10

TRUESSEC Trustworthiness Label Recommendations

Danny S. Guamán[1,7], Manel Medina[2,3], Pablo López-Aguilar[3], Hristina Veljanova[4], José M. del Álamo[1], Valentin Gibello[6], Martin Griesbacher[4] and Ali Anjomshoaa[5]

[1]Universidad Politécnica de Madrid, Departamento de Ingeniería de Sistemas Telemáticos, 28040, Madrid, Spain
[2]Universitat Politécnica de Catalunya, esCERT-inLab, 08034, Barcelona, Spain
[3]APWG European Union Foundation, Research and Development, 08012, Barcelona, Spain
[4]University of Graz, Institute of Philosophy and Institute of Sociology, 8010, Graz, Austria
[5]Digital Catapult, Research and Development, NW1 2RA, London, United Kingdom
[6]University of Lille, CERAPS – Faculty of Law, 59000, Lille, France
[7]Escuela Politécnica Nacional, Departamento de Electrónica, Telecomunicaciones y Redes de Información, 170525, Quito, Ecuador
E-mail: ds.guaman@dit.upm.es; medina@ac.upc.edu; pablo.lopezaguilar@apwg.eu; hristina.veljanova@uni-graz.at; jm.delalamo@upm.es; valentin.gibello@univ-lille.fr; m.griesbacher@uni-graz.at; ali.anjomshoaa@ktn-uk.org

The main goal of TRUESSEC project is to foster trust and confidence in new and emerging ICT products and services throughout Europe by encouraging the use of assurance and certification processes that consider multidisciplinary aspects such as sociocultural, legal, ethical, technological and business while paying due attention to the protection of Human Rights.

TRUESSEC's central recommendation to the European Commission (EC) is a label scheme that can suitably address found issues that is worth developing and testing. While actual software development is beyond the current scope of TRUESSEC, the remainder of this paper describes the characteristics of such a solution, allowing the EC to commission a working prototype should it wish to do so.

At the heart of the proposed solution is a set of prioritized survey questions that take into account a set of core areas of trustworthiness to produce both a visual "transparency" statement that is easy for the citizen to understand, and additionally provides a specific piece of code to enable machine-to-machine integration based on the policy settings of 3rd party users. In this regard, the Creative Commons licensing model[1] is analogous to our proposed solution.

10.1 Introduction

This paper provides a recommendation for a TRUESSEC labelling solution, aimed to show users the level of trustworthiness of applications and services, according to multi factor criteria.

The central task of the TRUESSEC project is to apply an interdisciplinary approach, encompassing ethics, sociology, law and technical engineering, to make recommendations to the European Commission for a certification and labelling of ICT products and services that will foster trust among citizens that use them.

Both the core areas that constitute "trust" (which spans cybersecurity through to branding and user experience), and the various potential fields of application (from web services to cyber-physical systems) means that the remit of this project is very broad indeed.

Nevertheless, the project team values this approach and, as background to our recommendation, has noted that good progress has been made with European legislation which, over time, is likely to enhance levels of citizen trust. Even though the Digital Single Market legal framework is still a work in progress, these advances have resulted in a strong legal foundation to protect the rights of EU citizens entrenched in the Charter of Fundamental Rights of the EU [1].

In addition, pan-European bodies, such as ENISA[2], are progressing well with security and privacy certification and codes of conduct in relatively

[1] https://creativecommons.org

[2] See the Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on ENISA, the "EU Cybersecurity Agency", and repealing Regulation (EU) 526/2013, and on Information and Communication Technology cybersecurity certification ("Cybersecurity Act"), COM/2017/0477 final - 2017/0225 (COD).

new areas, such as security in the Cloud and Internet of Things – although certification remains a voluntary responsibility of the online service providers with little legal implications.

Our research started "evaluating existing trustworthiness seals and labels" [2], and the analysis of these existing schemes showed a general lack of adoption and awareness, as well as poor transparency regarding what is being certified and under what conditions. In fact, citizens tend to employ other indicators of trust (3rd party payment systems, branding, user experience, and user-based recommendation engines) to make decisions about their use of a service, despite how little guarantees they actually offer.

TRUESSEC's research work also went beyond current business practices, technology and legislation to explore the social and ethical questions behind what constitutes trust from users. This is summarized by our criteria catalogue, which was published as deliverable [3].

Given these inputs, there are a number of issues with existing label schemes:

- There are too many labels to provide a common understanding for citizens or service providers
- Businesses tend not to understand the cost/benefits of using labelling
- They are not sufficiently flexible and updated to acknowledge relatively new legislation, such as the GDPR
- They are not inclusive enough to incorporate additional 3rd party certification
- They do not "go beyond the law" to enable service providers to demonstrate that they have taken an ethical, responsible and transparent approach
- They rarely encompass all major components of trust such as safety or "security by design", personal data protection and consumer rights enforcement
- They provide insufficient information on how they are awarded and on the safeguards offered

These shortcomings mean that current labels are often out of date, removed from best practices, poorly understood and therefore little known and used.

10.2 Interdisciplinary Requirements

TRUESSEC.eu Core Areas of trustworthiness are based on the findings from five support studies, **considering the European values and**

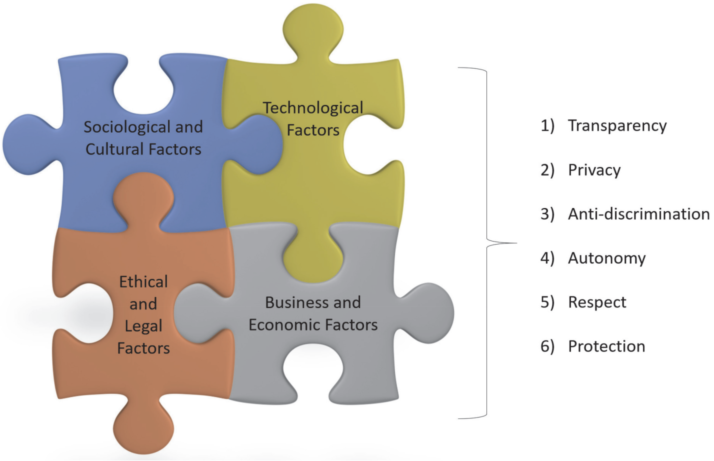


Figure 10.1 TRUESSEC.eu core areas of trustworthiness.

fundamental rights as well as following joint work among all disciplines represented in the TRUESSEC.eu project. Six Core Areas have been agreed upon, that set the stage for the search of the multidisciplinary criteria [4]. The Core Areas displayed in Figure 10.1 represent the reflections of the five disciplines: ethics, law, sociology, business and technology.

