

FP6-IST-2004-027510

ASSIS+

Association Studies assisted by Inference and Semantic Technologies

Specific Targeted Research Project

Integrated Biomedical Information for Better Health

D2.1 Ethics, information security and patient privacy manual

Due Date of Deliverable: 30/06/06

Actual Submission Date: 15/02/07

Revision: Final

Start date of project: January 1st 2006

Duration: 3 years

Organization name of lead contractor for this deliverable: Custodix

Authors: Filip de Meyer (Custodix)

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of Contents

Table of Content	2
1 Introduction	3
2 Scope and objectives	3
3 Legislative and regulatory references	4
3.1 Personal data privacy protection EU legislation/regulation.....	4
3.2 Clinical trials legislation/regulations	6
3.2.1 Definitions	7
3.2.2 Confidentiality requirements.....	8
3.3 Conclusions of the clinical trial references	12
4 Genetic exceptionalism.....	13
5 Cross border genetic testing.....	15
5.1 New uses of stored samples	16
6 Ethical Issues other than privacy	17
6.1 Patient related	17
6.2 Provision of publicly available electronic communications services.....	18
6.3 Ownership and IP rights	19
7 The Assist data resources	20
7.1 Availability and origin of the data	20
7.2 Information content of the gathered data.....	21
7.3 Recommendations on the collection of data.....	22
7.3.1 Re-use of existing data	22
7.3.2 Newly collected data or additional data.....	23
8 Privacy protection model for Assist Data	25
8.1 The personal data domain	25
8.2 The de-identified data domain	26
8.3 The de-identification services.....	27
9 A conceptual model for de-identification of personal data	28
9.1 Objectives of personal privacy protection.....	28

9.2	Personal data vs. de-identified data	28
9.2.1	Definition of personal data	28
9.2.2	The concept of identification	29
9.2.3	The concept of de-identification	30
9.3	Real world identifiability and anonymity	32
9.3.1	Rationale.....	32
9.3.2	Levels of anonymity.....	32
9.4	Remark on privacy threats.....	34
9.5	The pseudonymisation process.....	35
9.5.1	Entities in the model.....	35
9.5.2	Preparation of data	36
9.5.3	Processing steps.....	37
10	The ASSIST data collection model	38
11	Informed consent.....	39
12	Conclusions.....	41
13	References.....	42

1 Introduction

This document is the D2.1 deliverable called «ethics, information security and patient privacy manual».

The content of this document will serve as reference and input for:

- D2.2 (Final report on ethical conduct and security methods) with due date M36
- Other Internal and public deliverables that touch upon ethics or security issues, in particular WP4 (User requirements-system specification).

2 Scope and objectives

The objective of the security workpackage is to present a generic model for the privacy protection of research data, modelled after the experiences in ASSIST, in order to facilitate research through systems that virtually associate multiple patient record repositories, physically located in different centres.

It is targeted at the ICT professionals with the responsibility to include state-of-art privacy protection into data collection and processing solutions and who have to be informed about the implementation and usage

principles. Another major target group is the representatives of the data centres in the project and their respective review boards, in order to allow them to obtain a fair level of understanding of the quality and robustness of the proposed solutions within Assist and make a selection of privacy protection measures.

The term «ethics» covers more than only privacy issues. Some of the non-privacy issues will be discussed on the sections on informed consent. This deliverable will prove that nonetheless, the protection of the privacy of the trials subjects is the major ethical concern in this project.

3 Legislative and regulatory references

3.1 Personal data privacy protection EU legislation/regulation

The basic document in Europe concerning the privacy of personal data in Europe is the so-called Data protection directive [1].

This document is supplementary to other directives, namely the so-called electronic communications directory (directive 97/66/EC) of the European Parliament and of the Council of 15 December 1997 concerning the processing of personal data and the protection of privacy in the telecommunications sector.

The Council of Europe, Committee of Ministers, Recommendation No. R (97) 5 on the Protection of Medical Data (Feb. 13, 1997) is also of relevance as it deals explicitly with the protection of medical data [2].

A more recent document of relevance is the Regulation (EC) no 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data.

The aim of the directive is twofold: on the one hand to safeguard the free flow of goods, persons, services and capital within the EU, and on the other hand to guarantee sufficient protection of the data privacy when dealing with personal data. The Assist project will be a model research application in which at least EU based legal entities and natural persons will be involved from several countries. Should the case arise that trial samples are sent to laboratories outside the EU, the consequent issues will have to be dealt with as well.

It is clearly stated in the objectives of the directive that «Member States shall neither restrict nor prohibit the free flow of personal data between Member States for reasons connected with the protections afforded under paragraph 1.» Paragraph 1 states that the Member States shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data. This in turn has been translated into national legislation in the Member states whose adoption and implementation of the DPD may differ from other Member States. The project's intention is not to treat the national legislations of the countries or regions (as may be the case e.g. in Germany) in detail, unless specific requirements to do so arise during the course of the project.

As part of the implementation of the DPD, Member States have set up so called data controllers or data registrars.

The DPD formulates a number of conditions concerning the processing of personal data. Member States shall provide that personal data must be:

- (a) processed fairly and lawfully;
- (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide **appropriate safeguards**;
- (c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed;
- (d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;
- (e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

Article 8 in the DPD deserves a closer examination:

1. Member States shall prohibit the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life.

2. Paragraph 1 shall not apply where:

(a) the data subject has given his explicit consent to the processing of those data, except where the laws of the Member State provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject's giving his consent; or

(b) processing is necessary for the purposes of carrying out the obligations and specific rights of the controller in the field of employment law in so far as it is authorized by national law providing for adequate safeguards; or

(c) processing is necessary to protect the vital interests of the data subject or of another person where the data subject is physically or legally incapable of giving his consent; or

(d) processing is carried out in the course of its legitimate activities with appropriate guarantees by a foundation, association or any other non-profit-seeking body with a political, philosophical, religious or trade-union aim and on condition that the processing relates solely to the members of the body or to persons who have regular contact with it in connection with its purposes and that' the data are not disclosed to a third party without the consent of the data subjects; or



(e) the processing relates to data which are manifestly made public by the data subject or is necessary for the establishment, exercise or defence of legal claims.

3. Paragraph 1 shall not apply where processing of the data is required for the purposes of preventive medicine, medical diagnosis, the provision of care or treatment or the management of health-care services, and where those data are processed by a health professional subject under national law or rules established by national competent bodies to the obligation of professional secrecy or by another person also subject to an equivalent obligation of secrecy.

4. Subject to the provision of suitable safeguards, Member States may, for reasons of substantial public interest, lay down exemptions in addition to those laid down in paragraph 2 either by national law or by decision of the supervisory authority.

5. Processing of data relating to offences, criminal convictions or security measures may be carried out only under the control of official authority, or if suitable specific safeguards are provided under national law, subject to derogations which may be granted by the Member State under national provisions providing suitable specific safeguards. However, a complete register of criminal convictions may be kept only under the control of official authority.

Member States may provide that data relating to administrative sanctions or judgements in civil cases shall also be processed under the control of official authority.

6. Derogations from paragraph I provided for in paragraphs 4 and 5 shall be notified to the Commission

7. Member States shall determine the conditions under which a national identification number or any other identifier of general application may be processed.

3.2 Clinical trials legislation/regulations

Although this type of regulations focuses mainly on the development of pharmaceutical drugs and therapeutic practices, it clarifies the intention of informed consent and discusses the privacy of the trial subjects.

Terminology is not always consistent and the objectives of various legislations are different: most guidelines target «clinical trials» in which there is an active feedback channel towards the trial subject in the form of treatment of the trial subject. The trial subject actively participates in the trial during a defined period of time and often as part of treatment of a disease.

Assist is about «medical research involving human subjects» as is stated in the Helsinki declaration by the World Medical Association which can be considered as a reference document [8]. The Helsinki declaration recognises that «Medical progress is based on research, which ultimately must rest on experimentation involving human subjects». The declaration makes a distinction between basic principles for all medical research (section B) and medical research combined with medical care (section C). It is understood that Assist does not involve medical care.

In practice, because “informed consent” is given by participants in clinical trials, often, no further extra measures are taken to protect the confidentiality of the participants’ information. As the remainder of this section will show, that attitude is not in line with the privacy statements contained in the various documents.

This section probes this issue based on the following reference documents:

- Guidelines for good clinical practice (GCP) for trials on pharmaceutical products, World Health Organization WHO Technical Report Series, No. 850, 1995, Annex 3[3];
- Note for guidance on good clinical practice (CPMP/ICH/135/95), The European Agency for the Evaluation of Medicinal products [4];
- Directive 2001/20/EC of the European Parliament and the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use [5];
- International Ethical Guidelines for Biomedical Research Involving Human Subjects;
- 1991 International Guidelines for Ethical Review Of Epidemiological Studies[7].

3.2.1 Definitions

Informed consent [3]

A subject's voluntary confirmation of willingness to participate in a particular trial, and the documentation thereof. This consent should only be sought after all appropriate information has been given about the trial including an explanation of its status as research, its objectives, potential benefits, risks and inconveniences, alternative treatment that may be available, and of the subject’s rights and responsibilities in accordance with the current revision of the Declaration of Helsinki.

