INFRASTRUCTURE AND INTEGRATED TOOLS FOR PERSONALIZED LEARNING OF READING SKILL

D3.2 – Privacy and Security

Document identifier	**iRead_D3.2_PrivacyAndSecurity_final**
Date	**2017-05-02**
WP	**WP3**
Partners	**NTUA, KNOW, ULBS, PATAKIS, DOUK, BC, WISDOM**
WP Lead Partner	**UGOT**
Document status	**Final**

Grant Agreement number:
731724 — iRead H2020-ICT-2016-2017/H2020-ICT-2016-1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 731724

Deliverable Number	D3.2
Deliverable Title	Privacy and Security
Deliverable version number	Final
Work package	WP3
Task	Task 3.2 Privacy and security issues
Nature of the deliverable	Report (R)
Dissemination level	Public
Date of Version	2017-05-02

Author(s)	Antonios Symvonis
Contributor(s)	Chrysanthi Raftopoulou, Cantemir Mihu, Ioan Mihu, Dorin Sima
Reviewer(s)	Konstas Karpouzis

Abstract	In this document, an in-depth analysis of privacy and security risks for the iRead system is presented. More precisely, specific data protection and security issues that arise for such intelligent technologies are identified, owing to the fact that processing user's personal data is prerequisite for providing the desired services; since achieving an increased level of personal data protection is of fundamental importance, we investigate how such data protection issues can be effectively mitigated in the iRead system, from both the legal and technical aspect.

Keywords	Privacy, security, data protection, learner profile, stored data, consent, user authentication, access control, data protection impact assessment.

Document Status Sheet

Issue	Date	Comment	Author
v0.1	2017-04-03	Outline, Introduction, Data security requirements, Personal data protection framework, Personal data processing in iRead	A. Symvonis, C. Raftopoulou
v0.2	2017-04-10	Technologies for Security	C. Mihu, I. Mihu, D. Sima
v0.3	2017-04-20	Personal data protection and information security	A. Symvonis
v0.4	2017-04-24	Reviewing	K. Karpouzis
Final	2017-04-26	Final editing	C. Raftopoulou
Final	2017-05-02	Incorporating minor comments from partners	C. Raftopoulou

Table of content

1. Introduction 6

2. Data Security Requirements 7

2.1. Introduction 7
2.2. Security Objectives 7
2.3. Security Services and Mechanisms 8
2.3.1. Symmetric cryptography 8
2.3.2. Asymmetric (public key) cryptography 9
2.3.3. Cryptographic hash functions 9
2.4. Technologies for Security 10
2.4.1. Transport Layer Security 10
2.4.2. Authentication mechanisms 11
2.4.3. Access control 12
2.4.4. Data at rest encryption 13

3. Personal Data Protection Regulatory Framework 14

3.1. Introduction 14
3.2. The Data Protection Directive 14
3.2.1. Basic definitions 14
3.2.2. Basic principles 15
3.3. The National Data Protection Laws 16
3.4. The General Data Protection Regulation 18

4. Personal Data Processing in iRead 21

4.1. Introduction 21
4.2. Use Cases 21
4.3. Components of the iRead System 22
4.4. Language Support 22
4.5. Domain Model 23
4.6. Profile of Learner 24
4.7. Users of iRead 24
4.8. Stored Data 25
4.8.1. User details 25
4.8.2. iRead learner usage-data 27
4.8.3. iRead data access 28

5. Personal Data Protection and Information Security 29

5.1. Introduction 29
5.2. Legal Requirements – An Assessment 29
5.2.1. Legitimacy and purpose limitation 30
5.2.2. Data minimization 30
5.2.3. Consent 32
5.2.4. Special categories of data 33
5.2.5. Transparency and data subject's rights 33

5.2.6. Data transfer 33
5.2.7. Other data controllers' obligations 34
5.3. Management of Risks Relating to the Security of Personal Data 34
5.3.1. Security requirements of iRead system 34
5.3.2. User authentication 34
5.3.3. Confidentiality and integrity though transmission 36
5.3.4. Access control 37
5.3.5. Data at rest security 37
5.4. Data Protection Impact Assessment – Security Policy 38
6. Conclusions 39
References 40
Appendix 42
Current Technologies for Security 42
Acronyms 46

1. Introduction

In this document, an in-depth analysis of privacy and security risks for the iRead system is presented. More precisely, specific data protection and security issues that arise for such intelligent technologies are identified, owing to the fact that processing user's personal data is prerequisite for providing the desired services; since achieving an increased level of personal data protection is of fundamental importance, we investigate how such data protection issues can be effectively mitigated in the iRead system, from both the legal and technical aspect.

It should be pointed out that the European Personal Data Protection legislation will meet a significant change within 2018 – i.e. in the process of the iRead project; namely, the General Data Protection Regulation (Regulation (EU) 2016/679) will replace the existing Data Protection Directive (Directive 95/46/EC). **Therefore, in a proactively manner, the whole iRead system will take into account, from its early stages, not only the current legislation but also the new Regulation, so as to ensure full compliance with the legal requirements on personal data protection at any time.**

The structure of the document is as follows:

In Chapter 2, the basic definitions and concepts of information security are introduced. Several powerful security mechanisms and protocols are also described, since they provide resistance against specific security attacks that should be thwarted in the framework of the iRead operation.

Chapter 3 presents the basic elements of the European Personal Data Protection Legislation, including both the Data Protection Directive (Directive 95/46/EC) and the General Data Protection Regulation (Regulation (EU) 2016/679). This presentation is necessary for identifying the data protection risks that arise from the processing performed though the iRead system.

In Chapter 4, a detailed description of the iRead personal data processing is given. More precisely, the exact types of personal data that will be processed, as well as the exact kind and characteristics of their processing, are presented. Such a description is prerequisite for performing an assessment of the data protection issues that may arise.

Chapter 5 constitutes the core component of this document; in this Chapter it is described how personal data protection can be ensured for the iRead system, by adopting appropriate measures to confront the identified risks. These measures consist of both legal and technical controls that should put in place. In other words, this Chapter – in conjunction with Chapter 4 – corresponds to the first steps of a Personal Data Protection Impact Assessment. The output of this Chapter provides significant guidance for proper designing the iRead system so as to establish compliance with personal data protection legislation.

Finally, the conclusions of the above analysis are given in Chapter 6.

If needed, the present document will be appropriately updated during the iRead implementation period, since any Data Protection Impact Assessment is an ongoing process and, thus, several design decisions with regard to personal data protection and security will be validated and reviewed.

2. Data Security Requirements

2.1. Introduction

In this Chapter, we introduce the basic definitions and aspects covering the so-called *information security*. Although there are several definitions of the notion of information security (each of them highlighting a different aspect), it is commonly agreed that information security rests with the protection of data, either data at rest or data at motion.

In Information Technology (IT) systems, the IT security is often considered as strict technological issue, requiring appropriate technology-related solutions. However, several other perspectives also need to be considered. More precisely, data security requirements and countermeasures may lie in the following categories:

1. Technical (e.g., authentication, access control, data encryption);
2. Physical (e.g., issues such as physical access to systems, safeguards against environmental incidents etc.);
3. Organizational (appropriate controls relating to the people that use systems, the need for a security policy and/or a data recovery plan etc.)
4. Legal (the need to comply with relevant legislation, such as data protection law).

Data security is a main goal for the iRead project, especially in the light of the General Data Protection Regulation that is to be enforced in 2018 (see Chapter 3, page 18). This Chapter introduces the main definitions on data security, as well as the main technical security mechanisms, that will be considered in the design and operation of the iRead project.

2.2. Security Objectives

Information security is commonly associated with three main principles or characteristics: confidentiality, integrity, and availability (CIA triad) [21].

Confidentiality refers to the prevention of unauthorized information disclosure. In other words, only strictly authorized people (e.g. the legitimate recipients of a message) should have access to read the data. A loss of confidentiality is the unauthorized disclosure of information.

Integrity relates to the prevention of unauthorized information modification or destruction. Users must be able to trust their systems and be confident that the same information can be retrieved as was originally entered. Furthermore, data should only be modified by authorized parties in authorized ways. The term integrity is also referred to ensure information authenticity (the property of being genuine and being able to be verified and trusted). A loss of integrity is the unauthorized modification or destruction of information.

Availability relates to the need for data to be accessible and usable (by authorized parties) in a timely and reliable manner. This necessitates both the prevention of unauthorized withholding of information, as well as adequate safeguards against system failure. A loss of availability is the disruption of access capability or use of information or an information system.

In addition to the classic CIA triad, a fourth objective called **accountability** is often included. Accountability allows for tracing the actions of an entity, so as to support nonrepudiation,

deterrence, intrusion detection and prevention, and after-action recovery and legal action[1]. Without some form of accountability, it is impossible to directly attribute an action to an individual or to be able to prove that an individual did not perform a specific action. Accountability is usually achieved through appropriate log files.

2.3. Security Services and Mechanisms

Towards assessing the security needs, which is essential for choosing the proper security products and policies, a systematic way of defining the requirements for security and characterizing the approaches to satisfying those requirements is required. Such a systematic way is defined in the Recommendation X.800 (*Security Architecture for OSI).* The OSI security architecture focuses on security attacks, mechanisms, and services. These can be defined as follows [21]:

1. Security attack: Any action that compromises the security of information.
2. Security mechanism: A process (or a device incorporating such a process) that is designed to detect, prevent, or recover from a security attack.
3. Security service: A processing or communication service that enhances the security of the data processing. The services are intended to counter security attacks, and they utilize one or more security mechanisms to provide the service.

X.800 divides the security services into five main categories, namely authentication (the assurance that the communicating entity is the one that it claims to be), access control (the prevention of unauthorized use of a resource – i.e. this service controls who can have access to a resource, under what conditions access can occur, and what those accessing the resource are allowed to do), data confidentiality, data integrity and nonrepudiation (provides protection against denial by one of the entities involved in a communication of having participated in all or part of the communication). It should be pointed out that X.800 defines availability as the property of a system or a system resource being accessible and usable upon demand by an authorized system entity, according to performance specifications for the system – i.e. X.800 treats availability as a property to be associated with various security services.

There are several security mechanisms to address the aforementioned security services; cryptographic algorithms (including the notions of hash functions and digital signatures) constitute the main security mechanism covering confidentiality, integrity and authentication, whilst they may also play crucial role in access control.

2.3.1. Symmetric cryptography

In most circumstances, security of communication refers to protecting information from eavesdropping – that is the information exchanged is kept confidential and does not become known to anyone other than the sender and the receiver. To this goal, cryptography covers the procedure of appropriately disguising the data such that even if unintended parties are able to intercept the message they will not be able to read it. Hence, cryptography can be simply described as a transformation of a message (*encryption*) that makes the message

[1] E.g. notification of a data breach to a competent authority, under the corresponding GDPR provision (see Chapter 3, page 19).

incomprehensible to anyone who is not in possession of secret information that is needed to recover (*decryption*) the message to its initial form. The secret information is called the *key*.