Transparency. The TRUESSEC.eu Core Area transparency reflects the understandings of the five disciplines by having information in its focus. In this regard, the Core Area transparency evolves around the fulfilment of information duties related to personal data processing, but it also goes beyond that, as the business perspective shows. Overall, transparency can help to narrow down the existing informational gap and give users clearer answers to questions regarding their personal data and the products and services they purchase.

Privacy. The TRUESSEC.eu Core Area privacy is equally important in all disciplines. When users are provided with relevant information, this sets the ground for them to take control over their data. On the one hand, users must be able to make decisions regarding their personal data; on the other hand, providers must respect those decisions. The latter is a striking point, as providers have commercial interests in processing as many data as possible. Considering the economic relevance of data and the emerging data economy, it is crucial to ensure the protection of personal data. This includes considering aspects of privacy throughout the design and development of an

ICT product or service (privacy by design) as well as offering the privacy settings at a high level of privacy protection (privacy by default).

Anti-discrimination. This Core Area has a great relevance for trustworthiness. The need to formulate such a core area stems from the fact that discrimination concerning ICT products and services is present and it is very often hidden in decision-making carried out by algorithms and self-learning systems. This particularly relates to cases where parameters are included in the decision-making process, which go beyond the scope of the service or product in question.

Autonomy. The TRUESSEC.eu Core Area autonomy summarizes well the considerations of the five disciplines. Having access to and rights to use various ICT products and services brings up one very central issue, which is, the need for users to be given the opportunity to make decisions regarding their personal data. These decisions need to be well informed and free of manipulation and coercion.

Respect. The TRUESSEC.eu Core Area respect presents a transition from discipline-related understanding to a transdisciplinary one. It embodies the idea that based on societal, legal and ethical frameworks there are certain duties that arise for ICT providers that ground legitimate expectations on the side of users when dealing with ICT products and services. Legitimate expectations have three main hallmarks: they are predictive, prescriptive and justifiable. In the ICT context, this would suggest that users create expectations on what ICT providers will and should or should not do, or how they will and should operate. Whereby these expectations are justifiable, that is, users have justification or warrant for forming them in the first place. Example of such legitimate expectations is that ICT providers respect users' rights and freedoms.

Protection. The considerations of all five disciplines seem to be focused in the protection of individuals against any harms as well as the protection of their rights and freedoms. This has led us to formulate the TRUESSEC.eu Core Area protection as the sixth core area. In the context of ICT, protection relates to both safety and security thus encompassing risks of physical injury or damage and risks related to data such as unauthorized access, identity theft etc. In order to enable solid level of protection, compliance with already established safety and cybersecurity standards is essential. The aim is to hinder any harms that may be caused because of using ICT in the first place.

10.3 Criteria Catalogue and Indicators[3]

The TRUESSEC.eu Criteria Catalogue represents a constituent part of the TRUESSEC.eu work on labelling. It is a multidisciplinary endeavour to compile a list of criteria and indicators that could contribute towards enhancing the trustworthiness of ICT products and services. The development of the TRUESSEC.eu Criteria Catalogue consists of two phases: (a) development of the First Draft Criteria Catalogue, which includes only ethical and legal criteria and indicators, and (b) development of the multidisciplinary Criteria Catalogue, which builds upon the First Draft Criteria Catalogue, but it also includes sociological, business and technical input.

The basis for the Criteria Catalogue consists of the European values as stated in Article 2 of the Treaty of the European Union and the European fundamental rights, on the one hand, and the findings from the five support studies prepared in the first year of the project as well as some interdisciplinary work and discussion, on the other hand. It is from here that we extracted the hierarchical structure of the Criteria Catalogue. As depicted in Figure 10.2, we started with high-level concepts we called **Core Areas**. The very aim of the Core Areas is to provide a framework which in a next step could be broken down into elements that are more specific. In that sense, the Core Areas reflect the values that should be considered in the design and use of ICT products and services, and thus serve as an orientation tool when determining the criteria. Based on the Core Areas we then developed the **criteria**. The criteria show what requirements an ICT product and service should fulfil in order to be considered trustworthy. In the hierarchical structure, the criteria are less abstract than the Core Areas; however, they are still not concrete enough to be measurable. For that purpose, we formulated **indicators,** which could be measured. A set of indicators is determined for each single criterion. The aim of the indicators is to indicate the degree to which a particular criterion is met.

Based on the support studies and the interdisciplinary discussion we defined six TRUESSEC.eu Core Areas of trustworthiness: *transparency, privacy, anti-discrimination, autonomy, respect* and *protection* and provided a TRUESSEC.eu multidisciplinary understanding of each of them (see Table 10.1).

[3] For more on the TRUESSEC.eu Criteria Catalogue see Stelzer et al. "TRUESSEC.eu Deliverable D7.2: Cybersecurity and privacy Criteria Catalogue for assurance and certification," 2018, https://truessec.eu/library .