Direct Access[4]

Permission to examine, analyze, verify, and reproduce any records and reports that are important to evaluation of a clinical trial. Any party (e.g. domestic and foreign regulatory authorities, sponsor’s monitors and auditors) with direct access should take all reasonable precautions within the constraints of the applicable regulatory requirement(s) to maintain the confidentiality of subjects’ identities and sponsors’ proprietary information

Confidentiality [3]

Maintenance of the privacy of trial subjects including their personal identity and all personal medical information.

Sponsor [5]

An individual, company, institution or organization which takes responsibility for the initiation, management and/or financing of a clinical trial.

Investigator [5]

A doctor or a person following a profession agreed in the Member State for investigations because of the scientific background and the experience in patient care it requires. **The investigator is responsible for the conduct of a clinical trial at a trial site.** If a team of individuals at a trial site conducts a trial, the investigator is the leader responsible for the team and may be called the principal investigator.

3.2.2 Confidentiality requirements

The articles relating to confidentiality protection of trial subjects extracted from the references documents are given in this section. They will serve as the basis to define the requirement for the entities involved as will be explored in the next section.

Section 3.3-b on informed consent

“The subject must be made aware and consent that personal information may be scrutinized during monitoring, auditing or inspection of the trial by properly authorized persons, the sponsor or relevant authorities, and that participation and **personal information in the trial will be treated as confidential and will not be publicly available**”. National laws and regulations may modify this principle.

Section 3.4 of the GCP [3] on Confidentiality

“The investigator must establish secure safeguards of confidentiality of research data as described in the current revision of the International Ethical Guidelines for Biomedical Research Involving Human Subjects”.

Guideline 18 of these guidelines - bearing the title of “Safeguarding Confidentiality” states:

“The investigator must establish secure safeguards of the confidentiality of subjects research data. Subjects should be told the limits, legal or other, to the investigators’ ability to safeguard confidentiality and the possible consequences of breaches of confidentiality.”

This is elaborated in the commentary on guideline 18:

“Confidentiality between investigator and subject. Research relating to individuals and groups may involve the collection and storage of information that, if disclosed to third parties, could cause harm or distress. **Investigators should arrange to protect the confidentiality of such information by, for example, omitting information that might lead to the identification of individual subjects, limiting**

access to the information, anonymising data, or other means. During the process of obtaining informed consent the investigator should inform the prospective subjects about the precautions that will be taken to protect confidentiality.”

Section 7.2 of GCP reporting” further deals with confidentiality

“When reporting adverse events to the sponsor, the investigator should protect confidentiality by excluding the names of individual subjects, personal identification numbers (e.g. social security numbers) or addresses”.

Section 8 of GCP on record keeping and data handling states further:

“In the event of electronic data handling, confidentiality of the database must be secured by safety procedures such as passwords and written assurances from all staff involved. Provision must be made for the satisfactory maintenance of the database and for back-up procedures”.

Section 26 of the 1991 international guidelines for ethical review of epidemiological studies [7]

contains a complete section on confidentiality and identifiability:

Research may involve collecting and storing data relating to individuals and groups, and such data, if disclosed to third parties, may cause harm or distress. Consequently, investigators should make arrangements for protecting the confidentiality of such data by, for example, omitting information that might lead to the identification of individual subjects, or limiting access to the data, or by other means. It is customary in epidemiology to aggregate numbers so that individual identities are obscured. Where group confidentiality cannot be maintained or is violated, the investigators should take steps to maintain or restore a group's good name and status. Information obtained about subjects is generally divisible into:

- Unlinked information, which cannot be linked, associated or connected with the person to whom it refers; as this person is not known to the investigator, confidentiality is not at stake and the question of consent does not arise.

- Linked information, which may be:
 - **anonymous**, when the information cannot be linked to the person to whom it refers except by a code or other means known only to that person, and the investigator cannot know the identity of the person;

 - **non-nominal**, when the information can be linked to the person by a code (not including personal identification) known to the person and the investigator; or

 - nominal or nominative, when the information is linked to the person by means of personal identification, usually the name. Epidemiologists discard personal identifying information when consolidating data for purposes of statistical analysis. **Identifiable personal data will not be used when a study can be done without personal identification** - for instance, in testing unlinked anonymous blood samples for HIV in-



fection. When personal identifiers remain on records used for a study, investigators should explain to review committees why this is necessary and how confidentiality will be protected. If, with the consent of individual subjects, investigators link different sets of data regarding individuals, they normally preserve confidentiality by aggregating individual data into tables or diagrams. In government service the obligation to protect confidentiality is frequently reinforced by the practice of swearing employees to secrecy.

Article 2.11 of ICH [4]

“The confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s)”.

Article 4.8.10 of ICH

Both the informed consent discussion and the written informed consent form and any other written information to be provided to subjects should include explanations of the following:

n) That the monitor(s), the auditor(s), the IRB/IEC, and the regulatory authority(ies) will be granted direct access to the subject’s original medical records for verification of clinical trial procedures and/or data without violating the confidentiality of the subject, to the extent permitted by the applicable laws and regulations and that, by signing a written informed consent form, the subject or the subject’s legally acceptable representative is authorizing such access.

o) That records identifying the subject will be kept confidential and, to the extent permitted by the applicable laws and/or regulations, will not be made publicly available. If the results of the trial are published, the subjects’ identity will remain confidential.

3.3 Conclusions of the clinical trial references

The guidelines and other reference documents reveal that the main goal of 'informed consent' is to inform the trial subjects about the way the trial will be run, including their rights and risks but also on the way the confidentiality of their data will be handled.

"Informed consent" does not waive the privacy rights of the trial subjects, nor does it limit the responsibilities of the investigators and sponsors in case of a privacy breach because of insufficient privacy protection.

The use of identities and more in particular direct access to the medical data is restricted for specific purposes and to specific persons. The purpose should be control of the trial itself (both data and procedure) and access should be limited to authorised persons in charge of the verification. For running trials, identifiable information must be avoided and third parties must not have access to personal data.

The reference documents quoted above have been written in the 90-ies. Practical implementations of identity protection through privacy enhancing techniques such as electronic de-identification and pseudonymisation have only become available at the end of the 90-ies. Therefore the guidelines could not recommend these explicitly. Nevertheless they are clear about not using identities of trial subjects for the studies by advising the use of trial identities, which was the only alternative at the time of writing.

At the beginning of the 90-ies, trials identities based on a translation list were the only 'technology' available. Privacy risk analysis shows that trial identities based on a list can easily be cracked and the real identities be obtained, especially in electronic handling of trials. Subtasks in clinical trials are more and more outsourced to third parties (data collection, communication, statistical analysis, security,..) and consequently this increases the risk of unauthorised identification. Therefore, privacy protection should be applied as soon as the trial data leaves the computer system of the treating physician or at least when the data leaves the security domain for which the ICT department of his institution is responsible. Since 2000, pseudonymisation techniques that allow effective and efficient protection of trial subjects have become available. As privacy legislation states that protection should be commensurate to the risks, proper privacy protection should be based on state-of-the-art solutions.

The nature of the data being processed in the Assist project is of a highly sensitive nature as it comprises both health data and data related to the sex life of the data subjects. The purpose of the processing does not fall under the implicit exceptions stated in the DPD such as vital interest for the patient and necessity for the treatment of the patient. Therefore, it is highly recommended to obtain informed consent from the data subject, if not available yet.

Although the research involved in Assist is not a clinical trial as such, the analysis of privacy issues in the clinical trials domain can be easily transposed to the Assist context.

As will be explained in another section, Assist will make a distinction in two domains: one domain where personal data is collected and processed under the full control (and responsibility) of the data centre or investigators participating in the Assist association studies and another domain where only de-identified data will be used for processing.

European legislation on personal data protection is fairly comprehensive compared to e.g. the United States of America, where there is no clarity about the use of medical data or research data. In the U.S.A., the tendency is even to keep the collection of research data very far away from clinical data, as their use of genetic information in health insurance and on the workplace is known [13]. In fact, in Europe, all personal data is protected and the principles for collection, disclosure and use more or less defined. Health related data is even specifically covered. The situation in the EU can even be considered opposite from the situation in U.S.A., as in the EU researchers are trying to re-use data in research that has been previously collected for clinical purposes. The EU privacy protection legislation is valid for all personal information, regardless by whom it is collected. This is e.g. not the case in the U.S.A., where there is no legal requirement for companies to protect human subjects. Neither is there clarity about what happens to confidential information when a private biobank goes bankrupt.

4 Genetic exceptionalism

A recurring issue when discussing specifically genetic data is whether genetic data are different from other medical information. The point of view that it is, has been coined «genetic exceptionalism».

Both schools of thought are present in the medical and legal world. These discussions are particularly vivid in the context of human gene banks where the tensions between the drive for progress and benefit (and the associated business opportunities) and the protection of personal privacy can be very high.

In 2004, Directorate C (Science and Society), unit C3 (Ethics and Science) has issued a publication called «25 Recommendations on the ethical, legal and social implications of genetic testing» [9]. It is the result of a multidisciplinary group of experts that included stakeholders who were already involved or personally interested in the topic. Representatives came from the industry that produces or uses genetic tests, from NGOs (in particular, patient organisations with clear interests in the subject), and scientists and representatives from academic institutions with different backgrounds specialised in the field (law, philosophy, ethics, and medicine). The participants came from various national backgrounds within Europe and numbers were well balanced between men and women.