In cases that the same key is used from both the sender and the receiver, then the corresponding cryptographic algorithm is called *symmetric cryptographic algorithm* (*or symmetric cipher*). Symmetric cryptography is the most widely used of the two types of cryptography. There are two main categories of symmetric ciphers, the *block ciphers* and the *stream ciphers*, whilst the block ciphers can be used in various modes of operations. The **Advanced Encryption Standard (AES)** [8] constitutes the current approved cryptographic standard; it is a block cipher utilizing a 128-bit block size and a key size of 128, 192, or 256 bits.

2.3.2. Asymmetric (public key) cryptography

Asymmetric cryptography is a form of cryptosystem in which the encryption key is different from the decryption key. It is also known as public-key encryption. The underlying design idea of asymmetric cryptography is quite different from symmetric cryptography; although they can be theoretically used for ensuring confidentiality similarly to symmetric algorithms, their low performance impedes such a use. However, they are very good candidates for securely exchange the secret symmetric key[2] – and, thus, they can be used in conjunction with symmetric ciphers. Moreover, public-key algorithms allow for entity authentication via the so-called digital certificates, as well as for message authentication via the so-called digital signatures.

The most known public key algorithm is **RSA**, whilst the so-called **Elliptic Curve algorithms** (based on the algebraic structure of elliptic curves over finite fields) are also known to be significant public–key algorithms.

2.3.3. Cryptographic hash functions

A cryptographic hash function is any function that maps any message of arbitrary length into a fixed length hash value, or message digest, satisfying the following properties:

1. The function is one-way, i.e. it is computationally infeasible to generate a message from its hash value (although it is easy to compute the hash value for any given message),
2. It is computationally infeasible to find two distinct inputs which hash to a common value (i.e., two colliding inputs x and y such that $h(x) = h(y)$) and, moreover, given a specific hash-value y, it is computationally infeasible to find an input (pre-image) x such that $h(x) = y$.

The basic idea is that a hash-value serves as a compact representative of an input. The most common cryptographic use of hash functions rests with the digital signatures (since the hash-value of a message is digitally signed[3] and not the message itself) and data integrity. Known hash functions are **SHA-1**[4] and **SHA-2**, whilst the new standard is the **SHA-3** since October 2012 [11].

[2] The well-known Diffie-Hellman key exchange algorithm addresses exactly this issue, based on the general underlying idea of public key encryption.
[3] The signing procedure is generally performed through appropriate using a public-key encryption algorithm.
[4] A collision has been very recently found (Feb. 2017) for SHA-1 (https://shattered.io/); it should be pointed out that although SHA-1 was officially deprecated by NIST in 2011, it is still present in many applications.

Hash functions, as discussed above, do not involve any secret key. The so-called *Message Authentication Codes* (*MACs*) have somehow the same properties as hash functions in order to ensure data integrity, but they also involve a secret key for their operation. MACs usually include a hash function as a main building block; MACs suffice also to provide data origin authentication, under the assumption that the secret key has been securely exchanged between the two parties.

2.4. Technologies for Security

We next describe some main security technologies that fit well with the data processing in the framework of the iRead project and thus can serve as appropriate security solutions, as will be subsequently shown.

2.4.1. Transport Layer Security

The World Wide Web is fundamentally a client/server application running over the Internet. A relatively general-purpose security solution is to implement security just above Transmission Control Protocol (TCP)[5]. The foremost example of this approach is the Transport Layer Security (TLS) protocol, a successor of the Secure Sockets Layer (SSL) protocol, which is being considered as a somehow de-facto standard. Three versions of TLS have been standardized, namely 1.0, 1.1 and 1.2, while the last version of SSL (that is version 3) is still in use despite its known vulnerabilities. TLS is based on symmetric encryption for ensuring confidentiality, whereas the secret symmetric key is being interchanged via public key cryptographic algorithms. Therefore, block ciphers – and especially AES, the symmetric encryption standard – form the most common symmetric encryption scheme in TLS.

The main core of the TLS protocol consists of two phases: the connection setup (handshake protocol) and the steady-state communication (record protocol). During the handshake protocol, a negotiation takes place between the client and the server, in order to agree on algorithms and several security parameters. In this process, authentication of each party takes place (the client authentication is optional), based on exchanging digital certificates, whilst the symmetric cryptographic algorithm, as well as the keyed MAC, that will be subsequently used are also agreed. Therefore, public key (asymmetric) algorithms are utilized for checking the validity of the certificates, whilst symmetric algorithms are used for the encryption of the payload. All the necessary parameters for the various cryptographic schemes are being appropriately negotiated via the handshake protocol, such as to allow the two parties to verify that their peer has calculated the same parameters.

After the setup phase, the communication begins (record protocol). In this phase, the data is being split into packets, which can be optionally compressed, then the MAC is added for ensuring integrity and, finally, the packet is being encrypted, via a symmetric key cryptographic algorithm, and transmitted.

The TLS implements security that feeds upper-layer network protocols. Hence, HTTPS (HTTP over SSL/TLS) refers to the combination of HTTP and SSL to implement secure communication between a Web browser and a Web server.

[5] The TCP is a main protocol of the Internet Protocol suite, providing reliable delivery of data between applications running on hosts communicating by an Internet Protocol network.

For a more detailed technical description, as well as for description of other secure protocols that rely on TLS, see Appendix, page 42.

It should be pointed out that known attacks set specific requirements regarding the use of symmetric ciphers in TLS. More precisely:

- If block ciphers are being used, specific modes of operation should be avoided (see the padding oracle attack [6], the BEAST attack [10], the Lucky Thirteen attack [2])
- If the stream cipher RC4 is used, other types of attacks can be mounted, mainly due to inherent weaknesses of RC4 [1], [13], [17], [22].
- Even if the web server supports the last protocol version TLS 1.2 as the default option, simultaneous support of older versions of SSL (such as SSL v2.0) may lead to data leakage (see the Drown attack [4]).

As a consequence, the new version 1.3 of TLS, which is currently under development [18], does not support the CBC mode of encryption in block ciphers, whereas RC4 is also prohibited.

2.4.2. Authentication mechanisms

The front line of defense against intruders is the authentication system. The most common and easily deployable authentication mechanism is the usage of passwords – that is a multiuser system requires that a user provide not only a name or identifier (ID) but also a password. The password serves to authenticate the ID of the individual logging on to the system; in turn, the ID determines whether the user is authorized to gain access to a system, as well as the corresponding privileges assigned to the user (see access control in page 12).

Keeping the passwords secret is essential for ensuring a strong authentication mechanism. To this end, possible threats regarding password revealing should be considered; these include the following:

- Poor password choice, that can be guessed or inferred (e.g. password that is based on user's personal data such as, e.g., birthdate or data that the user may inadvertently disclosure in some contexts[6]);
- The chosen password can be revealed by appropriate password cracking tools: such tools allow for dictionary attacks (i.e. check the words of a dictionary in order to find out possible passwords lying in this list[7]) and/or brute force attacks (i.e. exhaustive search over all possible combinations of characters);
- The passwords are not securely stored in the system's database and, thus, accessing this database results in password revealing.

To address the first two threats, the users should appropriately choose passwords rendering these attacks ineffective and, thus, harmless; this is further analyzed in Chapter 5, where the management of risks related to the security of personal data is discussed (see page 35).

To address the third threat, an obvious approach is to forbid users – via appropriate access controls – having access to the passwords file. However, this is not adequate since the threat of

[6] For example, this is the case in the so-called social engineering attacks.
[7] Such dictionaries are quite "rich", since they include not only any possible English word but also strings that are known to be preferable by many users as a password choice, such as "abc123", "qwerty", "letmein" etc.

an unanticipated break-in should be never underestimated despite the security measures that are in place (i.e. an unauthorized attacker may succeed in obtaining this file bypassing the security controls), whilst there are always privileged users that necessarily have access to all system files (e.g. system administrators). Therefore, the passwords are never stored in clear but in a hashed form, via an appropriate cryptographic hash function; recalling the irreversibility of hash functions, this ensures that not even system administrators have knowledge of user's passwords. When a user attempts to log on, she provides an ID and a password; the system uses the ID to index into the password file and retrieves the hashed password, whereas it also computes the hash value of the password given at its input; if the result matches the stored value, the password is accepted as valid.

An additional security feature is to allow using a "salt" for each password [21]; the salt is a random string created by the system, which is somehow concatenated to the user's password. Therefore, the hashed value of the password also depends on the value of the salt. By this way, even if two users choose the same password, the corresponding salts will be different since theirs salts will be different and, thus, these duplicate passwords will not be visible in the password file. Moreover, the salt effectively increases the length of the password without requiring the user to remember additional characters; note that even if pre-computed hashes of possible passwords are available, such a list is not useful when the password management involves a "salting" process. The salt is also kept into the password file, since its knowledge is prerequisite for performing user authentication.

2.4.3. Access control

Authorization can be described as a definition of which access rights regarding a certain object (e.g. file, data) a certain subject (e.g. user, process) possesses. In this context, access control examines whether a subject possesses the authorization to access a desired object when an access request occurs; as a result, the action is granted or denied.

There are several access control models that can be applied to several domains [12] – i.e. to a database management system (DBMS) for database security or to an operating system for file system security. The same models can also be applied on application level (i.e. an application program can use an API to check user's permissions and act correspondingly).

A discretionary access control (DAC) constitutes the simplest form of access control, specifying the rules under which subjects can create and delete objects, as well as grant and revoke authorizations for accessing objects to others. More precisely, DAC is based on two principles: the first is the principle of ownership of information – i.e. the creator of an information item becomes its owner, who in turn may grant access to others and define the type of access – and the second is the principle of delegation of rights – i.e. a user who has been assigned a certain access right may be allowed to pass this right on to other users. DAC is implemented in most operating systems for file system security and database management systems (DBMSs). The access rules are usually stored by many software systems in an access control matrix or in an access control list.

The mandatory access control (MAC) is based on the fact that access control is based on a set of predefined rules, which are mandatory, rather than user definable. The most prominent example of MAC is the need-to-know model, which is based on the principle that a subject can access only those objects that are necessary for fulfilling his duties (i.e. he "needs to know" the content of the object).

The concept of role-based access control (RBAC) has evolved as de-facto and proposed NIST standard, which simplifies the administration of access permissions by removing the direct links between subjects (users) and objects. In RBAC, the access rights are determined by means of roles, which can be derived from the organizational structure encapsulating organizational functions. Although RBAC is different from MAC and DAC access control frameworks, it can enforce these policies without any complication.

