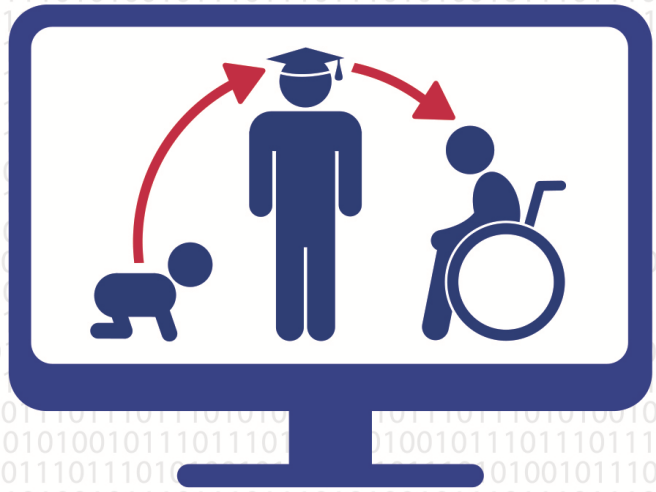


Readiness Analysis for the Adoption and Evolution of Privacy Enhancing Technologies

Methodology, Pilot Assessment, and Continuity Plan

APPROVED
VERSION 1.0
PUBLIC
DECEMBER 2015

www.enisa.europa.eu European Union Agency For Network And Information Security

About ENISA

The European Union Agency for Network and Information Security (ENISA) is a centre of network and information security expertise for the EU, its member states, the private sector and Europe's citizens. ENISA works with these groups to develop advice and recommendations on good practice in information security. It assists EU member states in implementing relevant EU legislation and works to improve the resilience of Europe's critical information infrastructure and networks. ENISA seeks to enhance existing expertise in EU member states by supporting the development of cross-border communities committed to improving network and information security throughout the EU. More information about ENISA and its work can be found at www.enisa.europa.eu.

Authors
Marit Hansen, Jaap-Henk Hoepman, Meiko Jensen

Editor
Stefan Schiffner

Contact
For contacting the authors please use pets@enisa.europa.eu
For media enquiries about this paper, please use press@enisa.europa.eu.

Acknowledgements
This report has been presented to the scientific community; parts have undergone blind peer review. We would like to thank these unknown reviewers. Furthermore, in the project's idea drafting phase and during the project's runtime, we have talked to many researchers to all of them our thanks; but, Carmela Troncoso and George Danezis need to be thanked explicitly for the valuable input.

ISBN 978-92-9204-151-9

Table of Contents

Executive Summary

This report aims at developing a methodology that allows to compare different Privacy Enhancing Technologies (PETs) with regard to their maturity, i.e., their technology readiness and their quality concerning the provided privacy notion. The report firstly sketches a methodology for gathering expert opinions and measurable indicators as evidence for a two dimensional rating scale. Secondly, this report reviews two pilots to test the proposed scales and methodology. The results of these pilots are presented in this study. Finally, a list of necessary steps towards a PET maturity repository is made available.

Target groups

This report is meant for **Data Protection Authorities (DPAs)**, which can adopt or adapt our results for defining how the legal obligations for "state-of-the-art technical and organisational measures" should be understood in the respective contexts. Similarly, groups such as the **Internet Privacy Engineering Network (IPEN)** could integrate the maturity information in their activities to build and maintain repositories for state-of-the-art technologies and best available practice advice. **Data controllers and data processors** and **developers of IT products**, **systems or services** may consult this report to understand what is expected from them with regard to privacy by design principles and can find support by choosing the right PETs for building in the desired privacy and data protection properties. **Researchers, educators and funding agencies** can use this method to identify their priorities in curricula and calls. **Standardisation bodies** could be interested in being aware of PETs before publication of their standards, as well as looking out for links to the standards they are working at. Finally, **policy makers** could interpret the data from PET assessments for a better understanding of the field and its evolution, drivers, and inhibitors.

Recommendations and Dissemination

Assessment of other PETs. The pilots demonstrated the practicality of the approach adopted in this document; however, we believe that additional test cases are needed to further develop and sharpen the methodology. Different kinds of PETs with varying complexity and expected maturity should be chosen to challenge the methodology and the assessment process. Thus, for the short term, we recommend that the topic technology maturity assessment should be included as subtopic of ENISA's efforts in the field of privacy and personal data protection. Also other stakeholders are invited to continue on the basis of our work.

Establishment of a structured assessment process. ENISA's efforts for continuation should concentrate on turning the methodology into a structured process. It is conceivable, and would be favourable, to develop tools that support a standardised step-by-step walk-through for the assessment of both readiness and quality of a PET.

Maintenance of a PET maturity repository. A community portal should be established that is used to publish tools and their assessment results. The European Commission should facilitate the forming of the portal; however, research communities, standardisation bodies, or a DPA board could be in charge of running it.

Dissemination. The project consortium has presented the (interim) results at international research events. Moreover, a presentation was given at a workshop of the Internet Privacy Engineering Network (IPEN). The dissemination needs to be extended towards Data Protection Authorities and their working groups, initiatives within standardisation bodies. In addition, certification bodies for privacy seals could be interested in combining their work with results from maturity assessment.

In conclusion the following actions are recommended; for a deeper discussion see Section 9.

The European Commission should support this research line with an appropriate mechanism, e.g. through a network of excellence in the field.

ENISA should form a consortium that prototypes such an assessment tool. The consortium needs to involve all relevant stakeholders.

The European Commission should mandate a supranational body to maintain a repository of best available techniques in the field of PETs. Without assuming a concrete structure for such a repository, the development and maintenance needs to be community driven, transparent, and independent form interests of a single stakeholder group.

1. Introduction

For decades, privacy enhancing technologies (PETs) have been playing an important role in the discussions on privacy and data protection. For example the New York Times Magazine wrote in 1994: "High-tech has created a huge privacy gap. But miraculously, a fix has emerged: cheap, easy-to-use, virtually unbreakable encryption."[1] And in fact cryptographers have contributed to this debate; most noticeable Diffie, who was described as "was always concerned about individuals, an individual's privacy [..]" or Schneier with his numerous books which aim to a wider public.

The research community on PETs has grown, demonstrators and pilots show that they can be employed for many use cases, some PETs can be found on the market. Although there are a few successful and widely distributed PETs, in general the adoption in practice is low (cf. [12]). The European General Data Protection Regulation will demand data protection by design (Art. 23 General Data Protection Regulation). While in the security domain catalogues on tools, components, and algorithms have been developed that help data processors to protect their assets, this has not happened yet in the domain of privacy and data protection. For data processors it is difficult to decide when a PET may be mature enough to implement it in a system. Similarly, important stakeholders such as politicians, supervisory authorities, funding agencies, or standardisation bodies currently lack an overview on the maturity of PETs which would have an impact on their work. For instance, the usage of sufficiently mature PETs may be demanded by law, standardisation or supervisory authorities much more than this is the case today. For promising PETs that are currently in a not so mature status it may be decided to invest more research and development work. This has motivated our research on the maturity of PETs.

In the following we will briefly summarise our findings, describe the relevance of our work to different stakeholders, and give an overview of this deliverable.