Their conclusions are that: «...the sentiment that genetic data are different from other medical information is inappropriate. Genetic information is part of the entire spectrum of all health information and does not represent a separate category as such. All medical data, including genetic data, must be afforded equally high standards of quality and confidentiality at all times. However, the current public perception that genetic information is somehow different is acknowledged by the Group. This perception is due to a number of factors. These include historical reasons (eugenics), the current predominance of predictive genetic tests for rare monogenic diseases which may give rise to particularly sensitive information affecting patients' relatives, the fact that no treatment is available yet for most monogenic diseases, potential loss of control over samples, plus a number of other reasons. Current efforts to establish guidelines, recommendations, rules, regulatory texts and laws that apply specifically to genetic testing and data handling should be viewed as an understandable response to specific public concerns. They are, however, only acceptable as a stepping stone to

more considerate and inclusive legal and regulatory frameworks that encompass all medical data and testing, and which reflect advancements made in healthcare provision...».

Their recommendations with potential relevance to ASSIST can be summarised as:

- “genetic exceptionalism” should be avoided, internationally, in the context of the EU and at the level of its Member States. However, the public perception that genetic testing is different needs to be acknowledged and addressed;
- all medical data, including genetic data, must satisfy equally high standards of quality and confidentiality;

On confidentiality, privacy and autonomy, the recommendations are:

- genetic data of importance in a clinical and/or family context should receive the same level of protection as other comparably sensitive medical data;
- the relevance for other family members has to be addressed;
- the importance of a patient’s right to know or not to know be recognised and mechanisms incorporated into professional practice that respect this.
- In the context of genetic testing, encompassing information provision, counselling, informed consent procedures, and communication of test results, practices should be established to meet this need;

There are of course other opinions that do not agree with this vision. One argument, for instance is that the information content of genetic information collected in the context of population research is not exactly known and that the information content of data in databanks or human tissue repositories may be more sensitive than currently can be assessed.

Other groups do not follow this line of thought, such as the so called «Montreux declaration» done at the 27th International conference of Data Protection and Privacy Commissioners in September 2005 where in the preamble it is stated that: «Aware that the fast increase in knowledge in the field of genetics may make human DNA the most sensitive personal data of all; Aware also that this acceleration in knowledge raises the importance of adequate legal protection and privacy».

The ethics task group want to stress that Assist is not involved in broad range population genetics research but is targeted at a well defined group (gender and age), with interest in specific polymorphisms, medical and behavioural data, although sensitive, with the aim to better understand the association of factors leading to cervical cancer.

Nevertheless it should be aware of existing legislation, guidelines, publications on related subject matter that often deals with the broader context of population genetic research and human tissue repositories. Assist has many commonalities (from an ethical point of view) with other gene related research. In generic terms, their goals can be described as either the identification of disease susceptibility genes or diagnostic biomarkers.

5 Cross border genetic testing.

The Assist project may require that samples from one institution be sent to another institution in another country to be (genetically) tested. It is assumed that the cross border testing is taking place in countries that fall under the EU DPD directive, and not, e.g. the United States of America, only to name one.

Cross border testing issues in research are often related to the availability of biological samples available in a centre whereas other centres are interested in doing research on these specimen. Consensus for new research on previously stored samples or on informing individuals about (or even providing them access to) the results of research carried out on their samples, or on intellectual property restrictions on bio specimens and data [12].

The OECD (Organisation for economic co-operation and development) has published a guide on «Quality Assurance and Proficiency testing for Molecular Genetic Testing» as the summary results of a survey of 18 OECD member countries. It does not focus on privacy or ethical issues, however, but is more concerning with the quality of the results of the testing.

In case cross border testing is done, the legislation of the country of the controller will be applicable. If the data controller is established in another country than the member state, the law of this member state can be applied by virtue of the international law (article 4, b of the Directive 95/46/CE). This may be the case when subcontracting is used.

Article 4 of the DPD states:

1. Each Member State shall apply the national provisions it adopts pursuant to this Directive to the processing of personal data where:

(a) the processing is carried out in the context of the activities of an establishment of the controller on the territory of the Member State; when the same controller is established on the territory of several Member States, he must take the necessary measures to ensure that each of these establishments complies with the obligations laid down by the national law applicable;

(b) the controller is not established on the Member State's territory, but in a place where its national law applies by virtue of international public law;

(c) the controller is not established on Community territory and, for purposes of processing personal data makes use of equipment, automated or otherwise, situated on the territory of the said Member State, unless such equipment is used only for purposes of transit through the territory of the Community.

2. In the circumstances referred to in paragraph 1 (c), the controller must designate a representative established in the territory of that Member State, without prejudice to legal actions, which could be initiated against the controller himself.

Therefore it is important that the centres involved in the Assist project each designate who is the controller of their data.

5.1 New uses of stored samples

Five countries (Estonia, Iceland, Norway, Sweden and the UK) have national legislation governing the collection, storage and use of biological samples but there is no uniform approach regarding consent for new uses of stored samples. Policies in Canada, Germany, Norway, the Netherlands and the U.S.A permit the use of stored samples without consent if the samples' subject is not identifiable.

Most of the regulations for the use of stored samples relate to tissue banks. The subject of the use of stored samples will be further elaborated in the final deliverable of this workpackages in order to extend Assist as a proof of concept into a generic approach.

From discussion with the data centres it looks like the samples in Assist have been specifically collected with Cervix cancer research in mind or that informed consent (cfr. The German data centre) has been obtained.

Therefore, wherever the bio-samples are available for research purposes, cervical cancer research is compatible with the original finality of the research.

6 Ethical Issues other than privacy

The focus of this document is personal privacy as, that will be the major concern of the consortium with regard to the collection and processing of data. The data will be obtained either indirectly from data collected earlier for other purposes (screening, diagnosis, treatment) or directly for research purposes. The issues covered with the related privacy are discussed in the preceding sections.

However, there are some other ethical and legal aspects that have to be considered as well. This section gives an overview and brief discussion.

6.1 *Patient related*

Apart from privacy concerns, there are also concerns on the physical and mental well-being of the trial subject. The trial subject should be informed and even be able to refuse before engaging in research activities. (S)He has to be aware of the risks involved in the trial, his/her rights and recourse to compensation, the guarantees to privacy and the ethical implications. Most legislations and regulations on this subject matter originates from the domain of clinical trials where trial subjects are actively involved in a treatment process and run considerable risks when consenting to participate because of the innovative and partly speculative nature of the trials being conducted. This will not be the case in the Assist project, as the patient does not undergo any treatment. In Assist the patient is seen as a data source.

In case additional data is needed from a patient within Assist, the risk for the patient is rather low, but still existent. Apart from the incentive that may be required to get the trial subjects involved (compensation for travel expenses), insurances should be taken by the trial centres, in case the trial subjects suffer any damage. Examples of damage could be:

- Traffic accident while going to or returning from a test centre;
- Haematomas or nerve damage following the taking of a blood sample;
- Infections following the taking of the blood sample.

Other types of damage that have to be considered are for instance the risk of social or psychological problems following the results of tests.

Therefore precautions should be taken if results are communicated back to trial subjects, in case those results are disturbing.

Communication, if at all, should be done by medically skilled professionals (e.g. genetic counsellors), possibly assisted by psychologists. This will allow to explain the consequences of the results on the health and life expectancy of the trial subject and to help in the mental processing of the results.

In most of the cases there will hopefully not be any (additional) problems, as either the results will not be thoroughly negative, or the patient will already have gone through the critical phase of the process.

The aforementioned precautions are mainly important when data from new trial subjects are taken into the project.

It should be emphasised that Assist is not a project where pharmaceutical research is being done or where other treatment methods are being developed and tested. It is an association study associating genotypic and phenotypic factors related to cervical cancer. In order to achieve this, Assist will resort to inference applied on real data from subjects. It is a project aimed at perfecting the efficient use of medical knowledge where cervical cancer serves as a practical example that could be extended to other types of cancers as well, especially cancers whose mechanisms are very complex and that cause a lot of casualties in target populations where it could be significantly reduced.

Wherever possible, existing data will be used. When the data is available but use for research purposes is not permitted, approval for use will be obtained.

In some cases, bio-samples (blood, saliva, PAP smears,...) may be available, but not the data in electronic form. In that case, the objective is to analyse the samples and consequently use the results for research.

When neither is available, the goal is to obtain samples directly from physical persons and analyse them according to the Assist objectives.

Normally this will be a one-time operation that will not require follow up of the patient. Follow-up is only necessary if the results were of such a nature that they require counselling and guidance of the patient.

When through participation in the Assist project, health risks, if any, are detected, the patient will be referred to a general practitioner or to a specialist, according to the custom way of working in the region or member state. Thereby, the patient is handed over from Assist to another care setting. The trial centre should of course commit to provide the care providers in that setting with sufficient information on the patient, possibly free of charge or as agreed in the informed consent procedure with the patient.

6.2 Provision of publicly available electronic communications services

Two directives may be of importance here and are often quoted in articles on privacy, including medical privacy.

- Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) (4) translates the principles set out in Directive 95/46/EC into specific rules for the electronic communications sector. In which Articles 5, 6 and 9 of Directive 2002/58/EC lay down the rules applicable to the processing by network and electronic communications services.

- Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available elec-

tronic communications services or of public communications networks and amending Directive 2002/58/EC.

Both directives are mainly intended for providers of tele- and data communication services such as telecom companies and Internet access providers.

It is unclear whether these directives apply to the Assist project, as they are mainly intended to retain data on the communication itself (and not on the content) for law enforcement and Intelligence (terrorism) purposes, but nevertheless do not specify clearly enough what is understood as 'publicly available electronic communication service'.