A more dynamical approach is the attribute-based access control (ABAC); in this case, access rights are granted to users through the use of policies which combine attributes together. Unlike RBAC, which employs pre-defined roles that carry a specific set of privileges associated with them and to which subjects are assigned, the key difference with ABAC is the concept of policies that express a complex Boolean rule set that can evaluate many different attributes [15]. The XML dialect for defining access control policies, XACML, is also based on user and object attributes [16].

2.4.4. Data at rest encryption

Encryption of data at rest, when used in conjunction with transport encryption and appropriate security policies that protect relevant passwords and encryption/decryption keys, further enhance the overall protection of data. It should be pointed out that, according to the General Data Protection Regulation (see Chapter 3, page 19), unauthorized access to personal data is not being considered as a data breach incident if these data are appropriately encrypted. To this end, several technological approaches are available; the main candidates for development in the iRead design are the following:

- Database encryption, to secure the actual data within the database (e.g. the tables of the database) and the backups. By these means, data remains protected even in the event of a data breach (under the assumption that proper key management practices are in place). Modern approaches to database encryption, such as the Transparent Data Encryption (TDE) architectures, make it easier to deploy database encryption because TDE does not require any changes to database applications.

- File encryption, to secure file/folders in a specific system/device (e.g. workstation, server, tablet). In this direction, a well-known approach of wide use is the so-called Pretty Good Privacy (PGP), which provides confidentiality and authentication services that can be used for file storage applications (as well as for e-mail). PGP is now on an Internet standards track (RFC 3156).

3. Personal Data Protection Regulatory Framework

3.1. Introduction

In this Chapter, the European personal data protection legal framework is presented. Since personal data processing is going to be held in the framework of the iRead project, special emphasis will be given on establishing a full compliance with this regulatory framework.

During the iRead implementation and evaluation, a significant change on the personal data regulatory framework is going to be held, namely the General Data Protection Regulation will replace the existing Data Protection Directive on May 25th, 2018. Therefore, both these legal frameworks will be described, since they both cover the whole time period that the iRead project will be active; it should be pointed though that since the forthcoming General Data Protection Regulation sets several additional requirements for lawful processing of personal data with respect to Data Protection Directive, the overall design of the iRead system will be based, at its very initial stage, on ensuring compliance with this Regulation too; besides, the evaluation phase of the iRead project will take place in 2019-2020 and, thus, the new Regulation will then apply.

3.2. The Data Protection Directive

In 1995, the EU adopted the Data Protection Directive 95/46/EC (Data Protection Directive – DPD) [9], forming the main foundation for data protection in Europe until 2016 (and will still remain in place until the mid of 2018, as described next in Section 3.3). Member states of the EU ought to have amended their national laws to conform to DPD within three years. The directive has the objective to provide for a high level of protection of the fundamental rights and freedoms of the individuals with regard to the processing of personal data.

3.2.1. Basic definitions

The DPD introduces the following definitions (Art. 2) with regard to personal data processing:

a) The term **"personal data"** refers to 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.
It should be pointed out though that, as stated in Recital 26 of the DPD, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller (as defined next) or by any other person to identify the said person; in other words, simply removing the identifiers of the data subjects or pseudonymizing their data does not ensure anonymization, which means that even in these cases the DPD may still apply (since such data are still considered as personal).

b) The term **"processing of personal data"** ("processing") refers to any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction;

c) The term "**controller**" (data controller) refers to the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data;

d) The term **"processor"** refers to any natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller;

e) The term **"third party"** refers to any natural or legal person, public authority, agency or any other body other than the data subject, the controller, the processor and the persons who, under the direct authority of the controller or the processor, are authorized to process the data;

f) The term "**recipient**" refers to a natural or legal person, public authority, agency or any other body to whom data are disclosed, whether a third party or not; however, authorities which may receive data in the framework of a particular inquiry shall not be regarded as recipients;

g) The **"data subject's consent"** means any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed.

Moreover, according to art. 8 of the DPD, personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, as well as personal data concerning health or sex life are being considered as **special categories of personal data** (also known as **sensitive data**).

3.2.2. Basic principles

The Data Protection Directive codifies basic privacy principles that need to be guaranteed when personal data are collected or processed, which include the ones listed here:

1. Legitimacy: Personal data processing has to be legitimate, i.e. data must be processed fairly and lawfully (art. 6 I (a));

2. Purpose specification and purpose binding (also called purpose limitation): Personal data must be collected for specified, explicit, and legitimate purposes and may not be further processed in a way incompatible with these purposes (Art. 6 I(b)). Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;

3. Data minimization: The processing of personal data must be limited to data that are adequate, relevant, and not excessive (art. 6 I (c)). Moreover, data should not be kept in a personally identifiable form any longer than necessary (Art. 6 I (e)). This principle is strongly related to the so-called proportionality principle and serves to motivate privacy-enhancing technologies – aiming to provide e.g. anonymity, pseudonymity etc. for data subjects, depending on whether the necessity of the processing is fully justified from the context of the purpose.

4. Consent: personal data may be processed only if the data subject has unambiguously given his consent (art. 7 (a)) – there are though some other exemptions for ensuring the legitimacy of the processing, without asking for the data subject's consent (e.g. if there is a legal obligation, or contractual agreement).

5. Additional safeguards for processing of special categories of data: According to Art. 8, the processing of so-called special categories of personal data is generally prohibited, subject to exemptions – a notable such exemption being the case that the data subject has given his explicit consent to the processing of those data. The DPD allows Member States for laying down additional exemptions to the processing of special categories of data.

6. Transparency and rights of the data subjects: The principle of transparency is of high importance and is prerequisite for the right of informational self-determination. Pursuant to Art. 10, individuals whose personal data are being processed have the right to information about at least the identity of the controller, their data processing purposes, and any further information necessary for guaranteeing fair data processing. If the data are not obtained from the data subject, the data subjects have the right to be notified about these details pursuant to Art. 11. Further rights of the data subjects include the right of access to data (Art. 12 a), the right to object to the processing of personal data (Art. 14), and the right to correction, erasure, or blocking of incorrect or illegally stored data (Art. 12 (b)). It is evident that these rights constitute specific obligations for the data controllers.

7. Security: The data controller should adopt appropriate technical and organizational security mechanisms to guarantee the confidentiality, integrity, and availability of personal data (Art. 17).

8. Independent supervisory authorities: Independent supervisory authorities (so-called data protection authorities) shall monitor compliance with the directive and the corresponding national law. Moreover, according to Art. 18 (1), Member States shall provide that the controller or his representative, if any, must notify the supervisory authority referred before carrying out any wholly or partly automatic processing operation or set of such operations intended to serve a single purpose or several related purposes.

9. Limitations to the data transfer to third countries: According to Art. 25 and art. 26 of the DPD, there are specific restrictions on data transfer to third countries which prevent data controllers to circumvent the relatively strict European data protection legislation by outsourcing the personal data processing to countries with no or with inadequate levels of data protection.

3.3. The National Data Protection Laws

The DPD has been transposed to Member States national legal systems over many years ago. With regard to the Member States employing in the iRead project (namely, Germany, Greece, Romania, Spain, Sweden and United Kingdom[8]), we identify the following national legal frameworks:

Germany:

The main legal source of data protection in Germany is the Federal Data Protection Act (*Bundesdatenschutzgesetz* in German) (BDSG) which implements the DPD. Additionally, each German state has a data protection law of its own. In principle, the data protection acts of the individual states intend to protect personal data from processing and use by public authorities of the states whereas the BDSG intends to protect personal data from processing and use by

[8] These are the Member States of the participants of the iRead who are going to perform personal data processing in the context of iRead (e.g. to perform learning trials/procedures).

federal public authorities and private bodies. Enforcement is through the data protection authorities of the German states. The competence of the respective state authority depends on the place of business of the data controller. Each individual German state has a Data Protection Authority which is responsible for the enforcement of data protection laws and competent in respect of data controllers established in the relevant state. For the special case of the iRead partner from Germany[9], the competent authority is the Baden-Württemberg Data Protection Supervisor (https://www.baden-wuerttemberg.datenschutz.de/).

Greece:

In Greece, the DPD has been incorporated into the Law 2472/1997, whilst the competent authority is the Hellenic Data Protection Authority (HDPA). This Law applies to any processing of personal data, provided that such processing is carried out by a controller or a processor established in Greek Territory or in a place where Greek law applies by virtue of public international law.

According to Law 2472/1997, the controller must notify the HDPA about the establishment and operation of a file or the commencement of data processing.

Romania:

The DPD has been implemented in November 2001 through Law no 677/2001 on the protection of individuals with regards to the processing of personal data and the free movement of such data ("Data Protection Law"). The competent Authority is the so-called National Authority for the Surveillance of Personal Data Processing (www.dataprotection.ro) – in Romanian, "Autoritatea Nationala de Supraveghere a Prelucrarii Datelor cu Caracter Personal" or ANSPDCP. Public and private entities processing specific types of personal data must notify ANSPDCP in respect of their personal data processing.

Spain:

The DPD has been implemented in November 1999 with the Special Data Protection Act 1999 (known as the "LOPD" in Spain). This Act updates an earlier Data Protection Act ("LORTAD") that was in place since 1992, which was consistent with most of the contents of the DPD. Enforcement of the LOPD is through the Spanish Data Protection Commissioner's Office (AEPD) (https://www.agpd.es/). The AEPD holds a detailed registry of databases containing personal information.

Sweden:

Sweden implemented the DPD in 1998 with the Personal Data Act (SFS 1998:204). The Data Inspection Board (DIB) is the supervisory authority under this Act (http://www.datainspektionen.se). All controllers except those whose processing falls under any of the exemptions in the Act, need to provide appropriate file notifications to the DIB.

United Kingdom:

[9] Duale Hochschule Baden-Wurttemberg (DHBW).

The United Kingdom implemented the EU Data Protection Directive 95/46/EC in March 2000 through the Data Protection Act 1998. Enforcement is through the Information Commissioner's Office ("ICO"), which constitutes the competent Authority (https://ico.org.uk/).

Data controllers who process personal data must inform the ICO so that their processing of personal data may be registered and made public in the register of data controllers.

3.4. The General Data Protection Regulation

The EU has created a new data protection regime, the **Regulation (EU) 2016/679** or **General Data Protection Regulation (GDPR)** [19], to replace the Data Protection Directive; the GDPR aims to harmonize data protection laws across Europe, by removing the need for national implementation. Moreover, the GDPR aims to further protect and empower all EU citizens' data privacy, whereas it introduces new obligations with which all data controllers must comply.

The GDPR entered into force on 24 May 2016 but its enforcement will not begin until 25 May 2018; therefore, the DPD still constitutes the current valid legal instrument. However, as illustrated in Section 3.1, it is crucial to take GDPR into consideration at the early stages of the design of the iRead.