1.1 Summary of results

We have developed a methodology that can provide comparable information on the maturity of different PETs. Our starting point was the discussion of the technology readiness level of a privacy enhancing technology based on objectively measurable indicators. However, one crucial finding in our work is the strong belief that a mere assessment of technology readiness may yield misleading results, i.e. a PET that is available and deployed, but shows severe shortcomings concerning its quality regarding privacy protection should not be preferred over a better privacy technology that — perhaps because of the predominance of the worse technology — scores lower on the readiness scale. For this reason we decided to pursue a two-fold strategy that tackles technology readiness as one dimension and privacy enhancement quality as a second dimension. The individual results are combined into an overall PET maturity score

1.2 Relevance to stakeholders

The methodology in this report is useful to the following stakeholders.

- Data Protection Authorities (DPAs) may use — i.e. adopt or adapt — our results for defining how the legal obligations for "state-of-the-art technical and organisational measures" should be understood in the respective contexts. Today, the terms "state of the art" or "current state of technology" are already part of Directives and Regulation (e.g. Art. 17 Directive

[1] Levy, Stephen (1994-07-12). "Battle of the Clipper Chip". New York Times Magazine

95/46/EC), and they are being used in Opinions of the Art. 29 Data Protection Working Party (e.g. [4, 5, 6]).

- Similarly, groups such as the Internet Privacy Engineering Network (IPEN), consisting of representatives from DPAs, researchers, and industry, could integrate the maturity information in their activities and give feedback.
- Data controllers and data processors would get help in understanding what is expected from them, especially when the demand for "data protection by design and by default" (Art. 23 of the proposed General Data Protection Regulation) will come into effect.
- Developers of IT products, systems or services would be provided with information about the maturity of PETs that will help them choose PETs for building in the desired privacy and data protection properties.
- Standardisation bodies could be interested in being aware of PETs before publication of their standards, as well as looking out for links to the standards they are working at.
- Teachers and trainers for privacy and data protection should use the information on PET maturity when educating Data Protection Officers or Privacy Engineers. This knowledge could be a valuable source for the standard curriculum for computer scientists or lawyers.
- Funding agencies could employ the data on maturity of PETs to decide on new calls, e.g. if PETs in a field are highly immature, or if they are almost ready for the market.
- Researchers may recognise fields where more work is needed and find research and development options.
- Since the maturity assessment would reveal the situation of PETs at a particular time, policy makers and researchers could interpret the data for better understanding of the field and its evolution — drivers and inhibitors may be identified more easily.
- Finally, all kinds of end-users — regardless of whether they are organisations or individuals — could profit from easily comprehensible PET maturity results when looking for PETs most suitable for their needs.

The relevance of PET maturity for a diverse set of stakeholders demands that the information is easily comprehensible by experts and laypersons; potential misinterpretation of the information should be prevented as far as possible.

1.3 Reading guide

The text is organised as follows: Section 2 introduces important terms and notions that be necessary to determine the scope of the project. An overview of related work concerning methods to measure technology readiness is given in Section 3. The requirements on the scale(s) and the methodology to be developed are presented in Section 4. The assessment framework is described in Section 5. The resulting assessment process is outlined in Section 6. Section 7 is dedicated to the Pilot PET Assessment, where we apply our methodology to assess the maturity of a concrete PET. Section 8 contains an evaluation of this practical application pilot. In Section 9 we discuss how to further develop, apply and disseminate our methodology. Finally, Section 10 summarises the findings and gives an outlook on future work.

2. Terminology

In this section, we introduce the basic terminology used throughout the report. We start with essential terms like "technology", "privacy", "data protection", and "privacy enhancing technology". Moreover, terms relating to the lifecycle of (privacy enhancing) technology are introduced. Also, we clarify the distinction between "readiness" and "maturity".

2.1 Defining privacy enhancing technologies

Privacy enhancing technologies, or PETs for short, play an important role in this report. Unfortunately, PETs are a fuzzy concept in practice, so it is important for us to define this concept more precisely here. We do so by first defining what we consider to be a technology, followed by a discussion on privacy and data protection (explaining that for the purposes of this report we consider them to be the same thing), after which we return to the definition of privacy enhancing technologies themselves.

2.1.1 Technology

Merriam-Webster defines Technology as

- "the practical application of knowledge especially in a particular area;
- a capability given by the practical application of knowledge;
- a manner of accomplishing a task especially using technical processes, methods, or knowledge;
- the specialized aspects of a particular field of endeavour".[2]

In this report, we will focus on the third meaning: "a manner of accomplishing a task especially using technical processes, methods, or knowledge". In particular, we will focus on software and hardware solutions, i.e. systems encompassing technical processes, methods, or knowledge to achieve specific privacy or data protection functionality or to protect against risks to privacy of an individual or a group of natural persons.

2.1.2 Privacy and data protection

Numerous definitions exist for "privacy" and for "data protection",[3] denoting sophisticated concepts that have developed over centuries. In the European Union, both terms are being used in the Charter of Fundamental Rights: Article 7 defines the "respect for private and family life" (i.e. "privacy"), Article 8 demands "protection of personal data" (i.e. "data protection"). Sometimes, the terms are used interchangeably; sometimes clear distinctions are pointed out. Usually, data protection deals with the organisational perspective: the European data protection framework specifically addresses organisations that have to comply with the law so that individuals are protected against misuse of their personal data by those organisations. In comparison, privacy often tackles the individual's perspective. It is related to the ability of the individual to protect herself or to fight against being informationally controlled by others.

The European legal data protection framework does not only address protection of individuals, but also free movement of personal data under defined conditions. Currently, the incentives for minimising processing of

[2] Source: Merriam-Webster Dictionary, http://www.merriam-webster.com/dictionary/technology

[3] We discussed the notion of privacy and data protection in more detail in our report on Privacy and data protection by design https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/privacy-and-data-protection-by-design

personal data as far as possible — a typical measure for protection individuals against misuse of their data — are low; the legal framework has even been criticised for assuming a "notion of the data controller as a trusted party" while the community of PET developers usually perceive it as an adversary [13].[4]

For the sake of simplicity and comprehensiveness we will not discuss an own category of "data protection (enhancing) technologies", but instead open up the established term of "privacy enhancing technologies" to cover both applications to be used by individuals to protect themselves and applications that can be used by organisations to support privacy and data protection for individuals.

2.1.3 Privacy enhancing technology

Privacy enhancing technologies (PETs) have been characterised in various ways:

- "The application of information and communications technologies (ICT) for the sake of privacy protection has become widely known under the name of Privacy Enhancing Technologies (PETs). PETs have been defined as a coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data; all without losing the functionality of the data system" [10]. While this definition focuses on data avoidance and data minimisation, the authors also acknowledge "other privacy-supporting technologies": "There are many other technologies that might also contribute to better privacy protection if PETs (...) cannot be applied effectively. This is certainly the case with the following data processing conditions derived from basic privacy principles: transparency, data quality, respect for the rights of parties involved, and security." [10].
- In the OECD Report on PETs in 2002, it is stated: "Privacy enhancing technologies (PETs) commonly refer to a wide range of technologies that help protect personal privacy. Ranging from tools that provide anonymity to those that allow a user to choose if, when and under what circumstances personal information is disclosed, the use of privacy enhancing technologies helps users make informed choices about privacy protection. PETs can empower users and consumers seeking to control the disclosure, use and distribution of personal information online. PETs can also aid businesses and organisations in enforcing their own privacy policies and practices." and "PETs vary widely in their functionality, capabilities, technical structure and usability. However, all PETs aim to give the individual user or technology manager the capability of controlling if, how much or under what circumstances information is disclosed." [31].
- The European Commission in its MEMO/07/159 considered a wider range of PETs that may support legal compliance with data protection regulation: "What are PETs? — The use of PETs can help to design information and communication systems and services in a way that minimises the collection and use of personal data and facilitate compliance with data protection rules. The use of PETs should result in making breaches of certain data protection rules more difficult and/or helping to detect them." [15].