It is the opinion of this task group that the Assist services will not fall under this definition, as it is not publicly available and not intended as a communication service.

6.3 Ownership and IP rights

Ownership and IP rights are still rather uncharted territory. Biological samples and information about samples' donors, genealogies are likely to have commercial value. The general principle in law is that there are no property rights on the human body. Nevertheless entities that store samples and genetic databases have asserted control over samples and databases and over the management of IP rights.

In January 2005 the UK biobank issued a draft IP and access policy. The principle is that when IP arises out of research using the UK biobank, it is vested in the researcher or his institution or their assignee.

Other projects do not confer ownership rights to researchers or to the biobank, but permit the parties to obtain intellectual property rights over inventions obtained from genetic information.

There is a growing need for more explicit enforceable and coordinated international policy guidelines. The final ethics deliverable will present an update on these findings.

7 The Assist data resources

7.1 Availability and origin of the data

There are various directives, regulations, guides that may contain something useful with respect to the Assist project. Therefore, it is important to highlight the issues that concern the Assist project and to gain consensus over them by the project consortium and more in particular by the Assist ethical board.

The Assist project is a research project specifically targeted to study the association of genotypic and phenotypic factors related to cervical cancer.

In order to populate its data model and to provide the Assist inference engine with input, data can be obtained from in the following ways:

1. Data collected earlier for purposes of diagnosis, screening or treatment
 - a. For which no consent has been obtained for further use for research purposes (or that are not compatible with Assist finalities)
 - b. For which consent has been obtained for further use for research purposes (that includes finalities compatible with Assist)
2. Data collected earlier for research purposes
 - a. The consent includes finalities compatible with Assist
 - b. The consent does not include finalities compatible with Assist or no consent has been obtained
3. Biological samples (in Assist this will mainly be blood samples) collected earlier for purposes of diagnosis, screening or treatment
 - a. Consent for analysis of the samples to be used for research purposes compatible with Assist is available
 - b. Consent not available
4. Biological samples available that have been collected earlier for research purposes
 - a. Consent valid for Assist?
 - b. Consent not valid for Assist?
5. Additional samples required from persons from whom data had already been collected for research of for screening, diagnosis or treatment reasons
 - a. Demographic data available?
 - b. Demographic data not available?

6. Additional information from known persons, that does not require taking samples
 - a. Demographic data available?
 - b. Demographic data not available?
7. Samples and information from new persons
 - a. Target population demographic known
 - b. Target population not known but source where target population can be obtained known (and accessible for Assist purposes)

It is important for the data centres participating in the Assist project that they define their status and planning with respect to the various situations depicted above. Each of the centres has to gather information concerning the procedures of their respective Institutional Review Boards (IRBs) with respect to their data.

Although outside the scope of this workpackage, the centres and the consortium should unambiguously agree on the type and amount of data that they will supply to the Assist project. Agreement with WP 6 (Inference Engine) and with WP5 (Knowledge base) will be required concerning the suitability of that data. (WP2 remains neutral in that discussion). WP2 does not make any assumptions about the quality of the data as such. Indirectly, this may be requested by the respective IRBs as obtaining or using biological samples for research that does not achieve specific quality standards is not considered justified.

Each of the centres should be able to present a scenario on their planning to obtain data. WP2 does not make any assessments whether the planning and data are sufficient for carrying out the Assist objectives, but it guards over ethical and privacy issues and gives guidance wherever possible and required.

WP2 will concentrate on the following issues:

- Formulating requirements concerning ethics and privacy issues within Assist and in a generic way that can be used for similar types of projects.
- Summary of existing legislations and regulations in relation to the Assist issues within Europe.
- Consent issues and targeting of populations for additional or new input.
- Basic discussion and guidance of cross board issues in the project.
- Technical solutions and services for the protection of personal data.
- Technical solutions and services for the de-identification of personal data for research purposes.

7.2 Information content of the gathered data

In order to properly describe all kind of issues with the Assist project, it is important to get a better understanding of the nature of the data that will be entered or investigated within the project and more in particular the relevance of the information on the following items:

- Consequences of the findings on the current and future state of the data subject with respect to cervical cancer.
- Consequences of the findings on the current and future state of the data subject with respect to other diseases than cervical cancer.
- Consequence of the findings on the current and future state of genetic relatives of the data subject with respect to diseases and disease probabilities.

Genetic specialists should be able to describe what polymorphisms are being investigated and what they can reveal on the disease status and propensity of the data subjects and its genetic relatives.

The ethics task group does not suggest that research done in the project should cover all the situations described above, but that it should clearly define what it will or will not do. In some cases, genetic counselling of the data subjects may for instance be appropriate.

One of the issues in the discussions on genetic exceptionalism is the information content of genetic data. What phenotypic information can be obtained from the collected genetic information and in what degree can this lead to an indirect identification of data subjects or infer certain characteristics about data subjects (without uniquely identifying them), such as HIV?

A description of the genetic data collected in the Assist project will be added in the final deliverable, as well as the phenotype-genotype correlation with respect to privacy issues.

7.3 Recommendations on the collection of data

This context covers two different situations that have one element in common: Identifiable data is being used.

- Either the medical data is present but it was collected for the treatment of patients or as part of a screening process.
- The data is not present or only partly present and requires the collection of new or additional data that has to be obtained directly from the patient. In both situations, personal databases will have to be compiled so that the patient groups involved can be addressed and invited on a voluntary basis to participate in the ASSIST study.

This context or phase of the project will require the use of identifiable data at certain stages and by certain actors.

7.3.1 Re-use of existing data

The ASSIST ethical and privacy workpackage and the associated ethical board of the project advises each of the participating data-centres to follow common practices and procedures that are in place at their institutions.

7.3.2 Newly collected data or additional data

For some data centres, basic data sets are available that have been collected within the context of treatment or screening. However, in order to contribute in the ASSIST association study, these data may have to be complemented by additional data that can only be collected directly from the patient.

The involvement of the patient may consist of a simple physical examination and/or filling in a questionnaire. In other cases, bio samples from the patients will be required (mainly blood).

In the case physical examination or the collection of samples (blood, saliva,...) is required, the institution responsible for collection of the data and introduction of this data has to follow the common procedures used within its institution. It is up to the medical centres and the data centres, in case these are different legal entities, to define in a bilateral agreement their relationship from a personal data processing point of view and to define who is the controller of the data.

In addition to requirements stated in the local ethical guidelines or if not already included in them, the project consortium ethical board recommends the following:

- Medical examinations of data subjects should only be carried out by authorised medical personnel.
- Personal Data collected will be stored and protected physically and electronically and under the responsibility of the local data controller (medical centre or data centre). The responsibilities of the data controller and other parties involved will be defined in a privacy policy, if not already available.
- The Assist system shall not contain personal information of data subjects but shall only contain anonymised information and its engines/processes shall not process personal data but only de-identified data.
- The anonymisation process will be done by a state-of-the-art trust service provider. Its operation will be described in a privacy policy. The overall ICT design of the ASSIST project will be done in such a way that from a privacy point of view, the personal data domain and the anonymous domain will be separated.
- All partners directly involved in the processing and communication of data will be informed of the importance of the separation of the personal data domain and the de-identified domain and shall conform to the privacy that will explicitly prohibit actions that are aimed at re-identifying data subjects without approval from the Assist ethical board.
- In the cases where re-identification of data subjects whose data is in the anonymised database is required, the re-identification process will be technically carried out by the Trust Service provider and according to the procedures defined in the privacy policy of the TSP. Deviations for cases not included in the privacy policy will have to be documented and decided on by the project's ethical board, who can even decide to get into contact with the IRBs of the institutions.
- A legitimate case for re-identification could be the need of additional information form-specific data subjects. During the processing of the request for further information, the ASSIST system will never

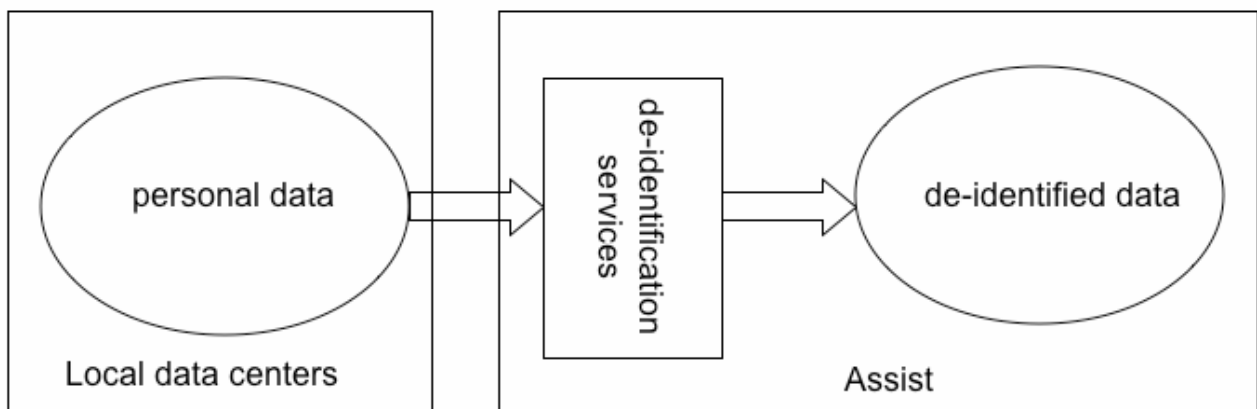
associate anonymous data with personal information. Instead, it will communicate its authorised request to the TSP, who will in turn pass the request to the local medical centre. Only the centre will be capable of identifying the patient. Even when identifying the patient, the local medical centre shall not be allowed to associate the personal data with de-identified data inside the Assist research registries or systems.