Some definitions in the DPD are re-formulated in the GDPR, whilst several new definitions are also introduced. These include – amongst others – the following:

a) The consent of the data subject means any freely given, specific, informed **and unambiguous** indication of his or her wishes by which the data subject, **either by a statement or by a clear affirmative action**, signifies agreement to personal data relating to them being processed (***Art.4 (11)***).

b) Data concerning health means **personal data relating to the physical or mental health of an individual, including the provision of health care services, which reveal information about his or her health status**. (***Art.4 (15)***). It should be stressed that the DPD does not explicitly define the notion «data concerning health».

c) "biometric data" means personal data resulting from specific technical processing relating to the physical, physiological or behavioral characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data (Art. 4); Processing of biometric data for the purpose of uniquely identifying a natural person constitutes a processing of special categories of data (Art. 9).

d) "Personal data breach" means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored or otherwise processed (***Art.4 (12)***).

Furthermore, the GDPR introduces several new obligations for data controllers, including – amongst others[10] – the following:

- the requirement to provide the data subject appropriate information on the data processing, not only in the case that the personal data have been obtained from the data subject (Art. 13(2)(f)) but also in the case that the personal data has not been obtained from the data subject (Ar. 14(2)(f));

[10] This list of new obligations does not constitute an exhaustive list, but it is a list that shall possibly apply to the specific data processing in the context of the iRead system

- the obligation to provide the data subject the personal data concerning him or her, which he or she has provided to the controller, in a structured, commonly used and machine-readable format, in case specific criteria on data processing are met (right to data portability) (Art. 20(1));
- the implementation by the controller of suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision, in cases of automated individual decision-making processing, including profiling (Art. 22(3));
- data protection by design and by default (Art. 25);
- the maintenance of a record of processing activities under data controller's responsibility (Art. 30);
- in the case of a personal data breach, the requirement to notify the breach to the supervisory authority, without undue delay and, where feasible, not later than 72 hours after having become aware of it, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons (Art. 33(1)). As stated in Recital 78, in setting detailed rules concerning the format and procedures applicable to the notification of personal data breaches, due consideration should be given to the circumstances of that breach, including whether or not personal data had been protected by appropriate technical protection measures, effectively limiting the likelihood of identity fraud or other forms of misuse.
- the requirement for the controller to carry out a Data Protection Impact Assessment (DPIA) where the processing is likely to result in a high risk to the rights and freedoms of natural persons (Art. 35).

Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data (Art. 7, par. 1). If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language (Art. 7, par. 2).

It should be stressed that the above does not constitute an exhaustive list of what GDPR brings, but it is a list that is related to the specific data processing in the context of the iRead system.

The GDPR also determines that, where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers (Art. 26); it should be noted that the DPD has not defined the term "joint controllers" – although, in practice, such joint controllers have been identified in several contexts (whilst the definition of data controller in the DPD states the controller may be "*alone or jointly with others (…)*"). According to the provisions of the GDPR, the joint controllers shall in a transparent manner determine their respective responsibilities for compliance with the obligations under GDPR, in particular as regards the exercising of the rights of the data subject and their respective duties to provide the necessary information to the data subject.

With regard to the security of the processing, Art. 31 states that "taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk, including inter alia

as appropriate: (a) the pseudonymizing and encryption of personal data; (b) the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services; (c) the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident; (d) a process for regularly testing, assessing and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing".

Moreover, according to Art. 89(1) of the GDPR, processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organizational measures are in place in particular in order to ensure respect for the principle of data minimization. Those measures may include pseudonymizing provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.

The GDPR allows the national legislation by each Member State to introduce additional grounds for processing personal data for research purposes (Art. 89(2)).

Although the GDPR has not been enforced yet, it is evident that any data controller should strive to ensure that the personal data processing is fully consistent with the provisions that the GDPR sets.

4. Personal Data Processing in iRead

4.1. Introduction

The central purpose of the iRead project is to assist and support children of primary schools who are trying to master the ability to read. Within the iRead system, the process of learning to read a language (either as a native speaker or as a foreign language), is viewed as a collection of learning-steps, where each step corresponds to a linguistic phenomenon the reader has to learn, or to a linguistic skill she has to master. Over the last years, many schools across Europe have adopted new technologies to support their teaching plan (access to internet, computers and tablets), since children are keener on learning if the content is appealing, pleasant and amusing.

In a high-level description, a child registered in the iRead system, will be able to practice and improve herself by playing interactive games, reading eBooks, undertaking activities and accessing appropriate reading material. The content presented to each child is personalized, i.e. appropriately selected to meet the current "needs" of the child.

To accomplish the aforementioned goals, the iRead system will model expert knowledge on the process of learning to read, and create and maintain a profile for each child that uses the iRead system, in order to reflect the individual needs and skills of the user.

Over the final two years of the project, a series of pilots will be conducted in order to evaluate the overall iRead system. Schools from European countries will integrate the iRead system as part of their learning activities. Teachers will be able to monitor the progress of their classroom and derive learning strategies, while feedback from the users and statistical analysis will allow experts to enforce and improve the deployed models. Finally, through an open online pilot more users (not necessarily children) will be able to use, test and evaluate the iRead system.

4.2. Use Cases

The main use case of the iRead system at its first stage includes a deployment of the iRead system in classrooms, in cases where the school has decided to use iRead system as part of the teaching plan. In such a case, a student plays an iRead game or uses an iRead Reader application, whereas the teacher monitors the student's progress; in this framework, the teacher has also the option to select appropriate learning material for a specific student. This use case includes several user roles (student, parent/guardian, teacher, etc.). The present document constitutes an analysis of data protection and security issues arising in this use case.

Several other use cases can be also examined, under a proper development of the iRead system. More precisely, the last pilot for the iRead evaluation will allow for unattended use of iRead through the internet. Moreover, the iRead user's profile could provide useful information to an external-application, so as to have more learning-related services to the user; for instance, a "personalized file-explorer" could sort documents based on their difficulty which is assessed according to the user's profile. Similarly, an eBook store or a publisher could suggest appropriate reading material for a student based on her user-profile. In addition, several other uses of the iRead system can be considered, such as the case that a school desires to have its own dedicated server installation, whilst other possible extensions of the iRead system will be also examined. A full list of use cases for the iRead system is provided in the D3.3 ("System Specifications").

4.3. Components of the iRead System

The iRead system supports learning through a set of educative components as well as monitoring progress and analytics display. All components that a user can interact with, will be stored as "authorized components" in the iRead system. The educative components are:

Literacy games

The learner will be able to practice and improve through a collection of literacy games. The games will offer a variety of levels both at word and sentence level, and covering the learning steps required for mastering the reading skill in a specific language.

Reader application

The Reader-app will offer text-to-speech capabilities and personalized text-annotation. The learners will have the opportunity to expose themselves to new vocabulary and to test skills that they practiced with other educative components with real text.

eBooks-activities

Interactive eBooks will allow the learner to explore by interacting with visual scenes and will provide training material for a specific language. The interactive eBooks will be able to upload information regarding the reader's performance to the iRead system.

Apart from the educative components, two more components will be developed within the iRead system.

Content classification

An online platform or desktop application will allow a user to compare texts based on her individual needs. Not all texts are appropriate for a learner, hence, based on scientific metrics, a learner (or a teacher) will be able to browse through text-material and select content that goes with a learner's skills.

Learning analytics tools

A web-based application will allow a learner to overview her progress, and a teacher to monitor the progress of an individual student or of an entire classroom. Based on the information provided, a teacher could modify her teaching plan in order to target particular difficulties. A similar tool could be used by linguistic experts and system engineers, in order to get a better understanding of the learning process, or to redesign a component (e.g. a literacy game).

4.4. Language Support

The iRead system is designed to assist learning to read in four native languages: English, Greek, German and Spanish, as well as English as a Foreign Language (EFL). Furthermore, particularly for English and Greek, the iRead system will support learners with dyslexia. Since the learning process of people with dyslexia differs from the process for non-dyslectic people, linguistic phenomena and skills have to be accordingly adjusted. In the following we list the different "languages" supported by the iRead system.

- Reading in English as a native language
- Reading in English as a native language for children with dyslexia
- Reading in Greek as a native language
- Reading in Greek as a native language for children with dyslexia
- Reading in German as a native language
- Reading in Spanish as a native language
- Reading in English as a foreign language

4.5. Domain Model

The knowledge provided by linguistic experts about the process of learning a language is referred to as **domain model**. The domain model for a language is a collection of learning-phenomena or skills that the reader has to master. Each such phenomenon or skill is called a domain model property, or property for short. The number of properties varies depending on the language. For example, the domain model for English as a native language for people with dyslexia developed in iLearnRW[11] contains more than 400 properties, while the corresponding model for Greek has fewer than 150 properties. Note that iRead will additionally support syntax-related properties that are currently under specification, as well as domain models for all languages supported by the iRead system.

All properties of a domain model are partitioned into several categories, again depending on the language. For example, in English the following categories are defined (syntax-related categories are not included):

- Consonants (49 properties)
- Vowels (71 properties)
- Blends and Letter Patterns (131 properties)
- Syllables (13 properties)
- Suffixes (92 properties)
- Prefixes (42 properties)
- Confusing Letters (15 properties)

When learning a language, properties of different categories have to be mastered in parallel, while other properties create chains of prerequisites, where properties to the end of the chain are locked until (almost) all previous properties are mastered. Therefore, properties are also partitioned into clusters. Each cluster contains properties that are somehow related to each other, allowing the learner to practice them in parallel. On the other hand, clusters create chains of prerequisites assuring that the learning process will evolve smoothly.

[11] EU FP7 ICT project iLearnRW - Integrated Intelligent Learning Environment for Reading and Writing (project number: 318803)

4.6. Profile of Learner

A basic principle in learning a language is that each learner has her own needs and particularities. The domain model is the representation of the linguistic experts' knowledge on the process of learning a language. In order to provide personalized-driven material for a learner, it is essential to track and record the specific properties of the domain model that each learner struggles with. This information is stored in the **user-model** of the learner or, as it is frequently referred to, the learner's profile. The user-model entries are in one to one correspondence with the properties of the domain model. For each property a value is stored, indicating the degree of effort required by the learner to master this property. Typical values could be integers from '0' to '3', with '0' indicating complete mastery of the property, and '3' low level of mastery.

The user-model of a learner can be initialized in two ways:

- upon registration of the learner a default user-model can be used or the learner could undertake some exercises and/or activities launched by the iRead system
- a teacher could hand out some offline tests to a student and record her answers into the iRead system manually.

After profile initialization takes place, a learner can log in the iRead system and use the educative components of the system. As the user plays, for example, the literacy games, she comes across several words and sentences that are related to properties of her user-model that need to be mastered. The more the learner practices, the more she improves her skills. This improvement is measured by the iRead system and the values in her user-model are updated accordingly.

Finally, another part of a learner's user-model is a set of preferences, mainly related to the appearance of iRead applications, e.g. font size, background color, etc.

