.g.* social registrations);
- Last known activities:
    - SMS sent at time XX.XX;
    - Weather data;
    - Activity tracker
- Chronic patient monitoring;
- Drug therapy.

This data coming from wearable devices are not directly health related but allow for health related conclusions after processing.

3.1.2 Metadata

All publications will be indexed by using Digital Object Identifiers or similar mechanisms to be discovered and identified. All papers in journals and magazines will use this identifier.

Concerning the naming convention, we will use the following: <<Dx.y Deliverable name _ date in which the deliverable was submitted.pdf>>.

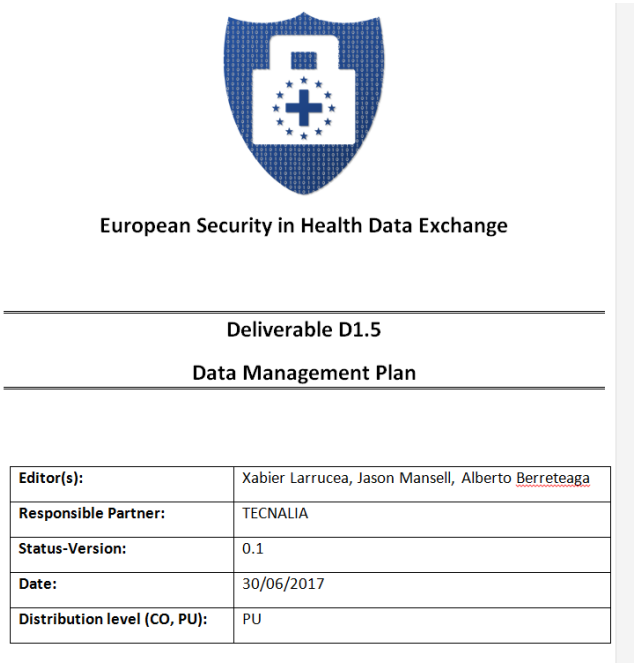
Each paper or deliverable contains a keywords section that can be used to optimize possibilities for re-use.

Each deliverable is tagged with a clear version number as indicated on Figure 1, Figure 2 and Figure 3. This is part of the metadata that each deliverable contains. Additionally

- Editor(s): who is/are the main leaders of this document
- Responsible Partner: who is/are the main responsible partner of this document
- Status-Version: draft, released, final
- Date: submission date
- Distribution level (CO, PU): confidential or public access according to SHIELD proposal
- Project Number: SHIELD project number
- Project Title: SHIELD title
- Title of Deliverable
- Due Date of Delivery to the EC: date to be sent to the European Commission (EC)
- Workpackage responsible for the Deliverable
- Editor(s):Who edit this deliverable
- Contributor(s): who have contributed
- Reviewer(s): reviewers
- Approved by: people who internally approved it to be submitted to EC
- Recommended/mandatory readers
- Abstract: it summarises this document
- Keyword List: a set of words which can provide an overview of the topic of this deliverable

- Disclaimer: copyrights if any

Each document registers its revision history: version number, data, reason for the modification, and by whom it is modified.



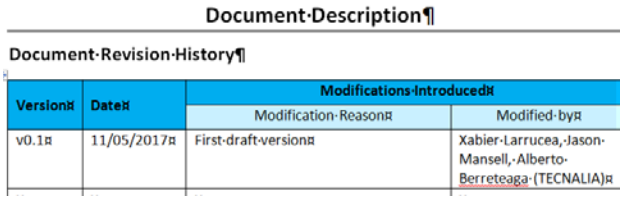
European Security in Health Data Exchange

Deliverable D1.5

Data Management Plan

Editor(s):	Xabier Larrucea, Jason Mansell, Alberto Berreteaga
Responsible Partner:	TECNALIA
Status-Version:	0.1
Date:	30/06/2017
Distribution level (CO, PU):	PU

Figure 3: Deliverable front page where version is shown



Document·Description¶

Document·Revision·History¶

Version¤	Date¤	Modifications·Introduced¤	
		Modification·Reason¤	Modified·by¤
v0.1¤	11/05/2017¤	First·draft·version¤	Xabier·Larrucea,·Jason·Mansell,·Alberto·Berreteaga·(TECNALIA)¤

Figure 4: Document description contains version number

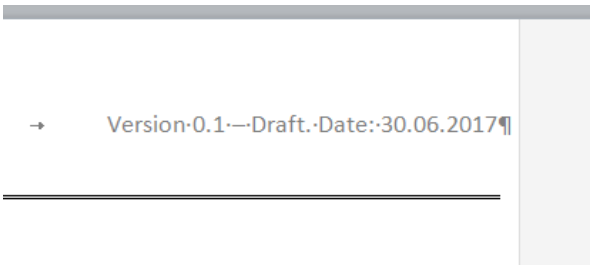


Figure 5: Page headers contains version number

3.2 Data openly accessible

Data related to the use cases are going to be accessible through the SHIELD deliverables which will be published on the website (http://www.project-shield.eu/). All deliverables include a set of keywords and a brief description that are aimed to facilitate the indexing and search of the deliverables in search engines. Scientific publications are going to be published as Open Data, we will use Open Aire [3] – compliant repositories. For example TECNALIA use its own repository, already indexed by Open Aire. There are other repositories such as Zenodo [4] that can be used. The deliverables will be stored at AIME's hosting provider, and for three years beyond the duration time frame of the project

All data produced will be made available through the use of deliverables, papers in journals/magazines/conferences, or repositories. The data used for proving functionalities are not real, and they are going to be distributed using open source repositories, which will be easily accessible by using a browser.
According to the SHIELD Grant Agreement (GA) page 15 "The SHiELD DevOps and solution will be as open source as possible (taking into account exploitation plans and the IPR issues that might arise from the usage of proprietary background)". But basically all tools are following a freemium licensing schema, where there is a public version that can be released as open source and a commercial edition. All these software will be released at the end of the project, because they are going to be mature enough.
At this moment, there is no specific arrangements, restrictions of use (apart from GA), there is no data access committee, and licenses depend on each tool used in SHIELD.