8 Privacy protection model for Assist Data

The data collected in the Assist Project consists of demographic, social, medical and genetic data.

There are two contexts that have to be considered from a privacy point of view in the data collection and data processing model: a context where the patient has to be identifiable by at least part of the actors in the process (this data is called «personal data») and a context where it is highly undesirable or even forbidden to be able to identify the data subjects whose data is being processed (this is called de-identified data). Data can only be passed from the personal data domain to the de-identified data domain by means of one or more de-identification services. The de-identification service ensures that its privacy policy for the project is enforced.

The overall privacy protection solution comprises both procedural and technical countermeasures.



8.1 The personal data domain

The personal data domain coincides with the data centres where the data is being collected. They are governed by their own institutional policies concerning privacy, running of trials and research and should take care of their own ICT-privacy policy enforcement. This will comprise traditional methods such as authentication and authorisation of users and access control to the data resources. According to privacy legislation the data controller (in the DPD sense) of the local institution is responsible for the privacy of the personal data of the trial subjects.

The Assist privacy protection model is independent on the source of the data that is being associated. This data can be present from earlier screening, diagnosis or treatment of patients, or specifically collected for the purpose of Assist or another research project.

The local institutions and data centres assume their responsibilities as data controller, principal investigator or whatever role applies in their context. They interface with their institutional review boards and follow research procedures, as is common practice in other studies than Assist.

The main issue is not if data that is available for research at the centres can be used within Assist for further research since no personal data should enter the Assist platform. Before entering the Assist platform, personal data will be filtered and de-identified.

The main issue will be if data that is collected for screening, diagnosis or treatment purposes can be used for research purposes. Theoretically, this should not pose a problem, as all privacy legislation relates to personal data, whereas Assist will only be using de-identified data. However, practice may show that institutions are unwilling to release such data for research purposes on the basis of a simple statement from the Assist consortium and will want their IRB to check for the privacy guarantees as well as for other sensitive issues for the use of data. Privacy concerns are not the only issued related to consent. It may be that some data subjects will be unwilling e.g. to let their data be used for research on birth control, etc.

It is unlikely, however, that the goal of cervical cancer research would be considered sensitive in those ethical terms and incompatible with the agreement of any sensitive human being. If they would object, most would not get involved in screening in the first place as screening is mostly done on a voluntary basis.

It is important to realise that article 4 of the DPD applies for each of the data centres and that a data controller of the data has to be identified. Article 4 states that national law is applicable.

8.2 The de-identified data domain

The de-identified data domain falls within the boundaries of the Assist platform. Data coming from the participating data centres has to pass through the anonymisation service (denoted in figure 2 in the description as «anonymisation layers»). The Assist platform will receive data from the data centres after it has been de-identified by the de-identification service(s). Per definition, the data is not considered personal data anymore, as discussed in the previous sections.

This does not mean that complementary protection has become superfluous. Privacy protection is an overall concern that does not end nor begin with a technical de-identification process. Moreover the data in the Assist platform should allow the grouping of data per case. This grouping can be done per data subject (remark that the data subject is not identifiable anymore but the system can tell which data belongs together on a per data subject basis, per data centre, etc.)

Protection of privacy is not the only concern; data integrity should also be safeguarded. Applying traditional security access mechanisms to the Assist associate databases can do this. Another concern is the protection of the property rights of the Assist assets.

Amongst the security precautions taken should be:

- Data shall only be accepted by the Assist platform if it has been de-identified by an authorised anonymisation service.
- Data shall only be accepted by the Assist platform if it originates from an authorised data centre (this is also checked by the anonymisation service before it accepts the data for de-identification).

- The mechanism for the grouping of the data (e.g. pseudonymisation) will be hidden from the users of the platform. These mechanisms or codes are intended for the query processes and for the data management functions within the system. In normal operating circumstances, data with fine granularity that can be linked will not be extractable by the users through the user interfaces.
- The use of the functions of the Assist platform will only be available to authenticated and authorised users. These will be known and registered at the Assist platform and their access to Assist resources defined. This countermeasure will be based on traditional security countermeasures such as authentication, access control to platform resources, etc.

8.3 The de-identification services

The anonymisation service(s) are the cornerstone of the privacy protection measures in the Assist platform. They are complemented, as has been mentioned in previous sections by traditional security mechanisms such as authorisation, authentication, access control of resources, VPN connections between entities to protect communication, etc.

The following section deals with the concept of de-identification.

The de-identification services are a separate independent entity. They are the gateway between the personal data domain and the Assist de-identified and consequently anonymised data domain.

The role of the de-identification services is not only a technical one. The goal of this service is enforcement of privacy policy statements that are aimed at de-identifying personal data. Other parts of the privacy policy will have to be enforced by the Assist platform by both technical means (as described in previous sections) and by procedural means (e.g. authorised users will be informed of the constraints to use the data due to privacy protection and what not to do).

Communication is usually one-way, from the personal data domain to the de-identified data domain. It is however possible and in some cases even desirable to apply communication in the other direction. Examples that require a reversal of a direction is e.g. a request from the Assist researchers to obtain additional data from specific data subjects or to double check suspicious values. Other reasons, slightly outside the scope of Assist would be to draw the attention of the local investigators to alarming findings about a data subject, or simply to systematically report back to participating data subjects, should this be useful and part of their «deal» for obtaining consent.

Without going into details it can be stated that an independent de-identification service is able to provide such a feedback channel without the service itself or its personnel being able to obtain the link with the data subject's identity.

It is important that in a study where potentially several data centres can be feeding data to the Assist platform the de-identification services are delivered by an independent trust service provider (TSP), also called TTP (Trusted Third Party).

9 A conceptual model for de-identification of personal data

9.1 Objectives of personal privacy protection

The objective of personal privacy protection is to prevent the unauthorised or unwanted disclosure of information about a person. Personal privacy protection is a sub-domain of generic privacy protection that per definition includes other privacy sensitive entities such as organisations. As personal privacy is the best-regulated and pervasive one, this conceptual model focuses on personal privacy. Protective solutions designed for personal privacy can be transposed for the privacy protection of other entities as well.

The scope of privacy protection of personal data is wider than only the protection of personal data but this model looks in particular at the protection of personal data through de-identification.

There are two strands in the protection of personal data: one that is oriented towards the protection of personal data in interaction with on-line applications (e.g. web browsing), and another strand that looks at the protection of collected personal data in databases.

The conceptual model is designed from the point of view that research data can be extracted from e.g. treatment or diagnostic databases, provided that the identities of the data subjects are not disclosed. Researchers work with “cases”, longitudinal histories of patients collected in time and/or from different sources. In its final form, the cases will be used in a completely anonymised way. For the aggregation of various data elements into the cases, it is however necessary to use a technique that enables aggregations without endangering the privacy of the data subjects whose data is being aggregated. This can be achieved by de-identification of the data subjects. Pseudonymisation is a type of de-identification.

9.2 Personal data vs. de-identified data

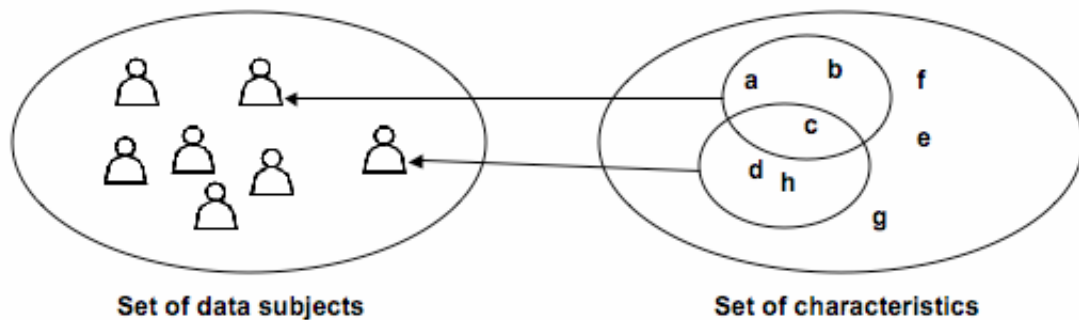
9.2.1 Definition of personal data

De-identification hinges around the notion of “personal data”.

According to the “Data Protection Directive (Directive 95/46/EC) of the European Parliament and of the Council of 24th October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (DPD), *'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.*

The DPD definition will be used as reference in this section of the document.

