

# **Improving research data quality by facilitating active curation : feasibility study for tabular data at Unisanté**

**Master's thesis submitted for the degree of  
Master of Science HES-SO in Information Sciences  
by  
Céline, RACINE**

Under the direction of :  
**Patrick RUCH, HES Professor and Head of Research**

**Geneva, August 15, 2022**

**Information Sciences  
Haute École de Gestion de Genève (HEG-GE)**

## Acknowledgments

I would like to thank Prof. Patrick Ruch for his advice and recommendations throughout this work and Mr. Sylvain Pradervand for his participation as an expert.

I would also like to address my gratitude to Loane Warpelin and Anne Niquille, heads of the research and cohort support sector as well as Aline Sager, head of the documentation and data unit, for their help and encouragement.

My grateful thanks are also extended to Carole Claire and Blaise Genton, in charge of the department of education, research and innovation (DFRI), for allowing me to do this master's degree and for supporting me in this process.

I wish to acknowledge the help provided by all the Unisanté collaborators who took part in this study, whether by participating in the survey or in interviews, or by sending messages of support.

Finally, special thanks should be given to Cristopher, for his patience and support and to my family and friends for their understanding and encouragement.

## Summary

To achieve a good level of quality when research data are shared, the FAIR principles have been established. They drive research teams to create data that are Findable, Accessible, Interoperable and Reusable. To implement these recommendations, data curation activities are performed at the end of the research project. They include data transformation, data enrichment, data anonymization or documentation. However, the concept of data quality is not restricted to data sharing. It is followed throughout the research project, by the implementation of data management best practices, by data preparation and data cleaning and by the compliance with various regulations, such as the Good Clinical Practices (GCP).

At Unisanté, a data curation workflow is processed on datasets shared on the institutional repository. The documentation and data unit (UDD) has identified that 40% of the datasets need additional treatment to reach a good level of quality before sharing, even if best practices seem to be applied.

This study aims to collect information on the use of best practices at Unisanté during a research project, on the data-related support wanted by the research teams and on potential possibilities of improvement in terms of data quality. The scope has been restricted to tabular data only, as free text data are handled differently. This information has been collected through a survey, interviews with research teams and interviews with Unisanté support services. From the analysis of this collected information, it can be concluded that the integration of active data curation into the research project would have a positive impact on data quality.

The results show that a form of support is wanted for some activities, mainly the data management plan, the anonymization and de-identification processes, the documentation, the data monitoring, the interoperability and data matching. The recommendations for Unisanté to support these activities would be the implementation of new services, such as an anonymization support service and a Data Protection Officer (DPO), the technical integration of interoperable pre-set fields into REDCap, the creation of a centralized data monitoring service or the improvement of available training courses.

This study also highlights the need for a better data governance at Unisanté and an improvement of the data-related infrastructure, mainly the archival system and the data repository.

**Keywords** : active data curation, data curation, data quality, research data management, research support services, Unisanté

# Table of contents

<b>Acknowledgments .....</b>	<b>i</b>
<b>Summary.....</b>	<b>ii</b>
<b>Table of contents .....</b>	<b>iii</b>
<b>List of Tables .....</b>	<b>vi</b>
<b>List of Figures .....</b>	<b>vii</b>
<b>1. Introduction.....</b>	<b>1</b>
<b>1.1 Context: Center for Primary Care and Public Health (Unisanté).....</b>	<b>1</b>
1.1.1 History.....	1
1.1.2 Missions.....	3
1.1.3 Organization.....	3
1.1.4 Research at Unisanté.....	5
<b>1.2 Problem statement.....</b>	<b>5</b>
<b>1.3 Objectives.....</b>	<b>7</b>
1.3.1 Objective 1: to establish a state of the art.....	7
1.3.2 Objective 2 : to evaluate practices and needs at Unisanté .....	8
1.3.3 Objective 3 : to facilitate data curation process and anonymization .....	8
1.3.4 Objective 4 : to integrate data curation and anonymization practices into the research project lifecycle at Unisanté .....	8
1.3.5 Objective 5 : to support change of practices in the research teams .....	8
<b>2. State of the art .....</b>	<b>9</b>
<b>2.1 Key concepts.....</b>	<b>9</b>
2.1.1 Scientific research.....	9
2.1.2 Research data.....	9
2.1.3 Tabular data.....	9
2.1.4 FAIR principles.....	9
2.1.5 Data quality .....	10
2.1.6 Personal and sensitive data .....	12
2.1.7 Identifiability .....	13
2.1.8 Re-identification risk.....	13
2.1.9 Privacy by design and by default.....	13
2.1.10 Anonymization.....	14
2.1.11 Pseudonymization .....	15
<b>2.2 Legal and regulatory framework .....</b>	<b>15</b>
2.2.1 Data protection laws.....	15
2.2.2 Federal Act on Research involving Human Beings.....	16
2.2.3 Good Clinical Practices .....	17
<b>2.3 Research data management.....</b>	<b>18</b>
2.3.1 Definition.....	18



2.3.2	The research data lifecycle .....	18
<b>2.4</b>	<b>Research data curation.....</b>	<b>20</b>
2.4.1	Definition.....	20
2.4.2	Activities.....	22
<b>2.5</b>	<b>Active data curation.....</b>	<b>24</b>
2.5.1	Definition.....	24
2.5.2	Activities.....	25
<b>3.</b>	<b>Results : data curation practices at Unisanté .....</b>	<b>28</b>
<b>3.1</b>	<b>Methodology.....</b>	<b>28</b>
3.1.1	Data collection .....	28
3.1.2	Sampling.....	30
3.1.3	Response rate.....	31
<b>3.2</b>	<b>Researchers' practices .....</b>	<b>32</b>
3.2.1	Planning research .....	32
3.2.2	Collecting data .....	36
3.2.3	Processing and analyzing data.....	39
3.2.4	Publishing, sharing, and preserving data.....	42
<b>3.3</b>	<b>Support services .....</b>	<b>45</b>
3.3.1	Biostatistics consultation unit .....	45
3.3.2	Survey Methodology unit.....	46
3.3.3	Research promotion unit .....	47
3.3.4	Research IT Services unit .....	48
3.3.5	Documentation and data unit .....	49
<b>4.</b>	<b>Recommendations and discussion.....</b>	<b>51</b>
<b>4.1</b>	<b>General discussion .....</b>	<b>51</b>
4.1.1	Terminology .....	51
4.1.2	Design of support services .....	51
<b>4.2</b>	<b>Recommendations for the research teams .....</b>	<b>52</b>
4.2.1	Data management plan.....	52
4.2.2	Methodology support.....	52
4.2.3	Interoperability with REDCap .....	53
4.2.4	Anonymization .....	54
4.2.5	Centralized data monitoring .....	55
4.2.6	Codes documentation and sharing.....	55
4.2.7	Training and guidelines .....	56
<b>4.3</b>	<b>Recommendations for the institution .....</b>	<b>57</b>
4.3.1	Data governance at Unisanté.....	57
4.3.2	Documentation.....	57
4.3.3	File organization and naming .....	57
4.3.4	Data repository.....	58

4.3.5	Archival system .....	59
4.4	Future work .....	59
5.	Limitations of the study .....	62
5.1	Definition of the project .....	62
5.2	Survey design .....	62
6.	Conclusion .....	64
	Bibliography .....	65
Annex 1 :	Unisanté research fields (in French).....	1
Annex 2 :	Sub-objectives of the study (March 2022).....	1
Annex 3 :	Structure of the table of data curation activities .....	1
Annex 4 :	Survey .....	2
Annex 5 :	List of question for researchers' interviews .....	51
Annex 6 :	List of questions for support services interviews .....	52
Annex 7 :	Data curation activities : UCB .....	53
Annex 8 :	Data curation activities : Survey Methodology unit .....	54
Annex 9 :	Data curation activities : UPR.....	55
Annex 10 :	Data curation activities : Research IT services unit.....	56
Annex 11 :	Data curation activities : UDD .....	57
Annex 12 :	Table of survey's comment transformations .....	58

## List of Tables

Table 1 : Data curation activities by Johnston.....	22
Table 2 : Response rate by department.....	31
Table 3 : Priority of the recommendations .....	59

## List of Figures

Figure 1 : Unisanté organigram 2021 (in French) .....	4
Figure 2 : Curation Lifecycle Model .....	6
Figure 3 : Data quality model characteristics from ISO 25012 .....	11
Figure 4 : Anonymization and Pseudonymization process.....	15
Figure 5 : The research Data Lifecycle .....	19
Figure 6 : The Biomedical Data Lifecycle (Harvard University) .....	20
Figure 7 : The Curation Lifecycle Model .....	21
Figure 8 : Mistake types categorized by research data management stage.....	24
Figure 9 : The survey project curation components performed by the survey research center .....	25
Figure 10 : The Research Data Lifecycle : steps included in the questionnaire .....	28
Figure 11 : Planning research : most implicated roles .....	33
Figure 12 : Planning research: activities .....	33
Figure 13 : Planning research : need of support .....	35
Figure 14 : Collecting data: most implicated roles .....	36
Figure 15 : Collecting data : activities .....	36
Figure 16 : Collecting data : need of support .....	38
Figure 17 : Processing and analyzing data : most implicated roles.....	39
Figure 18 : Processing and analyzing data : activities .....	40
Figure 19 : Processing and analyzing data : need of support .....	42
Figure 20 : Publishing, sharing and preserving data : most implicated roles.....	43
Figure 21 : Publishing, sharing and preserving data : activities.....	44
Figure 22 : Publishing, sharing and preserving data : need of support.....	45
Figure 23 : Diagram of the Project TIER folder structure .....	58

# 1. Introduction

## 1.1 Context: Center for Primary Care and Public Health (Unisanté)

### 1.1.1 History

The creation of the Center for Primary Care and Public Health (Unisanté) took place in a context of profound changes in healthcare systems and medical practices. To offer an equitable, viable and sustainable healthcare system to the population, it was necessary to improve coordination between the different actors in the healthcare system (Unisanté, 2022a).

To achieve this vision, a project called “*Alliance Santé*” started in 2017, with the idea of merging existing structures to create one multi-faceted center for primary care and public health. This new institution would be independent, with a multidisciplinary internal organization and strong collaboration between the different internal entities, with the aim of having a structure that could respond quickly to the health needs of the population but that also could easily adapt to new challenges and needs in general health context.

The conclusions of the work of Alliance Santé led to the creation of Unisanté on the 1<sup>st</sup> of January 2019, by merging four institutions described below, each one having a specific expertise in primary care or public health.

#### 1.1.1.1 Department of ambulatory care and community medicine

Located next to the University Hospital of Lausanne (CHUV), the Department of ambulatory care and community medicine (*Policlinique médicale universitaire (PMU)*) was a referral center for internal and general medicine. The PMU was a public institution, endowed with an official legal status under the supervision of the Canton of Vaud.

It offered a wide range of high-quality services in various fields (general medicine, ambulatory care, pharmacy, travel medicine...) for the people around Lausanne who could visit the PMU for any health problem either by appointment or in an emergency. In addition to the ambulatory-care center located next to the CHUV, the PMU also provided health care services to migrants, a walk-in clinic located in the city center and “health point” clinics on EPFL and UNIL campuses.

Furthermore, the PMU had a teaching activity in internal and general medicine, within the Faculty of Biology and Medicine (FBM) of the University of Lausanne (UNIL), for undergraduate and graduate students. Finally, in order to advance clinical knowledge, various professions were involved in research within the institution. (*Policlinique médicale universitaire (PMU)*, 2018)

#### 1.1.1.2 Institute of Social and Preventive Medicine

The Institute of Social and Preventive Medicine (IUMSP) (*Institut universitaire de médecine sociale et préventive (IUMSP)*) was an institute active in research, teaching and provision of public health expertise. Linked to both the CHUV and the FBM of the UNIL, the IUMSP brought together public healthcare specialists from 15 different disciplines.

Its principal mission was to develop appropriate responses to the population healthcare needs and support their implementation. It realized 3 types of activities: research, training of healthcare professionals and commissioned services.

The IUMSP was specialized in three main fields :

1. Epidemiology and prevention of chronic diseases
2. Organization of healthcare services
3. Quantitative methods in medicine and public health

Finally, in addition to “classic research projects”, the IUMSP had an activity of commissioned services covering diverse areas like health systems organization, prevention programs, HIV epidemiology, and sexual and reproductive health among various population groups (Institut universitaire de médecine, sociale et préventive (IUMSP), 2018).

#### **1.1.1.3 Institute for Work and Health**

The Institute for Work and Health (*Institut universitaire romand de santé au travail (IST)*) was a foundation under private law, supported in particular by the cantons of Vaud and Geneva. As a reference organization in its field, its missions were research, teaching, expertise and advice, as well as the promotion of occupational health.

As part of its public health mission, the IST was dedicated to develop the relationship between work and health. It strives to contribute to the promotion of favorable working conditions for workers, the economy and society.

Its research activity covered three main disciplines:

1. Exposure science
2. Environmental engineering
3. Workers' health

Moreover, the IST was a doctoral and post-doctoral formation center for occupational health and provided continuous education for health professionals and managers in companies.

The IST also had a medical activity, with multiple consultations on occupational health, respiratory diseases, pain at work, pregnancy at work and professional oncology.

Finally, the IST provided services for businesses on various areas like occupational medicine, hygiene, work psychology and ergonomics. (Institut universitaire romand de Santé au Travail (IST), 2018)

#### **1.1.1.4 Center for Health Promotion**

The Center for Health Promotion (*Association Promotion Santé Vaud (ProSV)*) was a cornerstone of health promotion and prevention in the canton of Vaud. ProSV had five principal activities (Promotion Santé Vaud (ProSV), 2018) :

1. Individual consultations
2. Professionals' training
3. Public sensibilization and information
4. Expertise
5. Support and labelling

After the initial merging in 2019, Unisanté has continued to grow. The Foundation for Cancer Screening (*Fondation vaudoise pour le dépistage du cancer (FVDC)*) has joined Unisanté in 2020, followed by the *Équipe mobile d'urgences sociales (EMUS)* in 2021.

### 1.1.2 Missions

*«The Center for Primary Care and Public Health, University of Lausanne, Switzerland, [...] is active in research, academic training, prevention and ambulatory care. Innovative and unique in Switzerland, it promotes interdisciplinarity and the pooling of skills committed to health issues »*  
(Unisanté, 2019, p. 4)

Unisanté has been created in 2019 to reach one goal: Promote and improve individual and population health in their environment, independently to their social status, by prevention, ambulatory care and public health measures, in an academic environment.

Each one of the institutions that have merged have a specific expertise in primary care or public health. To continue their activities and promote interdisciplinarity, Unisanté has decided to decline its mission in six main activities, representing the principal facets these two wide fields (Unisanté, 2022b).

1. Primary care

Primary care activities, particularly general and family medicine, nursing and pharmaceutical counseling, as well as access to care and orientation in the health care system.

2. Vulnerable populations

Interventions related to populations and patients in vulnerable situations or with special needs.

3. Health promotion and prevention

Primary and secondary health promotion and prevention services, as well as screening services.

4. Occupational health

Occupational health and medicine services, in particular in relation to general medicine and public health.

5. Health systems

Expertise and research in public health, in particular on the organization and funding of health systems, as well as population surveillance to better meet the health needs of communities.

6. Research and teaching

Academic research and teaching activities in general and community medicine, public health and occupational health.

### 1.1.3 Organization

Unisanté is an autonomous institution under public law with its own legal personality and governance. It carries out its missions within the framework of a service contract with the Department of Health and Social Action of the State of Vaud.

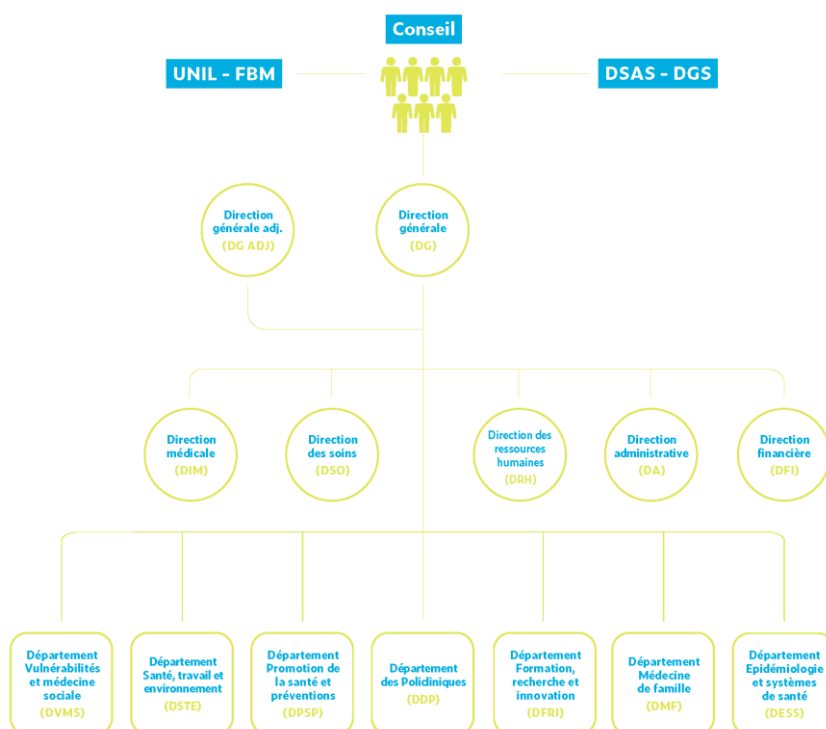
Its six activities, listed previously, are realized in seven departments, shown in Figure 1, and coordinated by the General Management (*Direction Générale*)<sup>1</sup> (Unisanté, 2022c)

---

<sup>1</sup> The organization of Unisanté has been modified on the 1<sup>st</sup> August 2022 but is not represented in this study, as it took place between February and July 2022

- **Department of ambulatory care (DDP)**  
General medicine, pharmacy, vaccination and travel health, ambulatory care and campuses' health points
- **Department of epidemiology and health systems (DESS)**  
Evaluation and expertise in public health, health economics, chronic diseases, health system and services, research in public health
- **Department of education, research and innovation (DFRI)**  
Quantitative research, community research, digital and global health, research and cohorts support, training and documentation
- **Department of family medicine (DMF)**  
Pregraduate and postgraduate education, primary care, FLON's walk-in clinic, family medicine research
- **Department of health promotion and prevention (DPSP)**  
Non-communicable disease prevention, cancer screening, community interventions, information and advocacy, cardiovascular prevention research
- **Department of occupational and environmental health (DSTE)**  
Occupational medicine, health at work, consultations / expertise, environment, services in companies
- **Department of vulnerable populations and social medicine (DVMS)**  
Medical expertise / consulting physician, prison medicine, community clinical activities, socio-medical coordination, care for migrants

Figure 1 : Unisanté organigram 2021 (in French)



(Adapted from (Unisanté, 2022c))



### 1.1.4 Research at Unisanté

As every mission need to rely on scientific evidence and progress, research is done in each department. This covers a lot of domains in general medicine and public health, as all the entities merged into Unisanté have kept their research areas. Research fields are presented in [Annex 1](#), in French.

As a consequence, research is realized through many different research approaches: interventional clinical trials, observational research projects involving human beings or data/biological samples collected from human beings but also studies on air quality, toxins, health systems organization or health related politics. Both quantitative and qualitative research is performed by Unisanté to reach their goal. A platform currently under development will allow Unisanté to have a precise tracking of all planned, ongoing and closed research activities. For the time being, ongoing research projects are published by researchers on Unisanté website, if desired.

Within research projects, some are mandates from external institutions. These commissioned projects are quite similar to research projects. The main difference is funding, which comes from institutions and not national funders, and publications, as the results will be presented in reports and not journal articles. However, for this thesis, mandates are considered as research projects because data management is identical.

Around 40 research projects and 26 mandates are in process in 2022, according to the website<sup>2</sup>. The research activity produces around 410 publications every year (Unisanté, 2021). These publications are journal articles, reports, conference articles, abstracts or posters, chapters and books.

## 1.2 Problem statement

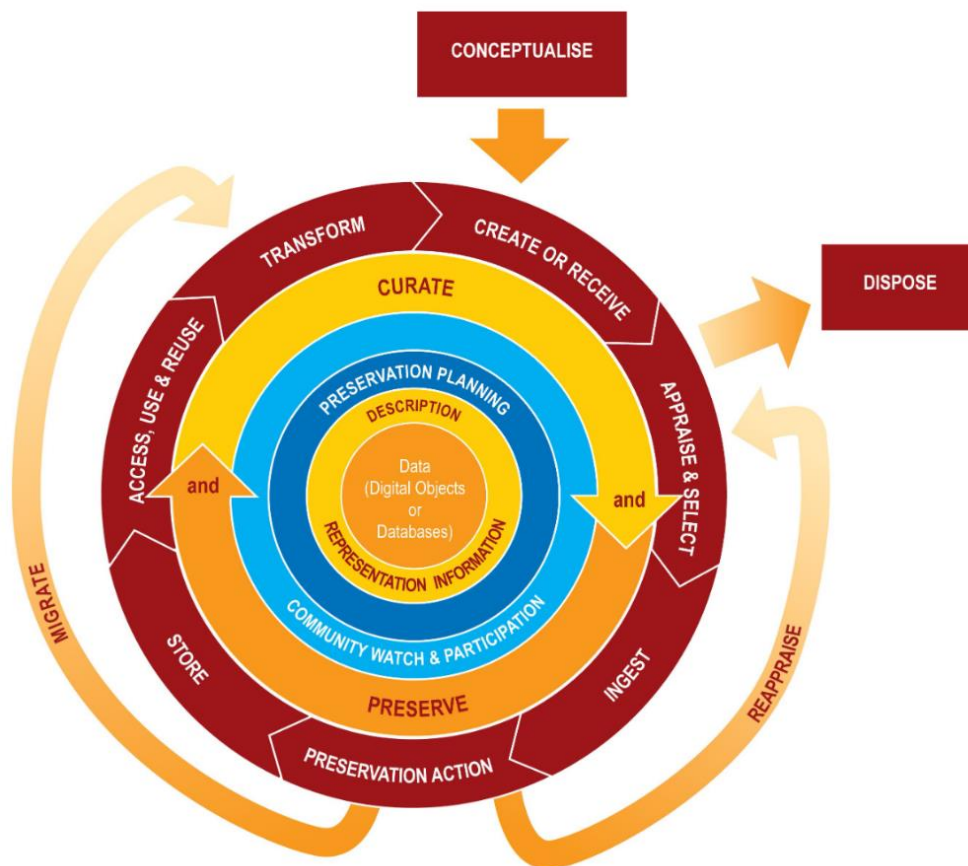
Research data management and data curation has been part of the research process for many years. Data curation is the process of ensuring data quality, integrity, interoperability, discoverability, and preservation. It has been modeled by the Digital Curation Center all along the research data lifecycle (see Figure 2). However, the curation activities described in the professional literature are focalized on discoverability and data preservation (Kouper et al. 2021). In the information science field, data curation is strongly associate with archiving (Pham, 2018).

But, being focused on the end of the research project cut the researchers out of the process of data curation. It has been seen that many errors in research data management occur during the research project (Kovacs, Hoekstra, Aczel, 2021) and could be avoided with the application of research data management best practices and active curation.

---

<sup>2</sup> <https://www.unisante.ch>

Figure 2 : Curation Lifecycle Model



(Higgins, 2008)

Moreover, the difference between data management and data curation is not clear (Pham, 2018). The concept of data curation is not properly defined and leads to a confusion. As presented in the [section 2.4](#) of this thesis, many definitions exist. For researchers, data curation is often assimilated to data cleaning. In the information science field, data curation is a process to add value and enable long-term preservation of the data (Henderson, 2017).

Data curation activities, in particular data cleaning, documentation and anonymization, are time-consuming. This could be a reason why some activities are partially done by research teams during the project. Regarding anonymization, in addition to the time-consuming aspect, the difficulty to understand and apply its definition can be highlighted. With health-related datasets, disclosure risk has to be evaluated for each case, as the risk depends on the collected variables. Therefore, there is a confusion between anonymization and pseudonymization, defined in [section 2.1.7](#) and [2.1.8](#). This confusion leads to potential issues, as some pseudonymized datasets could be shared in Open Access on a data repository, and participants to the study could be re-identified by data matching. Improving information, training and methods about anonymization could reduce this risk. However, it is rarely possible to properly anonymize personal data.

Some tools and process, for example YARD (Peer, Dull, 2020) or C<sup>2</sup>Metadata (Alter et al. 2021), have been developed to relieve researchers and curators with activities like data cleaning or documentation. Integrating data curation into the data lifecycle with light methods

and tools could ensure an ongoing and systematic research data curation, ensuring data quality and avoiding a large number of modifications at the end of the research project.

Research at Unisanté is supported by some services. One of them is the unit of documentation and data (UDD) which is the initiator of this study. The UDD has been offering broad support in research data management and a data repository to share research data since 2014. Its goal is to help researchers to manage their data according to the legal and regulatory framework in Switzerland and to the FAIR principles<sup>3</sup>.

One of its services is to curate data before sharing on the institutional data repository<sup>4</sup>. Based on the NADA software, it offers the same functionalities as other general data repositories (e.g., Zenodo), such as the attribution of a DOI, XML metadata (DDI schema) or various access modalities (Open Access, restricted access). A data curation process is performed on each dataset submitted to this repository. This process aims to ensure that file format is open and usable, to prevent re-identification of participants of a study and to provide sufficient metadata to be FAIR compliant.

It occurs that around 40% of the data which have been handled by this service needed a supplementary treatment before achieving the quality requirements in terms of documentation, file format or de-identification. Not all research teams use the UDD services, either because the teams hire data managers for their projects, because they do not have time or rely on their own research data management skills, or because they do not know the UDD services.

Moreover, Unisanté has no data policy or institutional directive related to data curation or data quality. Therefore, there is no standardized way to take care of data. This is due to the nature of Unisanté, which is the merging of multiple entities, each having their own habits and practices regarding data management, data curation and data quality.

This state-of-the-art leads to a risk that some data shared by Unisanté does not match FAIR principles, have issues with data protection and disclosure risk or does not reach the same level of quality than curated data.

## 1.3 Objectives

The general purpose of this master thesis is to identify all research data curation tasks realized in research project, to evaluate the potential improvement of these tasks and to propose a way to integrate active data curation to the research project lifecycle, in order to improve data quality and to reduce re-identification risk in Unisanté datasets. Five objectives have been set to achieve the main goal.

### 1.3.1 Objective 1: to establish a state of the art

The first objective is to make a state of the art on data curation tools and process, including anonymization. It requires first to define data curation and anonymization, and to identify all the activities related to these concepts. Then, methods and tools used in medical research will be listed, but also those which are new in the area and could be promising.

This objective will be performed with a literature review, detailed in the [Methodology](#) section, and its results will be presented in the [State of the art](#) section.

---

<sup>3</sup> <https://www.go-fair.org/fair-principles/>

<sup>4</sup> <https://data.unisante.ch>

### **1.3.2 Objective 2 : to evaluate practices and needs at Unisanté**

The second objective is to identify the actual practices and needs of the researchers at Unisanté in terms of data curation. As some support services already exist at Unisanté, the identification of their recommendations is also an important part of this objective.

This objective will be realized through a survey and interviews of research teams and support services. The data collection is detailed in the [Methodology](#) section, and the results are presented in the [Results](#) section, with a differentiation between [researchers' practices](#) and [support services](#).

### **1.3.3 Objective 3 : to facilitate data curation process and anonymization**

The third objective is to facilitate data curation and anonymization processes with dedicated tools or methods.

This objective will be achieved by the analysis of collected data (objective 2) and the comparison of data with results from the literature (objective 1). The results of this objective will be presented in the [Recommendations](#) section. A difference will be made between recommendations with a direct impact on research teams, and recommendations at institutional level.

### **1.3.4 Objective 4 : to integrate data curation and anonymization practices into the research project lifecycle at Unisanté**

This objective will result in recommendations for Unisanté, based on the results obtained previously. The recommendations will link to the research data lifecycle. Moreover, the recommendations could be scenarized within it, to illustrate how they could be integrated to actual practices. As an example, a scenario could be «XXXX, a statistician, use the software C2Metadata at the end of the "Processing and analyzing step" to document his cleaning code».

### **1.3.5 Objective 5 : to support change of practices in the research teams**

As implementing new activities may be difficult, a particular attention will be made to support change. This project will gather data on the preferences of the research teams in terms of type of support (e.g. : Trainings, in-person support, specific software). This information will be included into the recommendations, to allow the creation of user-friendly support at Unisanté.

Each of these objectives has been declined in sub-objectives and link to one or more data collection method. This information is available in [Annex 2](#).

## 2. State of the art

### 2.1 Key concepts

#### 2.1.1 Scientific research

Scientific research is a systematic acquisition of knowledge based on the collection of empirical data to describe, explain, predict or control phenomena (Fortin, Gagnon, 2016).

Data are at the center of scientific research. This fact is also seen through publication practices. Since 2014, publishers have encouraged researchers to share their data with their article (PLOS, 2014; Nature, 2021; Elsevier, 2022). Funders, like the Swiss National Science Foundation (SNSF), also consider data as fundamental in research and expect them to be shared as openly as possible (SNSF, 2022).

#### 2.1.2 Research data

Research data are defined by the Organisation for Economic Co-operation and Development (OECD) as :

*« [...] factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated. » (OECD, 2007, p. 13)*

Data are generated in various formats, depending on the subject, field or type of research. It can include documents (Text, PDF, Microsoft Word), spreadsheets (CSV, Microsoft Excel), laboratory notebooks, databases (Access, MySQL, Oracle), questionnaires, transcripts, codebooks, audiotapes, videotapes, protein or genetic sequences, raw data generated by software or sensors, images, videos, voice recordings... (Cox, Verbaan, 2018; Henderson, 2017).

#### 2.1.3 Tabular data

Tabular data are organized data in table with rows and columns (Zach, 2022). Tabular data also include relational database because it is composed by multiple spreadsheets that can be linked by a key. Tabular data can integrate metadata, like headings, variable names or code labels.

As several software programs can generate tabular data, multiple file format applies. The most commons are Comma Separated Values (csv), Excel spreadsheet (xlsx), Delimited Text (txt), Tab-delimited file (tab), MS Access database (mdb, accdb), OpenDocument Spreadsheet (ods), and any relational database (MySQL, dbf).

#### 2.1.4 FAIR principles

The international coalition FORCE 11 has put in place the FAIR principles which set a general rule: data must be findable, accessible, interoperable and reusable (FAIR). These principles are widely recognized and applied in all research fields. They aim to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of scientific data and their associated algorithms and workflows (FORCE11, 2014).

Each aspect of the FAIR principles has been declined in indicators to facilitate their application.

« *To be Findable:*

1. *(meta)data are assigned a globally unique and eternally persistent identifier.*
2. *data are described with rich metadata.*
3. *(meta)data are registered or indexed in a searchable resource.*
4. *metadata specify the data identifier.*

*To be Accessible:*

1. *(meta)data are retrievable by their identifier using a standardized communications protocol.*
2. *the protocol is open, free, and universally implementable.*
3. *the protocol allows for an authentication and authorization procedure, where necessary.*
4. *metadata are accessible, even when the data are no longer available.*

*To be Interoperable:*

1. *(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.*
2. *(meta)data use vocabularies that follow FAIR principles.*
3. *(meta)data include qualified references to other (meta)data.*

*To be Re-usable:*

1. *(meta)data have a plurality of accurate and relevant attributes.*
2. *(meta)data are released with a clear and accessible data usage license.*
3. *(meta)data are associated with their provenance.*
4. *(meta)data meet domain-relevant community standards. »*

(FORCE11, 2014)

### **2.1.5 Data quality**

Data quality is defined in multiple ways through ISO standards: ISO 8000-2, ISO 8000-8, ISO 25012 and ISO 15489.

The international standard 8000-2 on data quality vocabulary define data quality as « *degree to which a set of inherent characteristics of data fulfils requirements* » (International Organization for Standardization (ISO), 2020, p. 6). This norm also highlights data accuracy (chapter 3.8.10) and completeness (chapter 3.8.12) as important related concepts, to ensure that data value agrees « [...] *with the corresponding true value* [...] » and that the dataset include all the content « [...] *necessary for an intended purpose* [...] » (International Organization for Standardization (ISO), 2020, pp. 7-8).

For ISO 8000-8 (International Organization for Standardization (ISO), 2015), data quality is composed by syntactic quality, semantic quality and pragmatic quality. Data should conform to its specified syntax, should correspond to what it represents and should be appropriate and useful for its usage (Makhlouf-Shabou, 2021). This standard lists rules to check quality according to these definitions. For syntactic quality, the integrity should be verified and applied

to data and its environment. For pragmatic quality, data should be accessible and complete, flexible in its content and layout, secured and useful.

The ISO 15489-1, standard in Records Management, does not define data quality exactly but lists the important characteristics of a record. They can be transposed to data. The data (or record) has to be authentic (Authenticity), has to be trusted and accurate (Reliability), complete and unaltered (Integrity) and has to be «*located, retrieved, presented and interpreted within a time period deemed reasonable by stakeholders*» (Usability) (International Organization for Standardization (ISO), 2016, p. 5).

Finally, the ISO 25012, from the software engineering field, use 15 criteria, listed in Figure 3, to evaluate data quality. They resonate with the previous norms and the FAIR principles, as accuracy, completeness, accessibility, precision or availability.

Figure 3 : Data quality model characteristics from ISO 25012

Table 1 — Data quality model characteristics

Characteristics	DATA QUALITY	
	Inherent	System dependent
Accuracy	X	
Completeness	X	
Consistency	X	
Credibility	X	
Currentness	X	
Accessibility	X	X
Compliance	X	X
Confidentiality	X	X
Efficiency	X	X
Precision	X	X
Traceability	X	X
Understandability	X	X
Availability		X
Portability		X
Recoverability		X

(International Organization for Standardization (ISO), 2008, p. 5)

Through these definitions, data quality will be evaluated differently depending on the context, as the criteria varies. However, each one resonates with the other. It is therefore important to keep in mind all these standards to evaluate data quality.



These standards also highlight qualities like discoverability and availability, which are important aspects of the FAIR principles. Therefore, their application allows a minimum level of quality to be achieved.

### 2.1.6 Personal and sensitive data

Personal data are defined by the Federal Act on Data Protection (FADP) as « *all information relating to an identified or identifiable person* » (SR 235.1 - Federal Act of 19 June 1992 on Data Protection (FADP) 2019).

Sensitive data are defined by the same law as data on « *religious, ideological, political or trade union-related views or activities ; health, the intimate sphere or the racial origin; social security measures; administrative or criminal proceedings and sanctions* » (SR 235.1 - Federal Act of 19 June 1992 on Data Protection (FADP) 2019).

The Health Insurance Portability and Accountability Act of 1996 (HIPAA), a public law in the United States, provides a more accurate list of identifiers to remove in order to follow the “Safe Harbor” method. This list gives a more precise idea of personal and sensitive data.

- « *Names*
- *All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes*
- *All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older*
- *Telephone numbers*
- *Vehicle identifiers and serial numbers, including license plate numbers*
- *Fax numbers*
- *Device identifiers and serial numbers*
- *Email addresses*
- *Web Universal Resource Locators (URLs)*
- *Social security numbers*
- *Internet Protocol (IP) addresses*
- *Medical record numbers*
- *Biometric identifiers, including finger and voice prints*
- *Health plan beneficiary numbers*
- *Full-face photographs and any comparable images*
- *Account numbers*
- *Any other unique identifying number, characteristic, or code*
- *Certificate/license numbers* »

(Rights (OCR), 2012)



### 2.1.7 Identifiability

*«Identification results from the interpretation of identifiers present in the information or that can be deduced from such information or its context» (Jotterand, 2022).* A person can be identified, when her identity is clear and can be precisely stated for the data, or identifiable, *« when she cannot be clearly identified from the data alone, but can be identified from the circumstances, i.e. from the context of the information or on the basis of additional information » (Jotterand, 2022).*

The concept of identification is discussed in Switzerland and in the European Union, as it can be approached in two ways. The first one is the absolute approach, which means that data should be considered as personal *« if there is any theoretical possibility of identification by anyone » (Jotterand, 2022).* On another side, the relative approach will take in account the realistic chances of identification and the person who will try to identify the data.

### 2.1.8 Re-identification risk

A common example of re-identification is the “Bradley Cooper’s taxi ride”. A software developer has been able to re-identify all taxis in New York city by working on an “anonymous” database published online. In less than an hour, he was able to decrypt the taxis’ identification numbers. This database also included geolocation data (taxi’s route) and financial information (price and tip). With this information, it has been possible to match celebrities, such as Bradley Cooper, and taxis, based on paparazzi photographs. Therefore, the celebrities’ journeys have been published online, with information on the given tips (Johnston, 2015).

The re-identification risk is the evaluation of the possibility to identify a particular person into a dataset. It will be evaluated by examining the data, the context and environment of the data, the possible interaction between the dataset and external data. It will also consider which actors could try to re-identify people into the dataset (Jotterand, 2022). Depending on the identifiability definition chosen (absolute or relative), the evaluation of this risk will take in account the person who tries to re-identify data.

To perform the evaluation of the risk, the first option is to attempt to re-identify the data. This requires having the complete dataset, knowing the environment and possibilities of data matching, and to be able to realize such treatment. Another possibility is to evaluate the risk by a statistical approach. Various methods exist, such as K-anonymity (Olatunji et al. 2022). They will investigate the variables and produce a score of re-identification.

To evaluate the re-identification risk before collecting the data, the statistical approach can be used. In many cases, a researcher will have all the needed information to perform it, as he/she will know which variables will be collected, the estimation of the number of participants and the expected distribution. However, the possibility of data matching should still be considered and may require additional expertise.

### 2.1.9 Privacy by design and by default

To protect the identity of an individual, data can be collected according to the concept of “privacy by design”. It means that measures to protect the participants are taken proactively and are embedded in the data collection tool. The privacy by design will ensure the security of the data and privacy during all its lifecycle and its process will be transparent for all stakeholders (Cavoukian, 2011). As an example, in a research project, the concept of privacy by design could be performed by including thoughts on data protection at the beginning of a

project, by ensuring that only data needed to answer the research questions are collected, by evaluating the theoretical re-identification risk based on the variables planned, by modifying these variables to reduce the risk and by creating the data collection tool according to these modifications. Privacy by design is performed to protect the individual during all the data lifecycle. Applying this principle would protect them in case of a data leak.