Several researchers working in the field acknowledge the broad range of interpretations for the term "PETs" and therefore give a restricted definition in their paper (e.g. [13]). Others avoid exact definitions, but rather describe the privacy problem they want to solve by a technology-based solution (as recommended e.g. by [30]), thereby adding further dimensions to the fuzzy field of privacy enhancing technologies (e.g. [24]).

[4] This discussion will be taken further with respect to the interpretation of Article 23 "Data Protection by Design and by Default" of the upcoming European General Data Protection Regulation.

For this project, we aim at allowing a wide definition of PETs, encompassing all kinds of technologies (according to the definition in Section 2.1.1) that support privacy or data protection features (e.g. technologies that make use of privacy design strategies [21] or consider protection goals for privacy engineering [20]). Compared to a definition that restricts PETs to data minimisation, this approach provides greater flexibility and adaptability, although this adds complexity when statements on the privacy enhancement properties in various categories have to be elaborated.

2.2 The technology lifecycle

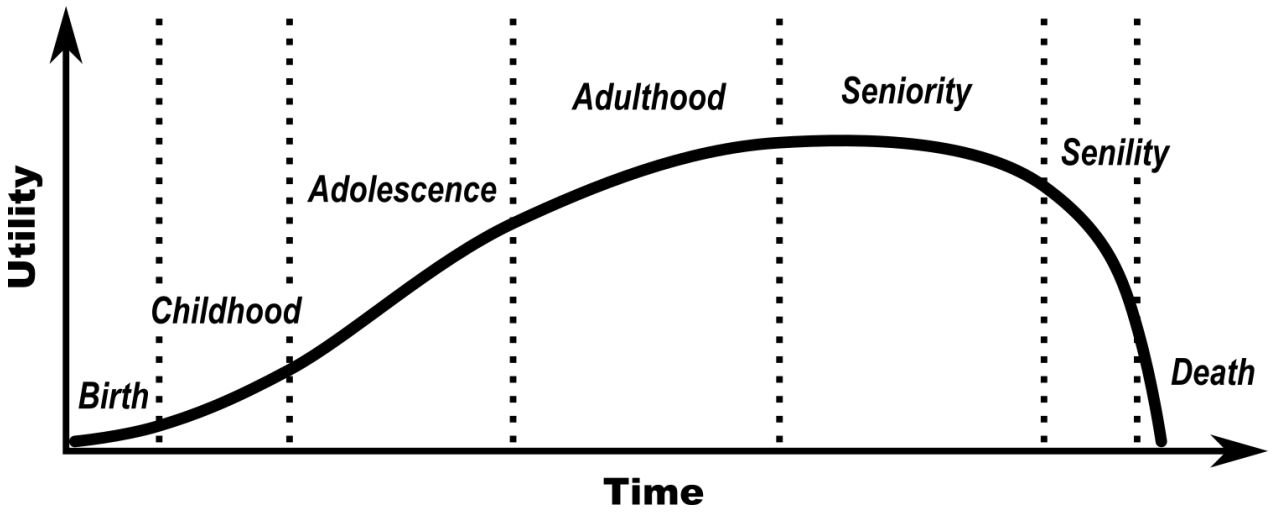


Figure 1. The lifecycle of a technology as adapted from [28].

For a readiness analysis, it is important to be aware of the development of a technology over time. This is not a speciality of privacy enhancing technology, but a characteristic for technologies in general.

We distinguish between seven different phases within the lifecycle of a technology, illustrated in Figure 1, as defined by William L. Nolte [28]. Initially, each technology starts off with an idea, its birth. Then, this idea is analysed preliminarily, elaborated on, and considered useful. Thus, in the next phase, the idea is discussed on a broad scale, e.g. within research and development communities. Yet, there is no working prototype, not even a demonstrator, so the correlated phase is that of childhood. At some point, a proof-of-concept is implemented in test environments under laboratory conditions, marking a progress towards adolescence level.

The next step is that of a real-world usage of the technology under non-laboratory conditions. Typically, this step is performed with the release of a first feature-complete implementation, or with the advent of early pilot implementations in real-world systems. Thus, the technology matures towards a state of adulthood.

Subsequently, the next remarkable transition is that of a full market participation of the technology, which is typically kicked off by advent of a ready-to-use product being sold (rented, consulted for, commercially supported for, etc.). This implies that the maturity of the technology has reached a point where it becomes feasible to gain profits from utilising the technology to such extent that a market emerges. The corresponding age is that of seniority.

Finally, the technology might become obsolete by technological evolution. For PETs, this could mean that devastating attack techniques render the technology useless in an irreparable way, or simply by the advent

of a superior technology that provides the same guarantees in a more favourable way. In each of these cases, the use of the technology decreases (into what we may call the senility phase), until it fades out of use, and reaches its final state of death.

How fast a technology can evolve through these phases depends to some extent on its complexity, but more important on the incentives to implement it. For example, the moon mission was barely a research idea when it was announced by Kennedy and it was tremendously complex, but was implemented in less than 10 years, from the moment the incentive was set.

2.3 PET maturity: technology readiness + privacy enhancement quality

As explained in the introduction, we regard PET maturity to be a combination of both technology readiness and (privacy enhancement) quality. The distinction is important because a particular PET may be quite developed (i.e. it is 'ready'), yet it may offer very little privacy protection. We therefore determine "PET maturity" as a result calculated from a "technology readiness" and a "privacy enhancement quality" scale.

These scales will be developed in the following sections. For now we would like to define the following terms to ensure that no confusion arises about them later.

Metric. A method to assign a score or level from a given scale to an item, in our case a PET, with the aim to compare items.
Level or score. The particular *level or score* on a metric (in our case readiness and quality), e.g. pilot as the value for the readiness level.
Scale. The set of *levels or scores* a certain metric can assume.
Indicator. A factor that may be meaningful for determining the level; input for the assessment.
Evidence. The set of indicators that support the assigned level for a metric.