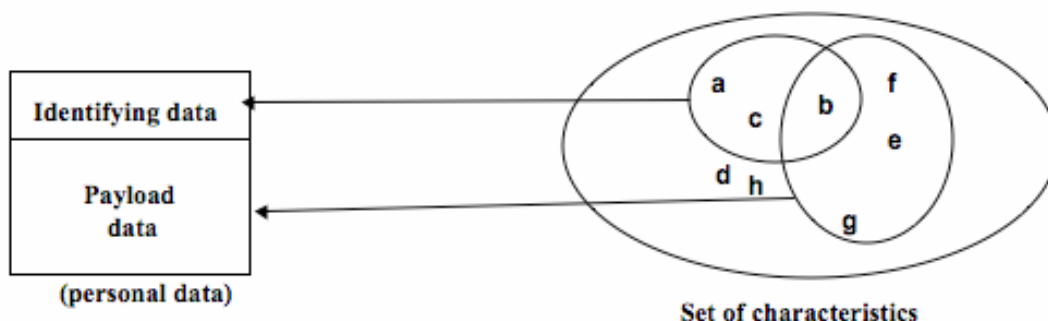
9.2.2 The concept of identification



A data subject has a number of characteristics such as name, date of birth. A data subject is identified within a set of data subjects if he can be singled out among other data subjects. That means that the set of characteristics associated with the data subject is unique. In some cases, only one single characteristic is sufficient to identify the data subject (e.g. if the number is a unique national registration number). In other cases more than one characteristic is needed to single out a data subject, e.g. when the address is used of a family member living at the same address. Some associations between characteristics and data subjects are more persistent in time (e.g. a national security number, date of birth) than others (e.g. an e-mail address).

Identity management comprises two complementary branches:

- In some applications a reliable linking of data of the same patients is important (e.g. for the continuity of care). A secondary (but equally important) requirement is that the personal privacy gets maximum protection. In these applications, the focus is on the ability to combine privacy protection with the ability to follow the patient throughout various processes and, wherever needed, to be able to identify the patient.
- In other applications, mainly for research purpose, the building of anonymous case libraries is important. From this perspective, the ability to identify patients is highly unwanted.



Personal data is a set of characteristics that can be associated with exactly one person.

A personal data set can be split up in two parts:

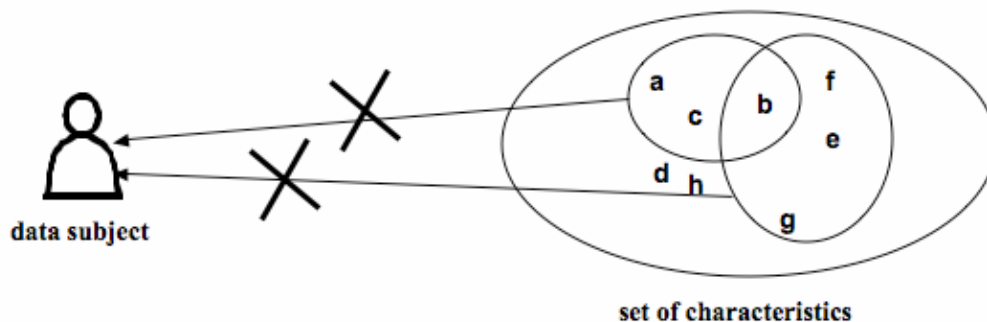
- A data part, containing characteristics that in itself do not allow unique identification of the data subject. This is commonly referred to as the “payload”. The payload contains anonymous data.
- An identifying part that contains a set of characteristics that allows unique identification of the data subject. (e.g. demographic data, social security number).

Normally the characteristics data set contains demographic data for one or more patient identifiers, whereas the payload contains medical or other care data.

9.2.3 The concept of de-identification

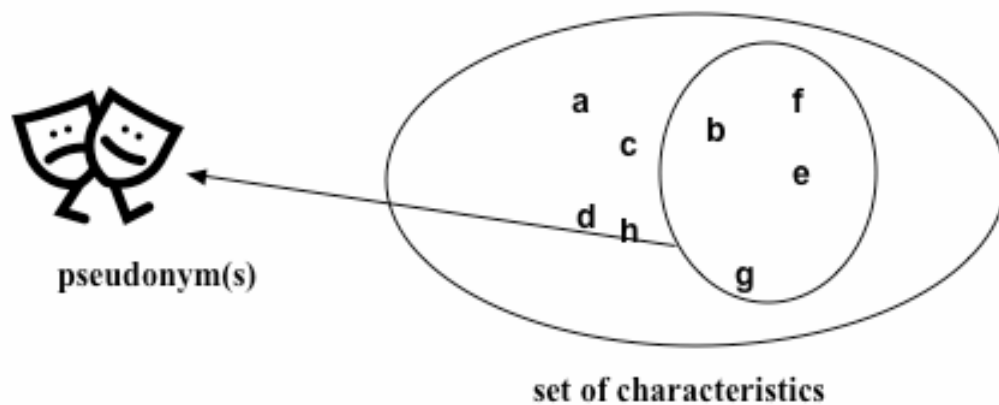
De-identification is a concept whereby the uniqueness of the link between the data set and the data subject is removed. This can be done in two different ways:

- by removing characteristics in the associated characteristics-data-set so that the association is not unique anymore and relates to more than one data subject.
- by increasing the population in the data subjects set so that the association between the data set and the data subject is not unique anymore.

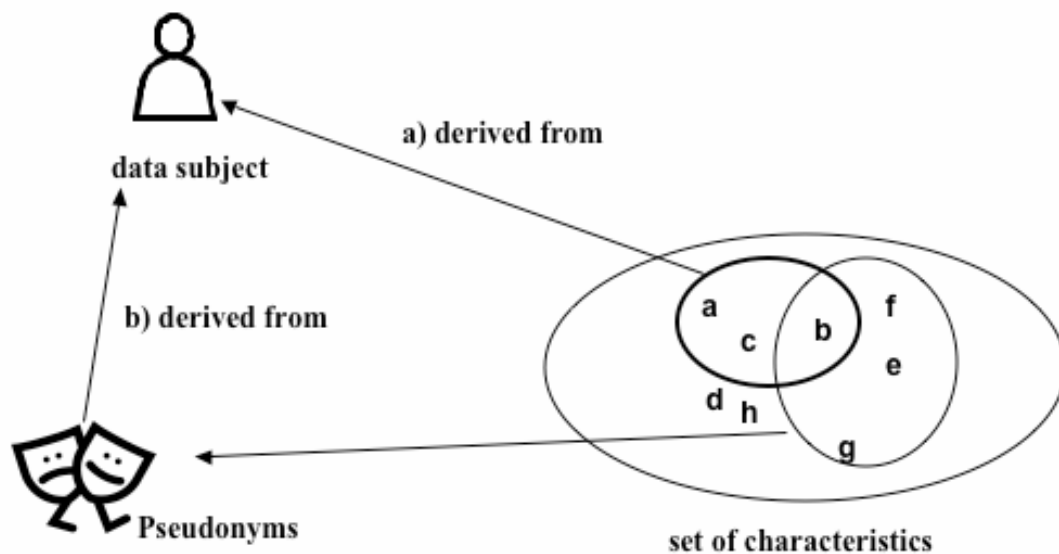


Pseudonymisation is a particular type of anonymisation that, after removal of the association with a data subjects, adds an association between a particular set of characteristics relating to a data subject and one or more pseudonyms.

In irreversible pseudonymisation, the conceptual model does not contain a method to derive the association between the data-subject and the set of characteristics from the pseudonym.



From a functional point of view, pseudonymous data sets can be anonymously grouped together as the pseudonyms allow associations between (anonymous) data sets, while disallowing association with the data subject. As a result it becomes possible, e.g. to carry out longitudinal studies in an anonymous way or to build cases from real patient data in an anonymous way.



In reversible pseudonymisation, the conceptual model includes a way of re-associating the data set with the data subject.

- a) derived from the payload,
- b) derived from the pseudonym or in lookup table.

It is understood that the reversal of the pseudonymisation can only be done by an authorised entity in controlled circumstances. It is outside the scope of this section of the document to enter into detail about policy aspects. This section only describes the conceptual approach of how it can be done.

9.3 Real world identifiability and anonymity

9.3.1 Rationale

The previous section depicts a conceptual approach where concepts such as “associated”, “identifiable”, “anonymous”, etc. are ideally defined.

In practice, identifiability (i.e. the ability to associate a data set with a data subject) is often difficult to assess. The introduction of privacy protection techniques is often countered by arguments on the concepts of identifiability and anonymity. Therefore the conceptual model should further refine these concepts.

Recital 26 of the DPD states that *“to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible”*.

The recital focuses, as the definition of personal data itself, on “identification”, i.e. the association between data and data subject.

As privacy protection measures described in privacy policies consist of a set of complementary organisational, procedural, infrastructural and ICT-security measures, the boundaries of the liability and the opportunities for violations of privacy legislation and ethics can be fairly well assessed for the controller. Common and contractual law towards the data subject often contractually binds the controller. The relationships between the controller, data subjects, processors of the data, etc. are known and they can be informed or should de-facto be aware of privacy ethics and legislation.

Statements such as “all the means likely reasonable” and “by any other person” are rather vague. Since the definition of “identifiable” and its pendant “anonymous” depend upon the undefined behaviour (“all the means likely reasonable”) of undefined actors (“by any other person”) the conceptual model in this document should include “reasonable” assumptions about “all the means” likely deployed by “any other person” to associate characteristics with data subjects.

The conceptual model will be refined to reflect differences in ‘anonymity’ and the conceptual model will take into account “observational databases” and “attackers”

9.3.2 Levels of anonymity

Current definitions lack precision in the description of terms such as anonymous, identifier, identifiable. It is unrealistic to assume that all imprecision can be removed, but precision can be greatly enhanced by introducing a classification of anonymity that takes into account the likelihood of identifying capability of data as well as by a clear understanding of the entities in the model and their abilities to identify.

The result of this refinement will be that privacy policies, for instance, will be able to define more precisely what is meant by “anonymity” and as a result legal issues will be easier to assess.