Privacy by default is a reactive concept. The anonymization process and the modification of the data will take place after the research project. This does not mean that the participants are not protected at the beginning of the project, but a re-identification risk remains in case of a data leak.

### **2.1.10 Anonymization**

Anonymization is a process applied to a dataset containing personal data. It aims to make it impossible to identify a person from a dataset (CNIL, 2020). This process involves a permanent and irreversible action, which will make re-identification impossible, or only with very intensive efforts (Stam, Kleiner, 2020). The identification is the possibility to retrieve a person's name or address, but also the potentiality to re-identify a person by singling out, linkability and inference (Data Protection Commission, 2019).

Many techniques exist to anonymize data and reach this goal. Research is currently performed in this area, and knowledge about anonymization and re-identification methods are constantly changing and evolving (Data Protection Commission, 2019). The main methods to anonymize data are : masking, randomization and generalization (Data Protection Commission, 2019; FORS, nd; Olatunji et al. 2022).

The first step towards anonymization is the removal of direct identifiers, such as names, address, or phone number. The indirect identifiers, such as age, profession or postal code, must also be controlled, recoded or deleted if necessary (FORS, 2021). This process is also known as masking.

Randomization is a method used to anonymize data, by altering them with various treatments. This will include transformations such as noise addition, which could be performed with statistical software, specific software (e.g., SDCMicro,  $\mu$ -ARGUS, DataSynthesizer) or generative neural network (Bae et al. 2019; Hundepool, 2012; Johnston, 2017; Ping, Stoyanovich, Howe, 2017; Yoon, Drumright, van der Schaar, 2020). Other transformations could be random small changes in the data, or permutation of columns or data. If an individual is too easily identifiable, data can be permuted. As an example, a person with a height of 120cm, who has a visible handicap and lives in a small town could be identified. To protect this person without modifying the values of the data, the data can be exchanged between participants (i.e. the city). This modification would keep the distribution of the data as originally. However, it would not maintain correlation. Therefore, permutation is not suitable in each anonymization cases.

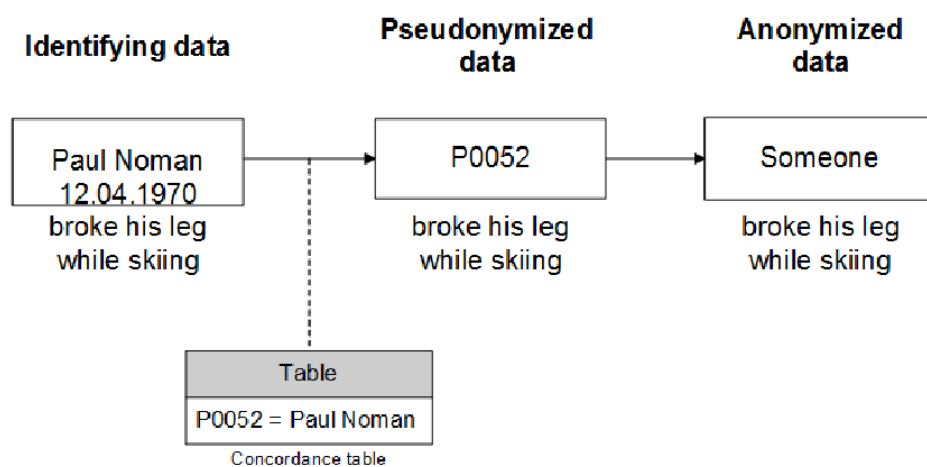
Generalization is a method that diminishes the granularity of the data to reduce the re-identification risk. The treatment performed on the data will be a (micro)aggregation. In other words, this process will create categories with variables (e.g., real age will be recoded in category 0-17, 18-25...). To find the right level of aggregation, various methods could be used, such as k-anonymity. This calculation will take in consideration all identifying variables (direct and indirect identifiers) and propose an aggregation in order to apply to reach the following goal : k individuals share the same attributes (Olatunji et al. 2022).

### 2.1.11 Pseudonymization

Pseudonymization is also a de-identification process. However, the link between the data and the identifiable person is not broken, as it will not go as far as anonymization. In many cases, it is possible to re-identify people from pseudonymized datasets using indirect identifiers (Data Protection Commission, 2019).

Pseudonymization will consist in replacing direct identifiers, like name or address, with a neutral code (FDPIC, 2015). A correspondence table is generated, to keep the link between the real data and the new code (see Figure 4). The pseudonymized data are still consider as “personal data” as long as the correspondence table exists (Stam, Kleiner, 2020). However, the deletion of this correspondence table is not sufficient to anonymize data. The people can still be re-identified by data matching or other techniques.

Figure 4 : Anonymization and Pseudonymization process



(FDPIC, 2015)

## 2.2 Legal and regulatory framework

The regulatory framework applicable to research data is wide (Santos, 2020) and will not be detailed in the thesis. Only laws and regulations included in risk evaluation during data curation will be quickly presented here. Santos' Master Thesis on regulatory framework for research data goes further in the analysis.

Since research at Unisanté also aims to explore health economics, ergonomics or exposure science at populational level, all projects are not necessarily ruled by the following regulatory framework.

### 2.2.1 Data protection laws

#### 2.2.1.1 Federal Act on Data Protection

In force since 1993, the Federal Act on Data Protection (FADP) aims to « *protect the privacy and the fundamental rights of persons when their data is processed* » (SR 235.1 - Federal Act of 19 June 1992 on Data Protection (FADP), 2019).

The law applies to the treatment of all personal data from physical or moral people, identified or identifiable, when treatment occurs in Switzerland. It does not apply for anonymized data, or data that cannot be linked to an identifiable person.

For scientific research purposes, it is recommended to use anonymized data when possible. If not, data should be anonymized as soon as possible or destroyed when the research project has achieved its goals. This does not apply to clinical trials, as the participants must stay identifiable for security purposes and audit trails (Sprumont, 2019). If identifiable data are used, the participant must be informed and must consent to the treatment of the data.

To achieve the same level of protection as the European General Data Protection Regulation (GDPR), this law is currently under modification. The new version should be enforced in 2023.

### **2.2.1.2 General Data Protection Regulation**

In force since 2018, the General Data Protection Regulation (GDPR) brings a common framework in personal data protection for all Europe. It aims to reinforce people's rights on their data, to reinforce sanctions, to split responsibilities between stakeholders and to facilitate circulation of data (Bayle, 2020).

GDPR applies for all personal data of European residents, but also for all personal data if the data treatment is done by a moral person based in the UE or if the service offered by the enterprise is available in the UE. For Unisanté, GDPR could apply in some research projects, when it is included in an European project.

The use of personal data in scientific research is defined in the GDPR.

*« Where personal data are processed for scientific research purposes, this Regulation should also apply to that processing. For the purposes of this Regulation, the processing of personal data for scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research »* (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016)

GDPR has two exceptions for scientific research (Santos, 2020), allowing the use of non-anonymized data.

- Anonymization is not possible to achieve the research purposes
- Appropriate safeguards to protect personal data are put in place through technical and organizational measures, in particular to ensure the principle of minimization of risks. For example, by using pseudonymization if the purposes can be achieved in this way.

### **2.2.2 Federal Act on Research involving Human Beings**

In force since January 2014, the Federal Act on Research involving Human Beings (HRA) aims to *«protect the dignity, privacy and health of human beings involved in research»* (SR 810.30 - Federal Act of 30 September 2011 on Research involving Human Beings (Human Research Act, HRA), 2021).

This law applies for health-related research at an individual level, except for IVG embryos and data that has been anonymized or collected anonymously. The research topics covered are :

- The understanding of human diseases
- The structure and function of the human body
- Public health

The anonymization definition in the law is the following :

*«Anonymized biological material and anonymized health-related data means biological material and health-related data which cannot (without disproportionate effort) be traced to a specific person» (SR 810.30 - Federal Act of 30 September 2011 on Research involving Human Beings (Human Research Act, HRA), 2021)*

It is stated that the Federal Council specifies the requirements for correct and secure anonymization. A guide exists to precise the technical and organizational measures to protect personal data (FDPIC, 2015). Unfortunately, this document has not been updated since 2015.

Like the previous laws, the consent of the participant is at the center of data treatment. The participants must be clearly informed and give their consent to the data treatment. They also keep the right to revoke it at any time.

### **2.2.3 Good Clinical Practices**

Good Clinical Practices (GCP) are an international standard for clinical trials (EMA, 2016). They aim to ensure *« ethical and scientific quality for the design, conduct, performance, monitoring, auditing, recording analyses and reporting of clinical trials »* (Vijayananthan, Nawawi, 2008, p. 1). These guidelines are used worldwide to protect and preserve human rights.

The GCP present the role and responsibilities of each stakeholder in clinical trials.

To ensure ethical and scientific quality during a clinical trial, the GCP ask for the implementation of a quality assurance and quality control system. They must assure that *« data are generated, documented (recorded), and reported in compliance with the protocol, GCP, and the applicable regulatory requirement(s) »* (EMA, 2016, p. 31).

To verify the application of these guidelines, an on-site monitoring should be performed. Moreover, a centralized data monitoring should be put in place, to remotely evaluate the quality of data. This centralized data monitoring must be supported by *«appropriately qualified and trained persons (e.g., data managers, biostatisticians)»* (EMA, 2016, p. 41).

The centralized data monitoring resonates with data curation, as its goal is to help distinguish between reliable data and potential unreliable data with the following statistical analyses:

- *« To identify missing data, inconsistent data, data outliers, unexpected lack of variability and protocol deviations*
- *To examine data trends such as the range, consistency, and variability of data within and across sites*
- *To evaluate for systematic or significant errors in data collection and reporting at a site or across sites; or potential data manipulation or data integrity problems*
- *To analyze site characteristics and performance metrics*
- *To select sites and/or processes for targeted on-site monitoring »*

*(EMA, 2016, p. 41)*

Data curators for the quality control step also perform these tasks. Therefore, centralized data monitoring could be assimilated to active curation in the specific field of clinical trials.

## 2.3 Research data management

### 2.3.1 Definition

*« Research data management (RDM) is about creating, finding, organising, storing, sharing, and preserving data within any research process. » (Cox, Verbaan, 2018, p. 4)*

Research data management (RDM) is a term that encompasses all activities around the lifecycle of research data (Borghi et al. 2018; Henderson, 2017; Tammaro et al. 2018),

Research data management is a process realized throughout a research project. It is put in action through multiple facets and activities.

*« Les principales pratiques de gestion des données concernent: la sécurité et le stockage des données ; l'organisation des fichiers ; la documentation des fichiers et des données ; le nettoyage et recodage des données ; la transcription des données ; l'anonymisation des données ; le consentement éclairé ; le copyright ; le partage des données. » (FORS, 2018, p. 1)*

Henderson lists various aspects of data management (Henderson, 2017), such as file naming, data access, data documentation, metadata creation, controlled vocabularies, data storage, data archiving and preservation, data sharing and reuse, data privacy, data rights and data publishing. Therefore, through all these activities, RDM appears like a very large domain.

More than best practices to apply, research data management has become mandatory, due to the exponential amount of data produced in research. Funders require a Data Management Plan which force the research teams to think and plan their data management before data collection. Sharing data at the time of the publication is also required by publishers, for reproducibility purposes, and by public funders, to prove a good use of the public funds.

### 2.3.2 The research data lifecycle

The research data lifecycle (Figure 5) created by the UK Data Service is one of the most used in research data management field. As research data lifespan often extends beyond the research project, its lifecycle includes the concepts of preservation and reuse.

Figure 5 : The research Data Lifecycle



(UK DATA SERVICE, 2019)

The research data lifecycle is sometimes adapted to a specific field, like for biomedical science. Here is the example of the Research Data Management Lifecycle Checklist created by the Longwood Medical Area (LMA) Research Data Management Working Group at Harvard University (see Figure 6). This lifecycle has been combined with research data management best practices to facilitate their understanding and application.

Figure 6 : The Biomedical Data Lifecycle (Harvard University)



(LMA Research Data Management Working Group, 2022)

## 2.4 Research data curation

### 2.4.1 Definition

The distinction between the terms “research data management” and “research data curation” is not clear in the information science field (Pham, 2018). The activities related to one or the other are not properly defined. This situation leads to a permanent confusion about data curation definition. For example, DoRANum defines data curation as the activities and operations required to manage research data through its lifecycle (Ministère de l’Enseignement Supérieur, de la Recherche et de l’Innovation, Inist-CNRS, GIS Réseau URFIST, 2022). This definition is quite similar to the previous ones on research data management. Literature offers more definitions, sometimes more precise.

« [...] digital curation is a multifaceted effort to ensure both current and long-term access to and use of digital content. » (DeRidder, 2018, p. 3)

« Work performed to ensure meaningful and enduring access to data » (Tammaro et al. 2018, p. 1)



« Curation encompasses the active policies and procedures needed to preserve research materials outside the framework of an original research project and beyond its original (short term) objectives. »  
(Rice, Southall, 2016, p. 31)

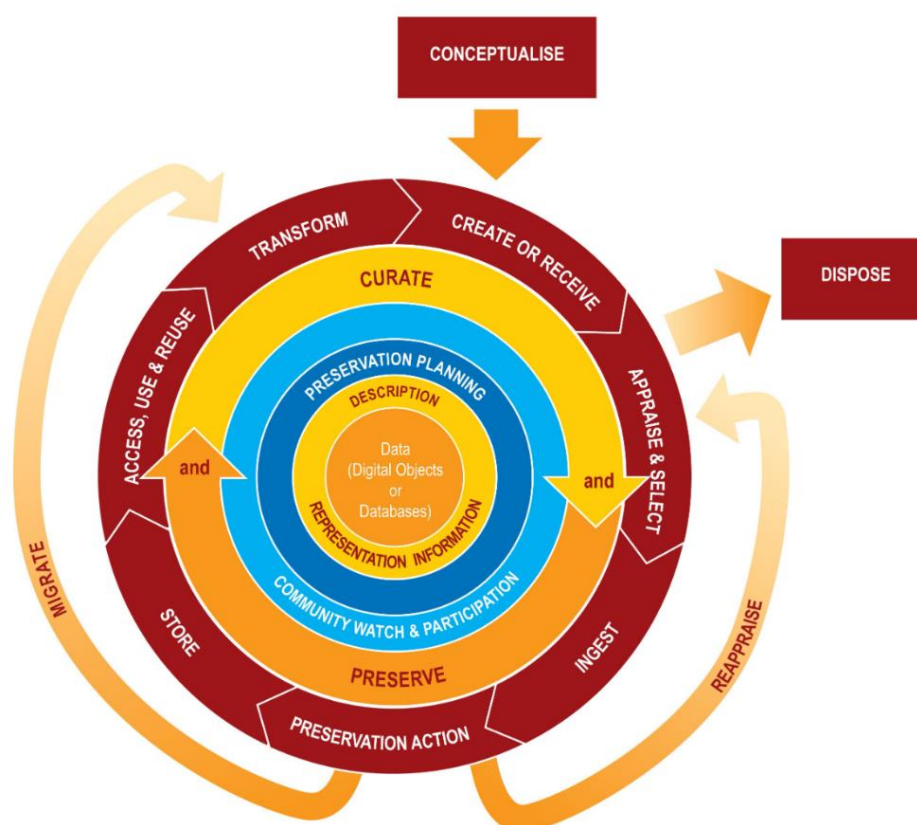
« La curation des données est une activité essentielle dans la pratique de gestion des données, car elle assure la pérennité des données sur le long terme, leur qualité et leur réexploitation. Elle s'avère toutefois difficile à définir, car sa pratique se situe très souvent à la croisée de différentes disciplines. Elle s'applique tout au long du cycle de vie de la donnée et intègre des tâches de nature parfois différentes comme la sélection, la vérification, la normalisation ou encore l'enrichissement nécessaires à la publication des données. »  
(Hadrossek et al. 2021, p. 12)

« Data curation is the active and ongoing management of data with the object of adding value to data to make it more usable in the future »  
(Henderson, 2017, p. 58)

« Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle »  
(Digital curation center, 2022)

The idea of long-term preservation is present in all these definitions. However, data curation is not only about preservation or archiving, even if they are important aspects of curation (Harvey, Oliver, 2016). Data curation is a way to care about data, to add value to it and to ensure its quality. It has been modeled by the Digital Curation Center to be performed during all the data lifecycle (see Figure 7).

Figure 7 : The Curation Lifecycle Model



(Higgins, 2008)

## 2.4.2 Activities

Data curation is often realized when data must be shared. Johnston explains the data curation activities by following the Data Curation Model (Johnston, 2017). For all steps, activities are presented and detailed with methods and advice. Table 1 summarizes the tasks.

Table 1 : Data curation activities by Johnston

Steps	Activities
Receive the data	<ul style="list-style-type: none"> <li>To securely transfer data to the repository</li> <li>To obtain information necessary to use and understand data (metadata and documentation)</li> </ul>
Appraisal and selection	<ul style="list-style-type: none"> <li>To choose the appropriate repository</li> <li>To define if curation is needed based on potential long-term value for reuse</li> <li>To consider any risk factors (sensitive information, copyright...)</li> <li>To inventory the submission (number of files, type, size)</li> <li>To capture organization of the files and technical information.</li> <li>To select : accept or reject the data</li> </ul>
Processing and treatment	<ul style="list-style-type: none"> <li>To create a copy of the data to secure the original</li> <li>To start a curation log</li> <li>To understand the file organization, relationship and file naming</li> <li>To review the content of the data files</li> <li>To check for quality : missing data, ambiguous headings, code execution failures</li> <li>To detect hidden documentation</li> <li>To verify all metadata and documentation</li> <li>To consider file format and transform</li> <li>To organize and rename files to optimize their meaning</li> </ul>
Ingest and store	<ul style="list-style-type: none"> <li>To maintain integrity and fixity (e.g., checksums) through Ingest process</li> <li>To store the data on a well-configured archival storage environment</li> </ul>
Descriptive metadata	<ul style="list-style-type: none"> <li>To create and apply descriptive metadata, including technical and provenance metadata</li> <li>To structure/present metadata in multiple schemas to facilitate discovery</li> </ul>
Access	<ul style="list-style-type: none"> <li>To determine access conditions and apply control access</li> <li>To apply Terms of Use, licenses, and copyright</li> <li>To contextualize the data in the discovery and access environment</li> <li>To generate a persistent identifier</li> </ul>
Preservation	<ul style="list-style-type: none"> <li>To rely on the plan for long-term reuse that anticipate format obsolescence and storage failures</li> <li>To actively monitor integrity and reusability of the data files</li> <li>To apply digital preservation strategies</li> </ul>
Reuse	<ul style="list-style-type: none"> <li>To monitor reuse with metrics on citation and downloads</li> <li>To provide ongoing support</li> </ul>

(Adapted from (Johnston, 2017))

The Data Curation Network has synthesized the important activities performed by a data curator in another way, without the data lifecycle, by creating the checklist CURATE(D) (Data Curation Network, 2022), which stands for :

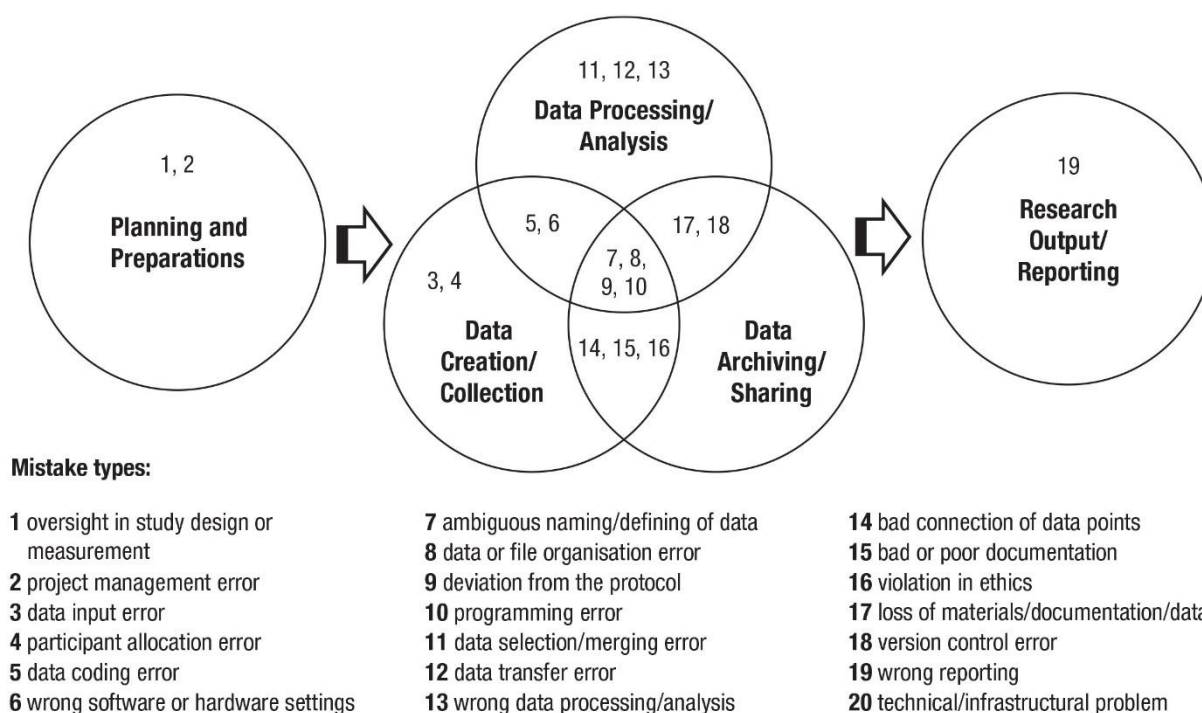
- « **Check** files and read documentation (risk mitigation, file inventory, appraisal/selection)
- **Understand** the data (or try to), if not... (run files/environment, QA/QC issues, readme)
- **Request** missing information or changes (tracking provenance of any changes and why)
- **Augment** metadata for findability (DOIs, metadata standards, discoverability)
- **Transform** file formats for reuse (data preservation, conversion tools, data visualization)
- **Evaluate** for FAIRness (licenses, responsibility standards, metrics for tracking use)
- **Document** your curation activities (Curator Log, correspondence)»

The checklist CURATED is less centered on preservation, like monitoring integrity and fixity, than the list of activity made by Johnston. However, there are common activities in these two visions of data curation. The first step is to obtain documentation and metadata to understand the data and its organization. The identification of the risk factors, like re-identification, is the second common step. Then, quality control is performed to identify issues such as missing data or ambiguous headings. Correction of these issues is realized either by the research team or the data curator, depending on the support service. The fourth common step is to complete the dataset with descriptive metadata. The file transformation is also a common step, to facilitate reuse and preservation of the data. Finally, evaluate FAIRness is the last activity in common.

With these activities, data curation is a complete process to ensure data quality and to prepare it to be shared. However, definitions indicate that data curation is an ongoing process, but key activities are still focused on discovery and data preservation (Kouper et al. 2021; Pham, 2018).

To ensure data quality and to reduce time of curation at the end of the project, it is recommended that data curation should be realized as soon as possible (Harvey, Oliver, 2016). This could prevent major mistakes identified in research data management practices (Kovacs, Hoekstra, Aczel, 2021) (see Figure 8). Integrating data curation practices, traditionally done before sharing, earlier in the process is called active curation. Some synonyms are used in literature, like “active data management” or “data preparation”.

Figure 8 : Mistake types categorized by research data management stage



(Kovacs, Hoekstra, Aczel, 2021)

## 2.5 Active data curation

### 2.5.1 Definition

Active curation is data curation that can be done during the research project. As this concept occurs in research teams, its tasks may be named differently: active research data management, data cleaning, data preprocessing, data processing or data treatment.

All these activities are part of the research project and are not always considered as curation. What will bring them into the “curation” appellation is their aims. If the purpose of these tasks is to improve quality, to reduce risks, to allow reusability or to facilitate archiving, they are completely under the curation definition.

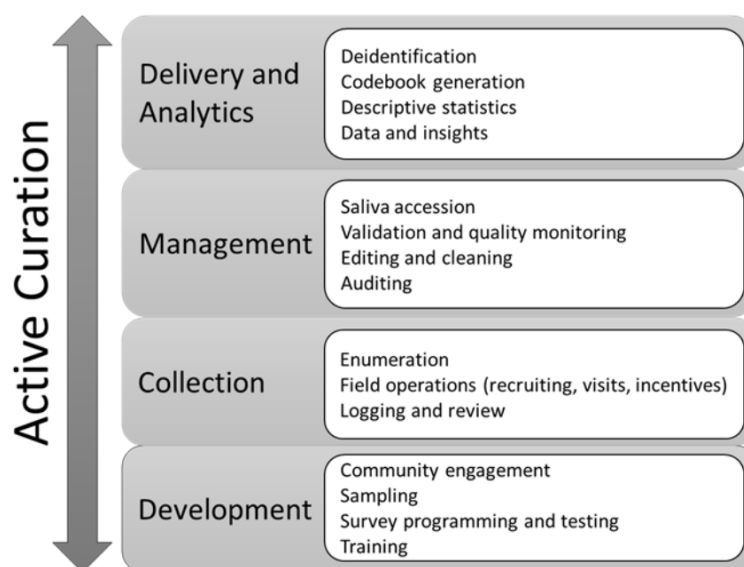
In the literature, active data curation is often focused on data preparation and cleaning, with the aim to improve data quality. The activity of data cleaning is known as an important step in a research project but is considered as time consuming. Therefore, research projects on data curation are interested in automation of data preparation. But active data curation should not be reduced to this vision.

Active curation has been modeled on a longitudinal survey case study (Kouper et al. 2021). This project extracted 15 tasks in 4 categories, covering all the data lifecycle (see Figure 9). The definition used for active curation in this project is the following :

*« We use the notion of active curation to describe activities and decisions about the data objects that are “live,” i.e., when they are still being collected and processed for the later stages of the data lifecycle. »*  
(Kouper et al. 2021, p.1)

These activities are not only done on research data, but also include training of the research team, sampling and recruiting. Active curation has to consider how data are produced to be able to improve its quality.

Figure 9 : The survey project curation components performed by the survey research center



(Kouper et al. 2021)

## 2.5.2 Activities

By parsing and merging data curation activities and research data management activities, it occurs that many of them are similar. Moreover, best practices in these two disciplines are to implement the recommendation as soon as possible, as active data curation. For example, the use of open software and file format are recommended since the beginning of the project, for all documents and data.

Some aspects of active data curation are recurrent in literature and should be highlighted as ways to improve data quality, FAIR compliance and facilitate preservation and sharing of the data.

### 2.5.2.1 Data collection

In an active curation context, data quality begins during the planning phase of a project. The transformation of research questions into variables is the first step to create an appropriate data collection tool. By improving the methodology to create a data collection tool and by keeping in mind the potential quality issues while creating it, data preparation and cleaning could be facilitated.

Then, to minimize errors in data capture, the design of the survey is important. To choose the right type of field and to set it up correctly can improve data quality by avoiding inconsistencies or out-of-range values.

The idea of interoperability must also be considered at this stage. The use of ontologies, thesaurus and controlled vocabularies for concerned field will help to standardize the dataset. Moreover, their usage is recommended to improve the FAIRification of the datasets, as long as this is well documented for a future reuse (Hank, Bishop, 2018). Multiples ontologies exist in medical field, like SNOMED CT<sup>5</sup> or LOINC<sup>6</sup> (Egli et al. 2020). The use of interoperability

<sup>5</sup> <https://bioportal.bioontology.org/ontologies/SNOMEDCT>

<sup>6</sup> <https://loinc.org/document-ontology/>

functionalities into the data collection tool is also a possibility to increase the data quality from the beginning of data collection. For example, in REDCap, the Clinical Data Interoperability Services (CDIS)<sup>7</sup> is a feature allowing a survey to be linked to an electronic health record. This kind of functionality may not be accessible in Switzerland, as the electronic health record is not standardized through all the country.

Finally, to anticipate the re-identification risk, only necessary data should be collected, especially socio-demographic data. Moreover, coding and aggregation of values could be integrated at this stage. Creating categories can reduce disclosure risk. However, a balance has to be found. If data are too categorized, it would be impossible to match them with other dataset (lack of interoperability).

### **2.5.2.2 Data preparation, cleaning and monitoring**

Through the literature, active data curation is commonly associated with data preparation and cleaning. These processes are already included in the research projects workflow and often associated with the data curation concept. Therefore, research is done on this domain to automate or to facilitate data cleaning activities (Álvarez Sánchez et al. 2019; Pezoulas et al. 2019; Corrales, Corrales, Ledezma, 2018; Konstantinou et al. 2019).

Data cleaning, automated or not, will improve data quality and integrity. The recurring tasks associated to it are the following ones, performed after data collection, but before analysis.

A first step is to create descriptive statistics. It is important to count records and compare it to the expected number. This can facilitate the identification of duplicates or missing data.

A second step is to compare codebook and dataset to ensure that all variables are correctly named and documented and to verify that data type in the dataset match the announced data type in the codebook. This comparison leads to more advanced verifications, such as identification of outliers' values. Depending on the design of the data collection tool, out-of-range values could occur, for example a 28 years old person in a study on 65+ years old people. These kinds of errors must be detected, documented and corrected if possible.

The identification of missing data is also an important step of data preparation. They should be identified and documented. In many studies, missing data are expected, therefore, it is not always suitable to correct them.

Finally, consistency checks will improve quality. This verification involves complex relationships among variables (ICPSR, 2022). For example, an inconsistent data would be a person declaring to have had no income in one variable, and then indicating in another variable to have declared 100'000 CHF to taxes.

After identifying these anomalies, researchers correct them by cleaning the data. This process can be realized with specific proprietary software like Altair, Paxata, SAP, SAS, Tableau, Talend, Trifacta (Hameed, Naumann, 2020) or Stata, with open source software like OpenRefine (Gaudinat, 2021; Kouper et al. 2021), YARD (Peer, Dull, 2020) or TAQIH (Álvarez Sánchez et al. 2019), or with scripts in R or Python (Hundepool, 2012; Corrales, Corrales, Ledezma, 2018; Pezoulas et al. 2019).

---

<sup>7</sup> <https://projectredcap.org/software/cdis/>

Some of these activities could be performed during data collection by using functionalities included in the data collection tool (e.g., REDCap) or by programming a monitoring workflow.

Outside of data cleaning, preparing a data set could include variables coding (ICPSR, 2022), data correction, harmonization and enrichment (including semantic web). These kinds of treatments could be done with a tabular software, with R or statistical software, with regular expressions, with command lines, with programming (Python) or with a specific software, such as OpenRefine, Trifacta Wrangler, Tableau Prep, Talend or Dataiku (Gaudinat, 2021).

### **2.5.2.3 Documentation**

Documentation is one of the most important tasks in data curation from the long-term preservation point of view. It allows data user to understand the data without any help through time. Moreover, documentation is a big step to FAIRification of the data, as it will improve its discoverability, its accessibility and its possibility to be reused.

There are different levels of documentation. The first one is contextual documentation. This type of documentation will describe the project, the stakeholders, the funding, the objectives and hypothesis (FORS, 2018; Harvey, Oliver, 2016; Henderson, 2017). This kind of information will mostly be found in the research protocol, other related documents and metadata.

Documentation on the data itself should be collected. A codebook or data dictionary can describe the variables and how they are set (FORS, 2018). This file can be created from scratch or generated with the data collection tool. Descriptive information about data, such number of records or number of missing values could also be included in this documentation.

Documentation on data treatment (modifications, preparation, analysis...) is very important for reproducibility. This kind of information is mainly recorded in the log files of the database, in electronic lab notebooks (ELN) or in the data cleaning script. This could also take the form of an audit trail.

Finally, technical and administrative documentation are important for long-term preservation. This will include information about provenance, copyright, files format, MIME-types, software used, versioning or relationship between objects. This kind of information is useful to share the data, but also to plan its preservation.

All this information allows potential users to understand the data and the context of creation and analysis. This documentation should be available in structured metadata to enable machines to access it. This improves FAIRification but also improves the development of discoverability software. For example, when the codebook is in XML or RDF, it is possible to explore it with a search engine. For metadata, different XML schemas apply (METS, PREMIS, DDI...), depending on the type of metadata : administrative, descriptive, technical, relational or structural.

Some information could be automatically extracted from the data collection tool or the statistical software (Johnston, 2017; DeRidder, 2018). If this documentation would be created and transformed in XML as soon as it is available, this would prevent a time-consuming information retrieval at the end of the project. Information on data preparation in particular could be required by the publisher to ensure a correct peer-reviewing. If this step is well documented, it would be a gain of time for publication.



## 3. Results : data curation practices at Unisanté

### 3.1 Methodology

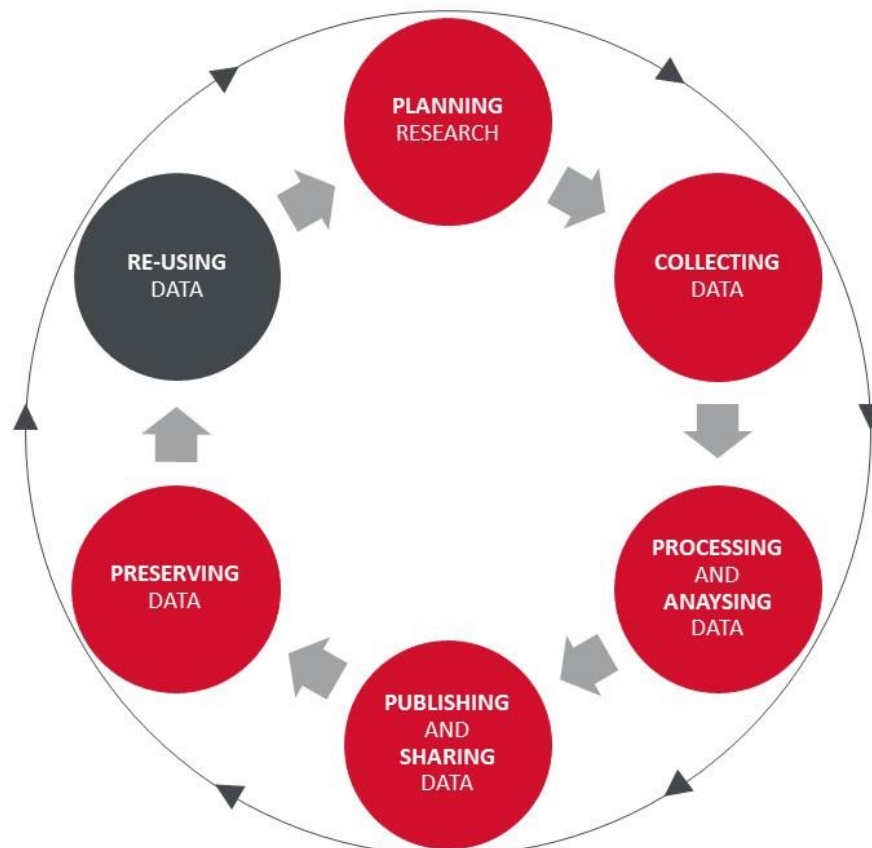
#### 3.1.1 Data collection

##### 3.1.1.1 Survey

A questionnaire has been created to collect quantitative data about research teams' practices and needs. It has been constructed on professional literature review. All activities and best practices related to data curation, data management or data quality have been listed. As the list of tasks was very long, groups of activities have been made to facilitate the questionnaire. Then all the methods or tools that could be used to realize these actions have been specified. The structure of the resulting table and the questionnaire are both available in [Annex 3](#) and [Annex 4](#). After user testing, the questionnaire has been reduced once again by removing some of the available answers and by adding free text field at some questions.

Tasks have been presented within the research data lifecycle into the survey, as this concept is usually used when presenting research data management to research teams (see Figure 10). Four groups have been created, to ask questions about the five main steps of the lifecycle : Planning research (group 1), Collecting data (group 2), Processing and analyzing data (group 3), Publishing, sharing, and preserving data (group 4).

Figure 10 : The Research Data Lifecycle : steps included in the questionnaire



(Adapted from (UK DATA SERVICE, 2019))



For each step of the data lifecycle, five main questions have been asked :

- Do you, or somebody else, realize this activity (related to the step) ?
- For which reason do you or do you not do this activity ?
- How do you perform this activity ?
- Would you like a support for this activity ?
- What kind of support would you prefer ?

The survey has been realized in French.

#### **3.1.1.2 Document review**

A document review has been realized to identify recommendations and processes already in place. The included documents were available on Unisanté public website, Unisanté internal website (intranet) and also as internal files that document processes within the organization (quality documents).

The results of this review have been used to prepare the interviews.

#### **3.1.1.3 Interviews**

The results of the questionnaire have been completed with semi-structured interviews.

For interviews with research collaborators, the purpose was to get qualitative insights from the data managers or researchers, who have advanced knowledge and practice in data curation. A list of questions has been created, available in [Annex 5](#), and was focused on which task they performed, how they did it and if they wanted support. The questions about the reason why these activities were realized have been avoided to gain time.

For interviews with heads of each support services, the goal was to acquire information on processes, tools or precise recommendations made to the researchers that were not documented. The list of questions is available in [Annex 6](#).

No exact transcription of the interviews has been made. The interviewer took notes during the interview. After the meeting, the interviewer sent a summary of the notes for reviewing and approval to the interviewee. All interviews have been conducted during face-to-face meetings.

All interviews have been realized in French, to facilitate communication and understanding. The results of the interviews were translated by the author of the present work.

#### **3.1.1.4 Literature review**

To provide a theoretical basis to this work, a literature review has been done. Articles, books, reports, presentations, video and websites have been harvested. The research has been made on Google, Google scholar, LISTA database, Renouvaud and Swisscovery. Documentation issued from previous searches has also been used (De la Lama, Racine, 2022; Racine, 2021).

The main terms for this review were “Data curation”, “Active data curation”, “Research data management”, “Researchers activities”, “Data preparation”, “Data cleaning”, “Research workflow”, “Curation workflow”.

The results of this literature review have been presented in the [State of the art](#) section of this thesis.

## **3.1.2 Sampling**

### **3.1.2.1 Research teams**

The first targeted population was the research collaborators. Two hundred and nineteen people have been identified as a research collaborator. This estimation comes from the Human Resources service (HR), completed with a listing from the DFRI. A research collaborator could have a full-time research activity, or a part-time research activity completed with a clinical, paraclinical or administrative activity.