4.7. Users of iRead

There are three types of users registered in the iRead system, based on the features of iRead that they make use of. The first type of users are learners, namely users who aim to improve their reading skills by using the personalized content-driven components of the iRead system (games, reader, eBooks and other material). As already mentioned, during the last two years, piloting evaluation will be held in several schools across Europe. Teachers belong to the second category of users, since they do not make explicit use of the iRead system, but are able to monitor the progress of their students by accessing the corresponding information and data stored in the system. The last type of users are linguistic experts and data analysts. Although the interaction of experts with the system is similar to the one of teachers, experts will be able to derive new knowledge based on the overall usage of the system.

The three types of users are described below.

Learner

In order to use iRead, a learner must be registered in the iRead system. Registration can be done directly by herself or by a teacher that is already registered in the iRead system. Afterwards, the learner can log in and start using the learning components of iRead. At any point, through a web application, the learner (or her parent) can access an overview of her progress by retrieving her data stored in the iRead system.

Teacher

Teachers as well as learners must be registered in the system in order to have access to iRead. Registration of teachers can be done by the consortium. Teachers can register students of their class and monitor their progress through a web application. Information on the whole class or on a particular student will be retrieved from the data stored in the iRead system. Based on their findings, teachers can adjust their teaching plan or make individual suggestions for further reading or exercises. Also, teachers will be able to assist students with setting or changing their preferences in iRead system.

Expert

Experts are registered in the iRead system by the consortium. Experts do not register other users in the iRead system, but they can access the stored data in the iRead system for scientific purposes. Overview of usage and statistical analysis of the data could lead to improvement upon the modeling of the learning process, or upon the design of educative components. At the end of the project, the evaluation of the outcomes will be held by experts.

4.8. Stored Data

Various data is stored in the iRead system. We distinguish three main categories of data, namely user details, learner's usage-related data and access-related data. In the following we describe each category and list the information stored in the iRead system.

4.8.1. User details

When a user is registered, her details are stored in a database along with her credentials. All user categories have similar attributes that will be stored, still, for learners more details are required. In the following table we list the attributes stored for a user, their purpose, and which type of user they apply to.

Attribute	Description	Learner	Parent / Guardian	Teacher	Expert
username	Used for log in.	YES	YES	YES	YES
password	Used for log in. Passwords are not directly stored in the iRead system. They are encrypted and their hashed key is stored.	YES	YES	YES	YES
first name	Allows for example a teacher to identify a learner/parent by her name.	YES	YES	YES	YES
last name	Allows for example a teacher to identify a learner/parent by her name.	YES	YES	YES	YES
e-mail address	Useful for strengthening the authentication procedure, as well as for exercising data subject's rights.	YES (optionally)	YES (optionally)	YES (optionally)	YES (optionally)
date of birth	Several linguistic phenomena are related to the age of the learner. For example, syntax-related skills can be mastered after a certain age. Also used in statistical analysis.	YES			
gender	Used in statistical analysis.	YES	YES	YES	
school	Required for pilots run at schools. Also used in statistical analysis.	YES		YES	
teacher	Required for pilots run at schools. Maps learners to teachers enabling monitoring by teacher.	YES			
classroom	Required for pilots run at schools. Maps learners to classrooms enabling monitoring of classroom by teacher.	YES		YES (many to one)	
language support	One of the seven languages supported by the iRead system. Used to select the appropriate learning-steps for the learner.	YES		YES	
mother language	Several learning skills are easier or more difficult to master depending on the mother language of the learner. Used in statistical analysis.	YES			
User - model	Contains information about the reading skills a user has mastered.	YES			

Table 1 - Personal data categories

It should be pointed out that each student will be uniquely associated with her/his parent guardian; therefore, appropriate linkage information between students and their parents/guardians will be necessarily stored in the iRead system.

Note that for experts, only the first two attributes are required (username, password) in order to access the iRead system, while for teachers, it is also important to store a mapping of teachers to classrooms or students, since a teacher could be responsible for more than one classroom.

4.8.2. iRead learner usage-data

Previously, when we described the user's user-model, we mentioned that the iRead system measures the improvement a learner has made while using the educative components of the iRead system. The most natural question that arises is how can this be accomplished. Every action of each learner is recorded in iRead logs. The amount of data stored and the details recorded are such that someone could simulate the learner's interaction with the educative components of the iRead system. In particular, for the educative components we store the following information:

Games

- login and logout
- starting a particular game
- content seen during the game
- success or failure on the game
- quitting or finishing a game
- time elapsed

Reader app

- login and logout
- opening a text
- applying highlighting rules
- looking up a word in the dictionary
- time elapsed

eBooks/other activities

- login and logout
- interaction with the eBook
- success or failure of exercises-activities
- time elapsed

4.8.3. iRead data access

Apart from user details and learner's data stored by the iRead system, data is also collected whenever a user or an application makes a request to the iRead system. This is used for security purposes, in order to assure that only authorized users or applications have access to a user's user-model. For the web application, logs are created whenever a user uses the application and queries through learner's data for viewing analytics or a user-model.

Web application

- login and logout
- learner viewing her user-model or statistics
- teacher viewing a student's user-model or statistics
- teacher viewing her classroom's statistics

Finally, all educative components of the iRead system have access to a learner's user-model.

Educative component

- login and logout of a learner, or a learner with teacher
- reading user-model
- updating user-model
- any other call to the server of the iRead system

5. Personal Data Protection and Information Security

5.1. Introduction

In this Chapter, an analysis of data protection and information security issues is presented, aiming to identify potential data protection risks that arise in the context of the iRead project operation as well as to derive appropriate mitigation controls for all these risks. To this end, the following principles should be explicitly studied:

a) Legal principles for personal data protection (specified, explicit and legitimate purpose, adequate, relevant and not excessive data, explicit unambiguous consent of the data subjects, clear and full information to data subjects, the rights of access and objection etc.)

b) Management of risks related to the security of personal data, so as to protect data – via appropriate organizational and technical measures – against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access and against all other unlawful forms of processing.

This approach is performed at the early stage of the design of the iRead system, such as to enable the determination of the necessary and sufficient controls to address any possible risks in a timely manner. It should be pointed out that such an approach is fully compliant with the so-called "data protection by design" principle (see Ar. 25 in the GDPR).

5.2. Legal Requirements – An Assessment

The personal data processing that is performed in the framework of the iRead is explicitly defined in Chapter 4. The iRead is generally bind to process nonsensitive data; however, as stated in Chapter 4, the iRead aims to support learners with dyslexia, in which case a somehow different user-model is utilized for the learning procedure. Although there is no any other personal data related to dyslexia (e.g. diagnosis, treatment, progress in the treatment process etc.), the fact that a specific user-model is activated in some cases results in characterizing such data as health data, i.e. special categories of data according to both the DPD and the GDPR.

The Members of the consortium of the iRead project who are going to process personal data (UCL, NTUA, UGOT, UOI, ULBS, DYSLEXIA, UB, DHBW DOUK, BC) constitute the joint controllers for the overall processing, since they jointly determine the purposes and means of personal data processing. The concept of controller is an essential element in determining which national law is applicable to a processing operation or set of processing operations; The main rule of applicable law under Article 4 (1)(a) of the DPD is that each Member State applies its national provisions to "the processing of personal data, where (…) carried out in the context of the activities of an establishment of the controller on the territory of the Member State". Therefore, the National Laws referred in Section 3.2 are appropriately applicable at the moment, whilst apparently GDPR shall commonly apply to all from May 2018.

We shall next identify and determine the controls (existing or planned) selected to ensure compliance with the legal requirements on the personal data protection.

5.2.1. Legitimacy and purpose limitation

As described in Chapter 3, the purpose of the data processing in the context of the iRead research project is fully described and well-determined, whereas the personal data are to be processed in a fair and lawful way. More precisely, the data are to be processed explicitly for scientific purposes, with the ultimate goal to assist and support children of primary schools who are trying to master the ability to read.

5.2.2. Data minimization

A detailed description of the type of personal data that will be processed is given in Chapter 3. These data are the following:

1. Learner's data: username, password, first name, last name, date of birth, gender, school, teacher, classroom, language support, mother language, user-model.

   As stated in Chapter 4 (see Table 1), these data are necessary for ensuring the proper implementation of the iRead system so as to satisfy the desired purpose. Each learner needs to have a dynamically updated personalized user-model, uniquely associated with her identity (the user-model entries are in one to one correspondence with the properties of the domain model, which in turn consists of a collection of learning-phenomena or skill that the reader has to master). The user needs to log into the iRead system with personal credentials (username and password). Since the password needs to be known only to the learner[12] to ensure her proper authentication, the password is not being stored in clear text but in an appropriate irreversible unintelligible form (see Section 5.3). The teacher is responsible for supervising the whole procedure, via registering her students into the iRead system and monitoring their progress – through retrieval of the data stored into the iRead system – with the aim to appropriately adjust the teaching plan or make individual suggestions. Therefore, the first name and the last name of the learners (students) are prerequisites for the teacher to identify them from the system's data. The name of the teacher and the classroom of each learner are also needed to map each learner to teachers and classrooms, thus facilitating the monitoring procedure by the teacher; these data are also important for the desired statistical analysis (i.e. to analyze separately the learning procedure for each class or for each teacher). The date of birth and the mother language are also crucial since linguistic phenomena are related to the age of the learner, whereas the mother language allows for proper evaluation of the learning skills (these skills are easier or more difficult to master depending on the mother language of the learner); moreover, these data are important for statistical analysis, whilst the date of birth – apart from its necessity for scientific purposes – can be also used to resolve possible conflicts (e.g. coincidences in both first and last names of students). Finally, the gender of the learner is also important for statistical analysis – i.e. to separately analyze the data depending on the gender of each student. The e-mail address is very useful for establishing a strong authentication procedure, as well as for exercising the rights with respect to personal data; since though there may be no any e-mail address available in several cases, this is an optional piece of personal information.

   Moreover, data referring to the usage of the iRead system are also being processed, namely login and logout information, any interaction (starting a particular game, content seen during the game, opening a text, applying highlighting rules, looking up a word in the dictionary,

[12] The teacher will possibly have appropriate privileges for resetting the password of the user.

success or failure on literacy games or on eBook exercises-activities, quitting or finishing) and time elapsed. These data are also learner's personal data, according to art. 2 of the DPD. The processing of these data does not violate the learner's privacy[13], whereas this piece of information is also essential for the learning process since such data can be used in providing feedback to the learner; a teacher can go over the played game (and replay it with students), pointing out problems, suggesting solutions and also providing positive feedback for successful playing.

2. Parent's/Guardian's data: username, password, first name, last name, gender.

Due to the age of the learners (6-12 years old), all their rights with respect to their personal data processing are being exercised by their parents/guardians. This implies that the above information is necessary for the iRead system, since each parent/guardian should have proper access to the iRead system with regard to her/his child.