3. Related Work

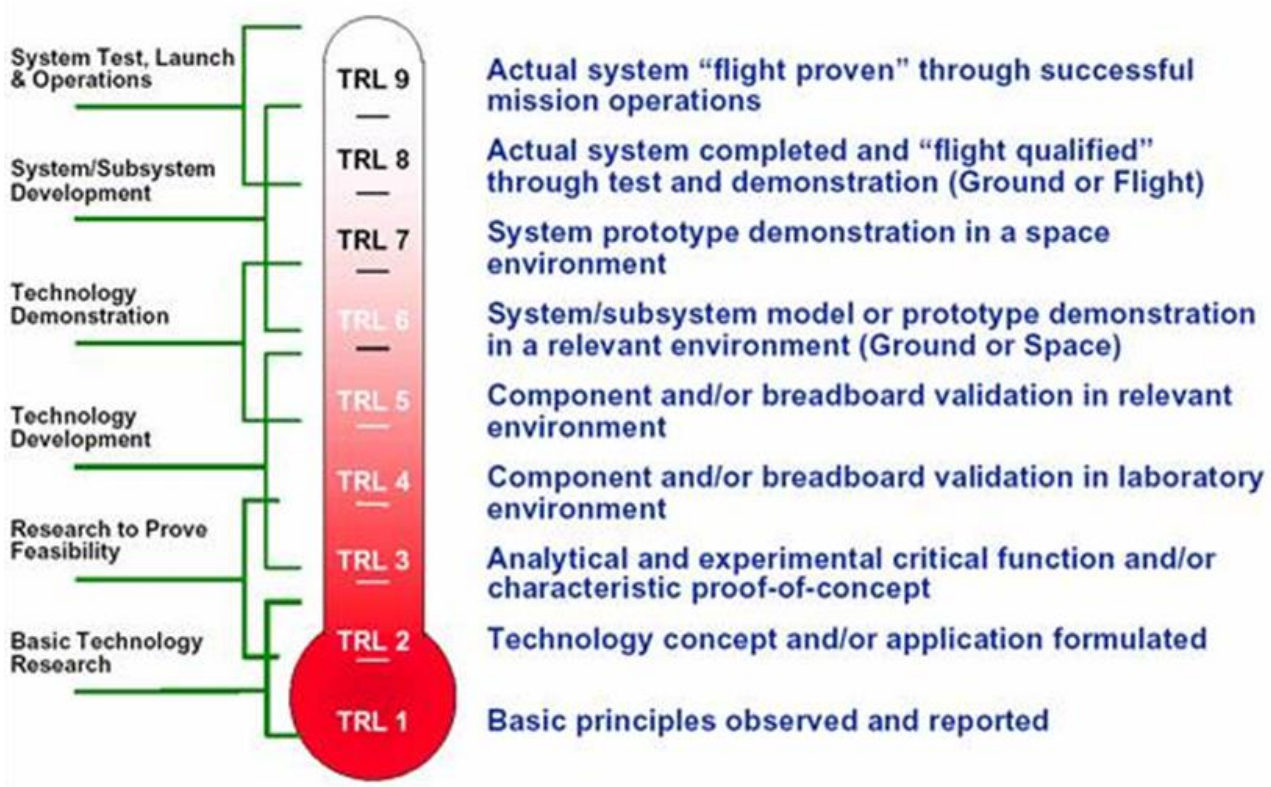


Figure 2. NASA's scale of technology readiness levels.

The starting point for our work are existing Technology Readiness Levels (TRLs) and further attempts to assess the maturity of technologies and systems. Since we are aiming at comparability between the assessment results, we have to think of scales for maturity. Foundations for scales have been defined in [38], but we also will take into account justified criticism concerning scale typologies, as e.g. presented in [40]. In particular, all assessments have to deal with potential misinterpretations of scales, mistakes in measurements, and data with different degrees of confidence and certainty. In the following subsections, we briefly show the mainly used TRLs for readiness analysis and related work on criteria for quality assessment of software and processes with regard to security.

3.1 Readiness – well-known TRLs

NASA uses a well-known and widely discussed readiness scale: the Technology Readiness Levels (TRL) 1-9 [25], see Figure 2. The NASA TRL scale is supported by extensive guidance reports, e.g. [7, 14, 27]. For instance, a specific TRL Assessment Matrix [27, 9, 28] has been developed for helping with the assessment of a technology. In particular, that matrix shows the need for addressing various kinds of subsystems to evaluate their technology readiness properties first, and then derive an overall TRL. Further, TRL Calculators were

developed that help collecting the necessary data in a spread sheet or database, taking into account whether hardware or software readiness is to be assessed and which use cases are chosen.

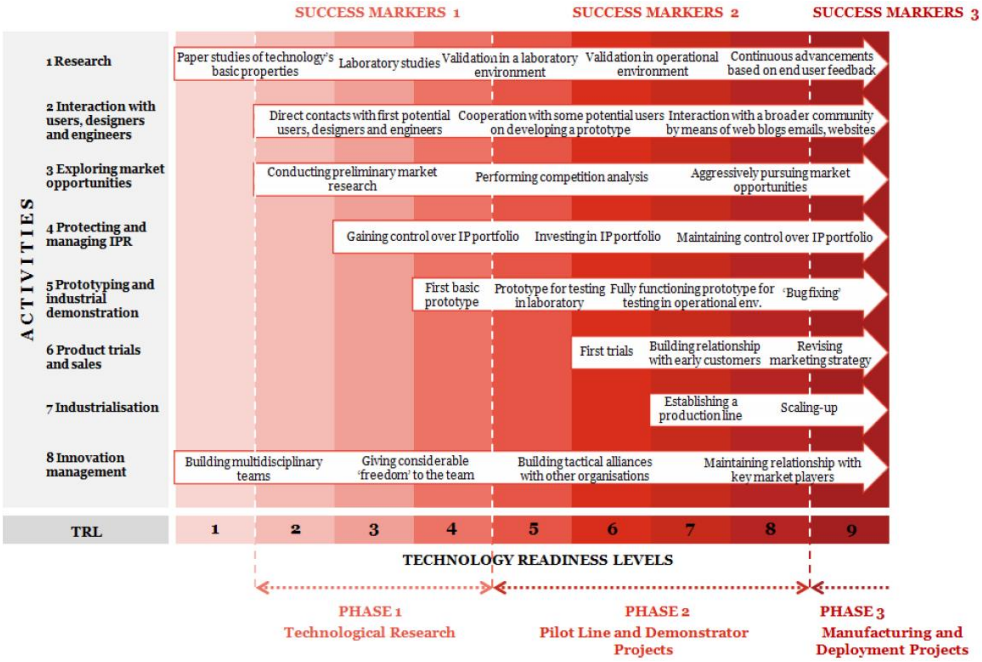


Figure 3. Success markers for Technology Readiness Levels as proposed in [17].

The European Commission itself uses the following nine Technology Readiness Levels in its funding programme Horizon 2020 [16] that are similar to the NASA TRL:

- **TRL 1:** basic principles observed
- **TRL 2:** technology concept formulated
- **TRL 3:** experimental proof of concept
- **TRL 4:** technology validated in lab
- **TRL 5:** technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies)
- **TRL 6:** technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies)
- **TRL 7:** system prototype demonstration in operational environment
- **TRL 8:** system complete and qualified
- **TRL 9:** actual system proven in operational environment (competitive manufacturing in the case of key enabling technologies; or in space)

The idea to use TRLs in funding programmes stems from an earlier analysis commissioned by the European Commission to bridge the gap between research and commercial success (e.g. for the innovation area of Nano sciences, Nanotechnologies, Materials and New Production Technologies (NMP) [17]). In this respect, indicators (so-called "success markers") of the TRLs have been discussed (see Figure 3).