To reach all this population, the distribution of the survey has been made through an institutional newsletter. This kind of sampling is defined as a probabilistic random sampling (Fortin, Gagnon, 2016), as the researchers had the possibility to accept or decline the participation to the survey. Then, recalls have been stratified to be more efficient. The administrative unit of each department has been contacted and have sent recalls to their own research community.

For the interviews with research collaborators, the survey integrated a specific question for those who would be interested in a meeting. To complete this method, some data managers at Unisanté have been contacted directly to get an interview. A snowball effect was expected, as they had the possibility to transfer the invitation email to somebody else in the institution, if they thought this was appropriate or if they could not participate. But this effect did not occur.

### **3.1.2.2 Support services**

The second targeted population was the head of the support services. As they help research teams during their projects, they recommend methods or tools. Therefore, they have a good vision of practices in research data management at Unisanté.

A non-probabilistic sampling has been made. The support services have been identified according to the following criteria :

- Is part of Unisanté
- Is available for Unisanté research teams
- Offers a theoretical or practical support on one or more research data management aspects

The support services available for Unisanté researchers dispensed by other affiliated or partner institutions (UNIL, CHUV, ...) have not been considered in this study.

Six support services have been identified :

- Biostatistics consultation unit (UCB)
- Survey Methodology unit
- Qualitative research platform
- Research promotion unit (UPR)
- Research IT Services unit
- Documentation and data unit (UDD)

The head of each service, or the contact person, has been reached by email to obtain an interview.

### 3.1.3 Response rate

#### 3.1.3.1 Survey

With 45 complete answers to the survey (out of 219 sent invitations), 20% of the targeted population has participated to the study (see Table 2).

Thirty incomplete answers have been received. The respondents have abandoned the survey after the 1<sup>st</sup> page (socio-demographic questions). As they did not contained information on data curation practices, they have been excluded from the analysis.

Table 2 : Response rate by department

Department	Researchers	Answers	Rate
Department of epidemiology and health systems (DESS)	87	20	23%
Department of education, research and innovation (DFRI)	47	11	23%
Department of occupational and environmental health (DSTE)	38	8	21%
Department of vulnerable populations and social medicine (DVMS)	17	2	12%
Department of ambulatory care (DDP)	12	6	50%
Department of family medicine (DMF)	11	3	27%
Department of health promotion and prevention (DPSP)	6	1	17%

This response rate is not as high as expected but can be explained by a few factors.

The first one is the real size of the population. The announced size of the sample is based on the HR listing. However, there is no distinction between researchers active in quantitative research projects or qualitative research projects. This separation is important, as this thesis is only interested in tabular data, and exclude transcriptions from interviews or focused groups. There is no estimation available on the number of concerned researchers. Therefore, the size of the targeted population hasn't been reduced, which explain the response rate.

The size of the survey could also explain the lack of answers. The newsletter indicated an estimated duration of 30 to 40 minutes. This could have discouraged researchers to fill in the questionnaire.

Another explanation to this response rate is the timing. The survey was active from the 15<sup>th</sup> of May to the 12<sup>nd</sup> of June. This time period is usually very busy, with research project and students mentoring. If the survey would have been done at another time, it could have harvested more answers. Moreover, the SNSF had an active call for proposal due to the 15<sup>th</sup> of June, therefore many researchers have not found time to answer the questionnaire due to their amount of work.

Finally, the status of the research project could have had an impact on the response rate. Two research teams have reached the UDD to indicate that their project was in a pilot phase and that they would not answer the survey, as they did not have data-related activities.

### 3.1.3.2 Interviews

The initial idea was to meet at least one person per department for an interview. However, only four interviews have been realized with research collaborators. As June was a busy month for researchers, some people could not find time to plan a meeting. Moreover, after two interviews, it occurs that the profiles and experiences of the researchers were too different to follow a similar pattern of questions. The interviews realized with the data managers lasted about two hours, whereas only one hour had been planned. This time extension is due to the topic, which is too wide to be covered in one hour. The decision to diminish the number of interviews has been made after this observation.

However, these interviews were able to harvest qualitative data on the main tasks performed during a project, the potential improvement that could be done at Unisanté and the institutional support that could be put in place.

For the interviews with head of support services, four interviews have been realized. The head of the qualitative research platform has not been met, due to the scope of their services, which was outside of this study. The documentation and data unit (UDD) has been added to the result of this thesis, as it offers support for research data management. However, the information was not gathered by interview, as the author of this thesis is a collaborator of the UDD and the main provider of research data management support.

## 3.2 Researchers' practices

This section presents the results of the objective 2 «to evaluate practices and needs at Unisanté» for the research teams at Unisanté. The data have been collected with a survey and interviews, according to the presented [methodology](#).

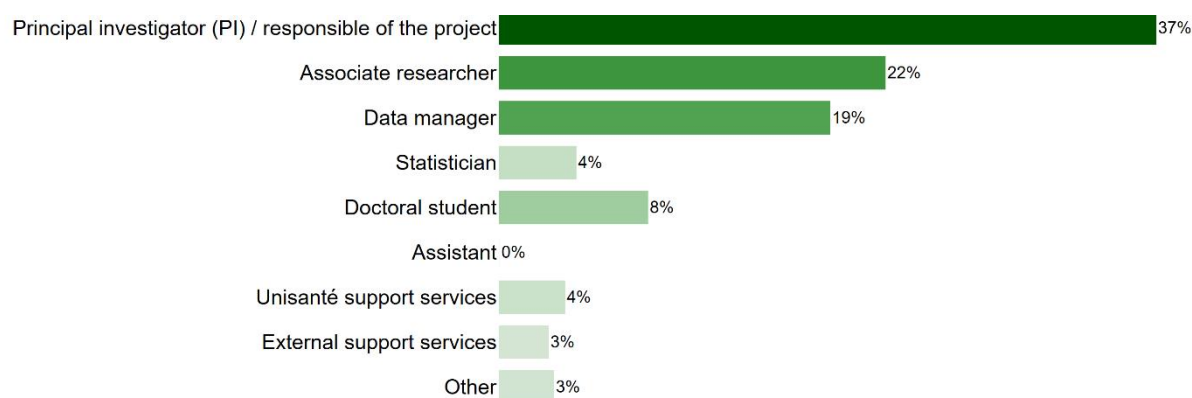
### 3.2.1 Planning research

#### 3.2.1.1 Roles

According to the survey, at the beginning of the project, the principal investigator (PI), or responsible of the research project, is the most implicated person (see Figure 11). This information has been extracted from the second question of the survey “What is your role in the research team ?” and from the answers to the questions “If you do not perform the task yourself, who do it ?”.

As an example, to create the variable “Principal investigator (PI) / responsible of the project” shown in the figure below, an addition has been done between the answers from a person with the role “PI”, who indicated doing the activity by herself or himself, and the ones from people with another role who indicated that the task was performed by the responsible of the research project.

Figure 11 : Planning research : most implicated roles

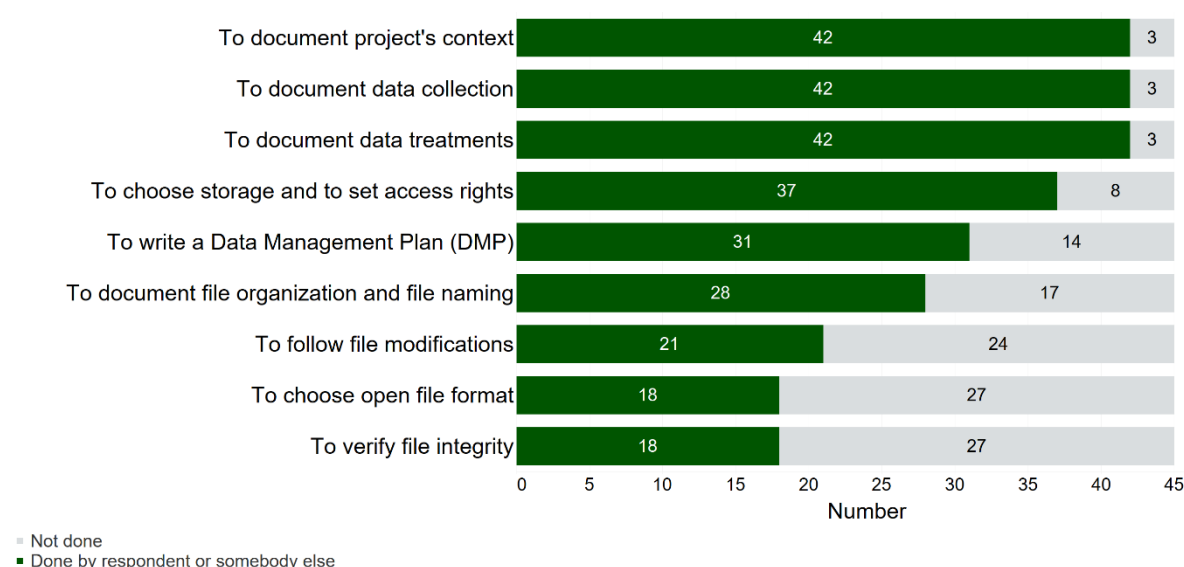


### 3.2.1.2 Activities

The question «Do you, or somebody else, realize this activity?», which appeared in each step, had three possible answers : «I do it myself», «Somebody else do it» and «I do not do it». To analyze these data, the two first answers have been considered as «task done». Therefore, if the respondent chose one of them, the activity has been considered as realized. This pooling has been done to facilitate the analysis of the data, as the important information to extract was the identification of the activities not done. Therefore, a merging of «I do it myself» and «Somebody else do it» was needed. This specific aspect of the survey is discussed in the [Limitations of the study](#), as it could be improved.

The following results could have been analyzed by department. However, due to the few answers in some of the (see [response rate section](#)), I decided to analyze the data at the Unisanté level. An exploration of the activities by department could be done after this thesis, by collecting more data.

Figure 12 : Planning research: activities



As presented in Figure 12, the main activities performed at the planning stage is the documentation of the project, by documenting its context, the planned data collection, and planned data treatments (42 answers). Documenting the organization of the files and their naming convention is also done, but by less people than the other documentation tasks (28 answers).

These activities are done to improve project and data quality. The documentation is done in a textual file (word or txt) by 64% of the respondents. Then, the tabular file (Excel or csv) is chosen by 27%. Finally, documentation with statistical software (Stata or SPSS), structured metadata (XML, JSON or RDF), scripts (R or Python) or REDCap is realized only by 2-3% of the respondents.

For the choice of storage and the choice of access rights, done by 37 people, the support and administrative services are strongly implicated (11 answers). This is due to the storage system used at Unisanté. Unisanté has a common server, hosted on the secured server of the University hospital CHUV, but managed by Unisanté. To create a folder with specific users' access on this server, research teams must require its creation to the Unisanté logistic service. The request can be made through the administrative respondent of the department or through the IT respondent of the department. To choose the right storage facility, if the common server is not the best option, a discussion occurs between the IT service and the research teams (8 answers).

Two thirds (31/45) of the respondents write a data management plan, because it is mandatory (12 answers) or to improve project or data quality (6 answers). Four people indicated that they do not write a DMP because it is useless, four others because they lack of knowledge about DMP, and two others because a DMP is incompatible with their project. Another explanation could be due to the funder of the study. The SNSF require a DMP, but not all projects at Unisanté are funded by it. Therefore, the DMP is not mandatory for all projects and is not part of the habits of the research teams. Sixteen respondents indicated that they write their DMP in file text (Word).

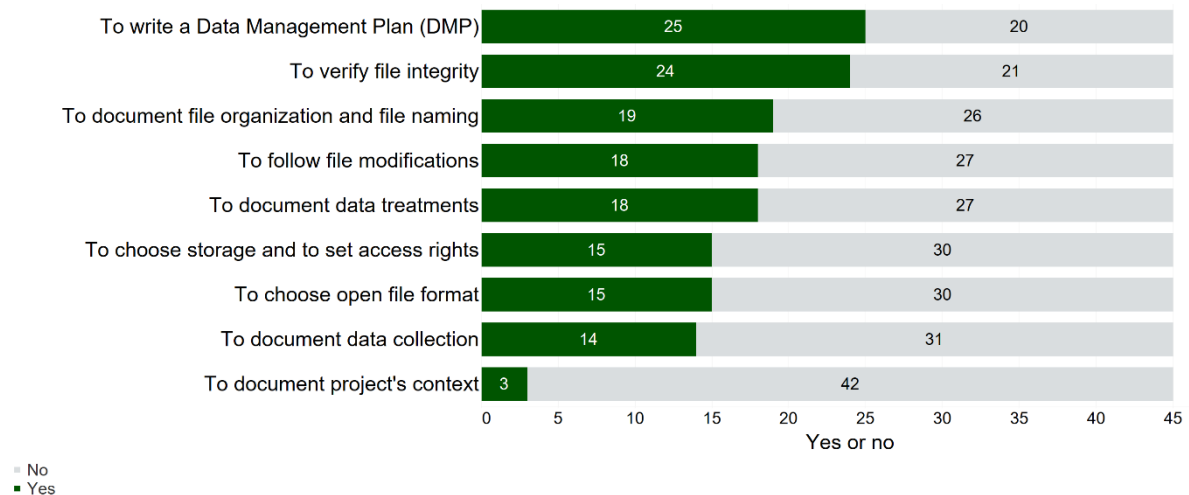
Finally, the less performed tasks are following the modifications of the files (not only data), using open file format or software, and implementing a system to verify file integrity. The main reason why these activities are not realized is the lack of knowledge (13 answers) on how to implement such verification systems. This result was expected. When the research teams use the common server, this storage facility is protected against attacks or corruption. Therefore, the research teams do not feel the need to implement additional security. For the research teams who put in place such verification systems, this is done through a textual file (2 answers), a tabular file (2 answers) or checksum (1 answer). For modification tracking, this is usually performed on data but not on the other files. The storage allows recovering of previous version of a file, which is sufficient in most cases for non-data files. For the research teams who track modifications since the beginning of the project, this is done through a textual file (6 answers) or a tabular file (3 answers).

Regarding the non-use of open format, the reason was different. The information system used at Unisanté is very restrictive and does not offer a lot of open software. Therefore, it is difficult to install an open software if it has not been approved by the University hospital IT service, as they are the provider of the server. Some research teams try to use more open format for their data, as the data managers and statisticians recommend csv for tabular data. However, this is not always possible, depending on the data and functionalities needed.

### **3.2.1.3 Need of support or improvement**

Through the survey, research teams have expressed their need of support (see Figure 13).

Figure 13 : Planning research : need of support



For the planning step, 25 respondents ask for a support for DMP generation. This support already exists in the institution, through the documentation and data unit (UDD).

The verification of file integrity (24 answers) and file modifications (18 answers) should also be supported, as they are one of the less performed tasks. A support for documentation of organization and file naming would be appreciate by 19 respondents and a support for data treatments documentation by 18 respondents.

Between 14 and 15 respondents asked for support on the choice of a storage, the use of open format and the documentation of data collection.

To provide the most useful support at Unisanté, the last question of the group was «What kind of support do you prefer?». This question was not specific to a selected task, therefore it is not possible to match a kind of support with an activity.

Twenty-five respondents would prefer a support through a dedicated person who could answer to question when needed, or could offer a practical help (20 answers). Fifteen respondents would appreciate to get a support through a training or a theoretical help (guidelines, manual...). The use of a dedicated tool (13 answers) or the realization of the task by somebody else (10 answers) would be less appreciated.

Finally, the comment section of this group highlighted that supports were already available at Unisanté (2 answers) for the Planning step. The research teams also took the opportunity here to indicate that some of the listed activities were not realized due to lack of time or money (2 answers).

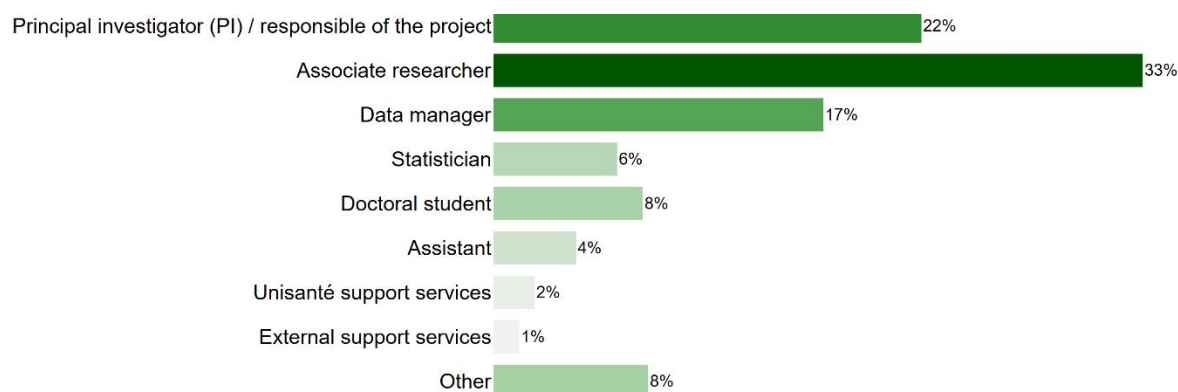
The need of an overall sensibilization, training or general guidelines on research data management in general has been highlighted in the comment section (2 answers) and in two interviews. The concerned people estimate that research data management is larger than just cleaning the data or write a DMP and wish to have more knowledge on how to manage their research data and where to obtain help and support.

## 3.2.2 Collecting data

### 3.2.2.1 Roles

According to the survey, when it comes to data collection phase (creating data collection method and collecting the data), the associate researcher is mainly active in this step. The PI or the data manager are also part of these activities, but in a less extent (see Figure 14).

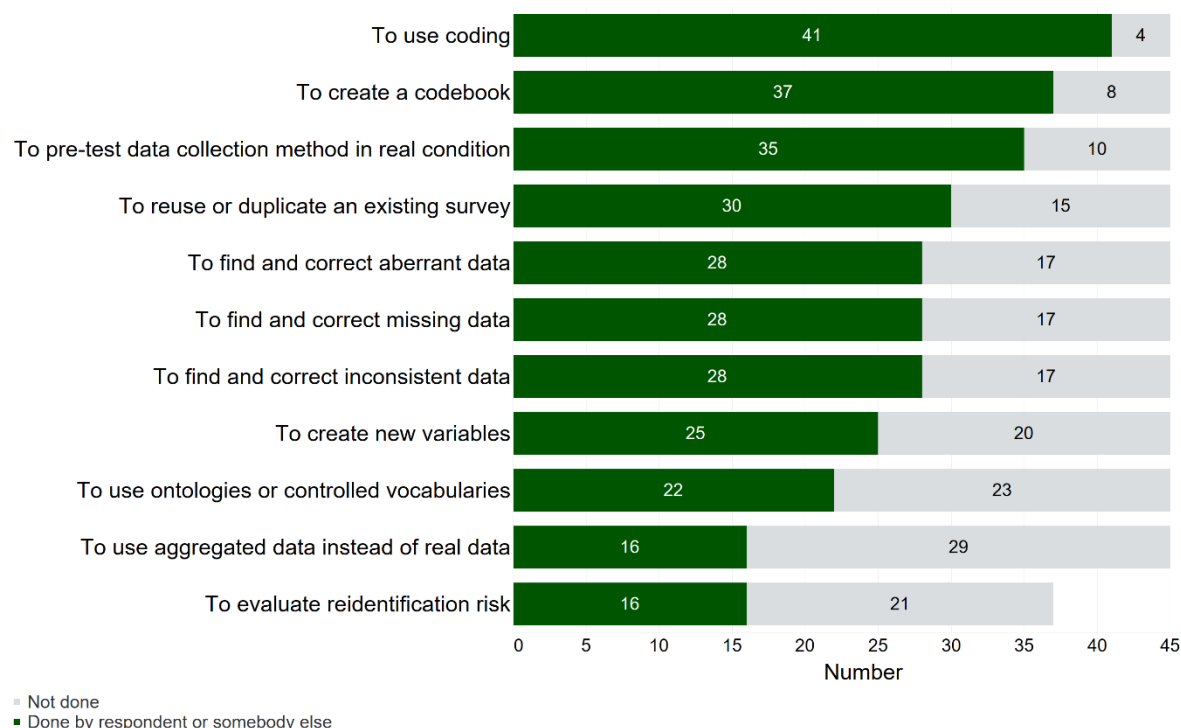
Figure 14 : Collecting data: most implicated roles



### 3.2.2.2 Activities

For this step, activities related to the creation of the data collection tool and to the data capture itself have been proposed (see Figure 15). Some of the vocabulary used has been confusing for researcher and more discussed during interviews.

Figure 15 : Collecting data : activities



Regarding the creation of the data collection tools, 41 people integrate coding into the data collection tool (e.g. men = m, female = f), then the creation of a codebook is done by 37 respondents, the pre-test of the collection method by 35 persons and the re-use of an existing



survey by 30 people. The reason behind these activities is mainly the project itself or the data quality.

Coding the values of a survey is mainly done through REDCap (12 answers), then in a specific software (11 answers), a tabular file (7 answers), through a statistical software (Stata or SPSS) (1 answer) or R (1 answer). On the other side, the creation of a codebook seems less dependent on the data collection software, as it is mainly created with a tabular file (Excel or csv) (11 answers), followed by REDCap (5 answers) or a specific tool (3 answers). The interviews underlined the importance of the codebook. This kind of documentation is considered as the most useful to understand the data. It helps to understand the data and how they were collected. This documentation is also useful in terms of interoperability, as it can explain if an aggregation of data has been decided and how it has been done, or if an ontology or a controlled vocabulary has been used. It will detail all the variables and facilitate their reuse and their matching with other data.

Three interviews highlighted the importance of a good setting of the data collection tool. The variables need to be set in the proper way, first by choosing the correct field type (e.g. a numerical field for age) and second by integrating simple quality control to the fields, such as range control. The interviewees indicated that a balance must be found between control and usability. For some projects, it is important to give some liberty to the person who will introduce data into the data collection tool. It is not always possible to force control on range or limit the possible answers. Data quality is strongly influenced by the data collection method, therefore its creation should be handled with care.

The less performed tasks realized when creating a questionnaire or a collection tool is the use of ontologies (22 answers) or aggregated values (16 answers), and the evaluation of re-identification risk (16 answers).

The use of ontologies or controlled vocabularies is not done by all research teams because 10 of them do not know what it is or 7 of the respondents find it useless or inapplicable. The interviews highlighted that those ontologies do not exist for all subjects treated at Unisanté, and are sometimes too detailed, or not enough, to be useful for the project.

The use of aggregated values, recommended to minimize re-identification possibilities, is not realized by 29 respondents. The main reason is that this is useless (14 answers) or that the respondent does not know what it is (6 answers). Another explanation has been provided in the comments of the «other» answers and interviews: The use of aggregated values diminishes the quality of the data. In a research project, collecting the real data is useful to answer correctly to the research questions. Moreover, collecting real data improve interoperability, as aggregation is not done in the same way through research projects.

Finally, the evaluation of the re-identification risk is not done by 29 people, 8 of them are not concerned as they do not work with personal data and have not seen the question. For the other ones, 11 people do not know how to realize this activity, 4 found this useless and 2 do not know what it is. Through the interviews and comments in the questionnaire, it has been highlighted that this risk is handled by the choice of variable, based on personal experience and good sense. For example, the research teams are aware that a person with a visible handicap in a small town is easily identifiable with only one or two variables. For the ones performing this risk evaluation, a specific software (3 answers), R (1 answer) or a tabular file (excel or csv) (1 answer) is used.

Regarding the quality control activities performed during the data collection, the two thirds of the respondents have integrated these tasks in their projects. The verification and correction of missing or inconsistent data is performed by 28 people, 27 people performed it for aberrant data, and 25 will create new variables during data collection. It is important to precise here that more than 50% of these answers were in the category «done by someone else». Therefore, it is possible that they are not done during the data collection, but after. Five people have indicated that these controls were done by REDCap, the others use a tabular file (6 answers) or a specific software (5 answers).

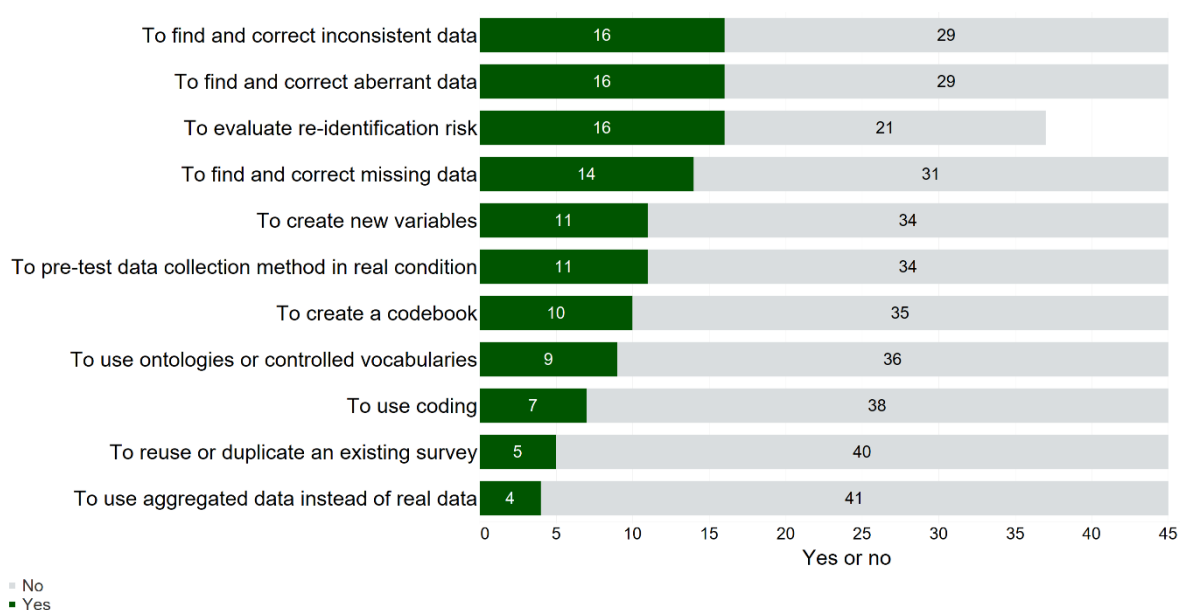
All these activities are documented by 24 respondents in a textual file (11 answers), tabular file (9 answers) or directly in the do files (Stata) or R scripts (4 answers).

### 3.2.2.3 Need of support or improvement

At this step, less respondents manifest their need for support, only 19 people tick the answer «I would like help or support for...».

The expressed need for support is focused on active monitoring (quality control during the data collection) for this step. The evaluation of the re-identification risk is also highlighted by 16 people (see Figure 16).

Figure 16 : Collecting data : need of support



The repartition of the preferred type of support is almost the same as in the Planification step. The support through a dedicated person who could answer to question when needed (19 answers) or could offer a practical help (18 answers) is still the preferred way to obtain help. Eleven people would appreciate to get a support through a training, and 10 people would like a theoretical help (guidelines, manual...). The realization of the task by somebody else (10 answers) would be less appreciated. The less preferred help here would be the use of a dedicated tool (6 answers).

Finally, the comments section of this group underlined that support would be needed punctually, depending on the project.

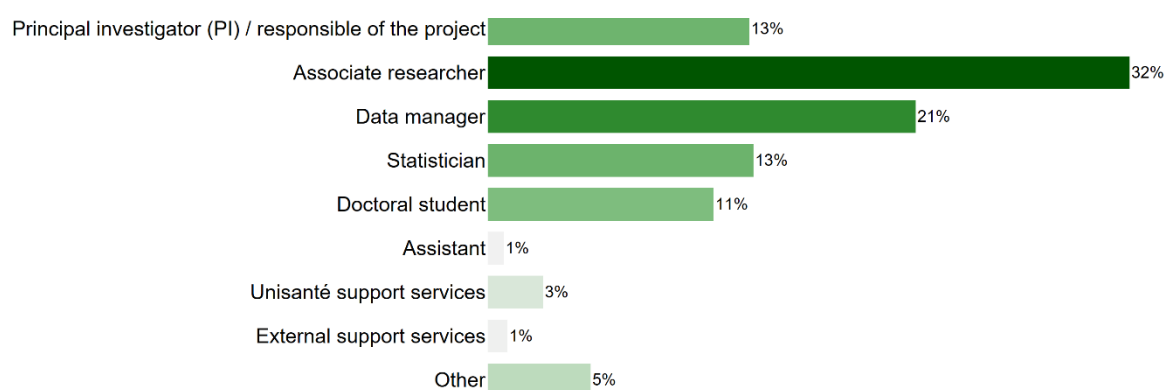
One interview with a researcher highlighted that a support to create data collection tool would be appreciated, moreover if it helps to be more interoperable. The idea of having a database of preset fields into REDCap or having standardized survey has been discussed and appreciated. However, the difficulty to put in place such development has been highlighted by the interviewee.

### 3.2.3 Processing and analyzing data

#### 3.2.3.1 Roles

At the «Processing and analyzing step», the associate researcher and the data manager are the ones handling the majority of the activities (see Figure 17). Statistician and doctoral students have also a major role here.

Figure 17 : Processing and analyzing data : most implicated roles

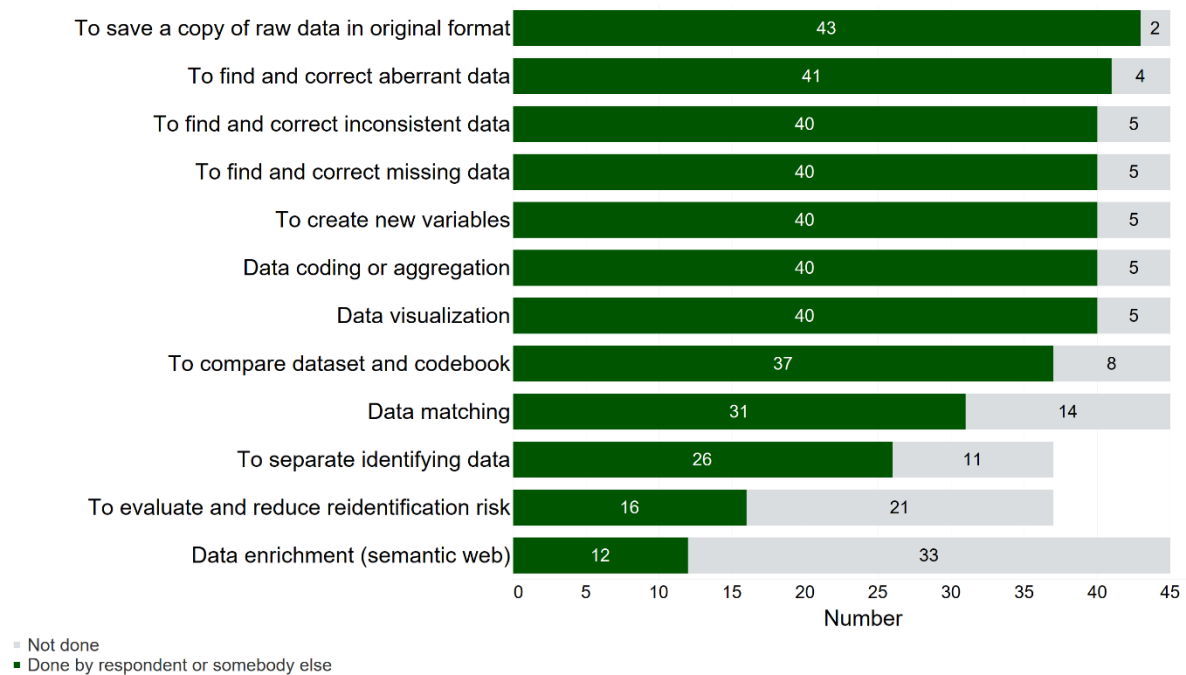


#### 3.2.3.2 Activities

For the processing and analyzing step, the majority of the listed activities are realized by the research teams, comparing to the previous stages.

As presented in the Figure 18, research teams at Unisanté save an original copy of their dataset before any modification in 95% of the cases. This is a very good practice, mainly done to assure data quality. As they use statistical software (16 answers), scripts (8 answers), or other specific software (4 answers), they are used to create a preparation and analysis code, apply it on data and then save the results under another version. This method ensure that original data are not modified.

Figure 18 : Processing and analyzing data : activities



For data preparation, the most frequent tasks are data quality control with the verification of aberrant, inconsistent or missing data (40-41 answers), the creation of new variables (40 answers), the aggregation of coding of values (40 answers), and the comparison between the dataset and the codebook (37 answers). These activities are performed to ensure data quality.

The repartition of tools and methods to realize these activities are the same between these tasks. Statistical software like Stata or SPSS are used by the majority of respondents (18 to 25 people), other will prefer the use of scripts in R or Python (6 to 8 people), and 4 to 7 people will use another software, like Excel (1 to 3 people), REDCap or Sphinx (1 to 3 people).

Data matching is less often performed by research teams at Unisanté, with 31 answers. The reason is the uselessness of this activity (6 answers), as not all research projects at Unisanté needs data matching, and lack of knowledge on how to do it (4 answers). Through the interviews, research teams have sometimes difficulties to match data from different sources due to their quality. Even between two Unisanté datasets, coding could be different for many variables, specifically socio-demographic data (gender, age...).

Finally, when research teams collect personal data, an important part of data preparation is to protect their participants. 8 respondents of the survey were not involved in human-related project, therefore only 37 people have seen these questions.

Twenty-six respondents separate identifying data from the dataset. Identifying data are names, telephone number, email address or any other direct identifiers. This activity is mandatory and performed for that reason. 11 people will not separate these data, because they estimate that this is useless (5 answers) or do not know how to do it (3 answers). As data could be collected without direct identifiers, it is possible that these eleven people were not concerned directly by this question.

The evaluation of re-identification risk is less realized. Twenty-one people on 37 will not evaluate the risk, mainly because they do not know how to do it (12 answers), because they find it useless (5 answers), because they did it previously (1 answer) or because of a lack of

time (1 answer). For the ones performing this activity, statistical software (4 answers) and R or Python script (3 answers) are used.

Finally, the less realized activity is the data enrichment (semantic web). Thirty-three respondents indicated that they do not realize this. For 14 people, they do not know what it is, 10 people find this useless and 5 do not know how to do it. This result was expected, as only a few cohorts at Unisanté need this kind of enrichment for now.

For the data analysis, only one question about data visualization has been asked in the survey. This is also one of the most preformed activity with 40 answers. Statistical software win again with 19 users, followed by R or Python scripts (9 answers), Excel (4 answers) and other software like REDCap or Sphinx (1 answer).

The question of data analysis was asked during the interviews, where researchers indicated that they mostly use Stata and R for their work. Excel is also used by some teams, as it is easy to interact with and offers quick visualizations.

The documentation of these activities is done by 21 people, in a text file (8 answers), directly in the preparation and analysis script (Stata or R) (7 answers) or in tabular file (3 answers).

Documentation of preparation and analysis codes has been widely discussed in interviews. The documentation is done directly into the code with simple manual comments. The content of this comments varies but the interviewees often indicate next to a function the reason of a decision behind a data modification and the date. Sometimes the comment goes further with the indication about the methodology used, a link to an article or a bibliography and a description of the function in natural language. In addition to the integrated comments, research teams create a text file to explain the preparation and analysis of the data, the different decisions made, the methodology used and the related bibliography.

This kind of documentation is important, as data manager and research teams tend to reuse their own codes for different studies. The documentation helps to understand the code and modify it quickly. It also ensure transparency, traceability and reproducibility of this process. Researchers sometimes share their work with other research teams inside Unisanté or externally on GitHub<sup>8</sup>.

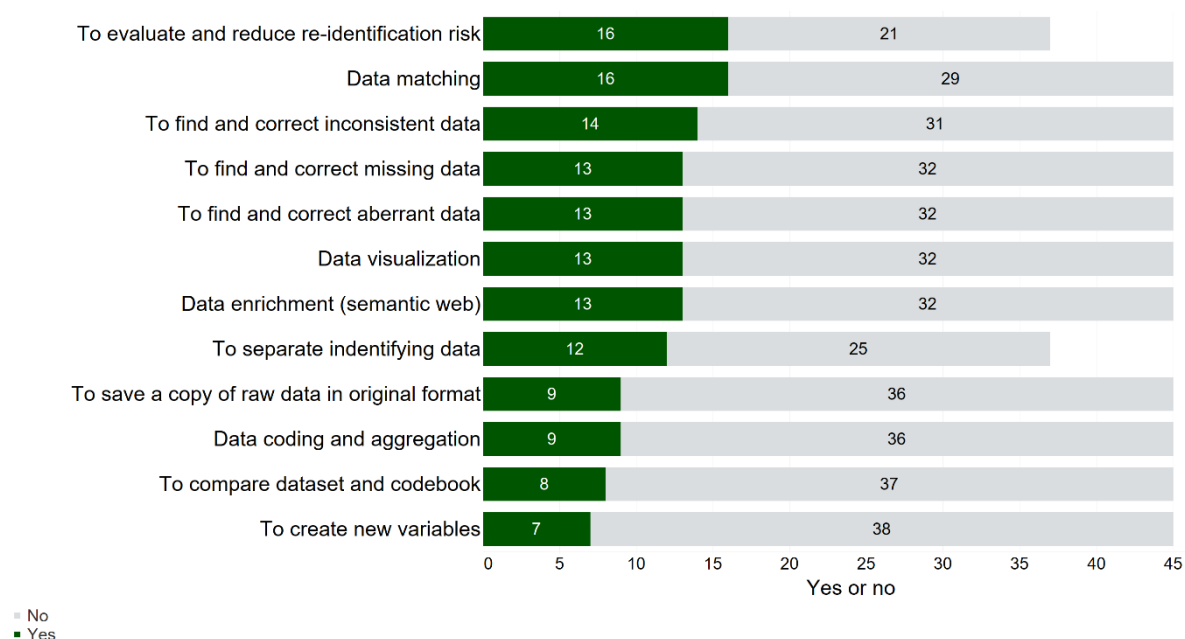
### **3.2.3.3 Need of support or improvement**

Just like the previous step, less need for support was expressed for the data preparation and analysis stage. This was expected as analysis is the heart of a research project and support is already available through different entities.

---

<sup>8</sup> <https://github.com/>

Figure 19 : Processing and analyzing data : need of support



The evaluation and diminution of the re-identification risk pops up again here. This thematic is trendy. As data protection laws are evolving and media are publishing re-identification stories, participants to research projects have more rights and more knowledge on data protection and re-identification. Therefore, researchers would like help to deal with anonymization and re-identification concepts and methods.