Note that Table 1 also refers to the e-mail addresses of students and parents, as an optional field; indeed, retaining these e-mail addresses seems to facilitate several important procedures, such as password assignment/creation/resetting, whilst the e-mail communication can be also used as a vehicle to exercise the data subject's rights (e.g. the right to object).

3. Teacher's data: username, password, first name, last name, gender, school, classroom, mother language.

As stated above, the role of teacher is crucial for the iRead system; each teacher needs to have personal credentials (username and password) for logging into the iRead system, with the ability to register and monitor learners (i.e. her students). Therefore, such personal data of the teacher are necessary for the proper data processing in the framework of the iRead project, whilst the mapping between teachers and schools/classrooms is also essential.

Similarly toy the case of learners (students), the passwords of the teachers are being stored in an appropriate irreversible unintelligible way. An email address is an optional field that can be used to facilitate password resetting and receiving notifications.

4. Expert's data: username, password, first name, last name, role/expertise.

Experts need also to have access to the iRead system's data for analyzing them for scientific purposes. Therefore, similarly to the previous cases, the credentials of each expert and her identity information are needed to be stored in the iRead system. There may be several roles for the experts, each of them associated with specific access privileges and, thus, the role/expertise of each registered expert needs also to be stored. As in the case of teachers, an email address is an optional field.

The retention period of the data is the period needed to achieve the desired purposes, i.e. the period that schools/teachers are willing to perform the educative trials. In case that any user revokes her consent for the processing – as described next – the personal data will be immediately deleted.

Issues to be addressed:

1. Data minimization also covers the aspect of revealing the data to authorized users of the iRead system only if this is absolutely necessary; namely, according to the so-called "need to know principle", a user should have access only to those personal data that are absolutely

[13] This information is fully transparent to the users, as described next.

necessary for fulfilling his duties within the overall framework of the data processing. For example, a teacher – being a user of the iRead system – should have access only to her student's data (with appropriate access permissions). Since the access control, in conjunction with authorization, generally addresses security requirements such as confidentiality and integrity, it will be overall discussed in the next Section, including the data minimization principle.

2. Depending on the role/expertise of the expert, their corresponding access rights should be appropriately adjusted so as to ensure a proportionate approach. For instance, experts performing statistical analysis should not have access to identification data of students and, thus, a pseudonymization of such data seems to be prerequisite in such cases.

5.2.3. Consent

During the project execution, a consent form should be signed by parents/guardians of a child participating in the learning procedure via the iRead system. This form will be signed not only in the case of children with dyslexia but also in any other case. The consent form will also provide detailed information on the purpose of the process and the methodology used, on the exact role of the teachers and experts, as well as on their rights, concerning the data related to the child. The provided form will be written in the mother language of the parent/guardian, to ensure an effective share of necessary information; by this way, it is ensured that such a consent coincides with a freely given unambiguous, specific and informed indication of the parent/guardian's wishes with regard to this process. Details about consent forms will be provided in D10.1.

It is essential that data will not be collected without the aforementioned consent. The user who proceeds with the registration of learners, i.e., a teacher, should be obligated to ensure, prior the registration process, the existence and validity of the corresponding signed form for any individual learner, unless (and subject to individual countries ethics approval procedures) the school have existing policies in place which allow them to provide consent on the parents' behalf. To this end, teachers should be informed of their responsibilities via a proper notification form; this form will describe all their responsibilities with respect to handling personal data of children, and should be signed by any teacher.

In the same line, the experts involved should be also informed of their responsibilities, namely that the data are confidential and should be processed only in the context of the project, as well as that the conclusions or the output of statistical analysis of the data can be announced/published only in a fully anonymized form. Therefore, an appropriate form should be signed by the experts.

It should be pointed out that that the above are also consistent with the definitions given in the GDPR with regard to the data subject's consent.

Issues to be addressed:

1. The templates of all the above forms (in English, German, Greek, Romanian Spanish, Swedish), fulfilling the aforementioned requirements, should be developed; they will be provided in D10.1.

2. The data controller needs to ensure that all consents have been asked and given in a proper way. Therefore, further controls should be considered to strengthen the assurance of the valid consent – e.g. the signed written consent forms could be electronically submitted to the iRead system, appropriate use of e-mails etc.

3. Regarding the scientific publication of the research results, emphasis should be given on adopting an appropriate anonymization technique (see [3]).

5.2.4. Special categories of data

As stated above, the cases of learners of dyslexia result in processing of special categories of personal data – that is health data. Hence, special safeguards need to be ensured that are present in the whole processing. More precisely, the legal requirements stemming particularly from this processing are the following:

1. The data subjects provide their explicit consent (actually, the explicit consent of their parents/guardians is provided, since the learners are below the age of 16 years). Any such explicit consent will be provided in a written form.

2. The data controller will notify the corresponding Data Protection Authority in writing about the establishment and operation of a file and the commencement of data processing. Moreover, the data controller – due to the existence of special categories of data – will ask the Authority to issue a permission for this processing, whenever such an obligation stems from the provisions of the national Law.

5.2.5. Transparency and data subject's rights

A comprehensive and detailed information on the data processing (the identity of the data controller, the purpose of the processing, the type of the processing, the exact type of personal data that will be processed etc.) will be provided to the data subjects. Such information will be included in the consent forms that are to be signed. It should be also stressed that such information will be also provided, on a permanent basis, in the project's web site.

The aforementioned information will also cover the data subject's rights with respect to personal data protection, that is the data subjects will be informed about the right to access their personal data, the right to object to the processing of their personal data (Art. 14), and the right to the correction of stored data. The right to object to the processing includes also the option of withdrawing the consent of the data, which means a full and irreversible deletion of any personal data of the data subject. It is evident that specific, well-determined procedures, need to be in place for allowing the data subjects to exercise their aforementioned rights. This procedure should be appropriately notified to the data subjects.

Issues to be addressed:

1. The templates of all the above forms (in English, German, Greek, Romanian Spanish, Swedish) fulfilling the aforementioned requirements with respect to information provided, should be developed; they will be provided in D10.1.

2. A well-determined procedure should be set up for allowing data subjects to exercise their rights (e.g. how can the parent/guardian withdraw her/his consent). This procedure will be also explicitly described in the project's web site.

5.2.6. Data transfer

There will be no personal data transfer, in the context of the iRead project, outside the European Union area.

5.2.7. Other data controllers' obligations

Any partner in the project that process personal data, being data controller, will perform all necessary actions (such as notifying the processing to the competent Authority and ask for a permission, in case of sensitive personal data processing) to ensure compliance with the corresponding National Law. Moreover, any data controller will proceed in all necessary actions to be compliant with the GDPR, as well as with any national Law that will possible be in force in the near future under the provisions of the GDPR.

Moreover, the respective responsibilities of each controller with respect to the data processing and compliance with data protection legislation will be fully transparent and documented.

5.3. Management of Risks Relating to the Security of Personal Data

5.3.1. Security requirements of iRead system

The requirements to security as described above are also applicable to the iRead system. In particular, the special requirements that should be considered during its design are the following:

- **Integrity of Data:** Data that are exchanged through the network are not altered.
- **Confidentiality:** Data can only be read by authorized devices and persons.
- **Authentication:** A secure mechanism exists for determining and verifying the identity of the source and the destination of a communication.

It should be pointed out that two possible modes of operation of iRead should be addressed – namely the full online mode, in which the learners utilize iRead through an Internet connection, and the offline mode, in which there is no Internet connection; although these two modes necessitate some different security mechanisms, the general security objectives as stated above are the same.

5.3.2. User authentication

With regard to the authentication procedure, the following desired characteristics should be taken into account:

- Any student in the class may have access to any tablet – although, it may be more appropriate in some cases to assign each table to a specific student.
- Once a student uses a tablet, there will be no other user using the same tablet at the same time.
- Each user of the iRead system (child, teacher, parent, expert/researcher) will be assigned to a different user role (with respectively different authorization rights).
- Appropriate user authentication should take place in both online and offline modes of operation of the iRead.

In an offline mode of operation (i.e. in case that there is no any Internet connection), each user should be able to be authenticated in a Local Login Mode of the iRead application. In this mode, the user will have access only to local data and, thus, appropriate resources need to have been

preloaded from the iRead server. In a full online mode of operation, the student – via the iRead application – will be directly connected, through the Internet, to the authentication server running on the web iRead server (that is a Server Login Mode).

There are several options for authenticating users, such as follows:

a) **Using fingerprint**. This should not be considered as an option, since fingerprint data are biometric data which in some cases, in the light of the GDPR, constitute special categories of data. Other more appropriate options should be considered.

b) **Using gestures.** It is not recommended because children usually forget it.

c) **Using password.** It is the most common approach and seems to be the most preferable option for the iRead, taking into account the students' age (6-12 years old).

Password creation analysis

Since the password authentication seems to be the most suitable approach for the iRead system, covering both the online and offline mode of operation, an analysis of appropriate creation/use of passwords follows. Based on the analysis presented in Chapter 2, any user should have a strong password. It is commonly known that strong passwords have the following characteristics (see, e.g., [20]):

- Contain at least eight alphanumeric characters.
- Contain both upper and lower case letters.
- Contain at least one number (for example, 0-9).
- Contain at least one special character (for example, !$%^&*()_+|~-=\`{}[]:";'<>?,/).

On the other side, poor, or weak, passwords have the following characteristics:

- Contain less than eight characters.
- Can be found in a dictionary, including foreign language, or exist in a language slang, dialect, or jargon.
- Contain personal information such as birthdates, addresses, phone numbers, or names of family members, pets, friends, and fantasy characters.
- Contain work-related information such as building names, system commands, sites, companies, hardware, or software.
- Contain number patterns such as aaabbb, qwerty, zyxwvuts, or 123321.
- Contain common words spelled backward, or preceded or followed by a number (for example, terces, secret1 or 1secret).
- Are some version of "Welcome123" "Password123" "Changeme123".

Passwords should never be written down. Instead, to create passwords that can be remembered easily, password creation could be based on a song title, affirmation, or other phrase. For example, the phrase, "This May Be One Way To Remember" could become the password TmB1w2R! or another variation. The latter one is a good approach that can be followed by any student in order to have a strong, but easy to remember, password.

Password assignment analysis

There are several options with regard to assigning passwords to users. An obvious option – which should be the default option – is to allow users (students or their parents) to choose their own passwords[14]; another option is to allow teachers[15] to determine the passwords of their students (i.e. in cases where the parents/guardians agree for such a procedure). In both cases, all involved entities (teachers, students, parents) should have explicit knowledge of the password creation rules determined above and, thus, an appropriate procedure for disseminating this information needs to be established (e.g. through the consent form).