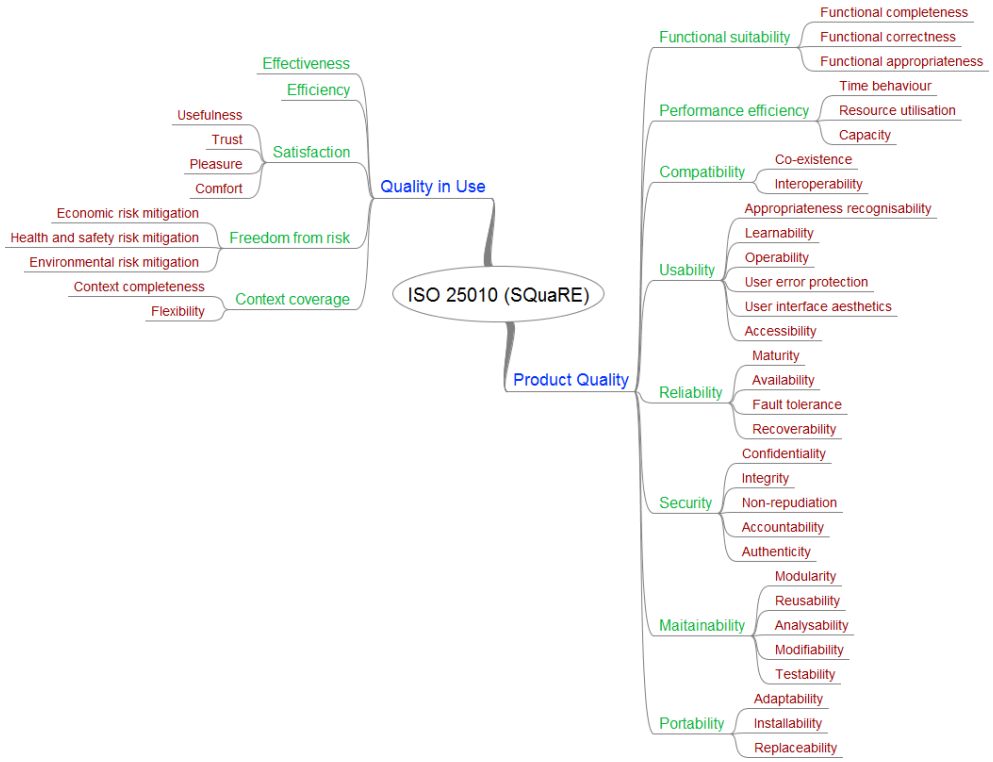


Figure 4. ISO 25010 – Systems and Software Quality Requirements and Evaluation (SQuaRE).

Since the proposal in 1995, the NASA TRL scale has been discussed and criticised, in particular by pointing out limitations and needs for a multidimensional approach [28]. A recent study comes to the conclusion that improvements of the process are necessary if the TRL assessment should be meaningful [29].

An important insight from the debate is the fact that readiness has to be understood in context and that it is usually not sufficient to assess "readiness" without regarding "quality", e.g. [37]. This has been taken up by a few researchers who combine readiness analysis with reliability features, e.g. [8].[5]

3.2 Quality – the SQuaRE approach

In the context of privacy and security this additional quality dimension is especially important because there are all too many examples of widely deployed technology that would score high on a pure "readiness" scale, which provide sub-optimal protection.

Since several privacy properties are related to security properties, the maturity assessment methodologies and criteria from this field will play a central role and have to be included, for instance ISO/IEC 27004 [1], NIST Special Publication 800-55 [11], and Control Objectives for Information and Related Technology (COBIT) [22].

Since in the area of PETs software- or algorithm-related criteria play a viable role, the recently released ISO/IEC standard 25010 on Systems and Software Quality Requirements and Evaluation (SQuaRE) turns out

[5] In our terminology this would be "maturity analysis", going beyond mere "readiness" assessment.

to be a valuable source for criteria and indicators, see Figure 1Figure 4. For an overview of older approaches see, e.g. [41, 26]), but we believe most of it is superseded.

However, the SQuaRE standard is not comprehensive for our needs, but extensions have been made, e.g. for green and reliability issues [18]. Other criteria may be more or less neglected for assessment of PET maturity since they most likely will not play a role. This will be further elaborated in the following sections.

Furthermore, criteria for process maturity might be necessary for specific PET types. In the literature we can find examples for such criteria, from the Capability Maturity Model Integration (CMMI) [36], or Software Process Improvement and Capability Determination (SPICE), standardised in ISO/IEC 15504 [3], systems readiness [35] or integration readiness [34].

4. Requirements

In this section we will discuss the criteria that we used to select the scales for our methodology. They are the basis for the selection of the scales in Section 5. Moreover, we present a set of criteria the methodology itself needs to satisfy. These criteria are evaluated after having gained experience from practical application of our methodology in a pilot PET assessment, see Section 7. The evaluation of the methodology is presented in Section 8.

4.1 Criteria for the effectiveness of the scale

For a scale to be useful in practice, it needs to be effective. The effectiveness of a scale depends on the following 4 factors defined below, namely its comprehensibility, its comparability, its scorability, and its reproducibility. We define these four criteria in the following paragraphs.

Comprehensibility. First of all a score should be easy to understand and to apply by users[6] looking for an appropriate PET to solve a particular problem in a certain context[7]. The meaning of a certain score should be intuitively clear.

Comparability. Similarly, comparing different scores should be straightforward. However, that can be hard if the score is multidimensional. Here an effective scale needs to come with a method to compare different scores and needs to describe pairs that are not comparable.

Scorability. Further, a particular PET should be easy to score objectively on the scale at hand by an evaluator. The score should be derived from clearly described indicators, which are easy to determine or measure for an arbitrary PET that is going to be evaluated. Moreover, it should be clear how a combination of values or appreciations for the different indicators should be combined into a score. Having said that, objectivity might be sometimes impossible to achieve, then a clear method to objectivise subjective answers is needed.[8]

Reproducibility. Finally, a score for a PET on some scale should be reproducible. This means that a PET should receive (almost) the same score, when independently scored by two or more evaluators. This further emphasises the objectiveness implicit in the definition of scorability.

4.2 Criteria for assessing the methodology

To appreciate the merit of the overall methodology, the following criteria are relevant.

Adequacy. The methodology should be adequate. It should deliver a sound and valid judgement of the maturity of a PET under evaluation.

Ease of use. The methodology should be easy to use. Instructions should be clear and unambiguous. Evaluation results (including intermediate ones) should be easy to record. An evaluation should be performable with little organisational overhead and should require little coordination among people involved.

[6] Most likely the developers.
[7] We note that in our methodology the application context of a PET is out of scope for determining its maturity, as explained further on in this report.
[8] This might be statistical methods or methods to achieve consent.

Effectiveness of the scale. The methodology should be based on a scale (or scales) that are effective, according to the criteria described in Section 4.1

Effort. The effort needed to perform an assessment of a PET needs to be reasonable. This means that the amount of effort should be commensurate to the value (in terms of significance and reliability) of the outcome the methodology produces.

5. The Assessment Framework: Scales and Evidence

In this section, we define the underlying concepts, scales, indicators, and evidence parameters utilised in the assessment process. We discuss the rationale for the scales of readiness and quality, analyse the tensions between measurable indicators and expert opinions, and set the basics for the assessment process defined in the next section.

5.1 A scale for readiness

We begin by defining a scale along which to express the readiness of a certain PET in line with the phases of the technology lifecycle described in Section 2.2, i.e. birth, childhood, adolescence, adulthood, seniority, senility, and death. Readiness of a PET expresses whether a PET can be deployed in practice at a large scale, or whether it can only be used within a research project to build upon to advance the state of the art in privacy protection. Readiness says something about the amount of effort, i.e. time, money, etc., still needed to allow the PET to be used in practice with a positive cost benefit balance. We favoured the following set of readiness levels over a linear scale to ensure comprehensibility, see Section 4.1.