A support would be appreciated for data matching (see Figure 19). As indicated earlier, researchers have been confronted to difficulties when they tried to merge data and had to perform more data cleaning to finalize this activity.

Data quality control activities could also be more supported according to these results (13-14 answers), the same as for data visualization (13 answers). Through interviews and comments, it occurs that associate researchers are not always confident on their own Stata or R skills and are very pleased when a statistician or data manager is present in the research team or available to help them with these activities.

During interviews, some people indicated that they did not think of doing some of these activities, such as comparison of dataset and codebook after collection or evaluation of re-identification risk. This fact is more present when research teams reuse data instead of collecting them. Therefore, some would appreciate to have a reminder on the most important activities to perform in order to have a good data quality.

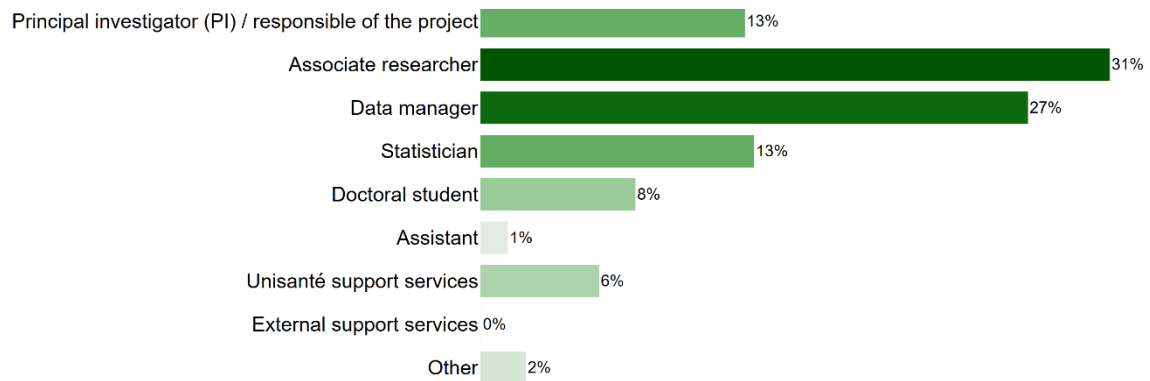
The type of wanted support stays the same as for the previous steps. Twenty respondents would prefer a person to answer question or give a practical help (16 answers). Then, 12 people would like a training, 9 a theoretical help or a person to do the task.

### 3.2.4 Publishing, sharing, and preserving data

#### 3.2.4.1 Roles

Like the previous step, the responsible of the research project is less involved at this stage (see Figure 20). The associate researchers and data managers are more implicated.

Figure 20 : Publishing, sharing and preserving data : most implicated roles



### 3.2.4.2 Activities

To share research data is not the most common activity. Only 55% (25 people) share their data at Unisanté, mainly on the institutional data repository (8 answers), but also through the article (4 answers), Maelstrom (3 answers), a dedicated website (2 answers), Zenodo (1 answer) or direct sharing by file transfer (1 answer).

The other 20 people do not share their data because it is forbidden by the sponsor or else (4 answers), because they do not know where to share (3 answers), have not had the occasion yet (3 answers), because the data is not the property of research team or Unisanté (2 answers), or because the data are sensitive and personal (1 answer).

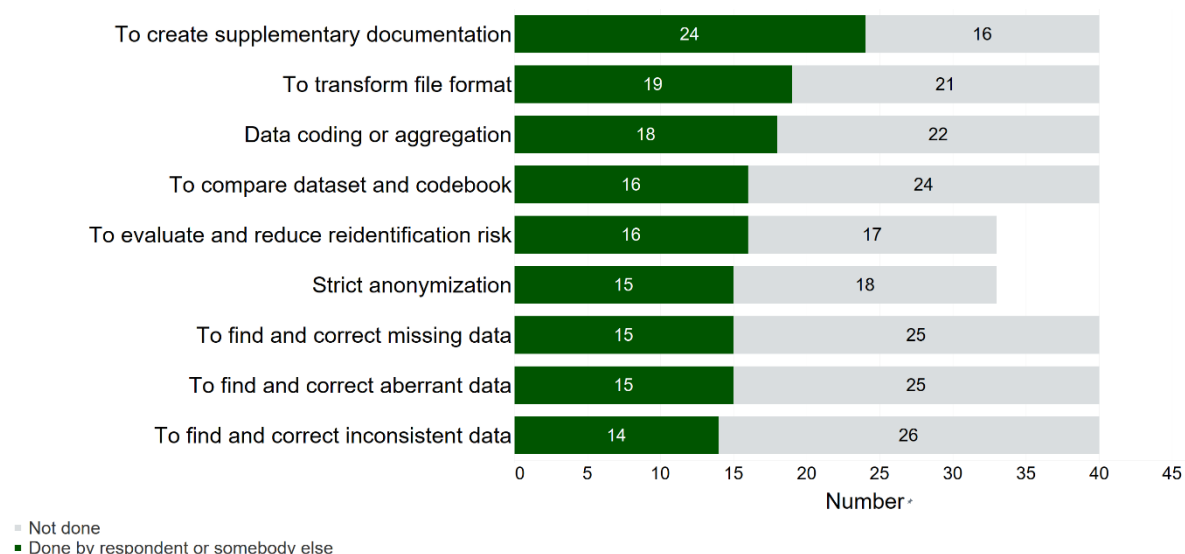
Archiving data is most frequent at Unisanté. 86% of the respondents archive their data, mostly on Unisanté common server (23 answers). 4 people have indicated that archiving was done by «letting the data where there are». Due to storage habits, it seems that the data are on the Unisanté common server. Two common servers are available at Unisanté. The first one is the common server where all departments store their data (administrative, research, etc.) and the second one is a more restricted one, dedicated to final archiving. However, it is not an archiving system, only a windows server with limited access right. No curation, no preservation actions, no file migration or integrity control are performed on this storage system.

The other respondent archive their data on a CHUV server (3 answers), on a data repository (2 answers), on REDCap (1 answer) or on personal computer (1 answer).

The duration of archival is not known by 13 people. For the other, 13 respondents archive their data from 6 to 10 years, 4 people from 11 to 15 years and 10 people archive the data for more than 16 years.

Best practices activities are less performed at this point of the research project compared to the previous ones. Five people were not concerned at all, as they indicated that they do not share nor archive data. Here 24 people or less realize these activities, against 42 people or less for planification activities, 41 people or less for data collection activities, 40 people or less for data preparation and analysis activities.

Figure 21 : Publishing, sharing and preserving data : activities



Twenty-four people will create supplementary documentation (see Figure 21), sometimes with the help of the UDD (3 answers). The documentation is created in text file (7 answers), through statistical software (3 answers), with tabular file (Excel) (2 answers), in XML metadata (1 answer) or R script (1 answer). The file format transformation is also realized by 19 people. Coding and aggregation of data can also occur at this point for 18 people.

The other actions are less performed, as the ones related to re-identification risk (16 answers) and anonymization (15 answers). Twelve people were not concerned by this question, due to their study design (8 answers) or by the fact that they do not share nor archive their data (5 answers). For the other, the reasons behind the non-realization of these activities are the lack of knowledge on how to do it (5 answers), its uselessness (5 answers) and the fact that it has been done previously (3 answers).

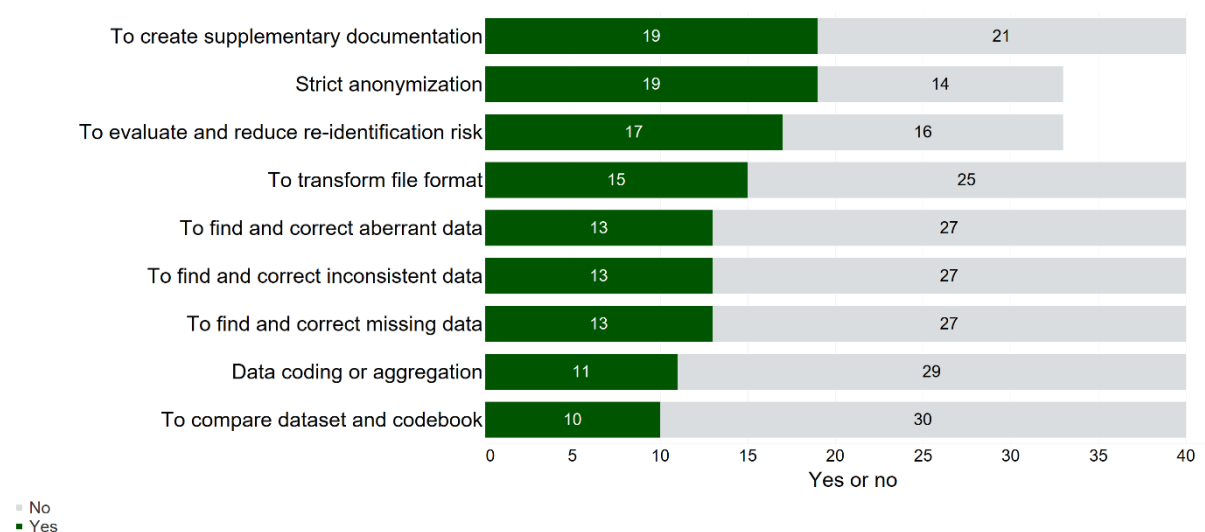
The less performed activities are the data quality controls such as comparison of the dataset and the codebook (16 answers), and finding and correcting missing, aberrant or inconsistent values (14-15 answers). The reason behind this is that these activities have been realized previously (7 answers). The uselessness of these tasks have been pointed here by 7 or 8 people (depending on the task).

Like the previous answers, evaluation of re-identification risk and data quality control is mostly performed with statistical software (6-8 answers).



### 3.2.4.3 Need of support or improvement

Figure 22 : Publishing, sharing and preserving data : need of support



As presented in Figure 22, the respondents would mostly like to obtain support for activities related to creation of supplementary documentation (19 answers), data protection (17-19 answers), following the needs expressed in previous steps, and file transformation (15 answers). Ten to thirteen people would also need support for data quality control or data treatment at the time of sharing or archiving.

Regarding the type of support, the tendency stays the same with a preference for a person of reference to answer questions (17 answers) and act if necessary (13 answers).

## 3.3 Support services

This section presents the results of the objective 2 «to evaluate practices and needs at Unisanté» for the support services at Unisanté. The head of each support service have been met during an interview, as described in the [methodology](#) section. These interviews have led to the following results.

### 3.3.1 Biostatistics consultation unit

The Biostatistics consultation unit (*Unité de Consultation de biostatistique (UCB)*) is part of the DFRI. Several statisticians, for a 1.6 FTE (full-time equivalent), compose the UCB.

All research teams at Unisanté can benefit from this service, if their projects do not need more than 40 hours of collaboration. For bigger projects, the UCB would recommend the hiring of a dedicated statistician in the team, and would stay as a support service for this person.

#### 3.3.1.1 Activities

The UCB aims to offer a support through all the data lifecycle. The visual representation of the tasks is available in [Annex 7](#).

At the beginning of a research project, the UCB can help research teams to precise and to formulate their research questions, and to calculate the minimal sample size necessary to guarantee sufficient statistical power while keeping the costs of the study within an acceptable margin. It can also provide methodological support to choose the more adapted study design to answer research questions, taking into account constraints related to cost, time, feasibility

or available population. The UCB will also collaborate with research teams to create a statistical analysis plan.

For the collection of data, the UCB can advise research teams to choose the more appropriate database and to develop it. It also uses its expertise to advise on survey design and field creation (coding, range controls...).

At the preparation and analysis stage, the UCB will clean the data to ensure its quality, perform the statistical analysis and realize data visualization. The statistical software Stata and the language R are mainly used for these activities, but other software could be used if necessary.

Finally, the UCB will support research teams in the redaction of their article, and will also answer to the possible questions from the reviewers.

All the UCB activities are documented, inside statistical codes and supplementary documentation, to ensure reproducibility and to facilitate reuse of statistical codes if necessary.

### **3.3.1.2 Recommendations and perspectives**

During the interview, the following question was asked «What would be your recommendations or perspectives of improvement to increase data quality at Unisanté?».

For the UCB, data quality issues could be anticipated by the formulation of the research questions and the choice of the study design. Therefore, if the research teams would come at the very beginning of their project, even before the protocol's redaction, some problems or additional work could be avoided.

Then, another recommendation to be highlighted according to the interviewee is the design of the survey and fields. The variables should be coded when possible. To choose the appropriate way of coding, it would be important to look at similar studies. Then, coding should be integrated into the data collection tool, such as REDCap. Fields could also include data quality control, such as range control, and be created in the right field format, for example age in a numerical format instead of text.

The final recommendation is the creation of a codebook. This kind of documentation contains all necessary information to understand the dataset. This is very useful for the research teams, the statistician, the reviewers of an article or future users of the dataset.

## **3.3.2 Survey Methodology unit**

The Survey Methodology unit (*Consultation en méthodologie d'enquête*) is a service proposed by the ESOPE Unit (*Enquêtes de satisfaction et d'opinion des patient·e·s et des employé·e·s*) and is part of the DESS. No FTE are exclusively attributed to the service, as the necessary resources are mobilized depending on the specific demand.

All research teams at Unisanté could submit a request to this service, if their projects involve a questionnaire about patients' satisfaction or experience, opinions or perceptions related to health.

### **3.3.2.1 Activities**

The Survey Methodology unit offers a customized support to research teams through all the data lifecycle. The visual representation of the activities is available in [Annex 8](#).

At the beginning of a research project, the unit will help research collaborators to define their research questions, to operationalize them into variables and to choose the appropriate data collection method. The consultation will also give information on data storage at this point.

Then, a support is offered to create and to evaluate a data collection tool. The advice will focus on the formulation of questions, the choice of possible answers, the design and the pre-test of the survey. A part of these advice is the sensibilization to the re-identification risk. The choice of variables and their possible answers will help to reduce this risk. The unit will also help research teams with recalls for their questionnaire, to reach the appropriate response rate.

During the analysis step, the unit will not perform the analysis but will advise researchers on the evaluation of data quality, data representativity, and data reliability. It will also provide support to aggregate data correctly.

Finally, the unit will participate to the redaction of articles for the methodology section and presentation of results.

### **3.3.2.2 Recommendations and perspectives**

At the question «What would be your recommendations or perspectives of improvement to increase data quality at Unisanté?», the unit highlighted the fact that data quality begins with the research question. The operationalization of research question into variables is an important step to create an efficient data collection tool. This step will help to design a survey which will collect the correct data to answer correctly to research hypothesis. The unit also indicated the importance to collect only the needed data, and not to add additional questions to a survey if they are not designed to answer the research questions.

### **3.3.3 Research promotion unit**

The Research promotion unit (UPR) is part of the DFRI. The UPR is composed by 4 collaborators, for a 1.9 FTE.

#### **3.3.3.1 Activities**

The UPR develops various activities related to research. The visual representation of these activities is available in [Annex 9](#). The first activity is a mission of support to researchers in the organizational and ethics/regulatory aspects of their research projects. In more details:

- The UPR is relaying information relating to internal and external tools and resources in connection with research using various communication channels (newsletters, web interfaces, research conferences and symposia). One of the internal tool is a Roadmap<sup>9</sup> giving to research teams a step-by-step guidance through their research project. The roadmap contains multiple resources, documentation and contact information to various support services at Unisanté.
- The UPR is supporting researchers with funding identification and applications.
- The UPR is answering questions related to regulatory framework, especially regarding the Federal Act on Research involving Human Beings, is helping to categorize research projects, advising on submission and reporting duties to authorities and clinical trials registrations.

---

<sup>9</sup> <https://www.unisante.ch/fr/formation-recherche/ressources-pour-recherche/roadmap>

- The UPR is engaged in the quality management system of the research projects by advising on document management, audits and inspections preparation and archiving of paper documents. It also provides monitoring services for clinical trials, according to the GCP guidelines, and should extend these services in the future to quality control for observational studies.

The second mission of the UPR is to facilitate access to resources and skills provided by other internal and external entities.

The third mission is to ensure that Unisanté has an institutional overview of all research activities including valorization and implementation of research results. In this regard, the UPR is currently developing a database of research projects, which will include all metadata at project's level. This database will help to have a risk-based management at Unisanté and will offer an exhaustive listing of all research projects within the institution.

### **3.3.3.2 Recommendations and perspectives**

The UPR already offers classical on-site monitoring for clinical trials, according to the GCP guidelines. It would like to go further, with the centralized data monitoring, a new activity referred by the GCP in response to advances in the use of electronic data reporting and implementation of improved and more efficient approaches. Centralized data monitoring processes provide additional monitoring capabilities that can complement and reduce the extent and/or frequency of on-site monitoring and help distinguish between reliable data and potentially unreliable data. Monitoring is mandatory for clinical trials, but the UPR would like to open it to other study designs (mainly cohorts studies), as a manner to improve and monitor research data quality. Since this new service is requiring data management and biostatistics skills, the UPR would collaborate with other existing support services (e.g. UDD or biostatistics consultation unit) to create a centralized data monitoring service.

### **3.3.4 Research IT Services unit**

The Research IT Services unit (*Unité de soutien IT à la recherche*) is part of the Information Technology and Digital Transformation sector (*Secteur Systèmes d'information et transformation digitale*), which is under to the Financial direction (DFI). Four informaticians and developers, for a total of 4 FTE, compose this unit.

This service aims to address the needs of research teams who manage projects with an IT component. Usually they have planned a budget for its realization.

#### **3.3.4.1 Activities**

The main service of this unit is the development of application or interface to collect or to present data. The backend software is often, but not only, REDCap, and the frontend software will differ depending on the needs of the research project. The research teams coordinate with the unit at the beginning of their projects to plan such development.

The unit also provides REDCap software and support the creation of collection tool through it, to ensure its architecture is suitable to the research design, and to answer questions related to field structure or setting. This support is done through direct advice and quality document, for example a checklist to verify that all variables are properly set or identifying data has been properly declared.

The unit can also offer advice on data cleaning but will not realize it. They will help research collaborators to parameter the quality control checks into REDCap or to develop descriptive dashboard to follow data collection.

Finally, the unit will participate to projects with the Information Technology and Digital Transformation sector to design and provide research applications or software. The visual representation of the activities is available in [Annex 10](#).

#### **3.3.4.2 Recommendations and perspectives**

Through this service, different challenges have been identified and indicated during the interview at the question «What would be your recommendations or perspectives of improvement to increase data quality at Unisanté?».

Some questionnaires are not well structured, many variables are dependent on others, some identifying variables are not flagged properly into REDCap, or some surveys have issues with access right. The unit would recommend more training and support on these themes.

Another development that could be done to improve data quality would be the creation of templates inside REDCap. To avoid access right issues, a project's template could be created and then duplicated to offer the right structure to research teams. Model surveys could also be created by thematic. For example, a survey on tabacology could be set into REDCap, with different variables related to the thematic. These standardized variables could be SPHN compatible, could already contains the correct ontologies and coding, and could be available in multiple languages. A similar project on interoperability is realized by a research team at Unisanté and should be involved in this kind of development.

A better data governance for Unisanté, where all stakeholders are included and have a defined role, would be appreciated by this unit.

Finally, the unit would like to propose a better research IT infrastructure. It would help to answers the growing needs of researchers in term of analysis, high performance computing and storage. The traceability of data exchange during research project has also been highlighted. The research teams transfer data inside the institution through multiple ways, therefore they can not be traced. A research infrastructure, like a data warehouse, could be part of the solution.

#### **3.3.5 Documentation and data unit**

The documentation and data unit (*unité documentation et données (UDD)*) is part of the DFRI. Five information specialists, for an 3.7 FTE, compose this unit. The research data management activities are performed by 1 FTE.

##### **3.3.5.1 Activities**

The documentation and data unit (UDD) provides research data management support to research teams through all the data lifecycle. The visual representation of the activities is available in [Annex 11](#).

At the beginning of a research project, the UDD will give advice on general data management and review data management plans (DMP). Its expertise in information retrieval allows it to help research collaborators to find secondary datasets. It will also answer questions related to

data protection and data anonymization, including the evaluation of existing datasets for an anonymized usage (strict anonymization).

The UDD is also present during a research project to give general support for research data management and to help with the creation of survey on the Limesurvey software. No service related to active curation is available at the moment.

At the end of research projects, the UDD offer a support to share research data. First, it answers questions and gives explanations about publishers' data policy and data access agreements. Second, the UDD manages the institutional data repository and help researchers to share their datasets according to the collected consents and applicable limitations.

For each deposit on the institutional repository, the UDD will perform a data curation workflow on each dataset, which will include data quality controls, file format transformation, encryption, de-identification (when necessary) and metadata creation in XML. Then the datasets are made available on the data repository and a copy is stored on the archival server at Unisanté. The UDD will also monitor manually the archived datasets to ensure their readability through time, with file format migration for example.

The UDD collaborates with the UPR to provide services to all the research community at Unisanté. It also participates to institutional projects, such as data governance project.

### **3.3.5.2 Recommendations and perspectives**

The UDD has identified issues with data quality at the end of research projects and wishes to provide help and support for active curation, in order to prevent such issues. This thesis is the result of that observation.

It also highlighted the management of data governance at Unisanté. There are multiple entry points to get advice on research data management. Therefore, research teams may be confused when they want to get an answer. For example, if researchers have a question about anonymization, some will go to the UPR, others to the UDD and some will go to the administrative management of Unisanté. All these entities could have the answers, but on a different level. A better data governance and communication towards research teams could be a benefit for Unisanté.

The UDD has also identified the archiving processes as an issue. As for the anonymization, when the research teams have a question, they will turn either to the UPR, the UDD, the administrative management, the Research IT Services unit or even to the CHUV IT, as they are the providers of Unisanté servers.

When it comes to electronic documents, including datasets, Unisanté has no precise procedures. Moreover, there is no real archiving system, OAIS<sup>10</sup> compliant, at Unisanté. Everything is stored on an server and left there. No preservation actions (file format transformation, integrity verification with checksums, additional documentation...) are done on the documents. The UDD identifies this state as a risk to loose data or have unreadable data in 5-10 years. It tries to prevent this risk with the data repository and parallel archiving, as described above, but knows that this is not a sustainable solution.

---

<sup>10</sup> <http://www.oais.info/>

## 4. Recommendations and discussion

### 4.1 General discussion

#### 4.1.1 Terminology

The terminology used in data curation does not have the same meaning if it is used in the Information Science field or in the research community. As seen in the [State of the art](#) section, the definition will differ. It would seem important to define more precisely the data curation perimeter, and to reach a consensus on definition in all the concerned fields. This would improve understanding between professions and comparison between research results from different research projects would be possible.

For this study, the confusion about terminology has been observed, as some people indicated into the survey that the definitions of tasks were not precise. For others, the helping texts integrated in the questions were not sufficient to understand the subject. To improve this, interviews could have been done before creating the survey. It has not been done for this thesis due to a lack of time. This process would have postponed the survey and would not have left enough time for the analysis.

During the interviews, it occurred that the term “active data curation” was not completely understood. The term “research data management” englobed this expression and the distinction between the two terms was not clear for the researchers. Depending on the activities, the participants would talk about data cleaning or preparation, and link them to data quality, research data management or active data management, but they never used the term “curation”. Therefore, when it comes to communication at Unisanté, the term “curation” does not seem appropriate.

#### 4.1.2 Design of support services

According to the survey, the preferred support is an “in-person” support, with somebody to answers questions, or somebody to act. This conclusion leads to a new question : are the support services adapted to the needs ?

Some universities discuss the idea of offering a funding for data management instead of providing a support service. The tasks performed by a data manager or the support needed in active data curation will differ from a project to another. A support service able to answer to all needs would have to be very performant, and would have to be composed by multiple professions. Therefore, this idea would be to create a fund for the research project to allow researcher to engage a data manager. Then, a more general support service could be created, to support data manager instead of participating into research projects. This kind of service design presents a risk, as there is no certainty that the allocated fund will be used for data quality related tasks.

Another design for support services is the network. Already in place in some universities, like at the university of Basel (De la Lama, Racine, 2022, p. 15), the organization in a network offers a large support for research teams through various services. This structure would be more appropriate at Unisanté, as several support services already exist and collaborate. Moreover, due to its link with the UNIL and the existent relationship between the UDD and



UNIRIS (*service des ressources informationnelles et archives*)<sup>11</sup>, some support services could be shared between the institution. The university could provide general support and Unisanté could be more specific to its field.

## **4.2 Recommendations for the research teams**

### **4.2.1 Data management plan**

The most wanted support requested by research teams in the survey is the support for data management plans. This support already exists at Unisanté. A person is available at the UDD to answer questions and to review DMPs. Moreover, a template is available in Word. This DMP template is based on the structure asked by the SNSF and is partly filled with Unisanté information on data storage, data sharing on the Unisanté repository and metadata.

This service does not seem to be known. Therefore, communication should be developed to promote it, through the intranet and internal newsletters. A presentation about DMP and the related services could also be organized in an institutional event.

The DMP support could also be improved to be more compliant with researchers' habits. A project to adapt the Unisanté template into the software DMPonline<sup>12</sup> is currently in process.

However, it could be interesting to evaluate the usage of the DMP. At the moment, DMPs are text of PDF files, created at the beginning of the project. For many researchers, the DMP will not be used after the redaction. As seen in the previous results, only 13% of the respondents use it to improve data quality or project quality. This state generates some questions : is the DMP in a good shape ? Would it be preferable to have a more useful DMP ? What system would be suitable for DMP ? These interrogations will be explored at Unisanté, to offer a better DMP services to the research teams.

### **4.2.2 Methodology support**

The definition of research questions and the ways to operationalize them is an important step at the planning stage. The declination of research question into variables will have a great impact on data quality, as identified by the support services during the interviews. This activity can help to anticipate interoperability issues and anonymization issues. The choice of the variables and the level of detail needed to answer the research questions have to be discussed at this point, to find the balance between the expected results, the protection of the participants and the data matching possibilities in the future.

The operationalization of the research questions can be considered as active data curation as it will help to plan a good data quality for the research project but also for the life of the data after the project. As this task is essential to realize the research project, it is important to provide a support to all research teams to ease this process.

The Survey Methodology unit offers a good service to support researchers at the beginning of the project and to help them with this task. However, this consultation is not free, therefore some research teams will not use it.

The roadmap created by the UPR already lists resources to help research collaborators with the research questions related activities. However, a supplementary support could be created,

---

<sup>11</sup> <https://www.unil.ch/uniris/home/menuintst/uniris-en-bref.html>

<sup>12</sup> <https://dmp.unil.ch>



in the form of an event or a training. This could provide basic methodological skills to researchers. Then, the advanced «in-person» service would be handled by the Survey Methodology unit as now.

### **4.2.3 Interoperability with REDCap**

It has been seen through survey and interviews that interoperability was a burning issue. Research teams face it when it comes to data matching and would like help with this task. Support services identify problems related to interoperability with the design of the surveys.

Anticipating interoperability problems as soon as the questionnaire is created is part of the scope of active curation. The choice and the settings of variables have a great place in this issue. Creating data collection tool with integrated range control, with ontologies or controlled vocabularies will help to capture good quality data. Moreover, if multiple research projects use the same variables with the same settings, this information could be matched easily. Similar projects are developed in more specific fields (Allie et al. 2021).

REDCap is a very used software at Unisanté. According to IT statistics, 50 new projects are created each year, two thirds including a survey. Therefore, this software could be used to promote interoperability. At the moment, some surveys integrate ontologies to REDCap through an API, but this is not a common practice. The creation of a database of Unisanté fields directly into REDCap or the creation of templates of surveys could help research teams when they create their data collection tool.

Such development would allow research teams to use the same variables, created and validated by Unisanté, and therefore be more interoperable between them, to ease data matching. This would also facilitate the creation of surveys, as the fields already exist.

The difficulty of this recommendation is to reach a consensus for the content of surveys and fields. The interviews highlighted the difficulties to agree on a common way of coding or a standard definition of concepts. The example of the gender has been given, as some will code gender with numbers (1, 2, 3...) or with letter (f, m, t...). The first step to go to this kind of development would be to create a working group with the goal to propose a collection of variables. This working group should include researchers, in particular the ones already involved in SPHN projects, and experts from Unisanté support services. As this is a huge work, it would be important to focus on new variables. The work done by SPHN could be reused, and existing ontologies such as SNOMED-CT could be recommended instead of rethinking the variables. Therefore, only the communication on the recommended variable and its integration into REDCap would be needed.

The integration of such development into REDCap would be easy, as a feature already exist to create a database of fields. This collection would be available through a button «import from field bank» when the researcher creates the fields of a survey.

If the option of creating templates of projects or surveys is preferred, the implementation would be quite easy too. Unisanté would have to create the survey and give access right to Unisanté users. Then the researchers would have to duplicate the wanted survey and then modify it according to their needs. User tests could be realized to identify which solution is more suitable at Unisanté.

The promotion of these developments could then be done by the creation of a specific training and the presentation of the functionality through institutional event and information support on the intranet.

#### **4.2.4 Anonymization**

The anonymization and de-identification related tasks have been highlighted in the survey. Many researchers would like to obtain help on these subjects.

Anonymization and de-identification practices aim to protect the research participants during the research project, but also after it, when the datasets are shared. An active data curation recommendation would be to find the right balance between participants protection and accuracy of the collected data. The concept of «privacy by design», which will be introduced in the revision of the Federal Act on Data Protection, could be promoted at Unisanté. The operationalization of the research questions is an important step to achieve this goal and could be supported through the [methodology support](#) presented earlier.

Other support regarding anonymization could be created at Unisanté, to improve data quality.

The first one could be the creation of a Data protection officer (DPO) position at Unisanté. This role is currently assumed by the administrative direction but having a specialized collaborator would be a benefit for research teams and support services. A DPO could be the person of reference for all data protection questions but could also support the UPR with the listing of the research projects and the risks management. This role would also be an asset for the UDD, with the evaluation of dataset anonymization. A DPO would have the appropriate knowledge and authority to validate this kind of evaluation.

A second recommendation regarding anonymization would be the sensibilization and training of research teams. It is important for the research teams to have a complete understanding of the possibilities of re-identification and their related risks. Moreover, the clarification of the difference between anonymization and pseudonymization could be an asset, as researchers often confuse these concepts (Stam, Kleiner, 2020). Events or training could be performed on regular basis, with a focus on :

- Consent and participants' rights
- Anonymization and re-identification concepts
- Privacy by design (creation of data collection tools)
- Anonymization and pseudonymization methods and tools
- Data protection laws and consequences

The UDD could provide such training and could collaborate with the other support services to go deeper into some aspects, such as regulatory framework applicable at Unisanté. A on stop shop, in the shape of a helpline or a chat could also be put in place, to quickly answer questions regarding anonymization and data protection. It would be interesting at Unisanté to have only one entry point for RDM and anonymization related questions.

A third service could be an anonymization and pseudonymization service for the sharing and archiving step. This kind of support already exist into the data curation processes at the UDD but could be developed to become more efficient. Two ways exist to offer this kind of support: a training or a person to realize it. This could involve the use of anonymization R package,

such as SdcMicro<sup>13</sup>, anonymization software, such as Amnesia<sup>14</sup>, or synthetic data generator such as DataSynthesizer<sup>15</sup> or generative neural networks (GAN). These methods have not been fully tested for this study but will be investigated by the UDD in 2022-2023.

It is important to highlight here that with data matching and big data analysis, re-identification becomes more and more easy. In consequences, collecting fully anonymized data, or anonymize them completely, is very difficult. Moreover, in the medical field, various rare diseases make impossible a true anonymization. The creation of a support service to perform a real anonymization could be quite difficult and expensive. An evaluation of the needs should be realized, to determine if such service would be necessary, as real anonymization is needed to share data in Open Access only. It would be therefore important to evaluate if this way of sharing is wanted at Unisanté, or if the restricted sharing would still be preferred in the future, for which only the pseudonymization support service would be needed.

#### **4.2.5 Centralized data monitoring**

The need for centralized data monitoring has been expressed by 16 people in the survey. The UPR also identified this need of improvement at Unisanté. The centralized data monitoring helps to control data quality all along the research project, which is also the one of the objectives of active data curation.

The implementation of active monitoring could be done at Unisanté in different ways. First, functionalities of the data collections tools (REDCap, Secutrial<sup>16</sup>) could be activated. The support service should train and support research teams to set their surveys correctly. The active monitoring cannot completely be realized through the data collection tool. If a centralized data monitoring service is put in place at Unisanté, it should include one or more person to perform monitoring outside of an automated framework.

As centralized data monitoring is mandatory for clinical research, this service would be used at Unisanté. However, the non-clinical researchers may not use this prestation, as it would slow down their processes. The UPR should therefore propose two levels of centralized data monitoring, a complete one, to be GCP compliant, and a light one, to only ensure data quality.

The centralized data monitoring is usually performed by statisticians. To evaluate the possibility to implement this prestation at Unisanté and to choose the structure to adopt, a benchmarking of similar services, for example the CTU Bern<sup>17</sup>, would be necessary. Due to the proximity with the CHUV, Unisanté could also consider a mutualized centralized data monitoring service.

#### **4.2.6 Codes documentation and sharing**

Through the interviews, the documentation of statistical codes and the possibility to share and reuse them has been highlighted.

---

<sup>13</sup> <https://cran.r-project.org/web/packages/sdcMicro/index.html>

<sup>14</sup> <https://amnesia.openaire.eu/>

<sup>15</sup> <https://github.com/DataResponsibly/DataSynthesizer>

<sup>16</sup> <https://www.secutrial.com/en/>

<sup>17</sup> [https://www.ctu.unibe.ch/services/quality\\_assurance\\_and\\_monitoring/index\\_eng.html](https://www.ctu.unibe.ch/services/quality_assurance_and_monitoring/index_eng.html)

The documentation is already done, but could be improved, to be more FAIR compliant and to increase reproducibility of research. The documentation is often integrated into codes through comments and described into additional text documents.

However, when it comes to publish an article, the statistical codes are not always shared. Researchers will describe in natural language the data preparation and analysis. This redaction asks for a supplementary effort and could be lighten.

A new software, C<sup>2</sup>Metadata<sup>18</sup>, allows the translation of statistical codes into natural language. The use of this software could be interesting when the statistical code cannot be shared. The translation in natural language of all steps of the code enable the reproducibility of the preparation and analysis process. Moreover, even if the code is shared with comments, its translation is very useful to those who do not know the statistical language.

C<sup>2</sup>Metadata is a promising software, as a future development could be the translation of statistical codes in another statistical language.

*« We also see a future for SDTL as an intermediary in translations between statistical languages »*  
(Alter et al. 2021, p. 11)

For example, a Stata code could be transformed into a R code automatically by the software. This future functionality would be very helpful at Unisanté, to allow the reuse of codes internally and for reproducibility proof. The evolution of C<sup>2</sup>Metadata should therefore be followed.

Another possibility to improve and document statistical code would be to share them with the community. The platform Galaxy<sup>19</sup>, specialized in the \*omics analysis, could be interesting for some research projects at Unisanté. This platform proposes a library of various analysis and preparation codes. Its interface is user-friendly and allows researchers to prepare and analyze their data without an advanced knowledge of statistical programming. However, the use of such platform would be a big change for Unisanté and will not be useful for a great number of researchers. Therefore, GitHub would be more suitable at the moment.

Finally, sharing codes internally at Unisanté would also be suitable, as many research project needs similar data preparation or analysis. A platform for sharing could be implemented, either with GitHub, through the data repository or with a new platform.

#### **4.2.7 Training and guidelines**

Training is not the preferred way to get help at Unisanté, according to the survey. However, this kind of support could be an answer to some of the information needs.

Support needs are different between roles. Unisanté should offer a support for all, but this cannot be uniformized due to different level of knowledge and practice. To be more user friendly, a level-support could be enforced (beginner, intermediate, advanced). This kind of level could be easily integrated into courses.

Therefore, Unisanté could offer trainings for beginner and intermediate in data management and curation. For the advanced support, a person could be in place to provide help or act directly on the data if necessary.

---

<sup>18</sup> <http://c2metadata.mtna.us/>

<sup>19</sup> <https://usegalaxy.org/>

In addition to trainings, guidelines and checklists could be created. This kind of documentation could provide a reminder of the important steps to perform in order to achieve a good level of data quality. The creation of this documentation should be done in collaboration with different support services, in order to have the most accurate advice.

## **4.3 Recommendations for the institution**

### **4.3.1 Data governance at Unisanté**

Interviews have highlighted the need for an institutional data governance. Support units offer services but are not always aware of the ones of the other units. More collaboration is essential to offer a better support. The point of view of the research collaborators was similar. They identify the multiple aspects of data governance, and some of the stakeholders, but do not see a common strategy to achieve a good governance.

The project launched by the administrative management of Unisanté (DA) should continue and induce more collaboration and discussion between the stakeholders. A formal data management network, including all research support services could be created to provide support for research teams. A data policy could be written when a general strategy would have been decided.

### **4.3.2 Documentation**

The documentation needed at the beginning of a research project is clear and well structured. However, documentation created during the project to explain data treatment or analysis, and the additional documentation needed to archive and share the data are not standardized.

As presented in the [state of the art](#) section, documentation is a main activity of active curation. Being able to document all aspects of the research project during its realization helps to get correct information and to gain time at the end of the project.

Unisanté could therefore defines which information is suitable at each stage and specify the format. Then the institution could provide templates to research teams, in order to facilitate the information collection. The automated extraction of information from institutional software, such as the metadata research project repository, project leaded by the UPR, should also be explored.