3.3 Data interoperable

Basically SHIELD project will produce a platform based on OpenNCP [6] which is interoperable with other software. The structures used for data exchange follow the eHealth DSI Interoperability Specifications [7]. Most of the vocabularies used follow the traditional software engineering artefacts descriptions, and for the eHealth domain we are using the HL7 [8] which specifications do not have a cost.

3.4 Increase data re-use (through clarifying licences)

Data stemming from the use cases will be delivered through the appropriate deliverables. Our approach is to extend a branch of the OpenNCP, and to add SHIELD functionalities. Once we have finalised the project we integrate these functionalities to the OpenNCP community, and this community will maintain this platform. At the time of writing, we do not envision any embargo on data.

4 Allocation of resources

SHIELD does not envision additional resources for handling data management. SHIELD will use open access repositories as much as possible for the following data:

- data related to the use cases
- data related to the meta-analysis
- data coming from publications
- data coming from public deliverables
- open source software

Obviously there is an indirect cost for making data FAIR in our project. But we consider as part of the activities of the SHIELD project. All partners in the SHIELD project are responsible for data management.

5 Data security

SHIELD will ensure that the General Data Protection Regulation (GDPR), which will enter into force in May 2018, is ensured, especially in regards to protection of private data. In addition SHIELD project provides the following key results dealing with data security:

- [KR03] Security requirements identification tool
- KR04] SHiELD open architecture and open secure interoperability API
- [KR06] Data protection mechanisms: a suite of security mechanisms that address data protection threats and regulatory compliance issues in end-to-end heterogeneous systems
- [KR07] Privacy tool: it monitors the data access attempts to ensure that only valid requests are accepted and only the data that is really needed is provided

6 Ethical aspects

The basis of ethical research is the principle of informed consent as stated in our proposal. All participants in SHIELD use cases will be informed of all aspects of the research that might reasonably be expected to influence willingness to participate. Project researchers will clarify questions and obtain permission from participants before and after each practical exercise (e.g. interview, co-creation session, etc.) to maintain on-going consent. Participants will be recruited by each organization leading the use cases (Osakidetza, FCSR, Lanc) and other supporting organizations (e.g. Ibermatica, Aimes) and will cover more than one type of citizens. If participants wish to withdraw from the participation in the use cases at any time, they will be able to do it, and their data, even the pseudo-anonymized data, will be destroyed.

In WP1 there is a task entitled as “Task 1.3 Ethical trials management” where we ensure that ethical principles are used throughout the use cases, which are clustered together for management purposes in the work package related to the use cases. Further explanations on ethical matters will be gathered in deliverable D1.6 Ethical protocols and approvals.

7 Conclusions

The document describes our SHIELD data management plan according to the established H2020 template for a Data Management Plan (DMP) [1]. This document is alive during the whole project, and it will be updated on a regular basis. Data Summary section indicates the purpose of the data collection and generation. This is a complex task, because there are several data which can be managed. Each data will be made FAIR (findable, accessible, interoperable and reusable). SHIELD project’s key results are briefly explains how the financial resources for this openness are envisioned at this stage to be allocated. Section 5 and 6 outline security and ethical aspects respectively, and finally Section 7 presents the conclusions and future work

8 References

[1] European Commission;, "Data Management," July 2016. [Online]. Available: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template. [Accessed 10 May 2017].

[2] e. Network, "European Commission," eHealth Network, 19 November 2013. [Online]. Available: http://ec.europa.eu/health//sites/health/files/ehealth/docs/guidelines_patient_summary_en.pdf. [Accessed 2017].

[3] OpenAire, "OpenAire," [Online]. Available: https://www.openaire.eu/. [Accessed 15 05 2017].

[4] Zendo, "Zenodo," [Online]. Available: www.zenodo.org. [Accessed 2016].

[5] S. G. Stage Gate, "Stage Gate," 2016. [Online]. Available: www.stage-gate.com. [Accessed 26 January 2016].

[6] AENOR;, "UNE-CEN / TS 16555-1 EX Innovation Management Part 1: Innovation Management System," Madrid, 2013.

[7] J. Finch, "The Vignette Technique in Survey Research," *Sociology,* vol. 21, pp. 105-14, 1987.

[8] European Commission;, "Guidelines on Open Access to Scientific Publications and Research Data," July 2016. [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. [Accessed December 2016].

[9] DIGIT, "DCAT application profile implementation guidelines," [Online]. Available: https://ec.europa.eu/isa2/solutions/dcat-application-profile-data-portals-europe_en. [Accessed March 2016].

[10 C. Richardson, "Microservice architecture patterns and best practices," 2017. [Online]. Available: https://microservices.io/. [Accessed March 2017].

[11 European Commission;, "ISA2: Interoperability solutions for public administrations, businesses and citizens," [Online]. Available: https://ec.europa.eu/isa2/home_en. [Accessed January 2017].

[12 SHIELD Consortium, "SHIELD Annex 1 - Research and Innovation Action - Number 727301," 2017.

[13 SHIELD Consortium, "SHIELD Annex 1 - Research and Innovation Action - Number 727301," 2017.

[14 SHIELD Consortium, "SHIELD Consortium Agreement v1.0," 2017.

[15 European Commission;, "Data Management," July 2016. [Online]. Available: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template. [Accessed 9 January 2017].