Idea. Lowest level of readiness. The PET has been proposed as an idea in an informal fashion, e.g. written as a blog post, discussed at a conference, described in a white paper or technical report.

Research. The PET is a serious object of rigorous scientific study. At least one, preferably more, academic paper(s) have been published in the scientific literature, discussing the PET in detail and at least arguing its correctness and security and privacy properties.

Proof-of-concept. The PET has been implemented, and can be tested for certain properties, such as computational complexity, protection properties, etc., i.e. “Running code” is available, but no actual application of the PET in practice, involving real users, exists, nor is the implementation feature complete.

Pilot. The PET is or has recently been used in practice in at least a small scale pilot application with real users. The scope of application, and the user base may have been restricted, e.g. to power users, students, etc.

Product. The highest readiness level. The PET has been incorporated in one or more generally available products that have been or are being used in practice by a significant number of users. The user group is not a priori restricted by the developers.

Outdated. The PET is not used anymore, e.g., because the need for the PET has faded, because it is depending on another technology that is not maintained anymore, or because there are better PETs that have superseded that PET.

These readiness levels relate to the technology lifecycle; a later evolutionary level does not necessarily mean that the PET is better, because the aging process may not improve the PET’s quality or its applicability when it becomes outdated. This readiness level indicates that the PET should no longer be used. The transition from one readiness level to the next is not as sharply delineated as the previous scale suggests. In fact, different PETs that belong to the same readiness level may differ significantly. Some barely made it the level assigned to them; others are about to enter the next level. To allow people to express these differences, a readiness level may be augmented with the next higher readiness level in the scale above. So, for example, a readiness level of pilot/product may be appropriate for a PET that has been used in several pilot programmes and is currently being beta-tested as a (commercial) general purpose product.

5.2 A scale for quality

Although quality is somewhat dependent on readiness, as a rolled out product has received so much more attention over the years than a concept still in its research stage, the quality of a PET is not only determined by its readiness. In fact, several PETs at the same readiness level may have varying levels of quality. As argued in the introduction, it is important to realise that sometimes a PET with high readiness may still have a low quality. We now turn to make this notion of quality more precise.

We base our approach on the ISO/IEC system and software quality models standard ISO 25010 [2], but adjust and refine it to our needs. ISO 25010 distinguishes the following eight quality characteristics, namely functional suitability, reliability, operability, performance efficiency, security, compatibility, maintainability and transferability. Not all of these characteristics are relevant for our purposes. Some characteristics are more important than others and therefore contribute more to the overall quality score.

For example, because we want the overall maturity scale to be independent of the particular context in which a PET is applied, characteristics like *functional suitability* are out of scope. We believe that a PET with limited functionality has the same quality as one with a larger or different functionality. Hence it depends on the application's requirements which PET to choose for a particular application context.

Similarly, *compatibility* is deemed a less relevant characteristic.

Since a PET is typically embedded into larger system, and not directly exposed to the user, we interpret *operability*, i.e. the degree to which a product is easy to learn and understand and its capability to attract users, to be directed at a system developer instead of an ordinary user.

The *security* characteristic is renamed to *protection*, and focuses on preventing privacy infringements. A separate characteristic *trust assumptions* is added to capture whether and if so how much trust in certain components and agents is assumed.

Also added are two other characteristics: *side effects* and *scope*. This brings us to define the quality scale as comprising the following nine PET quality characteristics, listed in decreasing order of importance

Protection. Protection should be understood as the degree of protection offered (in terms of for example unlinkability, transparency, and/or intervenability) to prevent privacy infringements while allowing access and normal functionality for authorised agents. Also depends on the type of threats and attacks against which the PET offers protection.

Trust assumptions. Trust assumptions are characterised by the technical components and/or human or institutional agents that need to be trusted, and the nature and extent of trust that must assumed in order to use the PET. The more components or agents need to be trusted, the lower the score. For example, whether the system assumes an honest but curious adversary, whether the system is based on a non-standard cryptographic assumption, whether it relies on a trusted third party, or whether a trusted hardware component is used. Standard assumptions, for instance that the software and hardware need to be trusted, are out of scope. Note that trust assumptions can also be legal, i.e. a juridical process is a critical part of the protection offered, or organisational, i.e. the protection offered depends on procedural safeguards.

Side effects. Side effects are the extent to which the PET introduces undesirable side effects. These effects include increased organisational overhead due to key management, increased use of bandwidth (without performance impact) due to cover traffic, etc. Assessing side effects depends on the composability, i.e. how easy it is to compose the PET with other components without negatively influencing these components, and on the number and severity of these side effects themselves.

Reliability. Reliability is the degree to which a system or component performs specified functions under specified conditions for a specified period of time. It is measured in terms of fault tolerance and recoverability, as well as in terms of the number of vulnerabilities discovered.

Performance efficiency. Performance efficiency is the performance relative to the amount of resources used under stated conditions. It is measured in terms of resource use, i.e., storage, CPU power, and bandwidth and speed, i.e., latency and throughput.

Operability. Operability is the degree to which the product has attributes that enable it to be understood, and easily integrated into a larger system by a system developer. It is measured in terms of appropriateness, recognisability, learnability, technical accessibility, and compliance.

Maintainability. Maintainability is the degree of effectiveness and efficiency with which the product can be modified or adapted to underlying changes in the overall system architecture. It is measured in terms of modularity, reusability, analysability, changeability, modification stability, and testability. Open source software typically scores high on this characteristic. Also, systems that have an active developer community, or that have official support, score high.

Transferability. Transferability is the degree to which a system or component can be effectively and efficiently transferred from one hardware, software or other operational or usage environment to another. It is measured in terms of portability and adaptability.

Scope. The scope refers to the number of different application domains the PET is applied in or is applicable to.

While each of these characteristics is relevant for a PET independent of its readiness level, the indicators that determine the score for each of the characteristics do depend on the readiness level. For example, the quality of a rolled out product depends on how well it is supported by a help desk, code updates, etc. These indicators are irrelevant for research level PETs. Here, the quality is determined by the quality of the research, e.g. the ranking of the venues in which the research is published.

For each of these nine characteristics, a PET can receive a score in the range {−− (very poor) − (poor) 0 (satisfactory) + (good) ++ (very good)}. The overall quality level also utilises this five-value scale, and is comprised of the nine individual scores, according to a specific quality evaluation function, as discussed in Section 6.5.2.