### **4.3.3 File organization and naming**

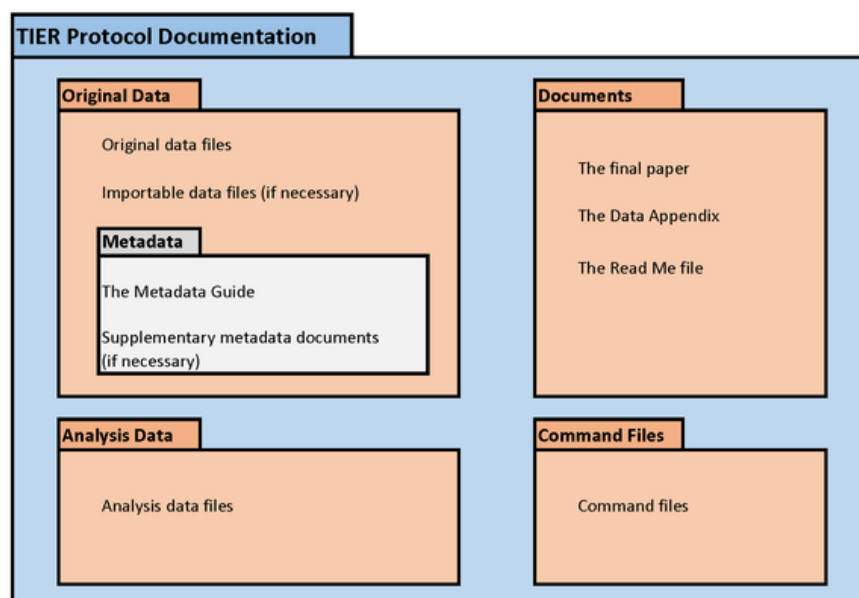
To facilitate the file organization, file naming and the documentation of these aspects, Unisanté could provide a standard file naming convention for research, and a standard file organization. Some internal procedures regarding file creation and naming are already available and should be promoted and implemented.

For the file and folders organization, Unisanté could be inspired by the TIER protocol<sup>20</sup>, which provide a file organization (see Figure 23) and specifies the contents and documentation necessary to replicate an article. This could include the documentation necessary to archive and preserve the data, in a text document or through an online form.

---

<sup>20</sup> <https://www.projecttier.org/tier-protocol/>

Figure 23 : Diagram of the Project TIER folder structure



(TIER Protocol 3.0 no date)

Providing a pre-structured storage area could help researchers to organize their data and documents during their project. From an archival perspective, a similar structure for all projects would also be an asset, as it ease the accessibility.

Finally, this structure could include an “Archive” file, where researchers could transfer their data, or include an “Archive” form to announce the end of the project and the necessity to archive the data.

#### 4.3.4 Data repository

The UDD and the IT units both identify the need of improvement when it comes to data sharing. The researchers also feel this need, as some do not know how or where to share their data, and some are not satisfied with the actual offer of the data repository.

The project of improving the institutional data repository has been planned for 2023 by the UDD, the manager of the repository. The actual repository has been created to answer a need for data sharing in 2014, but its functionalities have not evolved between its creation and now.

This project has to evaluate the necessity and the benefits of the institutional data repository. As various national solutions exist to shared data within the community (Olos<sup>21</sup>, SwissUBase<sup>22</sup>, SPHN<sup>23</sup>), the maintenance of the Unisanté repository should be questioned. Do these national repositories offer the expected functionalities and security, or do Unisanté still need a separate repository ?

If the repository is maintained, its specification should be discussed with researchers and support services to plan the appropriate improvements. Multiple modules or software related to data curation, for example YARD, or data visualization modules could be integrated or linked to a data repository. A replication service, to share preparation and statistical codes could also

<sup>21</sup> <https://olos.swiss/>

<sup>22</sup> <https://www.swissubase.ch/fr/>

<sup>23</sup> <https://sphn.ch/fr/home/>

be included to the repository, or an existing one could be recommended. A more powerful data catalogue could be created, to facilitate data discovery inside and outside of the institution. Finally, a multi-purpose solution could be design, to allow the storage of the data (data warehouse), their sharing among the community (data repository) and their archiving (archival system).

#### 4.3.5 Archival system

As presented previously, no electronic archiving system is available at Unisanté. Only a file storage system is available. This state presents a risk, as data are not maintained through time, they could be unreadable in 10 years. The other risk is the storage of forbidden data. Some data have to be deleted after a 10 year period. Without an archiving system with alerts and automated deletion, some data could be stored for a longer period than expected.

To improve the perennity of the data, Unisanté could put in place an OAIS compliant archiving system. This archiving system should integrate migration functionalities, data profiling tool, such as the File Information Tool Set (FITS)<sup>24</sup>, alerts, automatic deletion and all other basics functionalities of an archiving system. This system could be used for data but also for all electronic documents.

#### 4.4 Future work

This study offers recommendations about tabular data curation, but these recommendations should now be presented to the General Management at Unisanté. If some of these propositions are accepted, they should be developed and put in place. They have been prioritized and an estimation of involved support services has been done and is presented in the Table 3, to ease the decision of implementation.

Table 3 : Priority of the recommendations

Recommendation	Involved services	Priority
<b>4.3.1 Data governance</b>	Administrative management (DA) Research and cohort support sector (SSRC) Information Technology and Digital Transformation sector Biostatistics consultation unit Research IT Services unit Survey methodology unit UDD UPR	1 – Important and urgent
<b>4.3.5 Archival system</b>	Administrative management (DA) Research and cohort support sector (SSRC) Information Technology and Digital Transformation sector	1 – Important and urgent

<sup>24</sup> <https://projects.iq.harvard.edu/fits/tools>

<b>4.2.1 DMP</b>	UDD	2 – Important but less urgent
<b>4.2.3 Interoperability with REDCap</b>	Research IT Services unit Survey methodology unit UDD	2 – Important but less urgent
<b>4.2.4 Anonymization</b>	Biostatistics consultation unit UDD	2 – Important but less urgent
<b>4.2.5 Centralized data monitoring</b>	Biostatistics consultation unit UDD UPR	2 – Important but less urgent
<b>4.3.4 Data repository</b>	Research IT Services unit UDD UPR	2 – Important but less urgent
<b>4.2.2 Methodology support</b>	Survey methodology unit UDD	3 – nice to have
<b>4.2.6 Codes documentation</b>	Biostatistics consultation unit UDD UPR	3 – nice to have
<b>4.2.7 Training and guidelines</b>	Biostatistics consultation unit Research IT Services unit Survey methodology unit UDD UPR	3 – nice to have
<b>4.3.2 Documentation</b>	Administrative management (DA) UDD UPR	3 – nice to have
<b>4.3.3 File structure</b>	Administrative management (DA) UDD UPR	3 – nice to have

In order to improve this recommendation prioritization, a complete FAIR maturity model (Bahim et al. 2020) will be created. This driving tool could help the institution to prioritize the services and support to develop in terms of data curation and management. A maturity model for support service could also be created to plan their evolution (Cox et al. 2019).

To continue this study, the objective 4 could be completed by creating user scenarios with the implemented recommendations. The objective 5 could also be completed by creating training materials, workshops and exercises according to the implemented recommendations.

Another development to this study would be to perform a similar investigation for qualitative data to cover the full data types collected at Unisanté.



Finally, this works aimed to help other medical institutions to identify what curation tasks are performed by their research teams and help them to support this effort. By multiplying this research among other institutions, data curation could be better mapped, and recommendation could be improved. The questionnaire will be translated in English to be shared among the Information Science community. The structure of the survey will be available in a data repository.

## 5. Limitations of the study

### 5.1 Definition of the project

The first difficulty met in this study was the definition of the project. Data curation is a wide concept. In the Information Science field, data curation relates to archiving and preservation. In research, it can be linked to data preparation and cleaning. Due to the variety of terms and definitions of this field, the literature review has not been as deep as expected. If this project is done again, I would recommend a longer period of time, to improve the literature review, and a better definition of the research terms. I would also restrain the subject to concentrate on some data curation tasks instead of all. More focused research would have more applicable results.

The large definition of the project had also a big impact on the realization of its objectives. On the five defined objectives, only three and a half have been realized completely.

The fourth objective was the integration of recommendation into the research data lifecycle and the production of user scenario to illustrate it. Even if the recommendations issued from this study have been linked to the data lifecycle, no integration scenario has been created, due to a lack of time and data. To realize this objective, the recommended methods and tools should have been tested with real data and in real condition. The software YARD, C<sup>2</sup>Metadata, Amnesia and the package SDCMicro have been partially tested. Yard and Amnesia tests have not been completed due to the complexity of their installation. C<sup>2</sup>Metadata test has not been completed due to a lack of time to create the XML metadata required to process the statistical code. Finally, SDCMicro has been tested for some of its functionalities, but the testing was still in process at the end of this master's thesis.

The fifth objective initially aimed to create training material and exercises to support research teams in the change of practices. It has been quickly modified to harvest information on preferred support, as the creation of training material would have been possible only after the acceptance of the recommendations and completion of the software's tests. Moreover, the survey highlighted that training and theoretical help were less appreciated than a personal support. Therefore, the creation of trainings will be postponed to a future work at Unisanté, to target the real needs of the research teams.

### 5.2 Survey design

The survey design could have been improved to avoid some mistakes.

First, I could have used the Survey Methodology Consultation unit services to help me create the survey. With their expertise, the mistakes presented below could have been avoided. Moreover, the analysis would have been eased, as the chosen variables would have been match to my objectives.

The first improvement for the survey could be the main structure. All questions have been presented inside the data lifecycle. As this concept is widely used when it comes to research data management, this kind of presentation seemed to be appropriate. Moreover, after the first pre-test of the survey, this display of questions has been requested.

However, after analysis, it occurred that some researchers were confused with this presentation. Some of the repeating tasks were marked as «done at another step». Therefore,

the presentation of each task, independently to its place into the data lifecycle, could have improved the data collection. In this kind of structure, a question about the moment of realization of the task should have been added. This would also have improved the «how do you do it» questions, as more precise possible answers could have been proposed.

The second improvement concerns the first question of each group : «Do you realized the following activities ?». The respondent had three possible answers :

1. Yes, I do it myself
2. Somebody does it
3. No

This question should have been structured differently, as recommended by the mandator, to facilitate the interpretation of the results. If the question was limited to a Boolean answer «yes/no», it would have been simpler to map which activity was realized and by who. I did not follow this advice and I kept the three possible answers. If I had to do the study again, I would modify this question.

A third modification could be applied to the «reasons» questions. Data were collected to know why a certain task was realized. This information has not been really analyzed, as it did not add a value to the result. This question could have been removed. For the negative reasons, some possible answers were missing and have been added for the analysis from the comments. If this survey is done again, the following answers should be added : «Lack of time», «Not my responsibility».

The same observation has been made for the «how do you do it» questions. The respondents used the «Other» answer and wrote down the solution. This situation is due to the creation of the survey, where some answers were removed to improve usability. To correct this in the analysis, some variables have been created to complete the results. A grid has been made to match the comment to the new variable, to improve the reproducibility of this study. It is available in [Annex 12](#).

Regarding the survey distribution, the accompanying text indicated a 30 to 40 minutes duration time, but the statistics indicated that 25 minutes were sufficient to complete the questionnaire. The text could have been adapted. It could also have been more explicit in term of targeted population, to get more answers from some of the research's roles.

Finally, data related to the departments at Unisanté have been collected but not used in the analysis, except for the response rate. To obtain applicable results by department, more data should have been collected.

## 6. Conclusion

Data quality at Unisanté can be improved with active data curation.

The active support to research teams at the beginning of a project could be a benefit to increase the quality, with a support on methodology and survey creation. These recommendations could also improve interoperability of the Unisanté datasets. Providing help to document the project and the data, and to organize and name the files would also improve data quality for the long-term.

The recommended support during research project would also improve reproducibility. Contrary to the first ideas expressed in this thesis, the use of software like YARD does not meet the needs of the researchers and is not systematically recommendable. Its use would not adapt well to the practices and would be a too heavy change at Unisanté.

The implementation of a centralized data monitoring would be an asset first to the clinical trials, which are submitted to the GCPs, and second for general data quality, as this service would help to detect errors during data collection. However, the development of such service would be expensive and may not be used by the non-clinical researchers, as it is not mandatory. The documentation of statistical codes and their sharing would improve reproducibility of the research projects and would be a helpful resource to all research teams who would like to reuse some algorithms.

At the end of the project, allowing researchers to share their data on a secured platform, and ensure the perennity of the data with a good archival system would be the right end of the lifecycle of the projects.

Finally, the institutional framework is important to ensure the production of good data quality. The creation of an institutional data governance, involving all stakeholders would be an asset. The creation of a one stop shop (unique entry point) for all research data management and anonymization related questions could be an asset. This would ease the distribution of the issues through the different support services at Unisanté and allows collaboration between them to treat the questions.

Moreover, creating a data policy or a similar document to support the research teams and to transfer the institutional vision would be a benefit and would align Unisanté to other institutions ahead in data management, like the University of Geneva (UNIGE) (UNIGE, 2018), the University of Basel (UNIBAS) (UNIBAS, 2020) or the ETH Zürich (ETH Zürich, 2022).

## Bibliography

- ALLIE, Taryn, JACKSON, Amanda, AMBLER, Jon, JOHNSTON, Katherine, BRUYN, Elsa Du, SCHULTZ, Charlotte, BOLOKO, Linda, WASSERMAN, Sean, DAVIS, Angharad, MEINTJES, Graeme, WILKINSON, Robert J. and TIFFIN, Nicki, 2021. TBDBT: A TB DataBase Template for collection of harmonized TB clinical research data in REDCap, facilitating data standardisation for inter-study comparison and meta-analyses. *PLOS ONE* [Online]. 26 March 2021. Vol. 16, no. 3, pp. e0249165. [Accessed 3 April 2022]. Retrieved from: <https://doi.org/10.1371/journal.pone.0249165>
- ALTER, George Charles, GAGER, Jack, HEUS, Pascal, HUNTER, Carson, IONESCU, Sanda, IVERSON, Jeremy, JAGADISH, H. V., LYLE, Jared, MUELLER, Alexander, NORDGAARD, Sigve, RISNES, Ornulf, SMITH, Dan and SONG, Jie, 2021. Capturing Data Provenance from Statistical Software. *International Journal of Digital Curation* [Online]. 26 April 2021. Vol. 16, no. 1, pp. 14. [Accessed 28 May 2022]. Retrieved from: <https://doi.org/10.2218/ijdc.v16i1.763>
- ÁLVAREZ SÁNCHEZ, Roberto, BERISTAIN IRAOLA, Andoni, EPELDE UNANUE, Gorka and CARLIN, Paul, 2019. TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. *Computer Methods and Programs in Biomedicine* [Online]. 1 November 2019. Vol. 181, pp. 104824. [Accessed 3 March 2022]. Retrieved from: <https://doi.org/10.1016/j.cmpb.2018.12.029>
- BAE, Ho, JUNG, Dahuin, CHOI, Hyun-Soo and YOON, Sungroh, 2019. AnomiGAN: Generative Adversarial Networks for Anonymizing Private Medical Data. In: *Pacific Symposium on Biocomputing 2020* [Online]. WORLD SCIENTIFIC. 2 November 2019. pp. 563–574. [Accessed 7 August 2022]. Retrieved from: [https://doi.org/10.1142/9789811215636\\_0050](https://doi.org/10.1142/9789811215636_0050)
- BAHIM, Christophe, CASORRÁN-AMILBURU, Carlos, DEKKERS, Makx, HERCZOG, Edit, LOOZEN, Nicolas, REPANAS, Konstantinos, RUSSELL, Keith and STALL, Shelley, 2020. The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. *Data Science Journal* [Online]. 27 October 2020. Vol. 19, no. 1, pp. 41. [Accessed 3 March 2022]. Retrieved from: <https://doi.org/10.5334/dsj-2020-041>
- BAYLE, Aurélie, 2020. *RGPD et protection des données : panorama d'une Réglementation et ses impacts* [PowerPoint] . Course material : « RGPD et protection des données » course, Haute école de gestion de Genève, filière Information documentaire, academic year 2020-2022. 16 October 2020
- BOCCALI, Tommaso, SØLSNES, Anne Elisabeth, THORLEY, Mark, WINKLER-NEES, Stefan and TIMMERMAN, Marie, 2021. Practical Guide to Sustainable Research Data. *Science Europe* [Online]. 2 June 2021. [Accessed 16 June 2021]. Retrieved from: <https://doi.org/10.5281/ZENODO.4769703>
- BORGHI, John, ABRAMS, Stephen, LOWENBERG, Daniella, SIMMS, Stephanie and CHODACKI, John, 2018. Support Your Data: A Research Data Management Guide for Researchers. *Research Ideas and Outcomes* [Online]. 9 May 2018. Vol. 4, pp. e26439. [Accessed 1 December 2021]. Retrieved from: <https://doi.org/10.3897/rio.4.e26439>
- CAVOUKIAN, Ann, 2011. Privacy by design : The 7 Foundational Principles : Implementation and Mapping of Fair Information Practices. [online]. 2011. Information and Privacy Commissioner of Ontario. [Accessed 18 September 2022]. Retrieved from: [https://iapp.org/media/pdf/resource\\_center/pbd\\_implement\\_7found\\_principles.pdf](https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf)
- CNIL, Commission nationale de l'informatique et des libertés, 2020. L'anonymisation de données personnelles. *cnil.fr* [Online]. 19 May 2020. [Accessed 1 August 2022]. Retrieved from: <https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles>

CORRALES, David Camilo, CORRALES, Juan Carlos and LEDEZMA, Agapito, 2018. How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning. *Symmetry* [Online]. April 2018. Vol. 10, no. 4, pp. 99. [Accessed 3 April 2022]. Retrieved from: <https://doi.org/10.3390/sym10040099>

COX, Andrew M., KENNAN, Mary Anne, LYON, Liz, PINFIELD, Stephen and SBAFFI, Laura, 2019. Maturing research data services and the transformation of academic libraries. *Journal of Documentation* [Online]. 26 September 2019. Vol. 75, no. 6, pp. 1432–1462. [Accessed 18 March 2022]. Retrieved from: <https://doi.org/10.1108/JD-12-2018-0211>

COX, Andrew and VERBAAN, Eddy, 2018. *Exploring research data management*. London: Facet Publishing. ISBN 978-1-78330-278-9.

DATA CURATION NETWORK, 2022. CURATE(D) Steps and Checklist for Data Curation. *Data Curation Network : Ethical. Reusable. Better* [Online]. 2022. [Accessed 11 June 2022]. Retrieved from: <http://z.umn.edu/curate>

DATA PROTECTION COMMISSION, 2019. Guidance Note: Guidance on Anonymisation and Pseudonymisation. *Data protection Commission* [Online]. June 2019. [Accessed 2 February 2022]. Retrieved from: <https://www.dataprotection.ie/sites/default/files/uploads/2020-09/190614%20Anonymisation%20and%20Pseudonymisation.pdf>

DE LA LAMA, Dina and RACINE, Céline, 2022. *Les pratiques FAIR dans les Hautes écoles et institutions de recherche suisses : partage et accès aux données de la recherche*. Genève: Haute école de gestion de Genève.

DERIDDER, Jody L., 2018. *Digital curation fundamentals*. Lanham: Rowman & Littlefield. ISBN 978-1-5381-1121-5.

DIAZ, Pablo, 2022. Data protection: legal considerations for research in Switzerland. *FORS Guide* [Online]. January 2022. No. 17, pp. 12. [Accessed 23 April 2022]. Retrieved from: <https://doi.org/10.24449/FG-2022-00017>

DIGITAL CURATION CENTER, 2022. What is digital curation? *DCC : Because good research needs good data* [Online]. 2022. [Accessed 10 June 2022]. Retrieved from: <https://www.dcc.ac.uk/about/digital-curation>

EGLI, Adrian, BATTEGAY, Manuel, BÜCHLER, Andrea C., BÜHLMANN, Peter, CALANDRA, Thierry, ECKERT, Philippe, FURRER, Hansjakob, GREUB, Gilbert, JAKOB, Stephan M., KAISER, Laurent, LEIB, Stephen L., MARSCH, Stephan, MEINSHAUSEN, Nicolai, PAGANI, Jean-Luc, PUGIN, Jerome, RÄTSCH, Gunnar, SCHRENZEL, Jacques, SCHÜPBACH, Reto, SIEGEMUND, Martin, ZAMBONI, Nicola, ZBINDEN, Reinhard, ZINKERNAGEL, Annelies and BORGWARDT, Karsten, 2020. SPHN/PHRT: forming a Swiss-wide infrastructure for data-driven sepsis research. *Studies in Health Technology and Informatics* [Online]. 16 June 2020. Vol. 270, pp. 1163–1167. [Accessed 17 May 2022]. Retrieved from: <https://doi.org/10.3233/SHTI200346>

ELSEVIER, 2022. Research Data. *Elsevier.com* [Online]. 2022. [Accessed 11 December 2021]. Retrieved from: <https://www.elsevier.com/about/policies/research-data>

EMA, European Medicines Agency, 2016. Guideline for good clinical practice E6(R2) : Step 5. European Medicines Agency : Science medicines health [Online]. 14 June 2017 [Accessed 27 March 2022]. Retrieved from: [https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5\\_en.pdf](https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5_en.pdf)

ETH ZÜRICH, 2022. Guidelines for Research Data Management at ETH Zurich. *ETH Zürich* [Online]. 1 July 2022. [Accessed 3 August 2022]. Retrieved from: <https://rechtssammlung.sp.ethz.ch/Dokumente/414.2en.pdf>

FDPIC, Federal Data Protection and Information Commissioner, 2015. A Guide for technical and organizational measures. *Federal Data Protection and Information Commissioner* [Online]. 21 August 2015. Retrieved from: <https://www.edoeb.admin.ch/edoeb/en/home/data-protection/dokumentation/guides/technical-and-organizational-measures.html>

FORCE11, 2014. The FAIR Data Principles. *FORCE11* [Online]. 3 September 2014. [Accessed 10 December 2021]. Retrieved from: <https://www.force11.org/group/fairgroup/fairprinciples>

FORS, 2018. Guide n°3 : la gestion des données. *FORS : explore. understand. Share* [Online]. 2018. [Accessed 17 May 2021]. Retrieved from: <https://forscenter.ch/wp-content/uploads/2018/08/guide-3-gestion-des-donnees.pdf>

FORS, 2021. Préparer vos données pour leur dépôt dans SWISSUbase. *FORS : explore. understand. Share* [Online]. 2021. [Accessed 23 April 2022]. Retrieved from: [https://forscenter.ch/wp-content/uploads/2021/12/preparer-vos-donnees-pour-leur-depot-dans-swissubase-version-courte-6\\_fr.pdf](https://forscenter.ch/wp-content/uploads/2021/12/preparer-vos-donnees-pour-leur-depot-dans-swissubase-version-courte-6_fr.pdf)

FORTIN, Marie-Fabienne and GAGNON, Johanne, 2016. *Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives*. Montréal: Chenelière éducation,. ISBN 978-2-7650-5006-3.

GAUDINAT, Arnaud, 2021. *Data curation* [PowerPoint]. Course material : « Data curation » course, Haute école de gestion de Genève, filière Information documentaire, academic year 2020-2022. 2021.

GORDON, Ben, BARRETT, Jake, FENNESSY, Clara, CAKE, Caroline, MILWARD, Adam, IRWIN, Courtney, JONES, Monica and SEBIRE, Neil, 2021. Development of a data utility framework to support effective health data curation. *BMJ Health & Care Informatics* [Online]. 12 May 2021. Vol. 28, no. 1, pp. e100303. [Accessed 1 December 2021]. Retrieved from: <https://doi.org/10.1136/bmjhci-2020-100303>

HADROSSEK, Christine, JANIK, Joanna, LIBES, Maurice, LOUVET, Violaine, QUIDOZ, Marie-Claude, RIVET, Alain and ROMIER, Geneviève, 2021. *Guide de bonnes pratiques sur la gestion des données de la Recherche* [Online]. 8 December 2021. [Accessed 9 February 2022]. Retrieved from: [https://mi-gt-donnees.pages.math.unistra.fr/guide/guide\\_bonnes\\_pratiques\\_gestion\\_donnees\\_recherche\\_v1.pdf](https://mi-gt-donnees.pages.math.unistra.fr/guide/guide_bonnes_pratiques_gestion_donnees_recherche_v1.pdf)

HAMEED, Mazhar and NAUMANN, Felix, 2020. Data Preparation: A Survey of Commercial Tools. *ACM SIGMOD Record* [Online]. 17 December 2020. Vol. 49, no. 3, pp. 18–29. [Accessed 21 April 2022]. Retrieved from: <https://doi.org/10.1145/3444831.3444835>

HANK, Carolyn and BISHOP, Bradley Wade, 2018. Measuring FAIR Principles to Inform Fitness for Use. *International Journal of Digital Curation* [Online]. 22 December 2018. Vol. 13, no. 1, pp. 35–46. [Accessed 17 June 2021]. Retrieved from: <https://doi.org/10.2218/ijdc.v13i1.630>

HARVEY, D. R. and OLIVER, Gillian, 2016. *Digital curation*. Second edition. Chicago: ALA Neal-Schuman, an imprint of the American Library Association. ISBN 978-0-8389-1385-7.



HENDERSON, Margaret E., 2017. *Data management: a practical guide for librarians*. Lanham, Maryland: Rowman & Littlefield. Practical guides for librarians, no. 28. ISBN 978-1-4422-6438-0.

HIGGINS, Sarah, 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation [Online]*. 2 December 2008. Vol. 3, no. 1, pp. 134–140. [Accessed 17 June 2021]. Retrieved from: <https://doi.org/10.2218/ijdc.v3i1.48>

HUDSON-VITALE, Cynthia, HADLEY, Hannah, MOORE, Jennifer, JOHNSTON, Lisa, KOZLOWSKI, Wendy, CARLSON, Jake, BLAKE, Mara and HERNDON, Joel, 2020. Extending the Research Data Toolkit: Data Curation Primers. *International Journal of Digital Curation [Online]*. 30 December 2020. Vol. 15, no. 1, pp. 14. [Accessed 12 December 2021]. Retrieved from: <https://doi.org/10.2218/ijdc.v15i1.713>

HUNDEPOOL, Anco, 2012. *Statistical disclosure control*. Chichester, West Sussex, United Kingdom: Wiley. Wiley series in survey methodology. ISBN 978-1-118-34821-5.

ICPSR, 2022. Guide to Social Science Data Preparation and Archiving : Best Practice Throughout the Data Life Cycle: 6<sup>th</sup> Edition. *ICPSR [Online]*. 2022. [Accessed 30 March 2022]. Retrieved from: <https://www.icpsr.umich.edu/web/pages/deposit/guide/index.html>

INSTITUT UNIVERSITAIRE DE MÉDECINE and SOCIALE ET PRÉVENTIVE (IUMSP), 2018. Annual report 2017. *Unisanté [Online]*. 2018. [Accessed 23 June 2022]. Retrieved from: [https://www.unisante.ch/sites/default/files/upload/pdf/IUMSP\\_rapport\\_annuel\\_2017.pdf](https://www.unisante.ch/sites/default/files/upload/pdf/IUMSP_rapport_annuel_2017.pdf)

INSTITUT UNIVERSITAIRE ROMAND DE SANTÉ AU TRAVAIL (IST), 2018. Rapport d'activité 2017. *Unisanté [Online]*. 2018. [Accessed 23 June 2022]. Retrieved from: [https://www.unisante.ch/sites/default/files/upload/pdf/IST\\_Rapport\\_annuel\\_2017.pdf](https://www.unisante.ch/sites/default/files/upload/pdf/IST_Rapport_annuel_2017.pdf)

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO), 2008. *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model [Online]*. Geneva: ISO, December 2008. ISO/IEC 25012:2008. [Accessed 18 June 2022]. Retrieved from: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/57/35736.html>

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO), 2015. *Data quality — Part 8: Information and data quality: Concepts and measuring [Online]*. Geneva: ISO, November 2015. ISO 8000-8:2015. [Accessed 18 June 2022]. Retrieved from: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/08/60805.html>

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO), 2016. *Information and documentation — Records management — Part 1: Concepts and principles [Online]*. Geneva: ISO, April 2016. ISO 15489-1:2016. [Accessed 15 December 2021]. Retrieved from: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/25/62542.html>

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO), 2020. *Data quality — Part 2: Vocabulary [Online]*. Geneva: ISO, June 2020. ISO 8000-2:2020. [Accessed 18 June 2022]. Retrieved from: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/08/05/80543.html>

JOHNSTON, Anna, 2015. Bradley Cooper's taxi ride: a lesson in privacy risk. *Salinger Privacy [Online]*. 19 April 2015. [Accessed 18 September 2022]. Retrieved from: <https://www.salingerprivacy.com.au/2015/04/19/bradley-coopers-taxi-ride-a-lesson-in-privacy-risk/>

JOHNSTON, Lisa (ed.), 2017. *Curating research data. Volume two: A handbook of current practices [Online]*. Chicago, Illinois: Association of College and Research Libraries, a division



of the American Library Association, January 2017. [Accessed 11 November 2021]. ISBN 978-0-8389-8862-6. Retrieved from:

[https://renouvaud1.primo.exlibrisgroup.com/permalink/41BCULAUSA\\_LIB/1vikse1/alma991021107671202852](https://renouvaud1.primo.exlibrisgroup.com/permalink/41BCULAUSA_LIB/1vikse1/alma991021107671202852)

JOTTERAND, Alexandre, 2022. Personal Data or Anonymous Data: where to draw the lines (and why)? *Jusletter*. [Online]. 15 August 2022. No. 1119. [Accessed 24 August 2022]. Retrieved from: [https://jusletter.weblaw.ch/juslissues/2022/1119/personal-data-or-ano\\_173939252d.html](https://jusletter.weblaw.ch/juslissues/2022/1119/personal-data-or-ano_173939252d.html)

KONSTANTINO, Nikolaos, ABEL, Edward, BELLOMARINI, Luigi, BOGATU, Alex, CIVILI, Cristina, IRFANIE, Endri, KOEHLER, Martin, MAZILU, Lacramioara, SALLINGER, Emanuel, FERNANDES, Alvaro A. A., GOTTLOB, Georg, KEANE, John A. and PATON, Norman W., 2019. VADA: an architecture for end user informed data preparation. *Journal of Big Data* [Online]. December 2019. Vol. 6, no. 1, pp. 74. [Accessed 19 June 2022]. Retrieved from: <https://doi.org/10.1186/s40537-019-0237-9>

KOUPER, Inna, TUCKER, Karen, THARP, Kevin, BOOVEN, Mary Ellen van and CLARK, Ashley, 2021. Active Curation of Large Longitudinal Surveys: A Case Study. *Journal of eScience Librarianship* [Online]. 11 August 2021. Vol. 10, no. 3. [Accessed 3 March 2022]. Retrieved from: <https://doi.org/10.7191/jeslib.2021.1210>

KOVACS, Marton, HOEKSTRA, Rink and ACZEL, Balazs, 2021. The Role of Human Fallibility in Psychological Research: A Survey of Mistakes in Data Management. *Advances in Methods and Practices in Psychological Science* [Online]. 1 October 2021. Vol. 4, no. 4, pp. 1–13. [Accessed 3 March 2022]. Retrieved from: <https://doi.org/10.1177/25152459211045930>

LMA RESEARCH DATA MANAGEMENT WORKING GROUP, 2022. Biomedical Data Lifecycle. *Harvard medical school: Longwood Research data management* [Online]. 2022. [Accessed 22 April 2022]. Retrieved from: <https://datamanagement.hms.harvard.edu/about/what-research-data-management/biomedical-data-lifecycle>

MAKHLOUF-SHABOU, Basma, 2021. Data processing [PowerPoint]. Course material : « M7c Gouvernance des données » course, Haute école de gestion de Genève, filière Information documentaire, academic year 2020-2022. 18 October 2021.

MANU, T R and BHAKTI, Gala, 2021. Data Curation Activities in Research Data Repositories: Best Practice. *ResearchGate* [Online]. January 2021. [Accessed 29 October 2021]. Retrieved from: [https://www.researchgate.net/profile/Manu-T-R-2/publication/348873527\\_Data\\_Curation\\_Activities\\_in\\_Research\\_Data\\_Repositories\\_Best\\_Practices/links/60b8edb9458515218f89d75b/Data-Curation-Activities-in-Research-Data-Repositories-Best-Practices.pdf](https://www.researchgate.net/profile/Manu-T-R-2/publication/348873527_Data_Curation_Activities_in_Research_Data_Repositories_Best_Practices/links/60b8edb9458515218f89d75b/Data-Curation-Activities-in-Research-Data-Repositories-Best-Practices.pdf)

MCLURE, Merinda, LEVEL, Allison V., CRANSTON, Catherine L., OEHLERTS, Beth and CULBERTSON, Mike, 2014. Data Curation: A Study of Researcher Practices and Needs. *portal: Libraries and the Academy* [Online]. 2014. Vol. 14, no. 2, pp. 139–164. [Accessed 17 November 2021]. Retrieved from: <https://doi.org/10.1353/pla.2014.0009>

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION, Centre National de la Recherche Scientifique, INIST-CNRS, Centre National de la Recherche Scientifique and GIS RÉSEAU URFIST, Réseau des Unités Régionales de Formation à l'Information Scientifique et Technique, 2022. Glossaire. *DoRANum* [Online]. 25 March 2022. [Accessed 15 May 2022]. Retrieved from: <https://doranum.fr/glossaire-donnees-recherche>

NATURE, 2021. Reporting standards and availability of data, materials, code and protocols. *Nature Portfolio* [Online]. 2021. [Accessed 11 December 2021]. Retrieved from: <https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards>

OECD, ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding. *OECD : Better policies for better lives* [Online]. 2007. [Accessed 6 June 2022]. Retrieved from: <https://www.oecd.org/sti/inno/38500813.pdf>

OLATUNJI, Iyiola E., RAUCH, Jens, KATZENSTEINER, Matthias and KHOSLA, Megha, 2022. A Review of Anonymization for Healthcare Data. *Big Data* [Online]. 10 March 2022. [Accessed 24 June 2022]. Retrieved from: <https://doi.org/10.1089/big.2021.0169>

PEER, Limor and DULL, Joshua, 2020. YARD: A Tool for Curating Research Outputs. *Data Science Journal* [Online]. 15 July 2020. Vol. 19, no. 1, pp. 28. [Accessed 12 November 2021]. Retrieved from: <https://doi.org/10.5334/dsj-2020-028>

PEZOULAS, Vasileios C., KOUROU, Konstantina D., KALATZIS, Fanis, EXARCHOS, Themis P., VENETSANOPOULOU, Alik, ZAMPELI, Evi, GANDOLFO, Saviana, SKOPOULI, Fotini, DE VITA, Salvatore, TZIOUFAS, Athanasios G. and FOTIADIS, Dimitrios I., 2019. Medical data quality assessment: On the development of an automated framework for medical data curation. *Computers in Biology and Medicine* [Online]. 1 April 2019. Vol. 107, pp. 270–283. [Accessed 8 March 2022]. Retrieved from: <https://doi.org/10.1016/j.combiomed.2019.03.001>

PHAM, Amy, 2018. *Surveying the state of data curation: a review of policy and practice in UK HEIs* [Online]. Glasgow: University of Strathclyde. Master Thesis. [Accessed 1 December 2021]. Retrieved from: <http://eprints.rclis.org/33511/>

PING, Haoyue, STOYANOVICH, Julia and HOWE, Bill, 2017. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* [Online]. New York, NY, USA: Association for Computing Machinery, 27 June 2017. pp. 1–5. [Accessed 8 July 2022]. SSDBM '17. ISBN 978-1-4503-5282-6. Retrieved from: <https://doi.org/10.1145/3085504.3091117>

PLOS, 2014. PLOS' New Data Policy: Public Access to Data. *EveryONE* [Online]. 24 February 2014. [Accessed 11 December 2021]. Retrieved from: <https://everyone.plos.org/2014/02/24/plos-new-data-policy-public-access-data-2/>

POLICLINIQUE MÉDICALE UNIVERSITAIRE (PMU), 2018. Rapport annuel 2018. *Unisanté* [Online]. 2018. [Accessed 23 June 2022]. Retrieved from: [https://www.unisante.ch/sites/default/files/upload/pdf/PMU\\_RA\\_2018\\_web.pdf](https://www.unisante.ch/sites/default/files/upload/pdf/PMU_RA_2018_web.pdf)

PROMOTION SANTÉ VAUD (PROSV), 2018. Rapport d'activité 2018. *Unisanté* [Online]. 2018. [Accessed 23 June 2022]. Retrieved from: [https://www.unisante.ch/sites/default/files/upload/pdf/ProSV\\_rapport%20d%27activit%C3%A9%202018\\_planche.pdf](https://www.unisante.ch/sites/default/files/upload/pdf/ProSV_rapport%20d%27activit%C3%A9%202018_planche.pdf)

RACINE, Céline, 2021. *La qualité des données grâce à la curation : quand et comment l'intégrer au cycle de vie des données : Travail de réflexion*. Genève: Haute école de gestion de Genève.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). *EUR-Lex* [Online]. 2016. [Accessed 26 June 2022]. Retrieved from: <http://data.europa.eu/eli/reg/2016/679/oj/eng>

RICE, Robin C. and SOUTHALL, John, 2016. *The data librarian's handbook*. London: Facet Publ. ISBN 978-1-78330-047-1.