To further strengthen the unique association of a password with the corresponding user, the password assignment procedure may be appropriately bound to the e-mail address of the user (e.g. the mail of the parent/guardian); for instance, the activation of the password could be done only after the user follows a link that he/she receives into his/her e-mail address. By these means, it is ensured than only the user of a specific e-email address will be able to activate the account in the iRead system; moreover, this e-mail address can be also used for resetting the password in the future.

To ensure a proper choice of password, the iRead could employ a mechanism for estimating the password strength, so as to forbid uses from choosing passwords that they do not satisfy specific criteria. Moreover, appropriate mechanisms to handle the cases of forgotten or stolen passwords should also be in place (e.g. via utilizing the user's e-mail address).

Password storage analysis

As stated in Chapter 2, the passwords should be stored in a hashed form; in the iRead system, this requirement is further accentuated by the fact that, in an offline mode, the authentication procedure will be performed on the local device and, thus, the passwords file will be – with high probability – stored in the user's device.

Apart from the usage of an appropriate hash function, a "salting" procedure needs also to be in place (see Chapter 2). It should be pointed out that general hashing algorithms (MD5, SHA-1/256/512) are not recommended for password storage; instead, an algorithm specifically designed for the iRead could be used or, alternatively, we may use the new hash algorithm standard, namely SHA-3.

5.3.3. Confidentiality and integrity though transmission

In an online mode of operation, the iRead will be based on a web application. The best option to secure such a communication is the appropriate use of the TLS protocol, as stated in Chapter 2. By this way, confidentiality and integrity of the communication will be established, whereas the user/application can be sure that the communication link has been established with the right web server (a valid digital certificate needs to be in place for ensuring this).

To strengthen the overall communications security, the TLS protocol that will be implemented will satisfy the following (see also Chapter 2):

- Usage of the TLS v.1.2[16]; no earlier version of TLS/SSL will be supported.

[14] The role of the parents and the students are different and, apparently, a parent may get access only to her/his child's data; however, there may by the option to allow a parent/guardian to set the password of his/her child.
[15] Probably under a specific role of a somehow "local school administrator".
[16] In case that TLS v.1.3 becomes a standard, the iRead system will be appropriately adjusted.

- The encryption algorithm will be exclusively AES (with either 128 or 256 bit key-length) in Galois Counter Mode of operation.
- The digital certificate of our web server will be granted (and signed) by a trusted Certification Authority.

More technical description of TLS technologies is given in the Appendix.

Even though in offline mode of operation of iRead, it is crucial to consider communication security (e.g. it is needed to secure uploading/downloading of resources such as user-models, which are prerequisite for performing an offline mode of operation for the iRead). Moreover, exchanges of data between the different research sites may be also needed. To achieve security goals in all these cases, the secure FTP (sFTP) technology seems to be an appropriate solution; however, other options such as using PGP will be also considered.

5.3.4. Access control

The most appropriate model for access control, in the context of the iRead, seems to be the Attribute Based Access Control Model, owing to the fact that users with the same role may be associated with different access rights; for instance, two teachers, although they have the same roles, they should have access only to their students' data – and, thus, the first teacher should not have access to the students' data of the second teacher and vice versa.

The attributes corresponding to each user's access rights will be based on the "need-to-know" principle (see Chapter 2, page 12, and Chapter 5, page 31). In general, each student will have access only to her/his data, whilst the teacher (and relevant members of the school senior leadership team with reporting responsibilities) will have access only to the data of her/his students. The experts will have access to appropriate data depending on their specific role; all the access rules will be specifically determined and documented (see Section 5.4). These access control rules will be forced both to application users as well as to database users.

With regard to the implementation of the access control, OAuth 2.0, which constitutes a de-facto standard for managing distributed web authorization, will be considered.

5.3.5. Data at rest security

Appropriate encryption techniques will be also applied to crucial files of the iRead system that are being stored, in order to enhance confidentiality and confront the risks of a data breach. To this goal, database encryption will be performed on the database server, via an effective encryption algorithm (most database servers support such encryption technologies). Moreover, depending on other characteristics of the iRead system (e.g. the offline mode of operation), further encryption of files that are being processed locally in the student's tablets will be also considered in terms of mitigating the risks of a device lost.

Finally, other security measures will be also adopted in order to thwart several attacks that can be mounted on the system (especially network attacks, which are the most dangerous); possible security measures include the use of properly configured firewalls to exclude unauthorized access to the system, the use of intrusion detection system to identify abnormal behavior or traffic that could be a sign of an ongoing security attack, as well as the use of appropriate tools to identify malicious software. In addition, appropriate backup procedures will also be in place to protect the stored data and ensure the availability of the data.

Several available security technologies – not an exhaustive list though - are being presented in the Appendix.

5.4. Data Protection Impact Assessment – Security Policy

As stated in Chapter 3, a Data Protection Impact Assessment (DPIA) should be performed under the provisions of the GDPR, with the ultimate goal to output the appropriate measures that should be chosen and implemented to remedy the data protection risks for individuals. There are several approaches for conducting a DPIA – all of them though employ somehow the following steps[7]:

1. define and describe the context of the processing of personal data under consideration and its stakes;
2. identify existing or planned controls (to comply with legal requirements and to treat privacy risks in a proportionate manner);
3. assess privacy risks to ensure they are properly treated;
4. make the decision to validate the manner in which it is planned to comply with privacy principles and treat the risks, or review the preceding steps.

In this report, we have performed the aforementioned steps 1-3 of a DPIA (see Chapters 4 and 5); it remains though to make a decision on all explicit measures that will be adopted and validate them in terms of their effectiveness on treating the risks. A DPIA report will be provided at the end of this process.

With regard to security measures as discussed in Section 5.3, a security policy document needs to be developed, which will outline the main security goals, specific requirements or rules that must be met as well as the corresponding responsibilities of all parties involved. The security policy document will also explicitly describe the access rights for each role, as have been generally described above.

6. Conclusions

This document illustrates how data protection and security issues relevant to the iRead system can be addressed, via presenting applicable architectural patterns and controls pertaining to security and privacy. In this framework, this document constitutes an implementation of the first steps of a DPIA with regard to the iRead system. More precisely, two pillars have been considered:

1. Fundamental principles and rights, which are established by law and must be respected regardless of the nature, severity and likelihood of risks;
2. Management of data subjects' privacy risks, in order to determine the appropriate technical and organizational controls to protect personal data.

With regard to fundamental principle and rights, the new General Data Protection Regulation that will apply on May 25th 2018 has been also taken into account.

On this regard, the present document illustrates the following steps of a DPIA:

1. definition and description of the context of the processing of personal data under consideration;
2. identification of existing or planned controls to comply with legal requirements and to treat data protection risks in a proportionate manner.

Several specific decisions on the overall design of the iRead system will be based on the above analysis; an evaluation and review of all these decisions will be subsequently performed, which will lead to a DPIA report. In addition, a security policy document will be developed that will explicitly state all security goals and requirements, as well the corresponding responsibilities of all the involved parties; this document will also present the finalized access control mechanism, based on the data protection principles described in this document.

This document will constitute the basis for any possible use case of the iRead system as those described in Section 4.2 – i.e. all the data protection and security controls will be appropriately reflected to any possible use of the iRead system.

As stated in the Introduction, the present document can be appropriately updated during the development of the iRead system, if required.

References

[1] N. J. AlFardan, D. J. Bernstein, K. G. Paterson, B. Poettering, and J. C. N. Schuldt, *On the security of RC4 in TLS*, In Proceedings of the 22d USENIX Conference on Security, pp. 305-320, USENIX Association, August 2013.

[2] N. J. AlFardan and K. G. Paterson, *Lucky thirteen: Breaking the TLS and DTLS record protocols*, In IEEE Symposium on Security and Privacy, pp. 526-540, May 2013.

[3] Article 29 Working Party, *Opinion 5/2014 on Anonymization Techniques*, 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[4] N. Aviram et. al., *DROWN: Breaking TLS Using SSLv2*, In 25th USENIX Security Symposium, pp. 689-706, USENIX Association, August 2016.

[5] H. Brody, *How HTTPS Secures Connections: What Every Web Dev Should Know*, July 2013, https://blog.hartleybrody.com/https-certificates/, accessed April 6th, 2017.

[6] B. Canvel, A. Hiltgen, S. Vaudenay, and M. Vuagnoux, *Password interception in a SSL/TLS channel*, In Advances in Cryptology - CRYPTO 2003: 23rd Annual International Cryptology Conference, Lecture Notes in Computer Science, vol. 2729, pp. 583—599, Springer-Berlin, Heidelberg, August 2003.

[7] CNIL (Commission Nationale de l'Informatique et des Libert´es): *Privacy Impact Assessment: Methodology (how to carry out a PIA).* CNIL (2015). http://www.cnil.fr/fileadmin/documents/en/CNIL-PIA-1-Methodology.pdf

[8] J. Daemen and V. Rijmen, *The Design of Rijndael: AES – The Advanced Encryption Standard*, Springer-Berlin, Heidelberg, 2002.

[9] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L. 1995;281:31–50.

[10] T. Duong and J. Rizzo, *Here come the xor ninjas*. In Ekoparty Security Conference, 2011.

[11] FIPS 202, *SHA-3 Standard: Permutation-Based Hash And Extendable-Output Functions,* NIST.

[12] S. M. Furnell, S. Katsikas, J. Lopez, A. Patel, *Securing Information and Communications Systems: Principles, Technologies, and Applications*, Artech House, Inc. Norwood, MA, USA, 2008.

[13] C. Garman, K. G. Paterson, and T. V. der Merwe, *Attacks only get better: Password recovery attacks against RC4 in TLS*, in 24th USENIX Security Symposium (USENIX Security 15), USENIX Association, pp. 113-128, August 2015.

[14] M. Gregg, *Six ways hackers try to break Secure Sockets Layer-encrypted data*, http://searchnetworking.techtarget.com/tip/Six-ways-hackers-try-to-break-Secure-Sockets-Layer-encrypted-data, accessed April 6th, 2017.

[15] *NIST Special Publication 800-162. Guide to Attribute Based Access Control (ABAC),* NIST.

[16] OASIS eXtensible Access Control Markup Language Technical Committee, *eXtensible Access Control Markup Language (XACML),* http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml .

[17] K. G. Paterson, B. Poettering, and J. C. N. Schuldt, *Big Bias Hunting in Amazonia: Large-Scale Computation and Exploitation of RC4 Biases (Invited Paper),* in Advances in Cryptology – Asiacrypt 2014, pp. 398—419, Lecture Notes in Computer Science, vol. 8873, Springer-Berlin, Heidelberg, December 2014.

[18] E. Rescorla, *The transport layer security (TLS) protocol version 1.3*, draft-ietf-tls-tls13-latest, 2016.

[19] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) Official Journal L. 2016;119(1).

[20] SANS, *Password Construction Guidelines, 2014,* *www.sans.org/security-resources/policies/general/pdf/password-construction-guidelines* accessed April 20th, 2017.