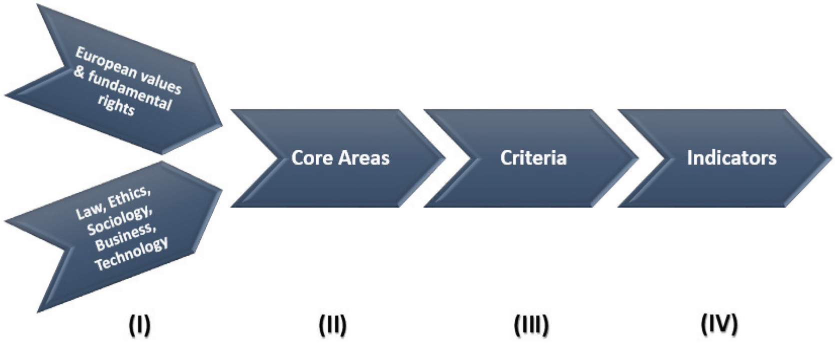

European values & fundamental rights
Law, Ethics, Sociology, Business, Technology
Core Areas
Criteria
Indicators
(I)
(II)
(III)
(IV)

Figure 10.2 Developing the criteria catalogue.

Table 10.1 TRUESSEC.eu Core Areas of trustworthiness

TRUESSEC.eu Core Areas	Multidisciplinary TRUESSEC.eu Understanding
Transparency	The ICT product or service is provided in line with information duties regarding personal data processing and the product/service itself.
Privacy	The ICT product or service allows the user to control access to and use of their personal information and it respects the protection of personal data.
Anti-discrimination	The ICT product or service does not include any discriminative practices and biases.
Autonomy	The ICT product or service gives users the opportunity to make decisions and respects those decisions. The ICT product or service also respects other parties'/persons' rights and freedoms.
Respect	ICT products or services are to be provided in accordance with the legitimate expectations related to them.
Protection	ICT products and services are provided in accordance with safety and cybersecurity standards.

To give a better understanding of the interdisciplinary nature of the Core Areas, we show some exemplary details on the discussion of transparency. From an **ethics** perspective, transparency relates to two aspects: (a) providing clear and sufficient information about products and services in general and (b) more specifically providing information to users regarding activities with their personal data. **Legally**, transparency can be understood as in information duties, laid down in the GDPR, the Directive on consumer rights or the e-commerce Directive. With respect to personal data, transparency is one of the core principles of data processing (Article 5 GDPR). From a **technology** perspective, transparency (in data protection) is defined as the property that all personal data processing can be understood (intelligible and meaningful) at any time by end-users (i.e., before, during, and after processing takes place). In the technical domain there is also a concept named ‘Service Level Agreement’, which describes technical specification of the service/product being used. You may think e.g. on a service availability, uptime, etc. These more normatively oriented definitions can also be complemented by a **sociological** perspective, which focusses on public opinion. Considering that currently (Eurobarometer data from 2015):

(a) only a minority of EU citizens reads privacy statements (less than 1/5),
(b) only about 4 out of 10 of internet users read the terms and conditions on online platforms,
(c) over 90 % want to be informed if their data ever was lost or stolen,

It can be assumed that there is a need for improvement in current information practices.

Having well-informed citizens, e.g. on the risks of cybercrime, also leads to improved cybersecurity behaviour, which emphasizes the importance of transparency and information. These interdisciplinary considerations can also be connected to a **business** perspective. Transparency includes a wide range of business processes which range from being clear about terms of use of the online service, through to publishing transparency reports about the passing on of user data to 3rd parties, such as law enforcement. Transparency of service and use of personal data is increasingly being perceived by business as a competitive advantage.

From the six Core Areas we extracted the following twelve criteria of trustworthiness:

- Information
- User-friendly consent
- Enhanced control mechanisms
- Privacy commitment
- Unlinkability
- Transparent processing of personal data
- Anti-discrimination
- Cyber security
- Product safety
- Law enforcement declaration
- Appropriate dispute resolution
- Protection of minors

It should be emphasized that the way the criteria are ordered in this list does not indicate their importance per se. Furthermore, we consider this list of twelve criteria to be the groundwork consisting of the most fundamental criteria in the context of ICT products and services. In that sense, the list is not complete from the simple reason that with the technological developments additional criteria might have to be added.

In what follows, we will choose one criterion from the list and use it as an example to elaborate our approach. Table 10.2 illustrates this example.

The Criteria Catalogue is represented in a tabular form. It consists of three columns. The middle column represents the criterion. The right column represents the indicators. As the table shows, to each criterion a set of corresponding indicators are assigned that, when checked, should show to what degree the criterion is fulfilled. In the column on the left, which is named '**Trustworthiness enhancer**', are represented the six Core Areas into six sections. By adding this column, we wanted to show the interrelation between the criterion in question and the Core Areas. In order to show this, we used a colour system. We divided each of the six sections representing the six Core Areas into three subsections where a colour can be applied that would indicate the degree to which based on our assessments the criterion addresses each Core Area. In that sense, one could apply colour to one, two

Table 10.2 Criterion – Information

TRUSTWORTHINESS ENHANCER				CRITERION	INDICATORS
Transparency	🟦	🟦	🟦	Information	i. Information is provided: a. In a user-friendly manner • In a plain language (understandable to lay persons) • As long as necessary and as short as possible (e.g. in a form of one pager) b. Relevant to the context c. Clearly visible and easy to locate d. In a structured machine-readable format. ii. Information is provided free of charge.
Privacy	🟥	🟥	🟥		
Anti-discrimination	🟪				
Autonomy	🟨	🟨			
Respect	🟧				
Protection	🟦				

or three boxes, with three meaning the criterion fully addresses and meets the particular Core Area. This proved to be, eventually, a very useful way to check whether the group of criteria we identified sufficiently addresses the identified six Core Areas [5].

In this is represented the criterion 'Information'. Our findings showed that information plays undoubtedly an important part in enhancing trustworthiness of ICT products and services. Having the relevant information allows one to make informed decisions and it also creates a climate of openness, and transparency. In general, information consists of two aspects:

(a) **content**, namely, *what* the user is informed about, and
(b) **form**, or *how* information is provided.

Since the first aspect, which is related to the content reappears as an indicator in few other criteria, we have not included it here. In that sense, this criterion was limited only to the *form* of the information provided to the user. As the table shows, the indicators we assigned to this criterion should check whether the information is provided in a user-friendly manner, which means that the information is provided in a plain language that is easily understandable also for laypersons, and that it is as long as necessary and as short as possible. Regarding the length, we suggested that information should be provided in a form of one pager. Additionally, the information should be relevant to the context, easy to locate by the user and it should be provided in a structured machine-readable format. Apart from the format, we also included here another indicator which should check whether the information is provided free of charge. This is just one example of how the Criteria Catalogue operated. The same logic was followed for the other eleven criteria.