RIGHTS (OCR), Office for Civil, 2012. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. *HHS.gov* [Online]. 7 September 2012. Last Modified: 31 May 2022. [Accessed 7 August 2022]. Retrieved from: <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

SANTOS, Anouk, 2020. *Données de la recherche: cadre juridique et licences* [Online]. Genève: Haute école de gestion de Genève. Master Thesis. [Accessed 6 May 2021]. Retrieved from: <https://doc.rero.ch/record/329704>

SNSF, Swiss National Science Foundation, 2022. Open Research Data. *Swiss National Science Foundation (SNSF)* [Online]. 2022. [Accessed 6 June 2022]. Retrieved from: <https://www.snf.ch/en/dMILj9t4LNk8NwyR/topic/undefined/en/dMILj9t4LNk8NwyR/topic/>

SPRUMONT, Dominique, 2019. *Protection des données, anonymisation et recherche, Lausanne, 3 October 2019* [PowerPoint]. Lausanne, Commission cantonale d'éthique de la recherche sur l'être humain, 3 October 2019. [Accessed 2 August 2022]. Lunch LRH. Retrieved from: <https://static1.squarespace.com/static/60b94bed393f8064950b2821/t/616e9f1dce9f7f7a05927376/1634639647288/Presentation+Protection+des+donnees+anonymisation+et+recherche+191003.pdf>

SR 235.1 - Federal Act of 19 June 1992 on Data Protection (FADP). *Fedlex : The publication platform for federal law* [Online]. 2019. [Accessed 6 June 2022]. Retrieved from: [https://www.fedlex.admin.ch/eli/cc/1993/1945\\_1945\\_1945/en](https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en)

SR 810.30 - Federal Act of 30 September 2011 on Research involving Human Beings (Human Research Act, HRA). *Fedlex : The publication platform for federal law* [Online]. 2021. [Accessed 14 May 2022]. Retrieved from: <https://www.fedlex.admin.ch/eli/cc/2013/617/en>

STAM, Alexandra and KLEINER, Brian, 2020. Data anonymisation: legal, ethical, and strategic considerations. *FORS Guide* [Online]. June 2020. No. 11, pp. 15. [Accessed 23 April 2022]. Retrieved from: <https://doi.org/10.24449/FG-2020-00011>

TAMMARO, Anna Maria, MATUSIAK, Krystyna K, CASAROSA, Vittore and SPOSITO, Frank Andreas, 2018. Findings from the IFLA LTR Project [Poster]. In: *Open Science Conference* [Online]. Berlin, Germany. 2018. [Accessed 10 June 2022]. Retrieved from: <https://www.ifla.org/fr/events/ifla-ltr-project-on-data-curation-at-the-open-science-conference-2018>

TIER Protocol 3.0, no date. *Project TIER | Teaching Integrity in Empirical Research*. Online. [Accessed 30 July 2022]. Retrieved from: <https://www.projecttier.org/tier-protocol/tier-protocol-version-history/specifications-3-0>

UK DATA SERVICE, 2019. Research Data Lifecycle [Video]. *Youtube* [Online]. 13 August 2019. [Accessed 10 June 2022]. Retrieved from: <https://www.youtube.com/watch?v=wjFMMQD3UA>

UNIBAS, University of Basel, 2020. Policy on research data management at the University of Basel. *University of Basel - Research Data Management-Network University of Basel* [Online]. 22 September 2020. [Accessed 3 August 2022]. Retrieved from: <https://researchdata.unibas.ch/en/services/policy-on-rdm/>

UNIGE, Université de Genève, 2018. Politique institutionnelles sur la gestion des données de recherche. *Université de Genève - Données de la recherche* [Online]. 25 June 2018. Last Modified: 19 July 2021. [Accessed 3 August 2022]. Retrieved from: <https://www.unige.ch/researchdata/fr/propos/politique/>

UNISANTÉ, Centre universitaire de médecine générale et santé publique, 2019. Rapport annuel 2019. *Unisanté* [Online]. 2019. [Accessed 20 May 2022]. Retrieved from: [https://www.unisante.ch/sites/default/files/upload/pdf/RA\\_2019\\_web.pdf](https://www.unisante.ch/sites/default/files/upload/pdf/RA_2019_web.pdf)

UNISANTÉ, Centre universitaire de médecine générale et santé publique, 2021. Rapport annuel 2020. *Unisanté* [Online]. 2021. [Accessed 5 May 2022]. Retrieved from: [https://www.unisante.ch/sites/default/files/upload/pdf/Unisante\\_RA2020-DIGITAL-OK.pdf](https://www.unisante.ch/sites/default/files/upload/pdf/Unisante_RA2020-DIGITAL-OK.pdf)

UNISANTÉ, Centre universitaire de médecine générale et santé publique, 2022a. Historique. *Unisanté* [Online]. 2022. [Accessed 20 May 2022]. Retrieved from: <https://www.unisante.ch/fr/unisante/propos/historique>

UNISANTÉ, Centre universitaire de médecine générale et santé publique, 2022b. Missions et valeurs. *Unisanté* [Online]. 2022. [Accessed 20 May 2022]. Retrieved from: <https://www.unisante.ch/fr/unisante/missions-valeurs>

UNISANTÉ, Centre universitaire de médecine générale et santé publique, 2022c. Rapport annuel 2021. *Unisanté* [Online]. 2022. [Accessed 2 August 2022]. Retrieved from: [https://issuu.com/unisante/docs/44646\\_insti\\_rapport\\_annuel\\_2021\\_web](https://issuu.com/unisante/docs/44646_insti_rapport_annuel_2021_web)

VENKATESAN, Aravind, KARAMANIS, Nikiforos, IDE-SMITH, Michele, HICKFORD, Jonathan and MCENTYRE, Johanna, 2019. Understanding life sciences data curation practices via user research. *F1000Research* [Online]. 11 September 2019. [Accessed 17 November 2021]. Retrieved from: <https://f1000research.com/articles/8-1622>

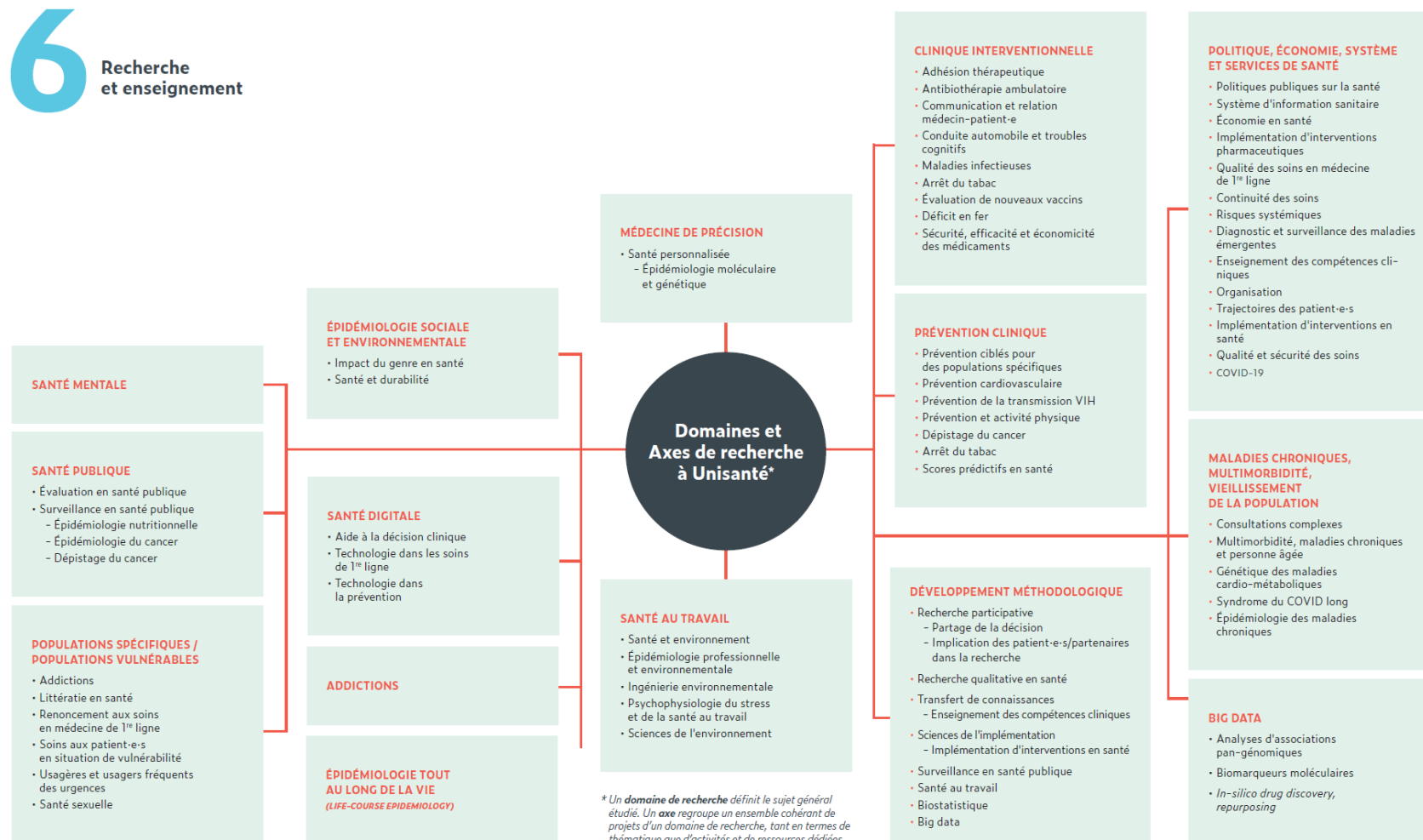
VIJAYANANTHAN, A and NAWAWI, O, 2008. The importance of Good Clinical Practice guidelines and its role in clinical trials. *Biomedical Imaging and Intervention Journal* [Online]. 1 January 2008. Vol. 4, no. 1, pp. e5. [Accessed 26 June 2022]. Retrieved from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3097692/>

WITT, Michael, CARLSON, Jacob, BRANDT, D. Scott and CRAGIN, Melissa H., 2009. Constructing Data Curation Profiles. *International Journal of Digital Curation* [Online]. 7 December 2009. Vol. 4, no. 3, pp. 93–103. [Accessed 17 November 2021]. Retrieved from: <https://doi.org/10.2218/ijdc.v4i3.117>

YOON, Jinsung, DRUMRIGHT, Lydia N. and VAN DER SCHAAR, Mihaela, 2020. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics* [Online]. August 2020. Vol. 24, no. 8, pp. 2378–2388. [Accessed 7 July 2022]. Retrieved from: <https://doi.org/10.1109/JBHI.2020.2980262>

ZACH, 2022. What is Tabular Data? (Definition & Example). *Statology* [Online]. 25 March 2022. [Accessed 18 April 2022]. Retrieved from: <https://www.statology.org/tabular-data/>

## Annex 1 : Unisanté research fields (in French)



(Unisanté, 2022c)

## Annex 2 : Sub-objectives of the study (March 2022)

Objective	Sub-objective	Method	Deliverable
<b>1. To establish a state of the art</b>	1.1 To define data curation and anonymization activities in scientific research	Literature review (books, database, ISO standards...) and analysis of results	Grid to present the literature review results  List of activities
	1.2 To establish a list of existing tools and methods	Literature review and analysis of results	List of tools
<b>2. To evaluate practices and needs at Unisanté</b>	2.1 Creation of data collection tools	Redaction	List of questions for interviews Questionnaire
	2.2 To list data curation activities performed in Unisanté research projects	Personal knowledge Survey Interviews	List of activities
	2.3 To evaluation needs of support and improvement potentials in terms of data curation and anonymization	Survey Interviews	List of needs and possibilities of improvement
	2.4 To identify recommendation and support already available	Interviews Documents review	List of prestation
<b>3. To facilitate data curation process and anonymization</b>	3.1 To connect identified tools to research activities	Data analysis from sub-objectives 1.1 and 1.2	Grid / matrix / table of results
	3.2 To connect identified tools to identified needs	Data analysis from sub-objectives 1.2 and 2.2	Grid / matrix / table of results
	3.3 To test identified tools	Tests with real data	Test report
<b>4. To integrate data curation and anonymization practices into the research project lifecycle at Unisanté</b>	4.1 To recommend tools or methods for Unisanté	Interpretation of objective 3	Master thesis
	4.2 To scenarize usage of recommended tools or methods within a project	Interpretation of objective 4	Schematization (processes flowchart), user scenarios

<b>5. To support change of practices in the research teams</b>	5.1 To inform teams on active data curation and available support	Redaction	Communication material (newsletter, web page)
	5.2 To train users to the recommended tools or methods	Redaction	Training material (courses, conference) and practical exercises



## Annex 3 : Structure of the table of data curation activities

Activity	Group	Tool or method	Source
Citation description the activity	Category or group	List of tools or methods indicated in the document	Citation of the source (ISO 690)
<i>Appraise : consider risk factors (LPD, sensitive info, copyright)</i>	<i>Re-identification risk</i>	<i>Disclosure risk : identity finder, bulk extractor, spider, find_ssn</i>  <i>Deidentify data : delete direct identifier (PII), redact sensitive info (OCR + script) do not perform in formation must stay identical, restrict access, store on HIPAA compliant repository</i>	(Johnston, 2017) (Data Protection Commission, 2019)

The full grid is available on demand, as it contains full citations, protected by copyright, and is too long to figure in this thesis.



# Annex 4 : Survey

## Gestion des données de recherche : activités réalisées à Unisanté

L'objectif de ce questionnaire est de mieux connaître vos pratiques en gestion des données de recherche à Unisanté, que ce soit au moment de créer vos outils de récolte, de nettoyer vos données, de les analyser, de les partager ou de les archiver.

Les résultats de ce questionnaire permettront à l'unité documentation et données (<https://unisanté/unites/7867>) de développer de nouvelles prestations pour correspondre à vos besoins en matière de gestion des données.

Répondez au questionnaire de la façon la plus sincère, au plus proche de vos pratiques actuelles en tenant compte des aspects suivants :

- **Ne répondez que si vous participez à des projets de recherche ou mandats réalisés à Unisanté**, où Unisanté est l'intervenant principal
- **Ce questionnaire ne concerne que la gestion des données tabulaires** (CSV, MS Access, Excel, etc...); les données sous forme d'image, les données d'entrevues ou de focus group, etc. seront prises en considération dans un second temps.
- **Si vos pratiques en gestion des données diffèrent d'un projet à l'autre** (notamment le traitement et le nettoyage des données), choisissez un projet précis que vous avez réalisé dernièrement.

Le questionnaire durera environ 30 minutes.

Merci de votre précieuse contribution.

Il y a 62 questions dans ce questionnaire.

## Informations personnelles

Afin de proposer des services de soutien adaptés à vos besoins, nous souhaitons en savoir plus sur vous.

### Dans quel(s) département(s) faites-vous de la recherche ? \*

① Cochez la ou les réponses

Veuillez choisir toutes les réponses qui conviennent :

- ☐ Département des polycliniques (DDP)
- ☐ Département épidémiologie et systèmes de santé (DESS)
- ☐ Département formation, recherche et innovation (DFRI)
- ☐ Département médecine de famille (DMF)
- ☐ Département promotion de la santé et préventions (DPSP)
- ☐ Département santé, travail et environnement (DSTE)
- ☐ Département vulnérabilités et médecine sociale (DVMS)

### Quel est votre rôle au sein de l'équipe de projet ?

Si vous remplissez le questionnaire par rapport à un projet en particulier : quel est votre rôle dans ce projet ?

① Cochez la ou les réponses

Veuillez choisir toutes les réponses qui conviennent :

- ☐ Investigateur.trice principal.e (PI)
- ☐ Chargé.e de recherche
- ☐ Data manager
- ☐ Data scientist
- ☐ Statisticien.ne
- ☐ Doctorant.e
- ☐ Auxiliaire

☐ Autre:

**Faites-vous partie d'un service de soutien à Unisanté ?**

Les services de soutien à la recherche à Unisanté sont :

- Unité documentation et données (<https://unisanté.unites/7867>) (DFRI)
- Unité promotion de la recherche (<http://unisanté.formation-recherche/prestations-unite-promotion-recherche-src.htm>) (DFRI)
- Consultation en biostatistique (<https://unisanté.formation-recherche/consultations/consultation-biostatistique.htm>) (DFRI)
- Soutien en méthodologie d'enquête (<https://www.unisanté.ch/fr/formation-recherche/ressources-pour-recherche/soutiens-methnologiques/methodologie-denquetes>) (DESS)
- Plateforme de recherche qualitative (<https://www.unisanté.ch/fr/formation-recherche/ressources-pour-recherche/soutiens-methnologiques/plateforme-recherche>) (DESS)
- Groupe Data Science (DESS)
- Secteur des systèmes d'information et transformation digitale (<https://unisanté.unites/6839>) (DFI)

Veuillez sélectionner une seule des propositions suivantes :

- ☐ Oui  
☐ Non

**Sur quel(s) type(s) de projet / design d'étude travaillez-vous ?**

Si vous remplissez le questionnaire par rapport à un projet en particulier : quel est le type de ce projet ?

📌 Cochez la ou les réponses

Veuillez choisir toutes les réponses qui conviennent :

- ☐ Essai clinique (OClin ou OClin-Dim)
- ☐ Etude observationnelle : Etude prospective impliquant des personnes (ORH 2)
- ☐ Etude observationnelle : Etude de réutilisation données ou échantillons (ORH 3)
- ☐ Etude observationnelle : Etude impliquant des embryons (ORH 4)
- ☐ Etude observationnelle : Etude impliquant des personnes décédées (ORH 5)
- ☐ Revue de la littérature (hors LRH)
- ☐ Rapport de série de cas (<5) (hors LRH)
- ☐ Enquêtes d'opinion (sans données médicales) (hors LRH)
- ☐ Enquêtes anonymes à la source (hors LRH)
- ☐ Etude de démarche qualité (hors LRH)
- ☐ Recherche sur des données agrégées (non personnelles) (hors LRH)
- ☐ Analyse de données anonymisées (stricte application) (hors LRH)

☐ Autre:

## Introduction au questionnaire

### Le cycle de vie des données

Ce questionnaire a pour but de connaître vos pratiques en gestion des données.

Vous allez passer à travers 4 pages, représentant plusieurs étapes du cycle de vie des données



### Planification du projet

La planification est la première étape du cycle de vie des données.

Elle couvre le début du projet, c'est-à-dire toutes les étapes réalisées avant la création des moyens de récolte des données



Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ?

Choisissez la réponse appropriée pour chaque élément :

	Oui, par moi	Oui, par quelqu'un d'autre	Non
Documenter le contexte du projet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Documenter la capture / collecte des données	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Documenter les traitements des données	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Documenter l'organisation et nommage des fichiers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rédiger un Data Management Plan	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mettre en place un suivi des modifications des fichiers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Choisir des formats de fichier ouverts	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- **Documenter le contexte du projet** : Par contexte, on entend décrire le projet, qui sont les investigateur.trice.s, le promoteur, le financement, etc...
- **Documenter la capture / collecte des données** : l'échantillonnage, les moyens de collecte utilisés, les logiciels utilisés, etc...
- **Documenter les traitements prévus des données** : le contrôle qualité, la préparation, le nettoyage, le codage, l'anonymisation des données
- **Documenter l'organisation et nommage des fichiers** : Mettre en place une structure des fichiers qui restera fixe et choisir la manière dont seront nommés les fichiers. Vous documentez peut-être ces aspects avec un fichier texte ou des métadonnées. Vous utilisez peut-être des régles de nommage (<https://vp.unil.ch/gedunil/2019/07/les-regles-de-nommage/>).
- **Rédiger un Data Management Plan** : décrire comment seront gérées les données tout au long du projet, comment seront gérés les aspects éthiques ou juridique, quelles sont les prévisions quant à l'archivage ou le partage de données...
- **Mettre en place un suivi des modifications**: qui a fait quelle modification, à quel moment, pourquoi...
- **Mettre en place un système pour vérifier l'intégrité des fichiers**. Est-ce que vous mettez en place un système, une méthodologie pour vérifier périodiquement que les fichiers ne sont pas corrompus, qu'aucune modification "non-autorisée" n'ait été faite, que les extensions correspondent bien aux types de fichier, etc. Vous utilisez un programme ou un script pour faire des vérifications régulières. Vous utilisez des logs ou un suivi de modification pour faire cette vérification.
- **Choisir un espace de stockage adapté et attribuer des droits** : Est-ce que vous choisissez un espace de stockage sécurisé, offrant des sauvegardes régulières ? Attribuer des droits d'accès et de modification selon les rôles et responsabilités définies pour le projet
- **Choisir des formats de fichier ouverts** : Est-ce que vous suivez la recommandation générale "choisir des formats libres et ouverts (exemple : fichier texte (.txt) ou Texte OpenDocument (.odt) plutôt que Word (.docx)". Vous utilisez peut-être des tableaux de correspondance (<https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving>), pour vous aider à choisir un format de fichier.

**Pour quelle(s) raison(s) effectuez-vous ces activités ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter le contexte du projet))

----- ou Scenario 2 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter la capture / collecte des données))

----- ou Scenario 3 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter les traitements des données))

----- ou Scenario 4 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter l'organisation et nommage des fichiers))

----- ou Scenario 5 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Rédiger un Data Management Plan))

----- ou Scenario 6 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Mettre en place un suivi des modifications des fichiers))

----- ou Scenario 7 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Mettre en place un système pour vérifier l'intégrité des fichiers))

----- ou Scenario 8 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Choisir un espace de stockage adapté et attribuer des droits))

----- ou Scenario 9 -----

La réponse était 'Oui, par moi' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Choisir des formats de fichier ouverts))

	C'est obligatoire	Qualité du projet ou des données	Autre
Documenter le contexte du projet	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter la capture / collecte des données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter les traitements des données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter l'organisation et nommage des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Rédiger un Data Management Plan	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mettre en place un suivi des modifications des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Choisir des formats de fichier ouverts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '7 [plandocwhy]' (Pour quelle(s) raison(s) effectuez-vous ces activités ? )

	Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?
Documenter le contexte du projet	<input type="text"/>
Documenter la capture / collecte des données	<input type="text"/>
Documenter les traitements des données	<input type="text"/>
Documenter l'organisation et nommage des fichiers	<input type="text"/>
Rédiger un Data Management Plan	<input type="text"/>
Mettre en place un suivi des modifications des fichiers	<input type="text"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="text"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="text"/>
Choisir des formats de fichier ouverts	<input type="text"/>

**Pour quelle(s) raison(s) ne réalisez-vous pas ces activités?**

\*

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter le contexte du projet))

----- ou Scenario 2 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter la capture / collecte des données))

----- ou Scenario 3 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter les traitements des données))

----- ou Scenario 4 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter l'organisation et nommage des fichiers))

----- ou Scenario 5 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Rédiger un Data Management Plan))

----- ou Scenario 6 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Mettre en place un suivi des modifications des fichiers))

----- ou Scenario 7 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Mettre en place un système pour vérifier l'intégrité des fichiers))

----- ou Scenario 8 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Choisir un espace de stockage adapté et attribuer des droits))

----- ou Scenario 9 -----

La réponse était 'Non' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Choisir des formats de fichier ouverts))

	C'est inutile	Ne sais pas comment faire	Ne sais pas de quoi il s'agit	Autre
Documenter le contexte du projet	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter la capture / collecte des données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter les traitements des données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter l'organisation et nommage des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Rédiger un Data Management Plan	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mettre en place un suivi des modifications des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Choisir des formats de fichier ouverts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '9 [plandocno]' (Pour quelle(s) raison(s) ne réalisez-vous pas ces activités? )

	<b>Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?</b>
<b>Documenter le contexte du projet</b>	
<b>Documenter la capture / collecte des données</b>	
<b>Documenter les traitements des données</b>	
<b>Documenter l'organisation et nommage des fichiers</b>	
<b>Rédiger un Data Management Plan</b>	
<b>Mettre en place un suivi des modifications des fichiers</b>	
<b>Mettre en place un système pour vérifier l'intégrité des fichiers</b>	
<b>Choisir un espace de stockage adapté et attribuer des droits</b>	
<b>Choisir des formats de fichier ouverts</b>	



**Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise ?**

\*

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter le contexte du projet))

----- ou Scenario 2 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter la capture / collecte des données))

----- ou Scenario 3 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter les traitements des données))

----- ou Scenario 4 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Documenter l'organisation et nommage des fichiers))

----- ou Scenario 5 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Rédiger un Data Management Plan))

----- ou Scenario 6 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Mettre en place un suivi des modifications des fichiers))

----- ou Scenario 7 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Mettre en place un système pour vérifier l'intégrité des fichiers))

----- ou Scenario 8 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Choisir un espace de stockage adapté et attribuer des droits))

----- ou Scenario 9 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '6 [plandocwhat]' (Parmi ces activités de début de projet (avant la récolte des données), lesquelles réalisez-vous ? (Choisir des formats de fichier ouverts))

	Investigateur.trice principal.e (PI)	Data manager	Chargé.e de recherche	Auxiliaire	Autre	Service de soutien Unisanté	Service de soutien externe	Autre
Documenter le contexte du projet	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter la capture / collecte des données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter les traitements des données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter l'organisation et nommage des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Rédiger un Data Management Plan	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mettre en place un suivi des modifications des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Choisir des formats de fichier ouverts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous dire qui d'autre réalise cette activité ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 2 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 3 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 4 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 5 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 6 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 7 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 8 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 9 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 10 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 11 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 12 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 13 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 14 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 15 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 16 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 17 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )  
----- ou Scenario 18 -----

La réponse n'était PAS " à la question '11 [plandocsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

	Pouvez-vous dire qui d'autre réalise cette activité ?
Documenter le contexte du projet	<input type="text"/>
Documenter la capture / collecte des données	<input type="text"/>
Documenter les traitements des données	<input type="text"/>
Documenter l'organisation et nommage des fichiers	<input type="text"/>
Rédiger un Data Management Plan	<input type="text"/>
Mettre en place un suivi des modifications des fichiers	<input type="text"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="text"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="text"/>
Choisir des formats de fichier ouverts	<input type="text"/>

### Comment réalisez-vous ces activités ?

Répondre à cette question seulement si les conditions suivantes sont réunies :

((plandocwhat\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1")) or ((plandocwhat\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1")) or ((plandocwhat\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1")) or ((plandocwhat\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1")) or ((plandocwhat\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1")) or ((plandocwhat\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1")) or ((plandocwhat\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1")) or ((plandocwhat\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1")) or ((plandocwhat\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1359/qid/32406) == "A1"))

	Fichier texte (Word, txt...)	Fichier tabulaire (Excel, csv...)	Métadonnées XML, JSON ou RDF	Autre
Documenter le contexte du projet	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter la capture / collecte des données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter les traitements des données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Documenter l'organisation et nommage des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Rédiger un Data Management Plan	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mettre en place un suivi des modifications des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Choisir des formats de fichier ouverts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

- Fichier texte : il peut s'agir du protocole, du data management plan, de la feuille d'information aux participant.e.s...
- Fichier tabulaire : il peut s'agir du codebook, d'un tableau, d'une base de données...
- Métadonnées : il s'agit de métadonnées structurées, interprétable par les machines

**Pouvez-vous indiquer comment vous réalisez ces tâches ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '13 [plandochow]' (Comment réalisez-vous ces activités ? )

	Pouvez-vous indiquer comment vous réalisez ces tâches ?
Documenter le contexte du projet	<input type="text"/>
Documenter la capture / collecte des données	<input type="text"/>
Documenter les traitements des données	<input type="text"/>
Documenter l'organisation et nommage des fichiers	<input type="text"/>
Rédiger un Data Management Plan	<input type="text"/>
Mettre en place un suivi des modifications des fichiers	<input type="text"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="text"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="text"/>
Choisir des formats de fichier ouverts	<input type="text"/>

Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ?

\*

Choisissez la réponse appropriée pour chaque élément :

	Oui	Non
Documenter le contexte du projet	<input type="radio"/>	<input type="radio"/>
Documenter la capture / collecte des données	<input type="radio"/>	<input type="radio"/>
Documenter les traitements des données	<input type="radio"/>	<input type="radio"/>
Documenter l'organisation et nommage des fichiers	<input type="radio"/>	<input type="radio"/>
Rédiger un Data Management Plan	<input type="radio"/>	<input type="radio"/>
Mettre en place un suivi des modifications des fichiers	<input type="radio"/>	<input type="radio"/>
Mettre en place un système pour vérifier l'intégrité des fichiers	<input type="radio"/>	<input type="radio"/>
Choisir un espace de stockage adapté et attribuer des droits	<input type="radio"/>	<input type="radio"/>
Choisir des formats de fichier ouverts	<input type="radio"/>	<input type="radio"/>

### Quel type de soutien préférez-vous ? \*

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Documenter le contexte du projet))  
----- ou Scenario 2 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Documenter la capture / collecte des données))

----- ou Scenario 3 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Documenter les traitements des données))

----- ou Scenario 4 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Documenter l'organisation et nommage des fichiers))

----- ou Scenario 5 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Rédiger un Data Management Plan))

----- ou Scenario 6 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Mettre en place un suivi des modifications des fichiers))

----- ou Scenario 7 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Mettre en place un système pour vérifier l'intégrité des fichiers))

----- ou Scenario 8 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Choisir un espace de stockage adapté et attribuer des droits))

----- ou Scenario 9 -----

La réponse était 'Oui' à la question '15 [plansupport]' (Souhaitez-vous un soutien / une aide pour réaliser cette tâche ? (Choisir des formats de fichier ouverts))

❗ Cochez la ou les réponses

Veuillez choisir toutes les réponses qui conviennent :

- ☐ Formation
- ☐ Aide théorique (guides, manuels...)
- ☐ Soutien pratique (quelqu'un pour intervenir quand j'en ai besoin)
- ☐ Outil ou logiciel
- ☐ Une personne ressource pour répondre à mes questions
- ☐ Une personne pour réaliser cette tâche à ma place

☐ Autre:

### Avez-vous d'autres remarques ou commentaires concernant les activités liées à la planification ?

Veuillez écrire votre réponse ici :

## Création et collecte

L'étape de création et de collecte des données contient généralement les tâches telles que :

- Création des moyens de collecte (questionnaire, base de données...)
- Collecte / acquisition de données
- Vérification périodique de la qualité des données pendant la collecte (monitoring)
- Création des premières métadonnées descriptives

Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? \*

Répondez à cette question seulement si les conditions suivantes sont réunies :

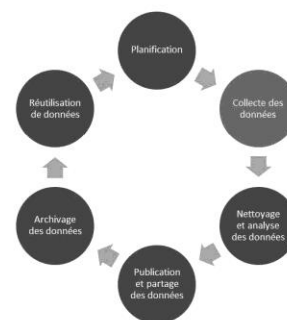
----- Scenario 1 -----

La réponse était 'Analyse de données anonymisées (stricte application) (hors LRH)' ou 'Recherche sur des données agrégées (non personnelles) (hors LRH)' ou 'Etude de démarche qualité (hors LRH)' ou 'Enquêtes anonymes à la source (hors LRH)' ou 'Enquêtes d'opinion (sans données médicales) (hors LRH)' ou 'Rapport de série de cas (<5) (hors LRH)' ou 'Revue de la littérature (hors LRH)' à la question '4 [persoinfodesign]' (Sur quel(s) type(s) de projet / design d'étude travaillez-vous ? Si vous remplissez le questionnaire par rapport à un projet en particulier : quel est le type de ce projet ? )

----- ou Scenario 2 -----

La réponse était 'Etude observationnelle : Etude impliquant des personnes décédées (ORH 5)' ou 'Etude observationnelle : Etude impliquant des embryons (ORH 4)' ou 'Etude observationnelle : Etude de réutilisation données ou échantillons (ORH 3)' ou 'Etude observationnelle : Etude prospective impliquant des personnes (ORH 2)' ou 'Essai clinique (OClin ou OClin-Dim)' à la question '4 [persoinfodesign]' (Sur quel(s) type(s) de projet / design d'étude travaillez-vous ? Si vous remplissez le questionnaire par rapport à un projet en particulier : quel est le type de ce projet ? )

Choisissez la réponse appropriée pour chaque élément :



	Oui, par moi	Oui, par quelqu'un d'autre	Non
Ré-utiliser / dupliquer un questionnaire existant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utiliser le codage	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluer le risque de réidentification	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Pré-tester le moyen de collecte en condition réelle	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Création d'un codebook / dictionnaire de variables	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- **Ré-utiliser / dupliquer un questionnaire existant** : par exemple pour une étude longitudinale, ou une étude similaire, si toutes ou une partie des variables collectées sont identiques, il est intéressant de dupliquer plutôt que de recréer la variable.
- **Utiliser des valeurs agrégées plutôt que réelles** : par exemple, choisir des catégories d'âges (30-39 ans) plutôt que l'âge réel (31 ans)
- **Utiliser le codage** : intégrer directement le codage dans le moyen de collecte. Par exemple, utiliser "1", "2" et "3" pour la nationalité plutôt que "Suisse", "Européenne", "Autre".
- **Utiliser des ontologies ou vocabulaires contrôlés** : Intégrer des termes validés par la discipline au moyen de collecte, plutôt que de créer une liste par soi-même. Par exemple SNOMED, UMLS (Unified Medical Language System) ou Gene Ontology
- **Evaluer le risque de réidentification** : En prenant en compte les variables collectées et l'échantillon prévu, il est possible d'évaluer si une personne sera réidentifiable ou non. Par exemple : je fais un projet sur le personnel d'Unisanté, que je demande le genre, le département, le secteur et si la personne porte des lunettes. Dans certains secteurs, il n'y a qu'un ou 2 hommes. En croisant ces données à l'annuaire sur Allegro, il est très facile de réidentifier quelqu'un, comme les photos y sont présentes. Le risque de réidentification est donc haut. Je peux le réduire en supprimant la variable "secteur", si celle-ci n'est pas utile pour mes résultats finaux.
- **Pré-tester le moyen de collecte en condition réelle** : tester le moyen de collecte avec une fraction de l'échantillon ciblé permet de détecter d'éventuelles erreurs ou incompréhensions
- **Création d'un codebook / dictionnaire de variables** : création d'un document permettant de décrire quelles variables sont collectées, de quel type elles sont (liste de sélection, question ouverte...), le label de chaque variable, etc.
- **Trouver et corriger les valeurs manquantes (missing values)**
- **Trouver et corriger les données aberrantes** : par exemple, pour une étude sur les EMS (65 ans et plus), un âge de "28" ans sera considéré comme aberrant (*Out-of-range, outliers*).
- **Trouver et corriger les valeurs inconsistantes** : une donnée inconsistante sera par exemple dans une étude sur le revenu, une personne indiquant dans une variable qu'elle n'a eu aucun revenu imposable en 2021, mais qui indique dans une seconde variable avoir déclaré 40'000 CHF aux impôts pour l'année 2021. (*inconsistent values*)
- **Création de nouvelles variables** : par exemple si l'indice de masse corporelle n'a pas été demandé, il est possible de générer cette variable à partir d'autres informations collectées

**Pour quelle(s) raison(s) effectuez-vous ces tâches ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenarior 1 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Ré-utiliser / dupliquer un questionnaire existant))

----- ou Scenarior 2 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser des valeurs agrégées plutôt que réelles))

----- ou Scenarior 3 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser le codage))

----- ou Scenarior 4 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser des ontologies ou vocabulaires contrôlés))

----- ou Scenarior 5 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Evaluer le risque de réidentification))

----- ou Scenarior 6 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Pré-tester le moyen de collecte en condition réelle))

----- ou Scenarior 7 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Création d'un codebook / dictionnaire de variables))

----- ou Scenarior 8 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)))

----- ou Scenarior 9 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)))

----- ou Scenarior 10 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)))

----- ou Scenarior 11 -----

La réponse était 'Oui, par moi' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Création de nouvelles variables PENDANT la collecte (monitoring)))

	C'est obligatoire	Qualité du projet ou des données	Autre
Ré-utiliser / dupliquer un questionnaire existant	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser le codage	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluer le risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Pré-tester le moyen de collecte en condition réelle	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création d'un codebook / dictionnaire de variables	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



**Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 2 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 3 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 4 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 5 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 6 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 7 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 8 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 9 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 10 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)  
----- ou Scenario 11 -----

La réponse n'était PAS " à la question '19 [createactivitieswhy]' (Pour quelle(s) raison(s) effectuez-vous ces tâches ?)

	Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?
Ré-utiliser / dupliquer un questionnaire existant	<input type="text"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="text"/>
Utiliser le codage	<input type="text"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="text"/>
Evaluer le risque de réidentification	<input type="text"/>
Pré-tester le moyen de collecte en condition réelle	<input type="text"/>
Création d'un codebook / dictionnaire de variables	<input type="text"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="text"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="text"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="text"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="text"/>

**Pour quelle raison ne faites-vous pas ces tâches ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Ré-utiliser / dupliquer un questionnaire existant))

----- ou Scenario 2 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser des valeurs agrégées plutôt que réelles))

----- ou Scenario 3 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser le codage))

----- ou Scenario 4 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser des ontologies ou vocabulaires contrôlés))

----- ou Scenario 5 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Evaluer le risque de réidentification))

----- ou Scenario 6 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Pré-tester le moyen de collecte en condition réelle))

----- ou Scenario 7 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Création d'un codebook / dictionnaire de variables))

----- ou Scenario 8 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)))

----- ou Scenario 9 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)))

----- ou Scenario 10 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)))

----- ou Scenario 11 -----

La réponse était 'Non' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Création de nouvelles variables PENDANT la collecte (monitoring)))

	C'est inutile	Ne sais pas comment faire	Ne sais pas de quoi il s'agit	Autre
Ré-utiliser / dupliquer un questionnaire existant	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser le codage	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluer le risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Pré-tester le moyen de collecte en condition réelle	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création d'un codebook / dictionnaire de variables	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 10 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

----- ou Scenario 11 -----

La réponse n'était PAS " à la question '21 [createactivitywhy]' (Pour quelle raison ne faites-vous pas ces tâches ? )

	Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?
Ré-utiliser / dupliquer un questionnaire existant	<input type="text"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="text"/>
Utiliser le codage	<input type="text"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="text"/>
Evaluer le risque de réidentification	<input type="text"/>
Pré-tester le moyen de collecte en condition réelle	<input type="text"/>
Création d'un codebook / dictionnaire de variables	<input type="text"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="text"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="text"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="text"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="text"/>

**Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Ré-utiliser / dupliquer un questionnaire existant))

----- ou Scenario 2 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser des valeurs agrégées plutôt que réelles))

----- ou Scenario 3 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser le codage))

----- ou Scenario 4 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Utiliser des ontologies ou vocabulaires contrôlés))

----- ou Scenario 5 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Evaluer le risque de réidentification))

----- ou Scenario 6 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Pré-tester le moyen de collecte en condition réelle))

----- ou Scenario 7 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Création d'un codebook / dictionnaire de variables))

----- ou Scenario 8 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)))

----- ou Scenario 9 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)))

----- ou Scenario 10 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)))

----- ou Scenario 11 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '18 [createactivities]' (Parmi ces tâches, lesquelles effectuez-vous lors de la création de vos moyens de collecte ou pendant la collecte ? (Création de nouvelles variables PENDANT la collecte (monitoring)))

	Investigateur.trice principal.e (PI)	Data manager	Chargé.e de recherche	Auxiliaire	Autre	Service de soutien à Unisanté	Service de soutien externe	Autre
Ré-utiliser / dupliquer un questionnaire existant	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser le codage	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluer le risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Pré-tester le moyen de collecte en condition réelle	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création d'un codebook / dictionnaire de variables	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous dire qui d'autre réalise cette activité ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 10 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 11 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 12 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 13 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 14 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 15 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 16 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 17 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 18 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 19 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 20 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 21 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 22 -----

La réponse n'était PAS " à la question '23 [createsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

	Pouvez-vous dire qui d'autre réalise cette activité ?
Ré-utiliser / dupliquer un questionnaire existant	<input type="text"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="text"/>
Utiliser le codage	<input type="text"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="text"/>
Evaluer le risque de réidentification	<input type="text"/>
Pré-tester le moyen de collecte en condition réelle	<input type="text"/>
Création d'un codebook / dictionnaire de variables	<input type="text"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="text"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="text"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="text"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="text"/>

# Comment réalisez-vous ces tâches ?

Répondre à cette question seulement si les conditions suivantes sont réunies :  
 ((createactivities\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1")) or ((createactivities\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gjid/1361/qjd/32438) == "A1"))

	Redcap (fonctionnalité intégrée)	Limesurvey (fonctionnalité intégrée)	Fichier tabulaire (Excel, csv...)	Logiciel spécialisé	Autre
Ré-utiliser / dupliquer un questionnaire existant	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser le codage	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluer le risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Pré-tester le moyen de collecte en condition réelle	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création d'un codebook / dictionnaire de variables	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

### Pouvez-vous dire comment ?

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 10 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 11 -----

La réponse n'était PAS " à la question '25 [createactivityhow]' (Comment réalisez-vous ces tâches ? )

	Comment réalisez-vous cette activité ?
Ré-utiliser / dupliquer un questionnaire existant	<input type="text"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="text"/>
Utiliser le codage	<input type="text"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="text"/>
Evaluer le risque de réidentification	<input type="text"/>
Pré-tester le moyen de collecte en condition réelle	<input type="text"/>
Création d'un codebook / dictionnaire de variables	<input type="text"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="text"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="text"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="text"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="text"/>

### Est-ce que vous documentez ces activités ?

Indiquez en commentaire la façon dont vous documentez le plus souvent.