[21] W. Stallings, *Cryptography and Network Security: Principles and Practice*, 6th ed., Pearson, 2014.

[22] M. Vanhoef and F. Piessens, *All your biases belong to us: Breaking RC4 in WPA-TKIP and TLS*, in 24th USENIX Security Symposium, pp. 97—112, USENIX Association, August 2015.

[23] *Website Password hacking using WireShark*, Never Ending Security, April 2015, https://www.blackmoreops.com/2015/04/11/website-password-hacking-using-wireshark/ , accessed April 6th, 2017.

[24] http://docs.oracle.com/javase/7/docs/technotes/guides/security/, accessed April 5th, 2017.

[25] https://docs.oracle.com/javaee/7/tutorial/partsecurity.htm#GIJRP, accessed April 5th, 2017.

[26] https://docs.oracle.com/javaee/7/tutorial/security-intro006.htm#BNBXW, accessed April 5th, 2017.

[27] https://www.instantssl.com/ssl-certificate-products/https.html, accessed April 6th, 2017.

Appendix

Current Technologies for Security

In the following, we present some indicative available technologies for security of client/server systems and computer networks in general. The exact security mechanisms that will be finally employed will be decided during implementation, since they depend on design decisions about the architecture of the software system.

- **TLS (Transport Layer Security):** Transport Layer Security (TLS) is a protocol widely used for authentication and encryption of communications between clients and servers. Transport Layer Security (TLS) and its predecessor, Secure Sockets Layer (SSL), are both frequently referred to as SSL; however, since there are known security weaknesses in all versions of SSL, we specifically refer to TLS so as to clarify that the most secure version of the protocol is being considered. The TLS runs above transport layer protocols like TCP/IP and below application level protocols (HTTP, IMAP) whilst providing authentication by employing public-key encryption. The TLS protocol comprises two sub-protocols, namely the Record protocol and the handshake protocol. The first one is responsible for defining and encapsulating the format of the data to be transferred, while the second one uses the Record Protocol for exchanging a series of messages between the client and the server upon establishment of the first connection. The handshake protocol is employed before any data transmission and allows the server and the client to authenticate each other therefore resulting in a successful TLS connection. The main security issues that TLS protocol addresses are server authentication, client authentication and encrypted connectivity.

- **TLS Server authentication mechanism:** A TLS enabled client application can employ known public encryption key techniques to confirm that the certificate and the ID of a server are valid and to check if the server's certificate authority is listed among the list of the client's trusted certificate authorities.

- **TLS Client authentication mechanism:** Optionally, a TLS enabled server application employs the same techniques as those used for the server authentication, in order to confirm whether the certificate and the ID of a client are valid and to check if the client's certificate authority is listed among the list of the server's trusted certificate authorities.

- **TLS Connection (Encrypted):** Confidentiality in a client-server communication is achieved by public-key cryptography: the sending application encrypts a message using the recipient's public key, while the receiving application decrypts the message using its paired private key. In a TLS connection a complementary mechanism exists, which automatically detects if the transferring data has been modified or altered.

- **SSH (Secure Shell):** In computing, Secure Shell or SSH is a set of standards and an associated network protocol that allows for the establishment of a secure channel between a local and a remote computer. Public-key cryptography is used for the user to authenticate the remote computer and (optionally) for the remote computer to authenticate the user. Confidentiality and integrity of data exchanged between the two computers is assured by using encryption and message authentication codes (MACs). Most commonly SSH is used to log into a remote machine and execute commands. However, other network services are also supported such as tunneling, forwarding arbitrary TCP ports and X11 connections; it can

transfer files using the associated SFTP or SCP protocols. An SSH server, by default, listens on the standard TCP port 22.

- **Secure HTTP:** The Secure HTTP (S-HTTP), is an extension to the HTTP protocol designed to enable the exchange of secure data over the Internet. S-HTTP is based on the underlying TSL protocol to implement a secure connection between the sender and the receiver, and encrypt the contents. In other words, whenever the TLS offers all its security features (confidentiality, integrity, authentication) to the HTTP, we refer to S-HTTP [5], [23], [14], [27].

- **FTP:** FTP or File Transfer Protocol is used to connect two computers over the Internet so that the user of one computer can transfer files and perform file commands on the other computer.

The *original* FTP specification lacks any specified method for encrypting transferring data, and is therefore an inherently insecure method for file transferring. This implies that under most network configurations, the entire conversation (user names, passwords, FTP commands and transferred files) can be "sniffed" or viewed by anyone on the same network who is using a packet sniffer. This problem is common to many Internet protocol specifications written before the creation of SSL/TLS, such as HTTP, SMTP and Telnet. The most common workaround to this problem is to use either SFTP (SSH File Transfer Protocol), or FTPS (FTP over TLS), which adds TLS encryption to FTP.

  - FTP over SSH**:** FTP over SSH refers to the practice of tunneling a normal FTP session over an SSH connection. FTP over SSH is sometimes referred to as secure FTP; this should not be confused with other methods of securing FTP, such as with SSL/TLS (FTPS). Other methods of transferring files using SSH that are not related to FTP include SFTP and SCP, where the entire conversation (credentials and data) is always protected by the SSH protocol.

  - FTPS: FTPS (commonly referred to as FTP/SSL) is a name used for a number of different ways in which FTP software can perform secure file transfers. Each way involves the use of a SSL/TLS layer below the standard FTP protocol in order to encrypt the control and/or data channels. It should not be confused with SSH file transfer protocol.

- **Firewall:** A firewall is a mechanism used to prevent uncontrolled access from an unsecured network to a private network. In more detail the firewall is a device that all traffic goes through, keeping unauthorized users out of the private network, and providing various kinds of protection. Mainly there are three types of firewalls, the Routers, the Application-level Gateways and the Circuit-level Gateways [21].

  - **Routers:** A typical router is configured to filter incoming and out coming packets (from and to the unsecured network). Packets are either forwarded or discarded by a set of rules the router applies. These rules are related to source and destination IP addresses as well as the TCP port number. The filtering process considers sets of rules based on matches in the IP and TCP headers. Whenever there is a match to one of the rules, then this rule is employed for deciding on whether to forward or to discard the packet. In the case where there is no matching to any rule, then default actions (discard or forward) are taken depending on the router's configuration.

  - **Application-level Gateway:** The application-level gateway can be thought as a switch in the middle of the application-level communication, i.e. between the two end-points. When the user contacts the gateway (using a TCP/IP application), the user is asked to

give the name of the remote host he wants to reach. If the authentication information provided by the user is valid, then the gateway contacts the remote host application and transfers the TCP segments containing the application data between the two end-points. In the case where the gateway is unable to implement the proxy-code for a specific application, then the service cannot be forwarded through the firewall. The advantage of the gateway over the rooter is that it examines only which applications are allowed to connect, while the router has to deal with a lot of possible combinations at the TCP and IP levels. The disadvantage of the application level gateway is that additional processing is required on each connection.

○ **Circuit-level Gateway:** Circuit-level gateways can be thought of as special function of application level gateways. This kind of firewall does not permit direct connections between the two end-points, instead it establishes two connections, one with the TCP user of the inside host and one with the user of the outside host. When these two connections are set up the gateway sends the TCP segments from the one connection to the other. The gateway does not examine the content of the data, while the security mechanism determines which connections will be allowed or not. Circuit-level gateways are typically used in cases where the system administrator trusts the internal users.

○ **Proxies:** A proxy device (running either on dedicated hardware or as software on a general-purpose machine) may also act as a firewall. It can respond to input packets (e.g. connection requests) in the same way as an application, while blocking other packets. With proxies, tampering with an internal system from the external network is far more difficult, and, as long as the application proxy remains intact and properly configured, possible misuse of an internal system does not necessarily cause a security breach that can be exploited from outside the firewall. On the other hand, intruders may hijack a publicly-reachable system and use it as a proxy for their own purposes. In such cases, the proxy masquerades itself as that system to other internal machines. While use of internal address spaces enhances security, crackers may still employ IP spoofing or other methods in an attempt to pass packets to a target network.

Java Security Mechanisms

The following sections discuss the characteristics of the common mechanisms that can be used to secure Java applications based on Java EE server(s) and Android, Java SE or Web clients. Java SE Security Mechanisms [24] are the following:

- **Java Authentication and Authorization Service (JAAS)** is a set of APIs that enable services to authenticate and enforce access controls upon users. JAAS provides a pluggable and extensible framework for programmatic user authentication and authorization.

- **Java Generic Security Services (Java GSS-API)** is a token-based API used to securely exchange messages between communicating applications.

- **Java Cryptography Extension (JCE)** provides a framework and implementations for encryption, key generation and key agreement, and Message Authentication Code (MAC) algorithms, includes interfaces and implementations of message digests and digital signature.

- **Java Secure Sockets Extension (JSSE)** is the Java version of the Secure Sockets Layer (SSL) and Transport Layer Security (TLS) protocols. Includes functionality for data encryption, server authentication, message integrity, and optional client authentication to enable secure Internet communications.

- **Simple Authentication and Security Layer (SASL)** is a framework for authentication and optional establishment of a security layer between client and server applications. SASL defines how authentication data is to be exchanged but does not itself specify the contents of that data. SASL is a framework into which specific authentication mechanisms that specify the contents and semantics of the authentication data can fit.

Java EE Security Mechanisms provide a robust and easily configured security mechanism for authenticating users and authorizing access to application functions and associated data at many different layers [25][26]:

- **Application-Layer Security**. In Java EE, component containers are responsible for providing application-layer security, security services for a specific application type tailored to the needs of the application. At the application layer, application firewalls can be used to enhance application protection by protecting the communication stream and all associated application resources from attacks. Java EE security is easy to implement and configure and can offer fine-grained access control to application functions and data. However, as is inherent to security applied at the application layer, security properties are not transferable to applications running in other environments and protect data only while it is residing in the application environment. In the context of a traditional enterprise application, this is not necessarily a problem, but when applied to a web services application, in which data often travels across several intermediaries, you would need to use the Java EE security mechanisms along with transport-layer security and message-layer security for a complete security solution. The advantages of using application-layer security include the following.
  - Security is uniquely suited to the needs of the application.
  - Security is fine grained, with application-specific settings.

  The disadvantages of using application-layer security include the following.
  - The application is dependent on security attributes that are not transferable between application types.
  - Support for multiple protocols makes this type of security vulnerable.
  - Data is close to or contained within the point of vulnerability.

Acronyms

DPD	Data Protection Directive
DPIA	Data Protection Impact Assessment
GDPR	General Data Protection Regulation
MAC	Message Authentication Code
TLS	Transport Layer Security
SSL	Secure Socket Layer
SHA	Secure Hash Algorithm
AES	Advanced Encryption Standard
ABAC	Attribute-Based Access Control
RBAC	Role-Based Access Control