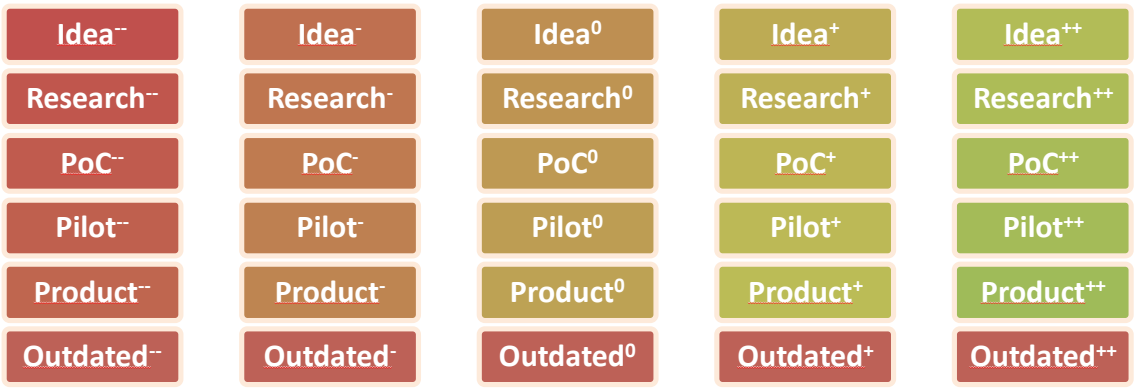


Figure 5. Overview of Possible PET Maturity Level Values.

5.3 Combining readiness and quality to express maturity

The scales for readiness and quality defined above allow us to define the real scale we are interested in: a scale for PET maturity. In fact this overall scale is simply the combination of the readiness level superscripted by the quality level.

$$readiness^{quality}$$

So for example a PET with readiness level $pilot$ and quality $+$ has an overall PET maturity level of $pilot^{+}$. Thus, the total set of potential PET maturity values spans from $idea^{--}$ and $idea^{++}$ to $outdated^{--}$ and $outdated^{++}$. The full set of possible values is illustrated in Figure 5.

5.4 Evidence: Measurable indicators vs. expert opinions

When assessing maturity of a PET, different experts may have different opinions with respect to its readiness and quality. Hence, each assessment approach that is solely based on expert opinions is likely to be affected by the choice of experts, and thus lacks reproducibility. Having the same PET assessed by different expert groups may lead to different assessment results, due to the different viewpoints and discussion dynamics among the chosen sets of experts.

In order to mitigate this biased assessment approach, it needs to have some indisputable parameters to be taken into account. Such parameters should be assessable in a way that is unambiguous, leading to the same parameter value (within a small range of deviation) and assessment indication no matter who performs the parameter assessment. We call these types of parameters measurable indicators, meaning that they indicate an assessment result based on objective evidence. As such, measurable indicators are robust against change of assessors, as different assessment instances of the same measurable indicator will always result in the same indicator values, and thus in the same assessment result.

In the following, we give examples for evidence to motivate the design of our methodology that is explained in Section 6 – with subsections on measurable indicators (Section 6.3) and on expert opinions (Section 6.4).

Examples for measurable indicators in the field of PET maturity assessment are:

- number of scientific publications referring to the PET to be assessed,
- number and type of audits/certifications performed for the PET,
- number of university courses covering the PET topic,
- number of commercially available products that use the PET as a component,
- number of hits when searching for the PET in online search engines,
- number of years since the PET was initially proposed.

As can be seen, each of these measurable indicators represents a certain characteristic with respect to the PET, and does so in an indisputable way. There can be no two different opinions on the total number of scientific publications referring to the PET, for example, at least not on a level of significance. Such a value is an objective evidence for a certain level of maturity of the PET.

However, though assessing these measurable indicators is feasible and quite robust, determining its implications with respect to the result of the assessment is more challenging. What does the number of search engine hits say about the maturity of a PET? What should be the impact of the existence of six different privacy certifications of a PET product? Each of these measurable indicators gives a small implication on the level of maturity the PET has probably reached. For instance, the existence of a substantial amount of competing products in the market of the PET to be assessed clearly implies that this PET has reached at least the pilot stage, more likely even the product stage of readiness. If there are no products in the market at all, this might indicate an earlier maturity stage, probably research, but it might also be the case that the PET itself is not suitable to be sold as a dedicated product. Nevertheless, it still could be utilised in many products out there, and still could be in the product readiness stage.

In conclusion, the measurable indicators are robust in assessment, but fuzzy in their implications to the result of the assessment. They need to be included in the overall assessment process, in order to mitigate the impact of assessor choices, but they are not precise enough to be used as the only, not even as the major base for a PET maturity assessment. Thus, we propose to utilise these indicators as input, but combine them with inputs from a dedicated board of experts.

6. The Assessment Process

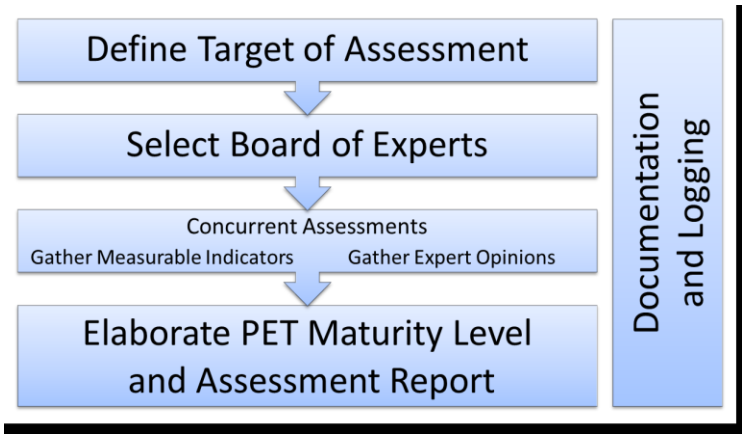


Figure 6. Overview of the PET Maturity Assessment Process.

In this section, we define the process of performing a PET maturity assessment, including a discussion on each of the individual process steps and the intermediate results of these.

6.1 Overview

The process of assessing PET maturity along the lines defined in this document involves four steps, as illustrated in Figure 6. The implicit initial step of an assessment consists in the determination of the assessor, as that is a very critical entity in performing the assessment. The assessor is the person responsible for performing the assessment. He/She needs to be an expert in the process of performing assessments. Beyond that, expertise both in terms of privacy and in the domain of interest the PET is assessed in would be beneficial. Moreover, the assessor needs to be unbiased, as far as possible, and objective in all decisions.

In the first explicit step of the assessment, it is necessary to select and precisely define the Target of Assessment (ToA), i.e. the concept, technology, or product that is to be assessed. Details on this step are given in Section 6.2.

Once the Target of Assessment is defined, the next step consists in gathering the board of experts to be asked for their opinion. Ideally, each expert should have expertise both in the application domain of the PET, and in privacy engineering. As with the assessor, it is necessary to gather an unbiased, objective, heterogeneous set of experts for this task, i.e. they should represent different perspectives which can yield more robust and comprehensive results. Though there is no upper bound on the number of experts, we propose a minimum of five experts to be involved in the board. This step also concludes the preparation phase of the assessment.

Based on the considerations described above, the best approach is to combine both the measurable indicators and the expert opinions approaches into a single, overarching assessment methodology.