One of the main features of the Criteria Catalogue is that it adopts a post-compliance or beyond compliance framework. This framework is very similar to the framework suggested by Luciano Floridi [6]. When analysing the Digital, Floridi distinguishes between hard and soft ethics. Hard ethics is, as he explains, *"what we usually have in mind when discussing values, rights, duties and responsibilities–or, more broadly, what is morally right or wrong and what ought or ought not to be done"* [6]. Soft ethics, on the other hand, is post-compliance ethics as it goes beyond the compliance level and hence beyond existing regulation. In that sense, the aim of the Criteria Catalogue is to address this post-compliance or beyond compliance, for the simple reason that compliance is a very important part in making sure that a business acts within the legal framework. Nevertheless, for enhancing trustworthiness and strengthening trust, which is the main focus of the TRUESSEC.eu project, that might not always be sufficient. With this in mind, in the Criteria Catalogue we provide Core Areas, criteria and indicators as possible ways to address the post-compliance level.

The development of the Criteria Catalogue also paved the way for the drafting of the TRUESSEC.eu recommendations.

10.4 Operationalization of the TRUESSEC.eu Core Areas of Trustworthiness

Using Core Areas of trustworthiness as a starting point, a potential set of ICT system properties and detailed operational requirements have been defined. They attempt to bring Social Science and Humanities requirements closer to the technical domain and analyse which of them have already covered by the state-of-the-art and which need more attention from stakeholders. ICT system properties are quality or behavioural characteristics of a system that, ideally, can be distinguished qualitatively or quantitatively by some assessment method. There are several ICT system properties already defined and studied in the technical realm (e.g. security and safety), so the knowledge base around them can be leveraged to analyse and identify the specific operational requirements that need to be met and assessed for a specific ICT product or service. Figure 10.3 provides an overview of how we have mapped the Core Areas (and criteria) in ICT system properties (details can be found in [7]).

Once identified the ICT system attributes, they can become the basis for carrying out an operationalization process and deriving a set of specific operational requirements that can be realised and assessed.

As depicted in Figure 10.4, operational requirements are requirements of capabilities that should be guaranteed by an ICT product or service to satisfy one or more of the aforementioned ICT system properties. Moreover, they can be used as a precursor to the selection of more specific measures or countermeasures that are known as controls. Controls can be of technical nature (i.e. functionality in hardware, software, and firmware), organizational nature (i.e. organizational procedures related to the system environment and people using it), or physical nature (i.e. physical protective devices).

Finally, controls are instantiated using one or more specific techniques, which are found adequate to fulfil requirements of controls.

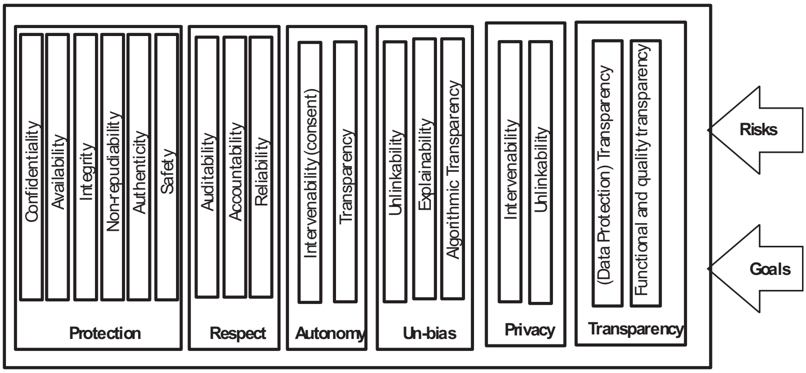


Figure 10.3 Core areas of trustworthiness and related ICT system properties.

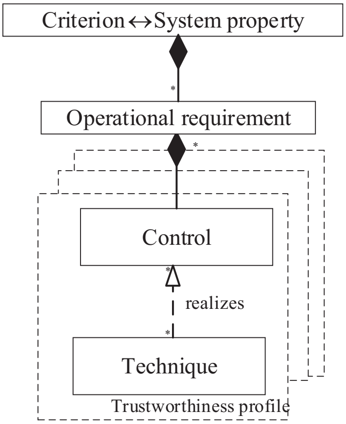


Figure 10.4 Guiding elements of the operationalization process.

It is worth noting the difference between operational requirements, controls, and techniques. Actually, both operational requirements and controls specify a system or organizational capabilities; however, an operational requirement recognises that a trustworthy capability seldom derives from a single control. In other words, one capability, depending on the context, may require several controls. On the other hand, while controls express what measure should be implemented, techniques indicate how it is implemented. Finally, it is important to mention that controls and techniques are context dependent, i.e. they are suitable for the specific context where a system is intended to work. Table 10.3 shows an example of the guiding elements of the operationalization process. Controls and a survey of the technical solutions for trustworthiness can be found in [8].

The state of the practice already includes plenty of controls contained within standard frameworks that, given the broad use of them during audits and certifications, enable to be closer to measurable (and assessable) factors and their corresponding evidence. Controls are widely used by the industry and the state of the practice shows hundreds of standards and certification schemes (around 290 according to ECSO[4]). Just to mention a few examples, security control frameworks include the ISO/IEC 15408 Common Criteria that contains a general catalogue of security requirements for ICT products, the ISO/IEC 27002 defines a set of organizational and technical controls intended to information security management, and the CSA Cloud Control Matrix (CCM) presents a catalogue of cloud-specific security controls. Privacy controls are defined, e.g. in the recent standard ISO/IEC 27018 that is intended to Cloud Service Providers (CSP) acting as Data.