The classification attempt below is partly based on existing practises, but further refinement is encouraged, especially since quantification of re-identification requires the establishment of mathematical models. The establishment of re-identification risk techniques is outside the scope of this document.

However, by applying definitions of levels that can be unambiguously defined, pointless discussions on ‘anonymity’ can be avoided and policy language can gain precision by shifting the boundaries of the vague.

9.3.2.1 Level 1 anonymity: Removal of clearly identifying data («rules of thumb»)

A first, intuitive level of anonymity can be achieved by applying rules of thumb. This method is usually implicitly understood when de-identifying data is discussed. In many contexts, this first level of anonymity may provide a sufficient guarantee and is even the only one that can easily be achieved.

As an example of a level 1 anonymity rule of thumb, the HIPAA rule is given. The HIPAA rules require that for data to be considered de-identified, the following elements should be removed from the data:

- Names (individual, employer, relatives, etc.)
- Address (street, city, county, precinct, zip code – initial 3 digits if geographic unit contains less than 20,000 people, or any other geographical codes)
- Telephone and Fax numbers
- Social Security numbers
- Dates (except for years)
 - Birth date
 - Admission date
 - Discharge date
 - Date of death
 - Ages >89 and all elements of dates indicative of such age (except that such age and elements may be aggregated into a category “Age >90”)
- E-mail addresses
- Health Plan Beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle Identifiers and Serial numbers (e.g., VINs, License Plate numbers)
- Device Identifiers and Serial Numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric Identifiers (e.g. finger or voice prints)
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

9.3.2.2 Level 2 anonymity: Static, model based re-identification risk analysis

The second level of anonymity takes into account the global data model and the data flows inside the model. This level includes a static risk analysis that checks for re-identification vulnerabilities by different actors. This level may e.g. include the removal of absolute time references. A reference time marker “T” is defined as e.g. the admission of a patient for an episode of care and other events, e.g. discharge is expressed with reference to this time marker.

In level 2 anonymity, the databases are not populated with data, hence the denomination of “static”.

An important element of anonymity levels 2 and higher is that ‘Attackers’ can be part of the model. Depending on the risk analysis method used, various assumptions about the attackers can be used.

An attacker is an entity that gathers data (authorised or unauthorised) with the aim of attempting to attribute the gathered data in an unauthorised way to data subjects and thus obtaining information that he is not legally entitled to. From a risk analysis point of view, data gathered and used by an attacker is called “observational data”.

Note that the disallowed or undesired activity by the attacker is not necessarily the gathering of the data, but the attempt to attribute the data to a data subject in an unauthorised way.

A risk analysis model may include assumptions about attacks and attackers. E.g. in some countries it may be possible to legally obtain discharge data by entities that are not implicitly involved in the care or associated administration of patients. The risk analysis model may take into account the likeliness of the availability of specific data sets.

From a conceptual point of view, an attacker brings data elements into the model that in the ideal world are not supposed to be there.

9.3.2.3 Level 3 anonymity: continuous re-identification risk analysis of live databases

The re-identification risk can be seriously influenced by the data itself, e.g. by the presence of outliers. A static model based risk analysis cannot quantify the vulnerability due to the content of databases; therefore running regular risk analyses on populated models provides a higher level of anonymity.

9.4 Remark on privacy threats

Privacy protection is about protecting characteristics of a data subject contained in personal information and not only about protecting the identity linked to a specific data item. A conceptual model for privacy protection should take into account the following threats:

- Identification or re-identification:
 - Given a data item, establish the link with the data subject in a set of data subjects.
 - Given a data subject, establish the data items that are associated with it in a set of data items.
- Inference of group characteristics
 - Given a data subject, verify if a set of characteristics is associated with the data subject.
 - Given a data subject, verify if a set of characteristics is not associated with the data subject.

The EU DPD focuses on the property of identifiability. However, the scope of privacy protection encompasses identification because inference of group characteristics can also reveal personal information that in itself is sensitive or that can lead to further identification of a data subject (e.g. the assessment if a data subject is or isn't HIV positive does not necessary imply that the personal data of this data subject has been found).

At the source side that is submitting personal data to the pseudonymisation service, data is split in an identifying part and in an anonymous part. What is considered identifying data and what is considered anonymous data depends on the target level of anonymity in the security policy of the data collection project. Preparation of the submission of the data to the service consists of:

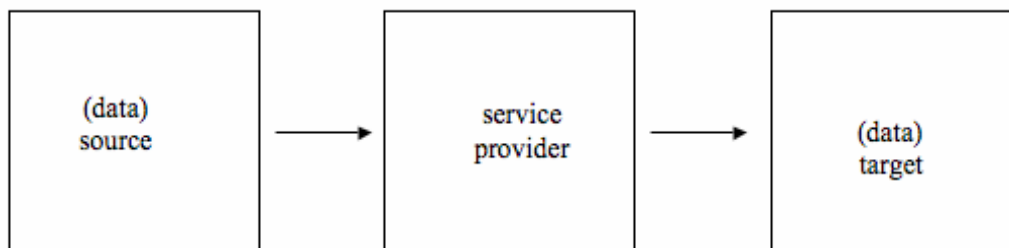
- Encrypting the payload data so that it passes through the pseudonymisation service in encrypted form so that it can only be decrypted at the target site
- Encrypting the identifying data to protect it during transport to the pseudonymisation service. It is decrypted at the pseudonymisation service site

9.5 The pseudonymisation process

9.5.1 Entities in the model

The pseudonymisation model contains three entities that are denoted as:

- source
- target
- de-identification service (service provider)



These entities can be complemented by e.g. authentication services, key escrow services or other services required by the process model.

A source is an entity that performs the following functions:

- Structure the data for submission to the de-identification service. The de-identification service has to know what it is expected to do with a data element. This can be done by either tagging the data elements or by positioning the data elements in defined location that will each be processed in a pre-defined way.
- Submit the data to the de-identification service. Calling a de-identification service client from the application running at the source side does this.
- Read and follow-up the result code from calling the de-identification service. This can consist of simply logging the result in case of success or of retrying or sending warnings in the case of failure and depending on the return information.

A target is an entity that receives de-identified data from the de-identification service and that takes care of the further processing of the data.

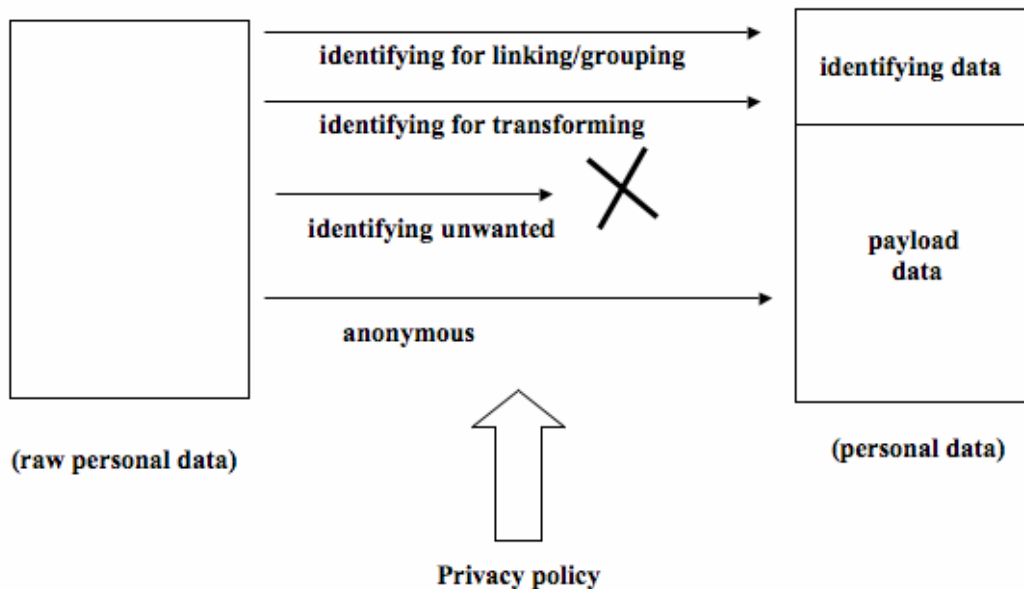
- decryption of the data received from the de-identification service.
- Insertion of the received data into the target repositories according to the rules of the system (checking for doubles, updates, etc...)

The de-identification service is the entity that performs the de-identification process. This entity will be defined more in detail in this document. Important to remark is that the pseudonymisation service does not store data. As a result, its operation is 'stateless' with respect to various sessions submitted to it by a data source. All information needed to base its policy decision upon during a session has to be present in the session data.

9.5.2 Preparation of data

Before data can be submitted to the de-identification service, it has to be prepared at the source side. The preparation is necessary in order to apply the privacy principles defined in the privacy policy.

The conceptual model for the use of de-identification services requires that the data be split up in a part containing identifying data and in another part containing nothing but anonymous data.