Exemple : fichier texte explicatif, protocole, fichier Excel, codebook, métadonnées en XML, JSON ou RDF.

Répondre à cette question seulement si les conditions suivantes sont réunies :

((createactivities\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1")) or ((createactivities\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1"))

⚠ Veuillez sélectionner une réponse ci-dessous

Veuillez sélectionner une seule des propositions suivantes :

- ☐ Oui  
☐ Non

Faites le commentaire de votre choix ici :

### Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ?

\*

Répondre à cette question seulement si les conditions suivantes sont réunies :

((persoinfodesign\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ012.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y")) or ((persoinfodesign\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y"))

Choisissez la réponse appropriée pour chaque élément :

	Oui	Non
Ré-utiliser / dupliquer un questionnaire existant	<input type="radio"/>	<input type="radio"/>
Utiliser des valeurs agrégées plutôt que réelles	<input type="radio"/>	<input type="radio"/>
Utiliser le codage	<input type="radio"/>	<input type="radio"/>
Utiliser des ontologies ou vocabulaires contrôlés	<input type="radio"/>	<input type="radio"/>
Evaluer le risque de réidentification	<input type="radio"/>	<input type="radio"/>
Pré-tester le moyen de collecte en condition réelle	<input type="radio"/>	<input type="radio"/>
Création d'un codebook / dictionnaire de variables	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)	<input type="radio"/>	<input type="radio"/>
Création de nouvelles variables PENDANT la collecte (monitoring)	<input type="radio"/>	<input type="radio"/>



### Quel type de soutien préféreriez-vous ? \*

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Ré-utiliser / dupliquer un questionnaire existant))

----- ou Scenario 2 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Utiliser des valeurs agrégées plutôt que réelles))

----- ou Scenario 3 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Utiliser le codage))

----- ou Scenario 4 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Utiliser des ontologies ou vocabulaires contrôlés))

----- ou Scenario 5 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Evaluer le risque de réidentification))

----- ou Scenario 6 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Pré-tester le moyen de collecte en condition réelle))

----- ou Scenario 7 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Création d'un codebook / dictionnaire de variables))

----- ou Scenario 8 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Trouver et corriger les valeurs manquantes PENDANT la collecte (monitoring)))

----- ou Scenario 9 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Trouver et corriger les données aberrantes PENDANT la collecte (monitoring)))

----- ou Scenario 10 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Trouver et corriger les valeurs inconsistantes PENDANT la collecte (monitoring)))

----- ou Scenario 11 -----

La réponse était 'Oui' à la question '28 [createsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser cette tâche ? (Création de nouvelles variables PENDANT la collecte (monitoring)))

❗ Cochez la ou les réponses

Veuillez choisir toutes les réponses qui conviennent :

- ☐ Formation
- ☐ Aide théorique (guides, manuels...)
- ☐ Soutien pratique (quelqu'un pour intervenir quand j'en ai besoin)
- ☐ Outil ou logiciel
- ☐ Une personne ressource pour répondre à mes questions
- ☐ Une personne pour réaliser cette tâche à ma place

☐ Autre:

### Avez-vous d'autres remarques ou commentaires concernant les activités liées à la création des moyens de collecte ou à la collecte des données?

Veuillez écrire votre réponse ici :

## Nettoyage des données et analyse

Le nettoyage des données et l'analyse des données sont le cœur de votre recherche.

Afin d'avoir des données de qualité, vous effectuez des contrôles de qualité sur vos données, appliquez un codage et faites du nettoyage. Vous pouvez également compléter ou enrichir vos données.

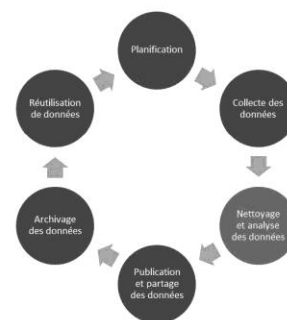
De nouvelles métadonnées se créent également à ce moment pour documenter ces traitements

**Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

((persoinfo\_design\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ012.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y")) or ((persoinfo\_design\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y" or persoinfo\_design\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/qid/1357/qid/32396) == "Y"))

Choisissez la réponse appropriée pour chaque élément :



	Oui, par moi	Oui, par quelqu'un d'autre	Non
Comparaison des données et du codebook	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs manquantes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les données aberrantes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs inconsistantes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Création de nouvelles variables	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Codage ou agrégation de valeurs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Enrichissement de données (web sémantique)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Appariement de données	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluer et diminuer le risque de réidentification	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Séparer les données identifiantes des autres données	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Visualisation de données	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- **Comparaison des données et du codebook** : permet de vérifier si toutes les variables collectées sont bien présentes et totalement documentées dans le codebook
- **Trouver et corriger les valeurs manquantes** (*missing values*)
- **Trouver et corriger les données aberrantes** : par exemple, pour une étude sur les EMS (65 ans et plus), un âge de "28" ans sera considéré comme aberrant (*Out-of-range, outliers*).
- **Trouver et corriger les valeurs inconsistantes** : une donnée inconsistante sera par exemple dans une étude sur le revenu, une personne indiquant dans une variable qu'elle n'a eu aucun revenu imposable en 2021, mais qui indique dans une seconde variable avoir déclaré 40'000 CHF aux impôts pour l'année 2021. (*inconsistent values*)
- **Création de nouvelles variables** : par exemple si l'indice de masse corporelle n'a pas été demandé, il est possible de générer cette variable à partir d'autres informations collectées
- **Codage ou agrégation de valeurs** : transformer les informations textuelles collectées en code (femme = 1, homme = 2) ou agréger des valeurs (transformer les âges en catégorie d'âges par exemple)
- **Enrichissement de données** : Aussi appelé *Linked data* ou *web sémantique* selon la technologie utilisée. L'enrichissement de données consiste à transformer les données pour les structurer à l'aide de référentiels ou d'ontologies. Cette enrichissement consiste généralement à faire des liens vers des références standardisées. Exemple : Wikidata, MeSH
- **Appariement de données** : fusionner les données collectées avec d'autres données existantes
- **Evaluer et diminuer le risque de réidentification** : En prenant en compte les variables collectées et l'échantillon prévu, il est possible d'évaluer si une personne sera réidentifiable ou non. Par exemple : je fais un projet sur le personnel d'Unisanté, que je demande le genre, le département, le secteur et si la personne porte des lunettes. Dans certains secteurs, il n'y a qu'un ou 2 hommes. En croisant ces données à l'annuaire sur Allegro, il est très facile de réidentifier quelqu'un, comme les photos y sont présentes. Le risque de réidentification est donc haut. Je peux le réduire en supprimant la variable "secteur", si celle-ci n'est pas utile pour mes résultats finaux.
- **Séparer les données identifiantes des autres données** : si des noms, adresses, numéros de téléphone ont été collectés en même temps que les autres données, il est d'usage de les séparer et de les stocker dans un fichier séparé ou de les supprimer si elles ne sont plus nécessaires.
- **Sauvegarder une copie des données brutes (non modifiées) dans leur format original** : Il est recommandé de faire une copie des données et d'effectuer le nettoyage sur cette copie, afin de garder une sauvegarde des données brutes
- **Visualisation de données** : créer des visualisations (diagrammes, histogrammes, cartes...) pour permettre d'avoir une première idée du contenu des données (avant analyse)

**Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Comparaison des données et du codebook))

----- ou Scenario 2 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 3 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les données aberrantes))

----- ou Scenario 4 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 5 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Création de nouvelles variables))

----- ou Scenario 6 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Codage ou agrégation de valeurs))

----- ou Scenario 7 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Enrichissement de données (web sémantique)))

----- ou Scenario 8 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Appariement de données))

----- ou Scenario 9 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Evaluer et diminuer le risque de réidentification))

----- ou Scenario 10 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Séparer les données identifiantes des autres données))

----- ou Scenario 11 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Sauvegarder une copie des données brutes (non modifiées) dans leur format original))

----- ou Scenario 12 -----

La réponse était 'Oui, par moi' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Visualisation de données))

	C'est obligatoire	Qualité du projet ou des données	Autre
Comparaison des données et du codebook	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de nouvelles variables	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Codage ou agrégation de valeurs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Enrichissement de données (web sémantique)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Appariement de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluer et diminuer le risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Séparer les données identifiantes des autres données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Visualisation de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 2 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 3 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 4 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 5 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 6 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 7 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 8 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 9 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 10 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 11 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )  
----- ou Scenario 12 -----

La réponse n'était PAS " à la question '32 [processwhy]' (Pour quelle(s) raison(s) effectuez-vous ces traitements ou vérifications ? )

	Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?
Comparaison des données et du codebook	<input type="text"/>
Trouver et corriger les valeurs manquantes	<input type="text"/>
Trouver et corriger les données aberrantes	<input type="text"/>
Trouver et corriger les valeurs inconsistantes	<input type="text"/>
Création de nouvelles variables	<input type="text"/>
Codage ou agrégation de valeurs	<input type="text"/>
Enrichissement de données (web sémantique)	<input type="text"/>
Appariement de données	<input type="text"/>
Evaluer et diminuer le risque de réidentification	<input type="text"/>
Séparer les données identifiantes des autres données	<input type="text"/>
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	<input type="text"/>
Visualisation de données	<input type="text"/>

**Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Comparaison des données et du codebook))

----- ou Scenario 2 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 3 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les données aberrantes))

----- ou Scenario 4 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 5 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Création de nouvelles variables))

----- ou Scenario 6 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Codage ou agrégation de valeurs))

----- ou Scenario 7 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Enrichissement de données (web sémantique)))

----- ou Scenario 8 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Appariement de données))

----- ou Scenario 9 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Evaluer et diminuer le risque de réidentification))

----- ou Scenario 10 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Séparer les données identifiantes des autres données))

----- ou Scenario 11 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Sauvegarder une copie des données brutes (non modifiées) dans leur format original))

----- ou Scenario 12 -----

La réponse était 'Non' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Visualisation de données))

	C'est inutile	Ne sais pas comment faire	Ne sais pas de quoi il s'agit	Autre
Comparaison des données et du codebook	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de nouvelles variables	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Codage ou agrégation de valeurs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Enrichissement de données (web sémantique)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Appariement de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluer et diminuer le risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Séparer les données identifiantes des autres données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Visualisation de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 10 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 11 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

----- ou Scenario 12 -----

La réponse n'était PAS " à la question '34 [processno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

	Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?
Comparaison des données et du codebook	<input type="text"/>
Trouver et corriger les valeurs manquantes	<input type="text"/>
Trouver et corriger les données aberrantes	<input type="text"/>
Trouver et corriger les valeurs inconsistantes	<input type="text"/>
Création de nouvelles variables	<input type="text"/>
Codage ou agrégation de valeurs	<input type="text"/>
Enrichissement de données (web sémantique)	<input type="text"/>
Appariement de données	<input type="text"/>
Evaluer et diminuer le risque de réidentification	<input type="text"/>
Séparer les données identifiantes des autres données	<input type="text"/>
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	<input type="text"/>
Visualisation de données	<input type="text"/>

**Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Comparaison des données et du codebook))

----- ou Scenario 2 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 3 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les données aberrantes))

----- ou Scenario 4 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 5 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Création de nouvelles variables))

----- ou Scenario 6 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Codage ou agrégation de valeurs))

----- ou Scenario 7 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Enrichissement de données (web sémantique)))

----- ou Scenario 8 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Appariement de données))

----- ou Scenario 9 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Evaluer et diminuer le risque de réidentification))

----- ou Scenario 10 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Séparer les données identifiantes des autres données))

----- ou Scenario 11 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Sauvegarder une copie des données brutes (non modifiées) dans leur format original))

----- ou Scenario 12 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '31 [processactivities]' (Parmi ces vérifications et traitements, lesquels appliquez-vous à vos données (avant ou pendant l'analyse) ? (Visualisation de données))

	Investigateur.trice principal.e (PI)	Data manager	Chargé.e de recherche	Auxiliaire	Autre	Service de soutien à Unisanté	Service de soutien externe	Autre
Comparaison des données et du codebook	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de nouvelles variables	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Codage ou agrégation de valeurs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Enrichissement de données (web sémantique)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Appariement de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluer et diminuer le risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Séparer les données identifiantes des autres données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Visualisation de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Répondre à cette question seulement si les conditions suivantes sont réunies :

La réponse n'était PAS " à la question '36 [processsomeone2] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS " à la question '36 [processsomeone2] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS " à la question '36 [processsomeone2] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS " à la question '36 [processsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS "à la question '36 [processsomeone2] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS "à la question '36 [processsomeone2] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise?)

La réponse n'était PAS à la question 36 [processsomeonez] (vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise ?)  
 ----- ou Scenario 8 -----

----- ou Scenario 9 -----

----- ou Scenario 10 -----

— ou Scenario 11 —

----- ou Scenario 12 -----

\_\_\_\_\_ ou Scenario 13 \_\_\_\_\_

----- ou Scenario 14 -----

----- ou Scenario 15 -----

----- ou Scenario 16 -----

La réponse n'était PAS "à la question '36 [processsomeone?] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise?)

La réponse n'était PAS " à la question '36 [processsomeone?] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise?)

La réponse n'était PAS "à la question '36 [processsomeone2]" (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS " à la question '36 [processsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS " à la question '36 [processsomeone2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS " à la question '36 [processsomeone2] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS " à la question '36 [processsomeone2] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

La réponse n'était PAS " à la question '36 [processsomeone2] (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

	Pouvez-vous dire qui d'autre réalise cette activité ?
--	---

Comparaison des données et du codebook	
Trouver et corriger les valeurs manquantes	
Trouver et corriger les données aberrantes	
Trouver et corriger les valeurs inconsistantes	
Création de nouvelles variables	
Codage ou agrégation de valeurs	
Enrichissement de données (web sémantique)	
Appariement de données	
Evaluer et diminuer le risque de réidentification	
Séparer les données identifiantes des autres données	
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	
Visualisation de données	



### Comment réalisez-vous ces tâches ? \*

Répondre à cette question seulement si les conditions suivantes sont réunies :

((processactivities\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2")) or ((processactivities\_SQ012.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A1" or processactivities\_SQ012.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1362/qid/32409) == "A2"))

	Logiciel statistique (Stata, SPSS...)	Script (R, Python...)	Logiciel spécialisé	Autre
Comparaison des données et du codebook	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de nouvelles variables	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Codage ou agrégation de valeurs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Enrichissement de données (web sémantique)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Appariement de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluer et diminuer le risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Séparer les données identifiantes des autres données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Visualisation de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous dire comment ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 10 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 11 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

----- ou Scenario 12 -----

La réponse n'était PAS " à la question '38 [processhow]' (Comment réalisez-vous ces tâches ? )

	<b>Veillez préciser comment vous réalisez cette tâche</b>
<b>Comparaison des données et du codebook</b>	<input type="text"/>
<b>Trouver et corriger les valeurs manquantes</b>	<input type="text"/>
<b>Trouver et corriger les données aberrantes</b>	<input type="text"/>
<b>Trouver et corriger les valeurs inconsistantes</b>	<input type="text"/>
<b>Création de nouvelles variables</b>	<input type="text"/>
<b>Codage ou agrégation de valeurs</b>	<input type="text"/>
<b>Enrichissement de données (web sémantique)</b>	<input type="text"/>
<b>Appariement de données</b>	<input type="text"/>
<b>Evaluer et diminuer le risque de réidentification</b>	<input type="text"/>
<b>Séparer les données identifiantes des autres données</b>	<input type="text"/>
<b>Sauvegarder une copie des données brutes (non modifiées) dans leur format original</b>	<input type="text"/>
<b>Visualisation de données</b>	<input type="text"/>

### Est-ce que vous documentez ces activités ?

Indiquez en commentaire la façon dont vous documentez le plus souvent.

Exemple : commentaire dans le script, fichier texte explicatif, protocole, fichier Excel, codebook, métadonnées en XML, JSON ou RDF.

Répondez à cette question seulement si les conditions suivantes sont réunies :

((createactivities\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1") or ((createactivities\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1361/qid/32438) == "A1"))

⚠ Veuillez sélectionner une réponse ci-dessous

Veuillez sélectionner une seule des propositions suivantes :

☐ Oui

☐ Non

Faites le commentaire de votre choix ici :

### Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ?

\*

Répondez à cette question seulement si les conditions suivantes sont réunies :

((persoinfodesign\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ012.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y")) or ((persoinfodesign\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y"))

Choisissez la réponse appropriée pour chaque élément :

	Oui	Non
Comparaison des données et du codebook	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs manquantes	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les données aberrantes	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs inconsistantes	<input type="radio"/>	<input type="radio"/>
Création de nouvelles variables	<input type="radio"/>	<input type="radio"/>
Codage ou agrégation de valeurs	<input type="radio"/>	<input type="radio"/>
Enrichissement de données (web sémantique)	<input type="radio"/>	<input type="radio"/>
Appariement de données	<input type="radio"/>	<input type="radio"/>
Evaluer et diminuer le risque de réidentification	<input type="radio"/>	<input type="radio"/>
Séparer les données identifiantes des autres données	<input type="radio"/>	<input type="radio"/>
Sauvegarder une copie des données brutes (non modifiées) dans leur format original	<input type="radio"/>	<input type="radio"/>
Visualisation de données	<input type="radio"/>	<input type="radio"/>

**Quel type de soutien préférez-vous ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Comparaison des données et du codebook))

----- ou Scenario 2 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 3 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Trouver et corriger les données aberrantes))

----- ou Scenario 4 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 5 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Création de nouvelles variables))

----- ou Scenario 6 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Codage ou agrégation de valeurs))

----- ou Scenario 7 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Enrichissement de données (web sémantique))

----- ou Scenario 8 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Appariement de données))

----- ou Scenario 9 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Evaluer et diminuer le risque de réidentification))

----- ou Scenario 10 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Séparer les données identifiantes des autres données))

----- ou Scenario 11 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Sauvegarder une copie des données brutes (non modifiées) dans leur format original))

----- ou Scenario 12 -----

La réponse était 'Oui' à la question '41 [processsupport]' (Souhaiteriez-vous un soutien / une aide pour la réalisation de cette tâche ? (Visualisation de données))

❗ Cochez la ou les réponses

Veuillez choisir toutes les réponses qui conviennent :

☐ Formation

☐ Aide théorique (guides, manuels...)

☐ Soutien pratique (quelqu'un pour intervenir quand j'en ai besoin)

☐ Outil ou logiciel

☐ Une personne ressource pour répondre à mes questions

☐ Une personne pour réaliser cette tâche à ma place

☐ Autre:

**Avez-vous d'autres remarques ou commentaires concernant la préparation et l'analyse de données ?**

Veuillez écrire votre réponse ici :

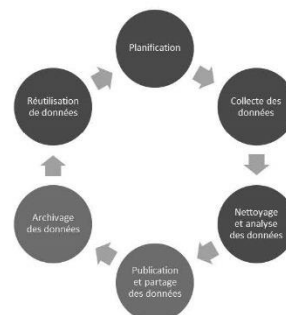
## Partage et archivage

A la fin de votre projet, vous publiez généralement un article, un rapport, une thèse ou une autre publication.

Les instances de financement ou les éditeurs demandent généralement de mettre à disposition les données probantes. Ceci permet de garantir la reproductibilité de la recherche et le bon fonctionnement du *peer-reviewing*.

Ce partage peut se faire en Open Access ou en accès restreint, selon la sensibilité de vos données.

Pour certaines études, les données ne sont pas mises à disposition, mais sont quand même conservées à des fins de reproductibilité et/ou sécurité pour les participant.e.s durant une dizaine d'années. On appelle ceci la préservation.



### Est-ce que vous partagez vos données ?

Si oui, précisez où

- dépôt de données Unisanté, autre dépôt de données, Zenodo, Maelstrom, data warehouse du CHUV...
- Matériel supplémentaire de l'article, site web dédié...
- Autre...

Si non, précisez pourquoi

- Interdiction de partager
- Ne sais pas où partager
- Ne veut pas partager

\*

❗ Veuillez sélectionner une réponse ci-dessous

Veuillez sélectionner une seule des propositions suivantes :

- ☐ Oui
- ☐ Non

Faites le commentaire de votre choix ici :

**Est-ce que vous archivez les données après le projet ?**

L'archivage peut être réalisé pour des raisons légales, pour garantir la reproductibilité de la recherche ou encore pour pouvoir assurer la sécurité des participant.e.s

**Si oui, précisez où**

- Dépôt de données
- Serveur Unisanté (L ou filearc)
- Je les laisse là où elles sont (serveur commun ou autre)
- Autre

**Si non, précisez pourquoi**

- Les données doivent être supprimées après analyse
- Les données sont gérées par quelqu'un d'autre

❗ Veuillez sélectionner une réponse ci-dessous

Veuillez sélectionner une seule des propositions suivantes :

- ☐ Oui
- ☐ Non

Faites le commentaire de votre choix ici :

**Combien de temps dure l'archivage ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

La réponse était 'Oui' à la question '45 [preserve]' (Est-ce que vous archivez les données après le projet ? L'archivage peut être réalisé pour des raisons légales, pour garantir la reproductibilité de la recherche ou encore pour pouvoir assurer la sécurité des participant.e.s Si oui, précisez où Dépôt de données Serveur Unisanté (L ou filearc) Je les laisse là où elles sont (serveur commun ou autre) Autre Si non, précisez pourquoi Les données doivent être supprimées après analyse Les données sont gérées par quelqu'un d'autre )

❗ Cochez la ou les réponses

Veuillez choisir toutes les réponses qui conviennent :

- ☐ 1 à 5 ans
- ☐ 6 à 10 ans
- ☐ 11 à 15 ans
- ☐ Plus de 16 ans
- ☐ Je ne sais pas

**Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Analyse de données anonymisées (stricte application) (hors LRH)' ou 'Recherche sur des données agrégées (non personnelles) (hors LRH)' ou 'Etude de démarche qualité (hors LRH)' ou 'Enquêtes anonymes à la source (hors LRH)' ou 'Enquêtes d'opinion (sans données médicales) (hors LRH)' ou 'Rapport de série de cas (<5) (hors LRH)' ou 'Revue de la littérature (hors LRH)' ou 'Etude observationnelle : Etude impliquant des personnes décédées (ORH 5)' ou 'Etude observationnelle : Etude impliquant des embryons (ORH 4)' ou 'Etude observationnelle : Etude de réutilisation données ou échantillons (ORH 3)' ou 'Etude observationnelle : Etude prospective impliquant des personnes (ORH 2)' ou 'Essai clinique (OClin ou OClin-Dim)' à la question '4 [persoinfodesign]' (Sur quel(s) type(s) de projet / design d'étude travaillez-vous ? Si vous remplissez le questionnaire par rapport à un projet en particulier : quel est le type de ce projet ? ) et La réponse était 'Oui' à la question '44 [share]' (Est-ce que vous partagez vos données ? Si oui, précisez où dépôt de données Unisanté, autre dépôt de données, Zenodo, Maelstrom, data warehouse du CHUV... Matériel supplémentaire de l'article, site web dédié... Autre... Si non, précisez pourquoi Interdiction de partager Ne sais pas où partager Ne veut pas partager )

----- ou Scenario 2 -----

La réponse était 'Analyse de données anonymisées (stricte application) (hors LRH)' ou 'Recherche sur des données agrégées (non personnelles) (hors LRH)' ou 'Etude de démarche qualité (hors LRH)' ou 'Enquêtes anonymes à la source (hors LRH)' ou 'Enquêtes d'opinion (sans données médicales) (hors LRH)' ou 'Rapport de série de cas (<5) (hors LRH)' ou 'Revue de la littérature (hors LRH)' ou 'Etude observationnelle : Etude impliquant des personnes décédées (ORH 5)' ou 'Etude observationnelle : Etude impliquant des embryons (ORH 4)' ou 'Etude observationnelle : Etude de réutilisation données ou échantillons (ORH 3)' ou 'Etude observationnelle : Etude prospective impliquant des personnes (ORH 2)' ou 'Essai clinique (OClin ou OClin-Dim)' à la question '4 [persoinfodesign]' (Sur quel(s) type(s) de projet / design d'étude travaillez-vous ? Si vous remplissez le questionnaire par rapport à un projet en particulier : quel est le type de ce projet ? ) et La réponse était 'Oui' à la question '45 [preserve]' (Est-ce que vous archivez les données après le projet ? L'archivage peut être réalisé pour des raisons légales, pour garantir la reproductibilité de la recherche ou encore pour pouvoir assurer la sécurité des participant.e.s Si oui, précisez où Dépôt de données Serveur Unisanté (L ou filearc) Je les laisse là où elles sont (serveur commun ou autre) Autre Si non, précisez pourquoi Les données doivent être supprimées après analyse Les données sont gérées par quelqu'un d'autre )

Choisissez la réponse appropriée pour chaque élément :

	Oui, par moi	Oui, par quelqu'un d'autre	Non
Codage ou agrégation de données	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Comparaison des données et du codebook	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs manquantes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les données aberrantes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs inconsistantes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluation du risque de réidentification	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Transformation du format de fichier	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Anonymisation au sens strict de la loi	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Création de documentation supplémentaire	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- **Codage ou agrégation de données** : transformer les informations textuelles collectées en code (femme = 1, homme = 2) ou agréger des valeurs (transformer les âges en catégorie d'âges par exemple)
- **Comparaison des données et du codebook** : permet de vérifier si toutes les variables collectées sont bien présentes et totalement documentées dans le codebook
- **Trouver et corriger les valeurs manquantes** (missing values)
- **Trouver et corriger les données aberrantes** : par exemple, pour une étude sur les EMS (65 ans et plus), un âge de "28" ans sera considéré comme aberrant (Out-of-range, outliers).
- **Trouver et corriger les valeurs inconsistantes** : une donnée inconsistante sera par exemple dans une étude sur le revenu, une personne indiquant dans une variable qu'elle n'a eu aucun revenu imposable en 2021, mais qui indique dans une seconde variable avoir déclaré 40'000 CHF aux impôts pour l'année 2021. (inconsistent values)
- **Evaluation du risque de réidentification** : En prenant en compte les variables collectées et l'échantillon prévu, il est possible d'évaluer si une personne sera réidentifiable ou non. Par exemple : je fais un projet sur le personnel d'Unisanté, que je demande le genre, le département, le secteur et si la personne porte des lunettes. Dans certains secteurs, il n'y a qu'un ou 2 hommes. En croisant ces données à l'annuaire sur Allegro, il est très facile de réidentifier quelqu'un, comme les photos y sont présentes. Le risque de réidentification est donc haut. Je peux le réduire en supprimant la variable "secteur", si celle-ci n'est pas utile pour mes résultats finaux.
- **Transformation du format de fichier** : modifier le format de fichier pour utiliser un format propice à la préservation ou l'archivage. Par exemple, un fichier Excel se transforme en CSV.
- **Anonymisation au sens strict de la loi** : Faire en sorte que les participant.e.s à l'étude ne puissent pas être réidentifié.e.s par tous les moyens à disposition. Ceci prend en compte l'appariement de données (y compris externe), le décryptage, etc. Des données anonymisées au sens strict ne sont plus considérées comme des données personnelles et peuvent donc être utilisées dans des projets hors-LRH.
- **Création de documentation supplémentaire** : création de fichiers ou de métadonnées spécifiques à l'archivage, qui n'auraient pas été créé.e.s à un autre moment du projet.

**Pour quelle(s) raison(s) effectuez-vous ces modifications ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Codage ou agrégation de données))

----- ou Scenario 2 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Comparaison des données et du codebook))

----- ou Scenario 3 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 4 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les données aberrantes))

----- ou Scenario 5 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 6 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Evaluation du risque de réidentification))

----- ou Scenario 7 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Transformation du format de fichier))

----- ou Scenario 8 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Anonymisation au sens strict de la loi))

----- ou Scenario 9 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Création de documentation supplémentaire))

	C'est obligatoire	Qualité du projet ou des données	Autre
Codage ou agrégation de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comparaison des données et du codebook	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluation du risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Transformation du format de fichier	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Anonymisation au sens strict de la loi	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de documentation supplémentaire	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



**Pour quelle autre raison effectuez-vous ces modifications ?**

\*

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '48 [sharewhy]' (Pour quelle(s) raison(s) effectuez-vous ces modifications ? )

	Pour quelle autre raison effectuez-vous ces modifications ?
Codage ou agrégation de données	<input type="text"/>
Comparaison des données et du codebook	<input type="text"/>
Trouver et corriger les valeurs manquantes	<input type="text"/>
Trouver et corriger les données aberrantes	<input type="text"/>
Trouver et corriger les valeurs inconsistantes	<input type="text"/>
Evaluation du risque de réidentification	<input type="text"/>
Transformation du format de fichier	<input type="text"/>
Anonymisation au sens strict de la loi	<input type="text"/>
Création de documentation supplémentaire	<input type="text"/>

**Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Codage ou agrégation de données))

----- ou Scenario 2 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Comparaison des données et du codebook))

----- ou Scenario 3 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 4 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les données aberrantes))

----- ou Scenario 5 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 6 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Evaluation du risque de réidentification))

----- ou Scenario 7 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Transformation du format de fichier))

----- ou Scenario 8 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Anonymisation au sens strict de la loi))

----- ou Scenario 9 -----

La réponse était 'Non' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Création de documentation supplémentaire))

	C'est inutile	Ne sais pas comment faire	Ne sais pas de quoi il s'agit	Autre
Codage ou agrégation de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comparaison des données et du codebook	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluation du risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Anonymisation au sens strict de la loi	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Transformation du format de fichier	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de documentation supplémentaire	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pour quelle autre raison n'effectuez-vous pas ces modifications ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :  
 ----- Scenario 1 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )  
 ----- ou Scenario 2 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )  
 ----- ou Scenario 3 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )  
 ----- ou Scenario 4 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )  
 ----- ou Scenario 5 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )  
 ----- ou Scenario 6 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )  
 ----- ou Scenario 7 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )  
 ----- ou Scenario 8 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )  
 ----- ou Scenario 9 -----

La réponse n'était PAS " à la question '50 [sharedataprepno]' (Pour quelle(s) raison(s) ne faites-vous pas ces traitements ou vérifications ? )

	Pour quelle autre raison n'effectuez-vous pas ces modifications ?
Codage ou agrégation de données	<input type="text"/>
Comparaison des données et du codebook	<input type="text"/>
Trouver et corriger les valeurs manquantes	<input type="text"/>
Trouver et corriger les données aberrantes	<input type="text"/>
Trouver et corriger les valeurs inconsistantes	<input type="text"/>
Evaluation du risque de réidentification	<input type="text"/>
Transformation du format de fichier	<input type="text"/>
Anonymisation au sens strict de la loi	<input type="text"/>
Création de documentation supplémentaire	<input type="text"/>

**Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise ?**

\*

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Codage ou agrégation de données))

----- ou Scenario 2 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Comparaison des données et du codebook))

----- ou Scenario 3 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 4 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les données aberrantes))

----- ou Scenario 5 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 6 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Evaluation du risque de réidentification))

----- ou Scenario 7 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Transformation du format de fichier))

----- ou Scenario 8 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Anonymisation au sens strict de la loi))

----- ou Scenario 9 -----

La réponse était 'Oui, par quelqu'un d'autre' à la question '47 [sharedataprep] (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Création de documentation supplémentaire))

	Investigateur.trice principal.e (PI)	Data manager	Chargé.e de recherche	Auxiliaire	Autre	Service de soutien à Unisanté	Service de soutien externe	Autre
Codage ou agrégation de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comparaison des données et du codebook	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Évaluation du risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Transformation du format de fichier	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Anonymisation au sens strict de la loi	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de documentation supplémentaire	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Pouvez-vous indiquer qui d'autre effectue ces modifications ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 10 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 11 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 12 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 13 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 14 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 15 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 16 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 17 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

----- ou Scenario 18 -----

La réponse n'était PAS " à la question '52 [shareprepsomebody2]' (Vous avez indiqué ne pas faire cette tâche vous-même, qui la réalise? )

	Pouvez-vous indiquer qui d'autre effectue ces modifications ?
Codage ou agrégation de données	<input type="text"/>
Comparaison des données et du codebook	<input type="text"/>
Trouver et corriger les valeurs manquantes	<input type="text"/>
Trouver et corriger les données aberrantes	<input type="text"/>
Trouver et corriger les valeurs inconsistantes	<input type="text"/>
Evaluation du risque de réidentification	<input type="text"/>
Transformation du format de fichier	<input type="text"/>
Anonymisation au sens strict de la loi	<input type="text"/>
Création de documentation supplémentaire	<input type="text"/>

### Comment réalisez-vous ces modifications ? \*

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Codage ou agrégation de données))

----- ou Scenario 2 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Comparaison des données et du codebook))

----- ou Scenario 3 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 4 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les données aberrantes))

----- ou Scenario 5 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 6 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Evaluation du risque de réidentification))

----- ou Scenario 7 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Transformation du format de fichier))

----- ou Scenario 8 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Anonymisation au sens strict de la loi))

----- ou Scenario 9 -----

La réponse était 'Oui, par moi' à la question '47 [sharedataprep]' (Est-ce que vous apportez une ou plusieurs modifications à votre set de données pour qu'il soit prêt au partage ou à l'archivage ? (Création de documentation supplémentaire))

	Logiciel statistique (Stata, SPSS...)	Script (R, Python...)	Logiciel spécialisé	Autre
Codage ou agrégation de données	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comparaison des données et du codebook	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs manquantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les données aberrantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Trouver et corriger les valeurs inconsistantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluation du risque de réidentification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Anonymisation au sens strict de la loi	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Transformation du format de fichier	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Création de documentation supplémentaire	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Comment réalisez-vous ces modifications ?**

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

----- ou Scenario 2 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

----- ou Scenario 3 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

----- ou Scenario 4 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

----- ou Scenario 5 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

----- ou Scenario 6 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

----- ou Scenario 7 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

----- ou Scenario 8 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

----- ou Scenario 9 -----

La réponse n'était PAS " à la question '54 [sharedataprephow] (Comment réalisez-vous ces modifications ?)

	Veuillez préciser comment
Codage ou agrégation de données	<input type="text"/>
Comparaison des données et du codebook	<input type="text"/>
Trouver et corriger les valeurs manquantes	<input type="text"/>
Trouver et corriger les données aberrantes	<input type="text"/>
Trouver et corriger les valeurs inconsistantes	<input type="text"/>
Evaluation du risque de réidentification	<input type="text"/>
Anonymisation au sens strict de la loi	<input type="text"/>
Transformation du format de fichier	<input type="text"/>
Création de documentation supplémentaire	<input type="text"/>

**Souhaiteriez-vous un soutien / une aide pour réaliser ces tâches ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

((persoinfodesign\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ012.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y") and (share.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1364/qid/32480) == "A1")) or ((persoinfodesign\_SQ001.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ002.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ003.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ004.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ005.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ006.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ007.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ008.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ009.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ010.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ011.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y" or persoinfodesign\_SQ012.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1357/qid/32396) == "Y") and (preserve.NAOK (/index.php/admin/questions/sa/view/surveyid/166551/gid/1364/qid/32481) == "A1"))

Choisissez la réponse appropriée pour chaque élément :

	Oui	Non
Codage ou agrégation de données	<input type="radio"/>	<input type="radio"/>
Comparaison des données et du codebook	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs manquantes	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les données aberrantes	<input type="radio"/>	<input type="radio"/>
Trouver et corriger les valeurs inconsistantes	<input type="radio"/>	<input type="radio"/>
Evaluation du risque de réidentification	<input type="radio"/>	<input type="radio"/>
Anonymisation au sens strict de la loi	<input type="radio"/>	<input type="radio"/>
Transformation du format de fichier	<input type="radio"/>	<input type="radio"/>
Création de documentation supplémentaire	<input type="radio"/>	<input type="radio"/>

**Quel type de soutien préféreriez-vous ? \***

Répondre à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui' à la question '56 [shareprepsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser ces tâches ? (Codage ou agrégation de données))

----- ou Scenario 2 -----

La réponse était 'Oui' à la question '56 [shareprepsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser ces tâches ? (Comparaison des données et du codebook))

----- ou Scenario 3 -----

La réponse était 'Oui' à la question '56 [shareprepsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser ces tâches ? (Trouver et corriger les valeurs manquantes))

----- ou Scenario 4 -----

La réponse était 'Oui' à la question '56 [shareprepsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser ces tâches ? (Trouver et corriger les données aberrantes))

----- ou Scenario 5 -----

La réponse était 'Oui' à la question '56 [shareprepsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser ces tâches ? (Trouver et corriger les valeurs inconsistantes))

----- ou Scenario 6 -----

La réponse était 'Oui' à la question '56 [shareprepsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser ces tâches ? (Evaluation du risque de réidentification))

----- ou Scenario 7 -----

La réponse était 'Oui' à la question '56 [shareprepsupport]' (Souhaiteriez-vous un soutien / une aide pour réaliser ces tâches ? (Anonymisation au sens strict de la loi))

① Cochez la ou les réponses

Veuillez choisir toutes les réponses qui conviennent :

- ☐ Formation
- ☐ Aide théorique (guides, manuels...)
- ☐ Soutien pratique (quelqu'un pour intervenir quand j'en ai besoin)
- ☐ Outil ou logiciel
- ☐ Une personne ressource pour répondre à mes questions
- ☐ Une personne pour réaliser cette tâche à ma place

☐ Autre:



Avez-vous d'autres remarques ou commentaires concernant le partage ou l'archivage de données ?