Processors, the NIST 800-53 Rev4 contains security and privacy controls meant to Information Systems and Organizations, and; the General Accepted Privacy Principles (GAPP). Safety requirements e.g. are defined in the IEC 61508-2, they are intended to electrical, electronic, and programmable safety-related systems. Similarly, in the literature we can find significant works that propose, e.g. taxonomies of requirements that can be leveraged to operationalize some of the ICT system properties defined in the section above (e.g. using a goal-oriented approach). For instance, intervenability property can be refined into two guidelines: Data Subject Intervention and Authority Intervention. The first one representing intervention actions for data subjects and the latter the intervention actions for supervisory authorities to intervene in the processing of personal data. Each guideline can be refined into one or more operational requirements that act as success criteria, being empirically observable and objectively measurable. Following up with the intervenability property, the possible intervention actions by data subjects (e.g. do not consent, withdraw consent, review, challenge accuracy, challenge completeness, and request data copy) and the required ICT systems capabilities (e.g. access, no processing, restricted processing, amendment, correction, erasure, data copy, and suspended data flow) may lead to the definition of specific intervention readiness operational requirements. For example, before

[4] European Cyber Security Certification, A Meta-Scheme Approach v1.0. December 2017. Available under: http://www.ecs-org.eu/documents/uploads/european-cyber-security-certification-a-meta -scheme-approach.pdf

Table 10.3 Example of the guiding elements of the operationalization process

Guiding Element	Example
Core Area/system property	Protection/Security (Authenticity)
Technical requirement	The system shall provide two-factor authentication for remote access by individuals.
Controls	The system implements multifactor authentication for network access to privileged accounts (NIST 800-53 R4 IA-2-1). The system implements multifactor authentication for network access to non-privileged accounts (NIST 800-53 R4 IA-2-2). The system implements cryptographic mechanisms during transmission (NIST 800-83 R4 SC-8-1).
Techniques	For NIST 800-83 R4 IA-2-1 and IA-2-2, a combination of the following authentication factors can be used: • *Something the principal knows,* such as a password, a personal identification number (PIN), a graphical password, and answers to a prearranged set of questions. A password can be either static or dynamic (e.g. One-Time-password). • *Something the principal has*, such as a digital certificate, smart cards, and mobile phone. More recently, smartphones are being a potential alternative as a key enabler of secure authentication. Some of the latest smartphones include important security components such as a Trusted Platform Module that is able to secure digital certificates and cryptographic keys used for authentication. • *Something the principal is*, such as static biometrics (e.g. fingerprint, retina, and face) or dynamic biometrics (e.g. voice pattern, handwriting characteristics, and typing rhythm). On the other hand, the NIST 800-83 R4 SC-8-1 controls can be realised by AES or Triple DES; two approved symmetric algorithms.

collecting personal data, the system shall provide data subjects with the option to 'consent' and 'do not consent' the [processing instance].

Finally, while it should be recognised that the state of the art already provides plenty of controls contained in standard catalogues and frameworks for other more mature properties (mainly in the cybersecurity realm), controls related to anti-discrimination or autonomy are scarce and only recently there are some efforts and initiatives to address them (e.g. the EC has released ethics guidelines for trustworthy AI on April 8, 2019 [5]).

10.5 Recommendations

The European and international landscape of labels/seals is heterogeneous, as there is a great variation around their core functional models, the criteria they assess, the assurance level they offer, etc., and they also present a number of issues that need to be addressed [1]. For example, most of labelling core functional models require a complex chain of trust involving several third parties throughout the labelling process (e.g. evaluation body, certification/declaration authority, and accreditation authority). This complexity often results in a lot of time (and effort) required in the preparation and assessment of an ICT product/service, as well as in affordability issues due to the high costs involved. These issues are exacerbated when an ICT product or service must pass through the same process several times (one for obtaining the label and some other for certifying specific properties), involving additional cost and time. The industry has also highlighted these matters and called to "*minimize the burden on providers/manufacturers with respect to assessment, costs and time to market while ensuring an adequate level of trustworthiness*" [2].

While the TRUESSEC.eu labelling proposal advocates for addressing the complexity and affordability issues by reducing the intervention of third parties as far as possible, it also recognises the relevance of pursuing the verifiability and credibility of the labelling process. Providing the necessary evidence to support what is claimed about an ICT product or service improves verifiability. In turn, adding an independent public or private authority responsible for defining and articulating the labelling governance framework enhances credibility.

In this context, the TRUESSEC.eu proposal advocates a labelling solution that includes the following key elements: a self-assessment questionnaire, a labelling portal, a transparency report plus a visual label, and a governance framework ruled by an authority. Figure 10.5 illustrates the labelling approach proposed by TRUESSEC.eu.

- *The self-assessment questionnaire* is based on the indicators defined in the Criteria Catalogue. It provides a set of yes/no questions for a service provider to determine its compliance with the Criteria Catalogue. A

[5] European Commision, "Ethics Guidelines for trustworthy AI", https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accesed 12 April 2019)

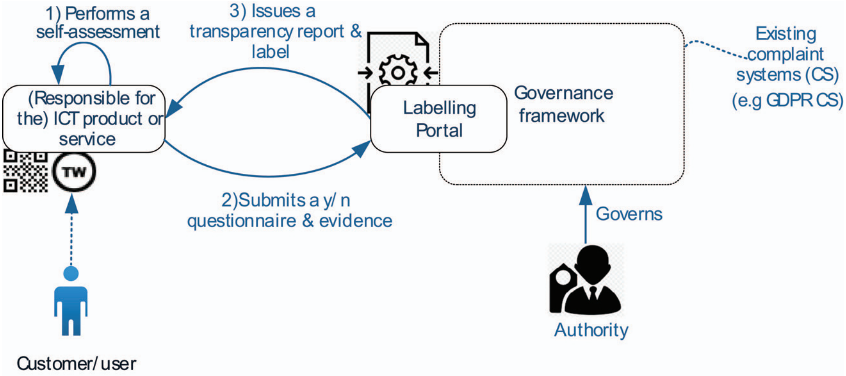

1) Performs a self-assessment
3) Issues a transparency report & label
Existing complaint systems (CS) (e.g GDPR CS)
(Responsible for the) ICT product or service
Labelling Portal
Governance framework
TW
2)Submits a y/ n questionnaire & evidence
Governs
Authority
Customer/ user

Figure 10.5 TRUESSEC.eu labelling proposal.

service provider performs the self-assessment and attaches the evidence of an indicator's fulfilment when the answer to a question is affirmative.