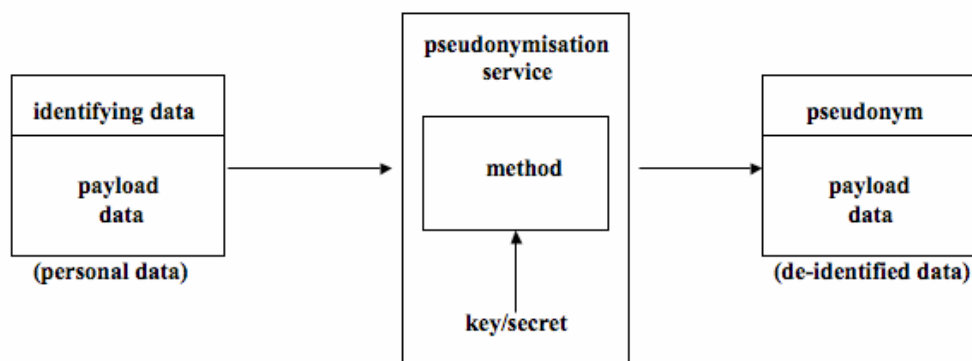
The structuring can be done by tagging the data elements, by creating a table with vectors to the data elements or by putting the data elements in a pre-defined location.

The following preparations can be distinguished in accordance with the conceptual model on the levels of anonymity:

- Data elements that will be used for linking, grouping, anonymous searching, matching, blinding, etc. have to be indicated and marked in such a way that the pseudonymisation service knows where to find them and how to handle them.
- Depending of the privacy policy, elements that need specific transformations, e.g. for changing absolute time references into relative time references, date of births into age groups, need similar marking.
- Identifying elements that, according to the privacy policy, are not needed in the further processing in the target applications shall be discarded.
- The anonymous part of the raw personal data is put into the payload part of the personal data element.

9.5.3 Processing steps

TTP based de-identification services can be implemented in several ways. Two typical implementations are the batch processing mode and the interactive mode. Hybrid forms are of course thinkable. As the Assist system takes form, an appropriate solution will be devised to meet the needs of the project.



The basis steps of a de-identification process consists of:

1. The TTP service parses the header and performs the tasks defined in its policy. This comprises the de-identification of data items, calculation of relative dates, removal of data items, encryption of specific data items as is defined in the privacy protection policy. Normally the data in the payload passes through the TTP service in encrypted form. Consequently, the content of the payload is not visible to the TTP. This is the preferable way of operation for the TTP. The policy can, however, define otherwise, such that the content is parsed as well. Typically the payload parsing is done to check for unwanted content (e.g. identifiers in the data items). All checking is done in a “stateless” way, without reference to previously received data. The TTP does not store data it has previously processed and therefore it cannot compare or check with that data. Only the current Session can be taken into account.

2. Once the processing is done, the TTP sends the de-identified data to the data repository through a secure channel. The channel has to be secure, not because of the risk to compromise the anonymity of the data but in order to safeguard the integrity of the data repository. One of the risks of data repository containing de-identified data is that wrong data is deliberately injected. When the repository application receives the data, it applies its business rules to it in order to incorporate the data into the repository. This may include checking for double entries, check for missing entries, required acknowledgement procedures, etc.

10 The ASSIST data collection model

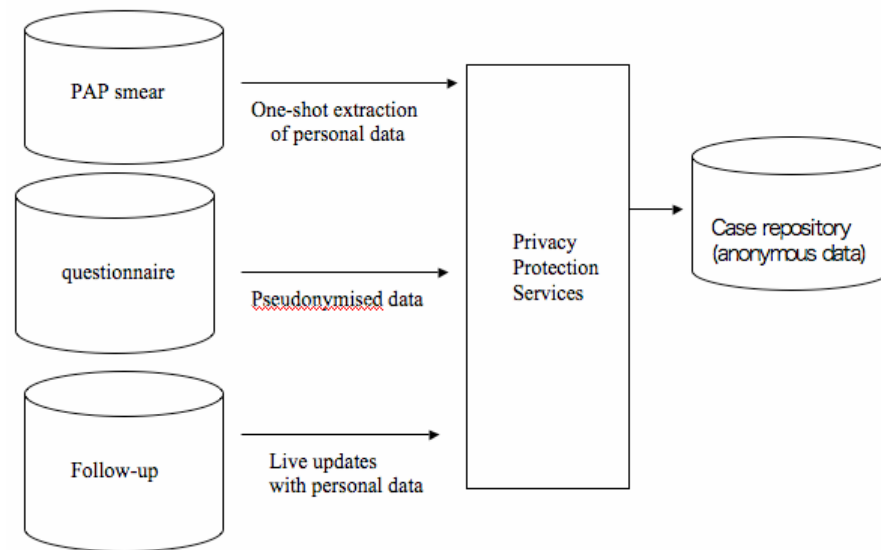
In the ASSIST project, three partners (Germany, Greece, Belgium) in three countries are expected to deliver data to the project for research purposes through their data centres.

Data of the data subjects may come from different data sources and will have to be linked while maximising their privacy protection. At the time of writing, the type and origin of data sources per country is not yet defined. The objective of ASSIST however is to provide a generic approach to the exploitation of research data. Therefore this document will assume three different data sources that differ in the gathering of the personal data and related finality and sensitivity:

- a PAP smear results database
- questionnaire results
- continuity of care/follow-up

Privacy protection services can reconcile the following conflicting requirements:

- protection of the privacy of the data subjects
- linking of data coming from different sources
- integration of new data without compromising the privacy of the data subjects



The end result is a live repository containing “cases” for longitudinal follow-up. This model is also applicable if the platform does not contain long-term repositories but only associates data «on the fly» for running hypothesis in the inference engine. In that case, the repository has a temporary lifetime, long enough for running the studies.

Anonymous questionnaires allow the gathering of highly sensitive data. It has not been decided yet if the use of anonymous questionnaires is required in this project. Their use may however motivate patients to participate, where otherwise they may be inclined to refuse participation or give false information because of the highly sensitive nature of it.

Patients included in the research will be registered in only one of the countries and therefore integrating data from the various centres will not be a requirement. This is not so much a technical issue, but rather a legal and ethical issue.

Even if Assist would not be using anonymous questionnaire, the option should be present in the proposition of generic solutions for similar types of projects.

11 Informed consent

This section explicitly deals with informed consent issues in the Assist project. The starting point is the existing situation in the three participating centres. It is recommended that centre follow their own internal IRB procedures concerning research and consent. However, the Assist ethical workgroups wants to summarise a number of generic principles that informed consent should provide answers to.

In Germany, Campus Benjamin Franklin, Campus Charité Mitte and Campus Virchow-Klinikum use a common consent form. In fact, consent is part of the patient agreement («Stationärer Behandlungsvertrag») in

which is stated that a University hospital has teaching and research assignments. It is stated that within this context, the patient implicitly agrees that his stored data can be consulted.

A separate paragraph deals with body samples that have been taken for the necessity of diagnosis or treatment. It states that as the result of that process «rest material» will be available. The patient is informed of this and an explicit «yes» or «no» is required. When consenting, the form states that restmaterial and related clinical data can be used anonymously for research, including by external organisations.

A further section is foreseen in the case the patient is treated for cancer, where is explicitly stated that personal clinical data can be sent to the official cancer centre, designated in the region of the patient. The patient has to accept or reject.

12 Conclusions

Assist clearly distinguishes domains with personal data and domains that only contain de-identified data. Research data shall only cross the boundary between the two domains in either direction by means of an independent Trusted Third Party in control of re-identification and de-identification. Privacy protection is not only a matter of technical measures, but above all a matter of privacy policy. The privacy policy shall be enforced by both technical and organisational measures. Traditional security services such as digital signatures, encryption for confidentiality, authentication, access control shall complement the deployment of state-of-the-art de-identification solutions.

Each of the research data providers shall observe the policies of its own IRB, be it for existing data or new data.

13 References

- [1] DIRECTIVE, concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), 2002/58/EC of the European Parliament and of the Council of 12 July 2002.
- [2] RECOMMENDATION, on the Protection of Medical Data, R(97)5 of the Committee of Ministers to Member States, Adopted by the Committee of Ministers on 13 February 1997 at the 584th meeting of the Ministers' Deputies.
- [3] GUIDELINES for good clinical practice (GCP) for trials on pharmaceutical products, World Health Organization WHO Technical Report Series, No. 850, 1995, Annex 3;
- [4] NOTE FOR GUIDANCE on good clinical practice (CPMP/ICH/135/95), The European Agency for the Evaluation of Medicinal products;
- [5] DIRECTIVE 2001/20/EC of the European Parliament and the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use;
- [6] GUIDELINES, International Ethical Guidelines for Biomedical Research Involving Human Subjects;
- [7] GUIDELINES, 1991 International Guidelines for Ethical Review Of Epidemiological Studies.
- [8] DECLARATION OF HELSINKI, Ethical Principles for Medical Research Involving Human Subjects, World Medical Association, Tokyo version, 2004.
- [9] RECOMMENDATIONS, 25 Recommendations on the ethical, legal and social implications of genetic testing, European Commission, Directorate General for Research, Directorate C- Science and Society, Unit C3: Ethics and Science, Brussels, 2004
- [10] Beskow, Laura, et.al., Informed Consent for Population-Based Research Involving Genetics, JAMA, November 14, 2001-Vol286 No.18, pp. 2315-2321
- [11] Reymond, Marc, et.al., Informed Consent for Molecular-Based Diagnostic and Prognostic Studies in the Cancer Patient, Digestive Diseases, Review article
- [12] Maschke, Karen J., Navigating an ethical patchwork-human gene banks, Nature Biotechnology, Volume 23, number 5, May 2005, pp. 539-545
- [13] Fuller, B.P, et.al., Ethics, Privacy in Genetic Research, Science 27 August 1999, Vol. 285, no 5432, pp. 1359-1361.