Veuillez écrire votre réponse ici :

## Fin du questionnaire

Avez-vous des commentaires / remarques / questions concernant la gestion des données de recherche ou le soutien offert à Unisanté ?

Veuillez écrire votre réponse ici :

Seriez-vous ouvert.e à participer aux tests d'outils et conclusions émises par le travail ?

Veuillez sélectionner une seule des propositions suivantes :

- ☐ Oui  
☐ Non

Il s'agirait soit de :

- Participer aux tests d'outils et de méthodes que recommanderaient ce travail
- OU
- Fournir des données pour que l'unité documentation et données réalise les tests

Seriez-vous d'accord de nous accorder une entrevue, afin de connaître plus en détail vos pratiques en gestion des données ?

Veuillez sélectionner une seule des propositions suivantes :

- ☐ Oui  
☐ Non

Merci pour votre participation.

Veuillez laisser votre adresse email afin de nous puissions vous contacter.

Répondez à cette question seulement si les conditions suivantes sont réunies :

----- Scenario 1 -----

La réponse était 'Oui' à la question '61 [enofsurveyinterview]' (Seriez-vous d'accord de nous accorder une entrevue, afin de connaître plus en détail vos pratiques en gestion des données ?)

----- ou Scenario 2 -----

La réponse était 'Oui' à la question '60 [endsurveytest]' (Seriez-vous ouvert.e à participer aux tests d'outils et conclusions émises par le travail ?)

Veuillez écrire votre réponse ici :

### Merci d'avoir pris le temps de répondre à ce questionnaire

Si vous gérez vos données différemment pour un autre type de projet, vous pouvez remplir à nouveau ce questionnaire (<https://survey.unisante.ch/index.php/239985?lang=fr>).

## **Annex 5 : List of question for researchers' interviews**

What data curation or data management activities do you perform (data cleaning, documentation, anonymization, integrity checking...) ?

How do you realize these activities ?

Would you like a support for some of these tasks ?

Do you have some difficulties related to data management or curation ?

Do you see a need of improvement for some activities at Unisanté ?

## **Annex 6 : List of questions for support services interviews**

What data curation or data management activities do you perform in your unit ?

How do you realize these activities ?

How many projects per year do you handle ?

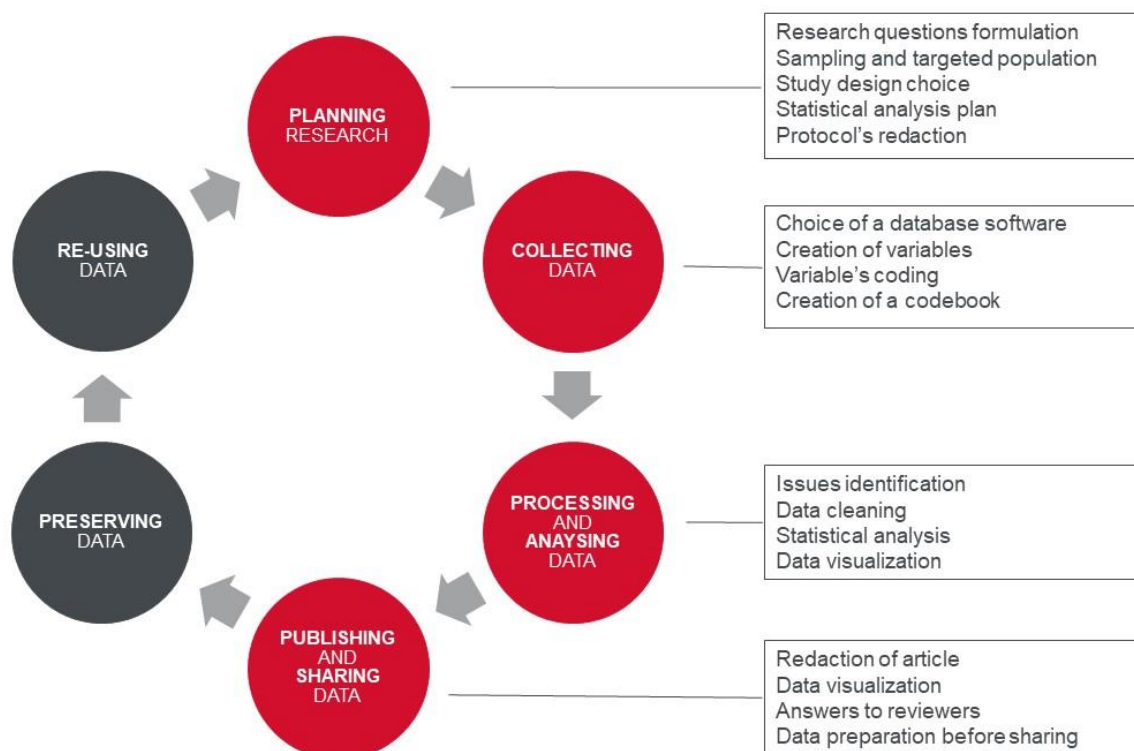
What are your main recommendations related to data quality, data curation and data management ?

What are the future perspectives for your unit ?

What kind of support or improvement for data quality would be ideal for Unisanté ?

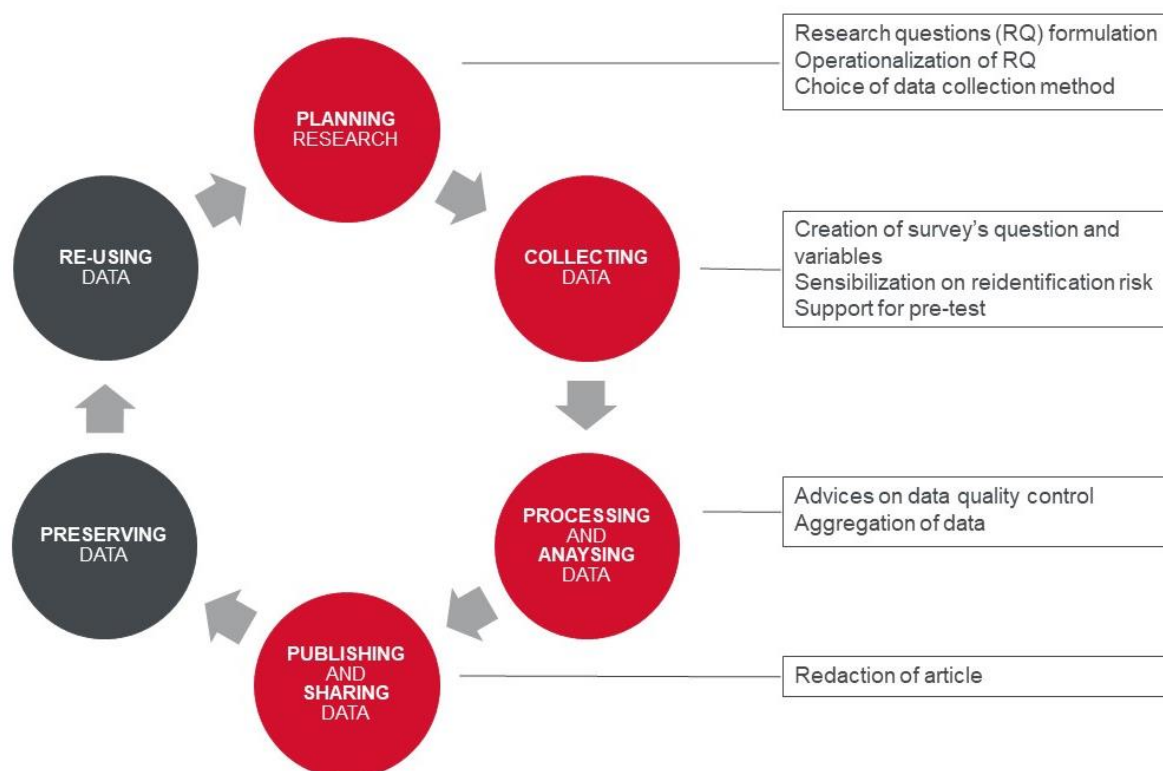
## Annex 7 : Data curation activities : UCB

Activities related to data curation performed by the Biostatistics Consultation unit



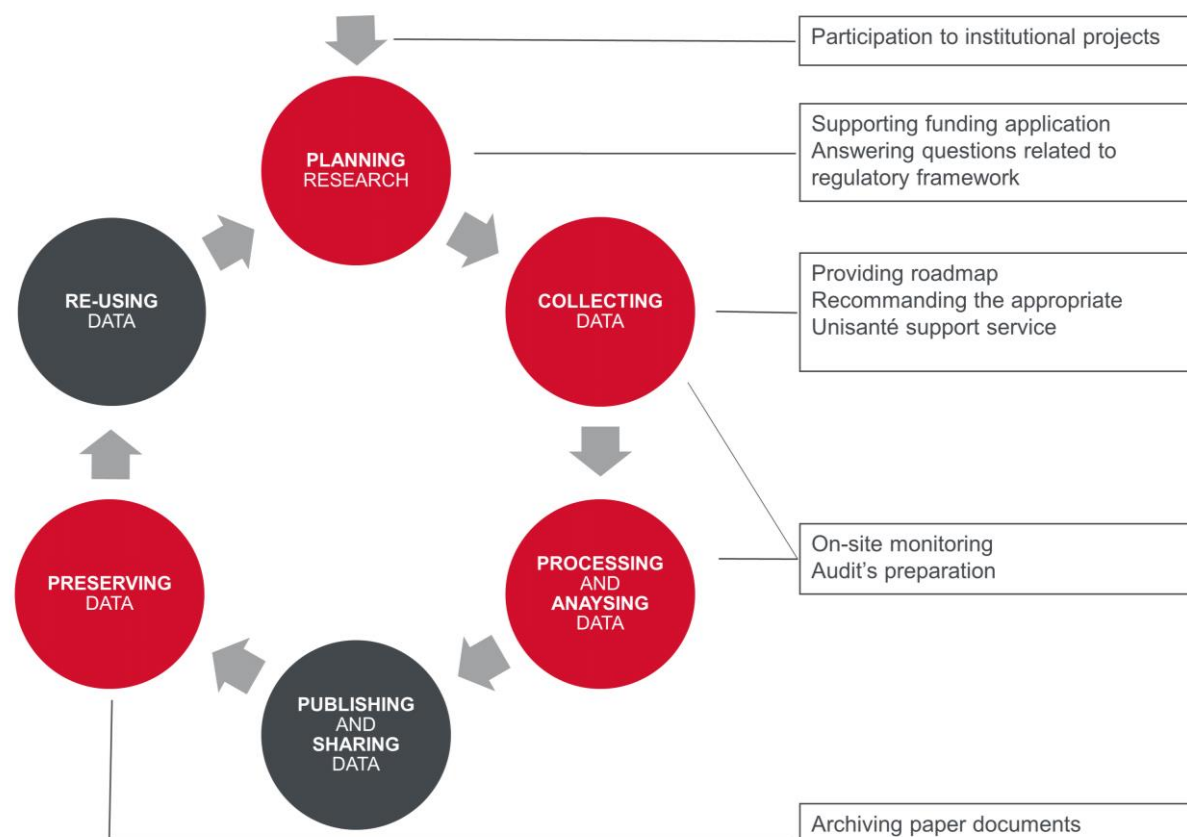
## Annex 8 : Data curation activities : Survey Methodology unit

Activities related to data curation performed by the Survey Methodology unit



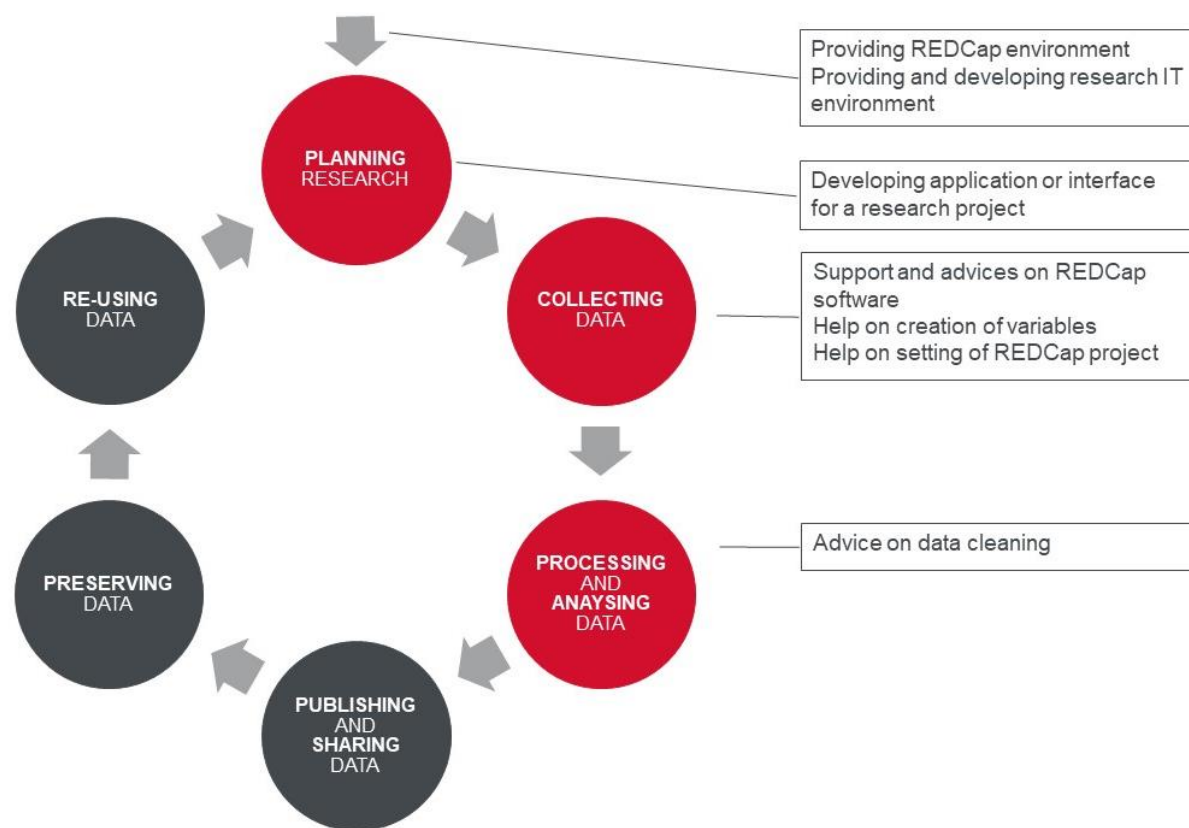
## Annex 9 : Data curation activities : UPR

Activities related to data curation performed by the Research promotion unit



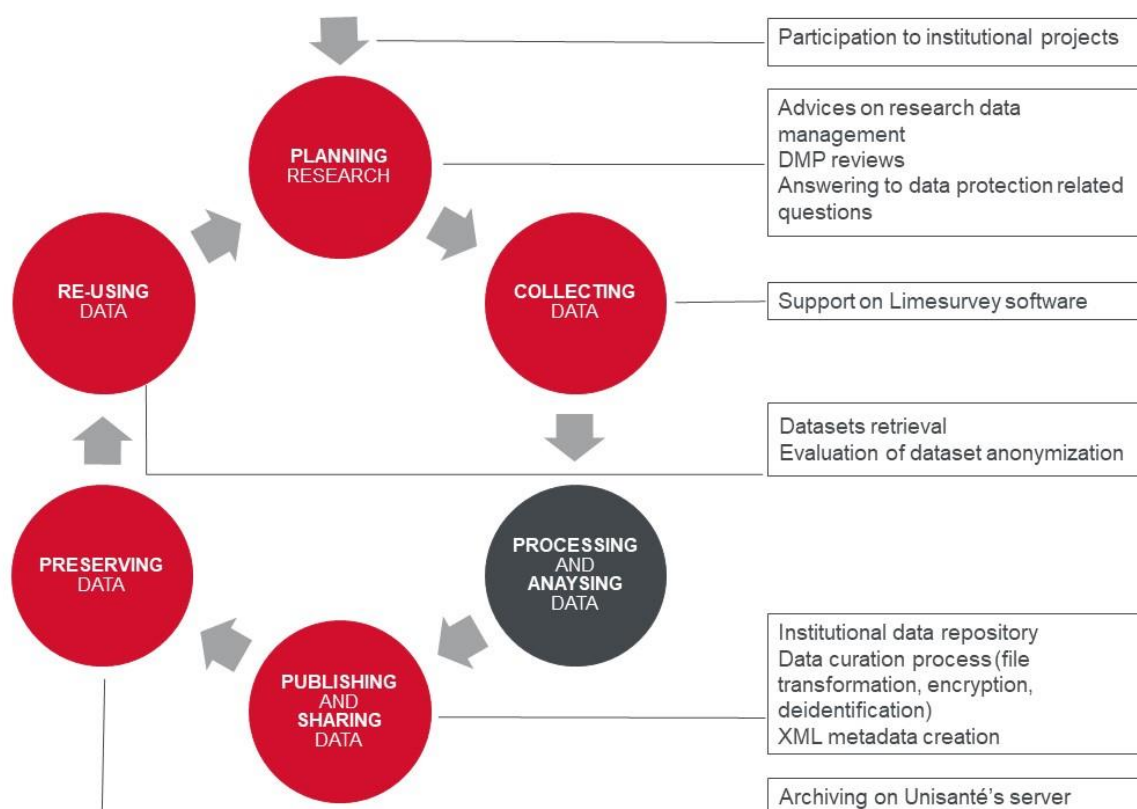
## Annex 10 : Data curation activities : Research IT services unit

Activities related to data curation performed by the Research IT Services unit



## Annex 11 : Data curation activities : UDD

Activities related to data curation performed by the documentation and data unit





## Annex 12 : Table of survey's comment transformations

A comment could appear multiple times, for multiple questions. To facilitate readability of this table, duplicates have been removed. Therefore, this table does not show how many times a comment has been written, nor the activity related to it.

The complete grid will be available with the data.

Question (in French)	Comment (in French)	Data transformation
<b>Planification</b>		
<b>Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?</b>		
	Par habitude	
	Requis par la réglementation	Answer modified to "c'est obligatoire"
<b>Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?</b>		
	À faire, mais pas le temps.	Creation of a new variable "Lack of time"
	À faire.	
	Aucune ressource pour effectuer ce suivi. Autres tâches de gestion des données dépassent déjà largement mon cahier des charges.	
	Ce n'est pas ma responsabilité, c'est à quelqu'un d'autre de le faire.	Creation of a new variable "Not my responsibility"
	Ce n'est pas systématique. Une partie en txt, mais sinon bcp de fichiers propriétaires.	
	Ceci a été fait par d'autres personnes avant moi	Answer modified to "oui fait par quelqu'un d'autre"
	connaissances insuffisantes dans ce domaine.	
	il manque souvent de temps pour cela; l'organisation idéale peut varier d'un projet à l'autre;	Creation of a new variable "Lack of time"

	il n'existe pas de règles claires à ce sujet	
	Je ne sais pas que c'est recommandé	
	L'équipe n'est généralement pas habituée aux formats de fichiers ouverts	
	Les études auxquelles je pense sont pilotes et ne nécessitaient pas de le faire; j'ai prévu de le faire dans le cas d'études cliniques de plus grande envergure	Creation of a new variable "Not applicable to my last projects"
	Manque de temps	Creation of a new variable "Lack of time"
	ne fait pas partie de mon expertise/tâches	Creation of a new variable "Not my responsibility"
	on le fait partiellement. Comment l'organiser de manière efficace?	
	Pas approprié aux données et analyses que nous souhaitons en faire	
	Pas d'actualité lors des derniers projets avec collecte de données	Creation of a new variable "Not applicable to my last projects"
	pas prioritaire, selon les outils disponibles dans l'institution (et installables sur poste HOS)	
	Système NAS du CHUV restrictif	
	Tâche attribuée à un collègue de l'équipe	Answer modified to "oui fait par quelqu'un d'autre"
	Uniquement les personnes sur le projet ont accès aux fichiers. Ainsi, chaque modification est suivie via des discussions à l'intérieur de l'équipe.	
<b>Pouvez-vous indiquer comment vous réalisez ces tâches ?</b>		
	Arborescence de répertoires réservés à	Creation of a new variable "Common server Unisanté :

	l'équipe sur le serveur de fichiers institutionnel	Exchanges with IT respondent, IT service or logistics"
	checksum	Creation of a new variable "Checksum"
	Collecte à partir de REDCap traitée ensuite dans STATA	Creation of a new variable "REDCap" + Creation of a new variable "Statistical software (Stata, SPSS...)"
	Dans mon logiciel de traitement de données (Stata ou SPSS)	Creation of a new variable "Statistical software (Stata, SPSS...)"
	dans un fichier .do (Stata)	Creation of a new variable "Statistical software (Stata, SPSS...)"
	Demande au RI	Creation of a new variable "Common server Unisanté : Exchanges with IT respondent, IT service or logistics"
	do file	Creation of a new variable "Statistical software (Stata, SPSS...)"
	enregistrements en format open	
	Fichier commun	Creation of a new variable "Common server Unisanté : Exchanges with IT respondent, IT service or logistics"
	FileMaker, gestion des accès sous Windows.	Creation of a new variable "FileMaker, Windows"
	J'ai demandé à ouvrir un nouveau dossier au service informatique	Creation of a new variable "Common server Unisanté : Exchanges with IT respondent, IT service or logistics"
	Je copie/colle les anciennes versions des documents que je mets dans un dossier "old" et que je date. Ainsi, j'ai toutes les versions au fur et à mesure.	
	Je n'ai pas besoin de fichier pour choisir.	
	Parfois certaines opérations sont indiquées	Creation of a new variable "Script"

	en commentaires de scripts	
	Redcap	Creation of a new variable "REDCap"
	Répondant informatique	Creation of a new variable "Common server Unisanté : Exchanges with IT respondent, IT service or logistics"
	sur base de données	
	sur email échange avec les personnes en charge dans l'institution, DSI, etc	Creation of a new variable "Common server Unisanté : Exchanges with IT respondent, IT service or logistics"
	sur fichiers scripts	Creation of a new variable "Script"
	sur le serveur Unisanté, par demande au RI ou à logistique	Creation of a new variable "Common server Unisanté : Exchanges with IT respondent, IT service or logistics"
	Sur REDCap	Creation of a new variable "REDCap"
	via des mails également	Creation of a new variable "Common server Unisanté : Exchanges with IT respondent, IT service or logistics"
<b>Pouvez-vous dire qui d'autre réalise cette activité ?</b>		
	Chargé.e de projet	
	Direction des ressources humaines	Answer modified to "Service administratif Unisanté"
	les droits sont demandés par les secrétaires et attribués par la DSI	Answer modified to "Service administratif Unisanté"
	L'investigateur clinique (du CHUV) responsable du projet	Answer modified to "PI Investigateur principal en [Investigateur.trice principal.e (PI) ou responsable de recherche]"
	Personne en charge du développement du code R de traitement des données	

	processus qualité déjà établi	
	responsable administrative de l'unité	Answer modified to "Service administratif Unisanté"
	Responsable de recherche	Answer modified to "PI Investigateur principal en [Investigateur.trice principal.e (PI) ou responsable de recherche]"
	Responsable informatique	Answer modified to "Service administratif Unisanté"
	université de Loyola (Chicago)	
<b>Collecte</b>		
<b>Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?</b>		
	Aide à l'exploitation des données	
	C'est le mandat auquel je pense en répondant qui veut ça: répétition de mesure à intervalles réguliers, donc avec le même questionnaire	
	C'est plus efficace	
	Dépend du projet - pas de standard, puisqu'il n'existe pas d'ontologie ou de voc contrôlé pour tout ...	
	gain de temps	
	Idem, question pas claire. est-ce en rapport avec un projet ou une analyse spécifique?	
	Je ne comprends pas trop votre question?? Comment définissez-vous un "questionnaire"?	
	On utilise en codage à chaque fois que l'on produit une analyse stat ...	
	Pour des questions pratiques	
	Pour la modélisation	

<b>Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?</b>		
	?	
	Car nous collectons des données réelles au niveau individuel	Creation of a new variable "Need the real data"
	Car nous devons toujours l'adapter au contexte de notre étude	
	Ce n'est pas ma responsabilité, c'est à quelqu'un d'autre de le faire.	Creation of a new variable "Not my responsibility"
	Cela diminue la qualité du processus et complique l'analyse des données	
	Cela ne s'applique pas aux études pilotes en cours	
	c'est une activité à faire à la fin de la récolte selon moi	Creation of a new variable "Done in following steps"
	DE manière peu formalisée	
	Elles sont réalisées par le responsable clinique	Answer modified to "oui fait par quelqu'un d'autre"
	Intégrer dans l'identification pour éviter la réidentification	
	J'ai accès des données collectées par d'autres (données de surveillance épidémiologiques)	Creation of a new variable "No data collection - data reuse"
	Je fais ce genre de correction après la collecte des données.	Creation of a new variable "Done in following steps"
	Je ne sais pas si cela a été fait lors de la récolte de données	
	Je n'en ai pas besoin	
	Je n'en ai pas besoin	
	je n'en vois pas l'utilité	Answer modified to "c'est inutile"

	J'utilise des données existantes et ne collecte pas de données de novo	Creation of a new variable "No data collection - data reuse"
	Les données réelles sont collectées pour avoir déjà une base de données, qui pourrait être ultérieurement agrégée avec d'autres bases de données.	Creation of a new variable "Need the real data"
	Manque de temps	Creation of a new variable "Lack of time"
	N/A	
	ne me concerne pas	Creation of a new variable "Not my responsibility"
	ne s'applique pas à mes recherches	
	Plus simple à traiter dans les analyses d'avoir les "vraies" données.	Creation of a new variable "Need the real data"
	questionnaire auto-administré, choix des personnes de ne pas répondre	
	Sera fait APRES la collecte	Creation of a new variable "Done in following steps"
<b>Pouvez-vous indiquer comment vous réalisez ces tâches ?</b>		
	?? comprends pas la question	
	Au laboratoire	Creation of a new variable "Laboratory test"
	Avant test en condition réelle, aussi entretiens cognitifs en face à face avec les patientes et patients (pour nouveaux items, ex. think aloud)	Creation of a new variable "User test"
	avec des utilisateurs du moyen de collecte	Creation of a new variable "User test"
	expertise	Creation of a new variable "Choice of variables / expertise"
	Le risque est évalué à priori, dans le choix des	Creation of a new variable "Choice of variables / expertise"

	variables identifiantes (adresses, sexe, âge,...)	
	Le test est fait en interrogeant des "yeux innocents" au DSTE, c'est à dire des personnes qui ne connaissent pas le projet et qui vont donner un avis sur la lisibilité du questionnaire, l'adaptation de son vocabulaire...	Creation of a new variable "User test"
	logiciel stat	Creation of a new variable "Statistical software (Stata, SPSS...)"
	logiciel stat - question pas claire	Creation of a new variable "Statistical software (Stata, SPSS...)"
	R	Creation of a new variable "R"
	recherche de la littérature - expertise du domaine	
	Sur papier	
	Test directement en laboratoire	Creation of a new variable "Laboratory test"
	Word	Creation of a new variable "Word"
<b>Pouvez-vous dire qui d'autre réalise cette activité ?</b>		
	Assistant.e médical.e, Chargé.e de projet	
	Assistants médicales, via volontaires	
	Elles sont réalisées par le responsable clinique	Answer modified to "PI Investigateur principal en [Investigateur.trice principal.e (PI) ou responsable de recherche]"
	Fournisseur externe des données	
	Gestionnaire de dossiers	
	La personne du charge du développement du code R de traitement/analyse des données	



	nécessite de bonnes connaissances techniques (high level data manager)	
	pas souvent de création de variables durant la collecte	Answer modified to "non, pas fait"
	pour tout ce tableau: le chargé de recherche est le data manager ... (qui est souvent statisticien ...)	
	Référent.e médical.e	
	Spécialiste en machine learning	
	université de Loyola (Chicago)	
<b>Nettoyage</b>		
<b>Pouvez-vous dire pour quelle autre raison vous effectuez ces activités ?</b>		
	C'est nécessaire pour exploiter les données	
	Seulement si nécessaire	
<b>Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?</b>		
	En général, je suis l'utilisateur final des données, je ne les prépare pas pour une utilisation par quelqu'un d'autre. Je pourrais le faire si nécessaire.	Creation of a new variable "Not my responsibility"
	Fait AVANT la collecte	Creation of a new variable "Done in previous steps"
	Jamais utilisé	
	Je ne corrige aucune données et les mets en valeurs manquantes	
	Je ne crois pas qu'il y ait de risque de ré-identification	
	Je ne crois pas qu'il y ait de risque de ré-identification + accès uniquement à des	

	personnes avec secret de fonction	
	nous utilisons des systèmes internationaux de codage des données cliniques	
	Pas fait formellement	
	pas le temps	Creation of a new variable "Lack of time"
	Pas utile pour le projet concerné	
	Toutes les données proviennent des bases.	
<b>Pouvez-vous indiquer comment vous réalisez ces tâches ?</b>		
	.csv	
	analyse de la qualité des données	
	Aussi des fichiers csv qu'on reçoit pour le projet (pas seulement les data générées par le projet)	
	Création de fichiers séparés	Creation of a new variable "Creation of a separate file"
	Enregistrement dans la baie de stockage du DSTE	
	excel	Creation of a new variable "Excel"
	Excel et Stata	Creation of a new variable "Excel" + Creation of a new variable "Statistical software (Stata, SPSS...)"
	informations entrées dans un fichier .xlsx séparé et protégé par mot de passe	Creation of a new variable "Creation of a separate file"
	J'enregistre le fichier et le place dans un dossier "raw data"	
	J'utilise aussi le logiciel R	Creation of a new variable "R"
	RedCap, Excel	Creation of a new variable "REDCap" + Creation of a new variable "Excel"
	RedCap, Excel, STATA	Creation of a new variable "REDCap" + Creation of a

		new variable "Excel" + Creation of a new variable "Statistical software (Stata, SPSS...)"
	requêtes	
	Sphinx	Creation of a new variable "Sphinx"
	Stockage de questionnaires papiers. Fichiers Redcap	Creation of a new variable "REDCap"
	Sur Excel (pour l'identification), puis dans REDCap (pour la correction)	Creation of a new variable "REDCap" + Creation of a new variable "Excel"
	utilisation des codes Meddra par exemple pour le codage des pathologies	
	XLS	Creation of a new variable "Excel"
<b>Partage/archivage</b>		
<b>Pouvez-vous dire pour quelle autre raison vous n'effectuez pas ces activités ?</b>		
	?	
	ça n'as plus de sens à ce stade du projet (analyses terminées)	
	Car déjà fait pour nous	
	Cela a déjà été fait avant l'étape de l'archivage	Creation of a new variable "Done in previous steps"
	Cela a déjà été fait avant l'étape de l'archivage. Je ne modifie pas le fichier à ce moment car il est archivé à la fin du travail et sa publication.	Creation of a new variable "Done in previous steps"
	cela ne rentre pas dans le cadre de mes projets	
	Déjà effectué et documenté dans les étapes précédentes	Creation of a new variable "Done in previous steps"
	Déjà fait dans le cadre du projet normalement	Creation of a new variable "Done in previous steps"

	Déjà fait lors de la collecte des données	Creation of a new variable "Done in previous steps"
	Déjà fait pendant la recherche	Creation of a new variable "Done in previous steps"
	Déjà réalisé dans les étapes précédentes du projet	Creation of a new variable "Done in previous steps"
	Etape effectuée avant	Creation of a new variable "Done in previous steps"
	Etape faite avant	Creation of a new variable "Done in previous steps"
	Je fais.	
	Je ne vois pas l'intérêt de cette transformation	
	Les données ne m'appartiennent pas	
	Manque de temps	Creation of a new variable "Lack of time"
	Ne sais pas comment faire	Answer modified to "Ne sais pas comment faire"
	non pertinent	
	Opérations déjà réalisées précédemment	Creation of a new variable "Done in previous steps"
	Pas assez de temps	Creation of a new variable "Lack of time"
	Pas forcément pertinent	
	Si, je le fais. Mais le questionnaire ne m'a pas proposé cette modalité de réponse (souci technique à votre questionnaire?)	
	Va au-delà des règles éthiques en la matière, ne permet pas d'ajouter des variables au cours de la période d'analyse	
<b>Pouvez-vous indiquer comment vous réalisez ces tâches ?</b>		
	Déjà fait à cette étape du processus	Creation of a new variable "Done in previous steps"
	Doc word avec les infos d'archivage	Creation of a new variable "Text file (Word, txt)"

	fait par data.unisante	Creation of a new variable "text file (Word, txt)" + Creation of a new variable "XML Metadata"
	fichier Excel	Creation of a new variable "Excel"
	fichier Word	Creation of a new variable "Text file (Word, txt)"
	Manuellement	
	Mes données sont codées mais je n'ai jamais anonymisé mes données au sens strict	
	Office	Creation of a new variable "Excel" + Creation of a new variable "Text file (Word, txt)"
	Sphinx	Creation of a new variable "Sphinx"
	txt	Creation of a new variable "Text file (Word, txt)"
	Word	Creation of a new variable "Text file (Word, txt)"
	Word, Excel	Creation of a new variable "Excel" + Creation of a new variable "Text file (Word, txt)"
<b>Est-ce que vous partagez vos données ? Si oui, précisez où</b>		
	Article	Creation of a new variable "Article"
	Dépôt de données Unisanté	Creation of a new variable "Unisanté data repository"
	dépôt de données Unisanté (avec toi d'ailleurs :-)	Creation of a new variable "Unisanté data repository"
	Dépôt de données Unisanté, DCC-SPHN, mandant (e.g. OSAV, OFSP).  Dans le cadre de collaborations scientifiques, via contacts e-mails, signature de DTA et envoi sécurisé de données..	Creation of a new variable "Unisanté data repository" + Creation of a new variable "DCC-SPHN + secured sending + Email"

	Dépôt de données/serveur Unisanté	Creation of a new variable "Unisanté data repository"
	Dépôt données Unisanté	Creation of a new variable "Unisanté data repository"
	Dépôt unisanté	Creation of a new variable "Unisanté data repository"
	Maelstrom	Creation of a new variable "Maelstrom"
	Maelstrom (métadonnées), Site web du projet (métadonnées)	Creation of a new variable "Maelstrom" + Creation of a new variable "Dedicated website"
	Matériel supplémentaire de l'article ou site web dédié (en développement)	Creation of a new variable "Article" + creation of a new variable "Dedicated website"
	Matériel supplémentaire de l'article, site web dédié (Github)	Creation of a new variable "Article" + creation of a new variable "Dedicated website"
	Répertoire dédié au projet	
	SI. de l'article publié	Creation of a new variable "Article"
	zenodo pour les données en dehors des essais cliniques	Creation of a new variable "Zenodo"
<b>Est-ce que vous partagez vos données ? Si non, précisez pourquoi</b>		
	Données de surveillance cantonales et fédérales (demande de réutilisation de données auprès de CER-VD)	
	Données personnelles et confidentielles	
	interdiction	Creation of a new variable "Forbidden"
	Interdiction de partager la plupart du temps	Creation of a new variable "Forbidden"
	Interdiction de partager  Les données ne m'appartiennent pas.	Creation of a new variable "Forbidden"
	J'analyse les données. Ce n'est pas à moi de les partager. Cependant, je	Creation of a new variable "Not my responsibility"

	peux le faire si on me le demande.	
	Je souhaite partager, mais n'ai pas le temps de préparer les données pour cela	Creation of a new variable "Lack of time"
	Le mandant reste propriétaire des données et interdit le partage	Creation of a new variable "Forbidden"
	Les données ne m'appartiennent pas	
	Ne sais pas où partager	Creation of a new variable "Don't know where to share"
	Ne veut pas partager	
	Pas de demande de données	
	Pas encore eu l'occasion	
	Pas encore! Jusqu'à date je n'ai pas créé des données suffisamment importantes pour justifier l'effort que ça prend pour partager des données.	
	Prend du temps, perte de contrôle	Creation of a new variable "Lack of time"
<b>Est-ce que vous archivez les données après le projet ? Si oui, précisez où</b>		
	Dépôt de données et archivage physique des documents papier	Creation of a new variable "Data repository"
	Dépôt de données Serveur Unisanté (L ou filearc)	Creation of a new variable "Common server Unisanté" + Creation of a new variable "Data repository"
	Interne / Filearc  Externe : données doivent être supprimées	Creation of a new variable "Common server Unisanté"
	Je les conserve sur mon disque dur C:\	Creation of a new variable "C Disk"
	Je les laisse là où elles sont	Creation of a new variable "I let the data where they are"

	Je les laisse là où elles sont (serveur commun ou autre)	Creation of a new variable "I let the data where they are"
	L	Creation of a new variable "Common server Unisanté"
	Le projet REDCap est archivé, et les données sont laissées sur le disque commun	Creation of a new variable "REDCap" + Creation of a new variable "Common server Unisanté"
	oui, mais de façon non systématique et insuffisante.	
	Serveur CHUV	Creation of a new variable "CHUV Server"
	Serveur CHUV, Unisanté et Switch drive	Creation of a new variable "Common server Unisanté" + Creation of a new variable "CHUV server" + Creation of a new variable "Switch drive"
	Serveur commun	Creation of a new variable "Common server Unisanté"
	Serveur Unisanté	Creation of a new variable "Common server Unisanté"
	Serveur Unisanté (L ou filearc)	Creation of a new variable "Common server Unisanté"
	Serveur Unisanté, espace dédié archives	Creation of a new variable "Common server Unisanté"
	serveurs DSI	Creation of a new variable "CHUV server"
<b>Est-ce que vous archivez les données après le projet ? Si non, précisez pourquoi</b>		
	fait à l'université de Loyola	
	interdiction de les conserver	Creation of a new variable "Forbidden"
	Les données sont gérées par quelqu'un d'autre.	