- *The labelling portal* processes the questionnaire answers and issues a transparency report and a visual label according to the level of conformance achieved for each of the twelve criteria included in the Criteria Catalogue.
- *The transparency report and the visual label* deliver a two-layer trustworthiness declaration. The visual label provides the first layer as it is easy to understand. The transparency report further details the assessment results in both text and machine-readable format, thus providing the second layer. The label and the report are both multi-dimensional (twelve criteria for trustworthiness) and multi-level (several levels of conformance for each criterion).
- *The governance framework* sets the fundamental rules the label must follow.

The following sections further elaborate on these elements.

10.5.1 Questionnaire

The questionnaire contains a set of yes/no questions, each asking whether an *indicator* of the Criteria Catalogue is met. The answer to each question allows to objectively determine which indicators are met and, ultimately, to what extent an ICT product or service meets a criterion for trustworthiness. Thus, we envisage a self-assessment and a yes/no questionnaire whereby providers/manufacturers reveal which *indicators* for trustworthiness they comply with, attaching the corresponding evidence when applicable. *Indicators* act as checkpoints, so they should be **empirically observable** (i.e. through evidence) and **objectively measurable** (i.e. a measurable element should be clearly defined in the indicators' description).

In our context, evidence refers to the information used to support the assessment and compliance of the *indicators.* Some evidences can refer to the implementation/realization of a given technique (e.g. the fifth indicator of the '*user-friendly consent' criterion: users are given the option to opt-out from data processing* can be supported by a centralized privacy control panel that includes opt-out options). Other evidences can describe organizational means to meet an indicator (e.g. those related to the *appropriate dispute resolution criterion*). Yet other evidence can be supported/provided by third parties who already performed an assessment on the subject-matter of the labelling e.g. through a certification or audit process. In this way, we prevent an ICT product or service from going through the same process several times (one for obtaining the label and some other for certifying specific properties). For example, a provider/manufacturer can link the certificate issued by a trusted third-entity as evidence of meeting the second indicator of the *'Cybersecurity' criterion: the ICT product or service is compliant with relevant [security] standards.*

An *indicator* should also include a measurable element easy to justify with evidence, calculate and understand. This measurable element should be clearly identified in the *indicator*'s description along with the corresponding measurement scale, which may be one of the following:

- **Nominal scales** are applicable for mapping values (without an intrinsic order) to categories, and only equality operation is allowed. The nominal **dichotomous** scale only has two categories and can be used to express whether a feature is present or not. In the Criteria Catalogue, several measurable elements are dichotomous in nature. For instance, the second *indicator* (ii) of the '*Cybersecurity' criterion* encloses a dichotomous measurable element with true or false as possible values. An evaluator will check whether the *ICT product or service is compliant with relevant [security] standards.* A provider/manufacturer can provide the certificate issued by a third trusted entity as evidence. Similarly, this can be applied for the first *indicator* of the *Privacy 'Commitment' criterion,* which states that "*The ICT provider clearly states its commitment to the GDPR in the form of a declaration*".
- **Ordinal scales** allow to sort or rank two or more categories, and equality and inequality operations are allowed. This may be applicable to, e.g.

the *'Enhanced control mechanisms' criterion.* The first *indicator* of this criterion states that *means to deletion of personal data should be provided.* In this respect, the *level of recovery* may be a measurable element intended to assess the difficulty (or easiness) to recover supposedly deleted data. For example, based on the guidelines and techniques presented into the NIST SP 800-88, three values on the ordinal scale can be abstracted:

- Level 1 (Clearing) – Deletion is done using overwriting software not only on the logical storage location but on also all addressable locations, so data cannot be easily recovered with basic utilities but could be possible with laboratory attacks.
- Level 2 (Purging) - Deletion is done using sophisticated sanitization techniques, so data cannot be possible at all.
- Level 3 (Destroying) – The media is destroyed (physical destruction).

Therefore, this measurable element can have three different ordinal levels, and the assurance of a given *level of recovery* can be an *indicator* attached to a particular *level of conformance*. For example, ensuring the Level 1 (Clearing) can be a criterion of the Level of Compliance 1, and the corresponding successive levels.

- **Interval/ratio scales** have numerical values and allow obtaining the difference or distance between them allowing be comparing and ordering. This may be applicable to measurable elements that have continuous numerical values. For example, the *period for the disposal of personal data once they have been processed for the purpose consented to* be another relevant, measurable element of the criterion mentioned in the previous paragraph. This period may have a continuous and infinite range of values, e.g. 1 day, 30 days, 365 days, etc. These quantitative, measurable elements can then be embedded in dichotomous (yes/no) *indicators* in terms of intervals or thresholds. As a matter of example, an *indicator* belonging to an advanced *level of conformance* may state that personal data are automatically deleted as soon as they are not used (0 days), while an *indicator* of *a basic/entry level of conformance* may state that personal data are deleted within 15 to 30 days.

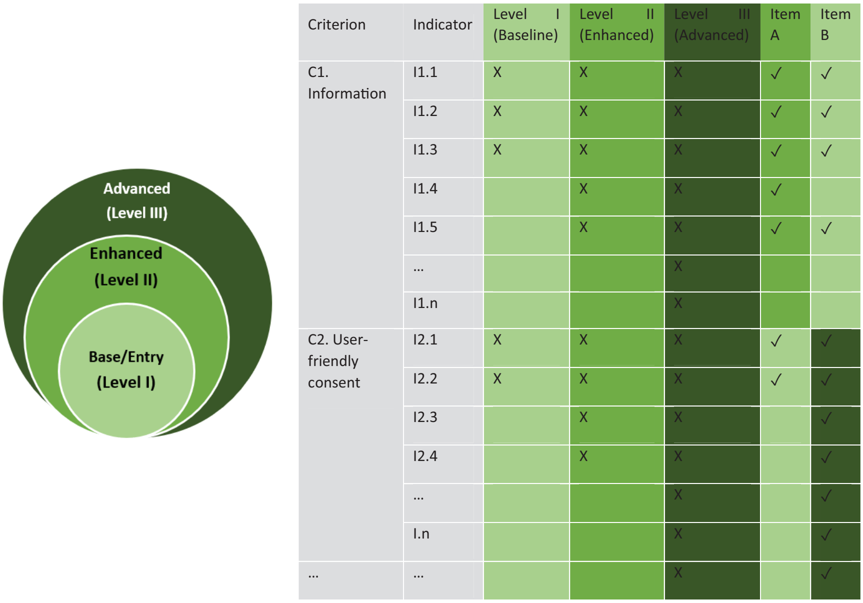
10.5.2 Labelling Portal

Based on the answers submitted by a provider/manufacturer, the labelling portal issues a transparency report and a visual label conveying the *level of* *trustworthiness* of the ICT product or service assessed. The notion of *level of trustworthiness* must be understood neither an absolute "yes/no trustworthy" nor as a single scalar "75.5% trustworthy", but as the extent to which the twelve criteria for trustworthiness defined in the Criteria Catalogue are fulfilled. This "extent" corresponds to one of the *levels of conformance* defined in the labelling scheme. To illustrate the notion of *level of conformance* and supported by the levelling structure defined in and [9], the following levels have been defined: Basic/Entry (Level I), Enhanced (Level II), and Advanced (Level III).

We advocate for an assessment based on groups of *indicators*, where each group is associated with a qualitative *level of conformance*. As illustrated in Figure 10.6, the *indicators* of each criterion are divided into subsets. Each subset is assigned to a particular Level of Conformance. For an ICT product or service to reach a superior *level of conformance* in any of its criteria, it must necessarily comply with all the *indicators* of the previous levels. Therefore, a criterion that has a Level I means that it complies with all the *indicators* belonging to Level I, Level II implies that a criterion complies with both Level I and Level II *indicators*, and Level III implies that a criterion complies with Level I, Level II, and Level III *indicators*.

An ICT product or service can have different *levels of conformance* for each of the twelve Criteria for trustworthiness. Figure 10.6(b) further depicts two items (ICT products or services) in its last two columns. On the one hand, the item A complies with Level II for criterion C1 (Information) and with Level I for criterion C2 (User-friendly consent). On the other hand, item B complies with level I for criterion C1 and level III for criterion C2. Also, note that item B does not conform to level II in criterion C1 because it fails to meet *indicator* I1.4. In this example, it can also be noted that if an ICT product or service is not able to comply with a single *indicator* for some level, it does not conform to that level.

The decision to define different levels for the Criteria is supported by different legislation; for example, the European GDPR (General Data Protection Regulation) defines different degrees of sensitivity of personal information, each requiring different privacy controls to protect them. Therefore, different privacy protection controls could be mapped to different subsets of *indicators*, each assigned to a respective *level of conformance*. Similarly, the Cyber Security Certification Framework by European Commission defines three Assurance Levels, each assigned to different subsets of requirements/criteria in terms of the risks involved.



Criterion	Indicator	Level I (Baseline)	Level II (Enhanced)	Level III (Advanced)	Item A	Item B
C1. Information	I1.1	X	X	X	✓	✓
	I1.2	X	X	X	✓	✓
	I1.3	X	X	X	✓	✓
	I1.4		X	X	✓	
	I1.5		X	X	✓	✓
	…			X		
	I1.n			X		
C2. User-friendly consent	I2.1	X	X	X	✓	✓
	I2.2	X	X	X	✓	✓
	I2.3		X	X		✓
	I2.4		X	X		✓
	…			X		✓
	I.n			X		✓
…	…			X		✓

(a) (b)

Figure 10.6 (Illustrative) Levels of conformance.

10.5.3 Transparency Report and Visual Label

The trustworthiness of an ICT product or service is expressed as twelve dimensions (criteria) each with a level of conformance depending on the subset of indicators met. These are conveyed to the label consumer through a two-layer declaration:

- The first layer shows a ***visual label*** that is easy for users to understand. It shows the extent to which each trustworthiness criteria is fulfilled (i.e. criterion plus its level of conformance).
- The second layer shows a ***transparency report*** in both text and machine-readable format. This should provide further details, i.e. criteria, *indicators* fulfilled, evidence provided (if applicable), and the individual *levels of conformance*. The machine-readable *transparency report* enables machine-to-machine integration based on e.g. the users' policy settings as set in their user agents such as a web browser. This may facilitate the automation of products and services trustworthiness comparison and assessment.

Both the transparency report and the visual label should highlight the date of the last update and should clearly specify which components of the product (modules/functionalities) or service (operations) are part of the labelling. In addition, in order to verify the authenticity of a label the following measures need to be considered:

- The *labelling portal* (who issues the *transparency report* and the *visual label*) should publicly provide a list of issued labels, including the two-layer information above described.
- The *visual label* also should integrate a link to forward the user to the *labelling portal*, which provides information about the corresponding ICT product and service.
- The authenticity of the *labelling portal* should also be ensured.
- The Criteria Catalogue should be easily accessible to the public, i.e. freely downloadable from a public website.

10.5.4 Governance and Authority

Having an independent third party managing the verification of criteria/ indicators and subsequent declaration increases the credibility and ultimately the degree of user confidence in a labelling scheme, since, e.g., fraudulent behaviour or user complaints are managed by these independent entities. This is supported by previous findings [1] which suggest that (i) the schemes operated by public bodies or foundations were found to be the most transparent, comprehensive and, trustworthy; and, (ii) labelling schemes have poor longevity unless they are backed by public authorities or large operators. Thus, the TRUESSEC.eu labelling solution advocates for a governance framework ruled by a public or private authority that will be responsible for:

- Creating the yes/no questionnaire.
- Deciding the number of *levels of conformance.*
- Assigning indicators to each level of conformance.
- Setting a validity period for the *transparency report* and the *visual label*. It should be considered that the Cybersecurity Act states that certificates shall be issued for a maximum period of three years and may be renewed, under the same conditions (Article 48). The same is stated by the GDPR (Article 42). However, the ‘lightweight nature’ of the proposed labelling solution allows re-issuing the *transparency report* and *visual label* in shorter time thus increasing the credibility of the approach. Therefore, we recommend a 12-month expiry date from the last update.

- Defining the terms and conditions on the use of a label. This should include penalty rules in case of cheating or non-compliance as well as supervision mechanisms to ensure the validity of the label (e.g. random audits or complaint channels). In this sense and aligned with our "re-use and no-burden approach":
  - ○ We recommend that penalty and complaint approaches already defined in other close legislation, e.g. GDPR, are considered and articulated with the labelling system here proposed. Some 'Core Areas of Trustworthiness' fall within already regulated areas (e.g. privacy and security). Therefore, considering, e.g. that most of the *indicators* in the Criteria Catalogue are covered by the GDPR, its complaint and penalty regime (GDPR CHAPTER VIII: Remedies, liability and penalties) should be articulated with the labelling system. Thus, e.g. a GDPR breach will trigger a re-issue of the *transparency report* and *visual label* (in this case, even the basic/entry level would not be met).
  - ○ Non-compliance with a criterion should not necessarily result in the revocation of the label, but its update to reflect a new *level of conformance*. Revocation should only be performed when at least the basic/entry level is not met.

10.6 Conclusions

The current world scenario shows that the users feel unable to recognise the level of trustworthiness of applications and services, and not even identify which characteristics should they have or show, depending on the confidentiality or sensitivity of the process the user is intending to perform with them.

This makes users feel helpless facing the dilemma "to trust or not to trust".

In this scenario, the trust labels appear to be the solution, i.e. the users could look at the label issuer, and ask its experts to take a decision on their behalf, or at least make some assessment of the level of trust on the application the user could make, in one or several of the criteria identified in TRUESSEC.eu.

This scenario is a somewhat utopic for several reasons:

- There are not well recognised trustworthiness labels, so the users don't know about its existence
- Which ones they should trust more, based on the specific user requirements and expectations about the behaviour of a specific application.

- Which levels of trust and on which areas should the user request from the application or service provider.
- Who evaluates the level of trust of the applications and on which criteria, to assess the level of trust, so that the users could be confident that the assessment itself is trustworthy.

In order to change this pessimistic scenario, the first thoughts of the project in order to propose a roadmap for the implementation of a trustworthy widely adopted trust label (or set of), are taking into consideration the following ideas:

1. Involvement of well-known and authoritative stakeholders, like ENISA, FRA or other European Union institutions, issuing and supporting recommendations to launch and promote the adoption of the trustworthiness label(s).
2. Encourage organisations active in the cybersecurity awareness, like APWG.eu and most of the EU Member States N/G CERTs, to disseminate and make the citizens aware of the existence and advantages of using those trustworthy labels for their own cyber-safety.
3. Define a methodology to allow application developers and service providers to self-assess the trustworthiness of their applications in some or all the criteria identified in TRUESSEC.eu. This approach is aligned with the policy adopted by ENISA in the PET assessment tool. Adoption of this strategy by application developers and service providers will be proportional to the effective demand expressed by the users in the Market.
4. National and/or European authorities should appoint a supervisory authority that could validate the accuracy of the self-assessment statements made by developers and service providers, in order to provide the required trustworthiness to the whole assessment schema. Optionally the assessment criteria could be upgraded to standard and be evaluated by an independent laboratory or trusted third party, which would provide an additional level of trust on the label by the citizens.

Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731711. The first author gratefully acknowledges his sponsor, Escuela Politécnica Nacional.

References

[1] V. Gibello, "TRUESSEC Deliverable D4.1: Legal Analysis," 2017. [Online]. Available: https://truessec.eu/content/deliverable-41-legal-analysis.

[2] V. Gibello, "TRUESSEC Deliverable D7.1: Evaluation of existing trustworthiness seals and labels," 2018. [Online]. Available: https://truessec.eu/content/deliverable-71-evaluation-exiting-trustworthi ness-seals-and-labels.

[3] H. Stelzer, E. Staudegger, H. Veljanova, V. Beimrohr, and A. Haselbacher, "TRUESSEC Deliverable D4.3: First draft Criteria Catalogue and regulatory recommendations," 2018. [Online]. Available: https://truessec.eu/content/d43-first-draft-criteria-catalogue-and-regulatory- recommendations.

[4] D. S. Guamán, J. M. Del Alamo, H. Veljanova, S. Reichmann, and A. Haselbacher, "Value-based Core Areas of Trustworthiness in Online Services," in IFIP International Conference on Trust Management, 2019, Springer, Cham.

[5] D. S. Guamán, J. M. Del Alamo, H. Veljanova, A. Haselbacher, and J. C. Caiza, "Ranking Online Services by the Core Areas of Trustworthiness", RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação, 2019.

[6] L. Floridi, "Soft Ethics: Its Application to the General Data Protection Regulation and Its Dual Advantage," Philos. Technol., vol. 31, no. 2, pp. 163–167, 2018.

[7] D. S. Guamán and J. M. del Alamo, "TRUESSEC Deliverable D5.2: Technical gap analysis." [Online]. Available: https://truessec.eu/content/deliverable-52-technical-gap-analysis.

[8] D. S. Guamán, J. Del Álamo, S. Martin, and J. C. Yelmo, "TRUESSEC Deliverable D5.1: Technology situation analysis: Current practices and solutions," 2017. [Online]. Available: https://truessec.eu/content/delive rable-51-technology-situation-analysis- current-practices-and-solutions.

[9] European Cyber Security Organisation, "European Cyber Security Certification: A meta-scheme approach," WG1 – Standardisation, certification, labelling and supply chain management, 2017. [Online]. Available: http://www.ecs-org.eu/documents/uploads/european-cyber-security-certification-a-meta-scheme-approach.pdf. [Accessed: 05 April 2018].