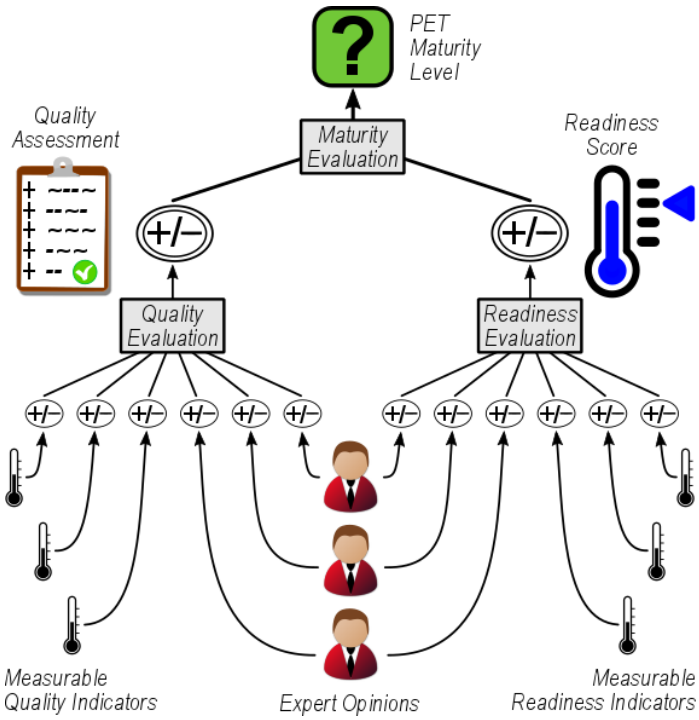


Figure 7. PET Maturity Assessment Methodology.

As shown in Figure 7, our methodology is based on both types of input, collected for both readiness and quality assessment. More precisely, the measurable indicators are collected and normalised according to reasonable individual scales, depending on the ToA. This step, which typically would be performed by the assessor, is described in Section 6.3.

In the next step, the expert opinions are collected by means of dedicated forms, consisting of both a scale-based assessment and a detailed opinion comment part. This part of the overall process is dealt with in Section 6.4.

After that, all of these inputs are processed by the assessor to gather two separate intermediate results: a Readiness Score (see Section 6.5.1) and a Quality Assessment (see Section 6.5.2).

Finally, both of these are combined into the final PET Maturity Level (see Section 6.5.3). In other words, the assessor performs the following steps:

1. determination of the level of technology readiness of the PET, according to the scale defined in Section 5.1,
2. assessment of the overall quality of the PET, according to the quality characteristics described in Section 5.2, and
3. aggregation of these two intermediate assessments into the final PET maturity level, as discussed in Section 5.3.

Finally, the documentation and logging inputs, which were collected throughout the other steps of the assessment, need to be aggregated, and comprise a PET Maturity Assessment Report accompanying the PET maturity level achieved. Once the final PET maturity result is obtained, and the PET Maturity Assessment Report is completed, the assessment process concludes.

6.2 Defining the Target of Assessment

The initial step of assessing a given PET's level of maturity is the precise definition of the Target of Assessment (ToA). Depending on its phase in the technology lifecycle as outlined in Section 2.2, a PET may consist of a few lines of demonstrator source code only, or may already have been implemented in a set of software products being sold and bought in a dedicated market of its own. Thus, the definition of the correct ToA can be quite tricky.

If a PET is in one of its early stages of evolution, where it merely is made up by a concept outline or a rough set of ideas, the ToA typically consists of the major concept of the PET, as outlined by its maintainers. Being a theoretical concept without even a basic implementation, measurable quantitative indicators like market share, lines of source code, etc., are not available, and thus cannot be used for maturity assessment. Available measurable readiness and quality indicators for this stage of maturity can only be found in the research and discussion domain, such as number of research papers published that refer to this PET.

The ToA can be narrowed down to the scope of a specific implementation, If a well-maintained implementation of a PET already exists, but no commercially available product along this implementation, such as a software product, consulting services, support desk is found in the open markets. Whenever a precise condition of the PET in question is required within the assessment, the concept is evaluated according to the details found in this implementation. Also, measurable readiness and quality indicators from the source code realm (like lines of code, amount of source code documentation, etc.) can be used based on the numbers available for the existing implementation.

If a dedicated market for solutions utilising this PET already is in place, the ToA can no longer be defined as the (single) concept or implementation of the PET. Given that different products and different domains of application may result in differing privacy guarantees, the ToA in this case has to be narrowed down to one of the existing products or implementations only. This is due to the fact that different implementations of the same PET may have different characteristics, different levels of completeness, and different levels of quality. Thus, an assessment should focus on a single product or implementation only, potentially relating it to other products of the same category for comparison, but fixing the ToA on the product, not on the theoretical concept beneath. Measurable indicators for such a level of readiness or quality may range from market share data to sales numbers, active developer community sizes, and total amount of financial capital allocated to utilisation of the PET, among others.

For defining the ToA, the following information is necessary:

- A description of the PET including a documentation (an academic paper, a hardware, a software, etc.; if possible, with a link to the source of a published paper or source code or product including a version number);
- Information on the privacy preserving goals this PET aims for (including the adversarial model);
- Information on the context (application domain if applicable, demands for the technical, organisational, or legal environment such as a technical platform, a system infrastructure, organisational processes, assumptions concerning the jurisdiction or contracts to be closed etc.).

Changes in the definition of the ToA can yield very different assessment results. Further, it would be pointless to add unrealistic or surprising restrictions in the definition of the ToA. Thus, proper diligence is necessary for this step. As soon as a maturity assessment has been conducted for several PETs, the ToA definitions for further comparable PETs should follow the previous descriptions.

6.3 Gathering measurable indicators

The term 'measurable indicator' refers to every type of information that helps with determining PET maturity and that can be assessed in an objective manner. In our methodology, this task is performed by the assessor, who selects and assesses all measurable indicators relevant to the ToA. Concerning the use of the results of this assessment as input to the methodology, we divided the total set of measurable indicators into two different categories:

Threshold indicators are used to assess the basic readiness level of the ToA, as described below. There are no threshold indicators for the quality assessment.

Soft indicators are assessed and provided to the experts as an additional information that may help them in judging on the particular readiness and quality characteristics of a ToA.

6.3.1 Indicators for technology readiness

According to above definition, assessment of the measurable readiness indicators is performed in two tasks.

Threshold readiness indicator assessment

First, the threshold indicators are assessed. Therefore, the assessor needs to assess the threshold indicator for each pair of adjacent readiness levels, in order to decide whether the ToA meets the threshold for the next readiness level or not. Once a ToA does not meet a certain threshold indicator, that indicator becomes decisive for the basic readiness level of that ToA. The threshold levels are defined as follows:

Idea→research. This threshold indicator is met if there exists at least one scientific publication that focuses on the ToA. The publication has to be published in a scientific, peer-reviewed context, such as conference proceedings, journals, or similar. The assessor has to validate that at least one such publication exists, and has to document his findings accordingly.

Research→proof-of-concept. This threshold indicator is met if there exists at least one working implementation (e.g. laboratory prototype, open source project, proof of existing code, or similar, that compiles and executes, and implements the ToA). The assessor has to validate that at least one such implementation exists, and has to document his findings accordingly.

Proof-of-concept→pilot. This threshold indicator is met if there exists at least one real-world utilisation of the ToA, with non-laboratory users, performed in a real-world application context. The assessor has to validate at least one such pilot use case scenario is performed currently or has been performed in the recent past, and has to document his findings accordingly.

Pilot→product. This threshold indicator is met if there exists at least one product available in a business market, or in a context in which the utilisation of the ToA happens in a real-world business context with transfer of value (typically: money). The use of open source project code in a productive business environment, e.g. within a product suite, along with (paid) consulting services on the use of the project code also fulfils this definition. The assessor has to validate at least one such product or paid use case exists, and has to document his findings accordingly.

Product→outdated. This threshold indicator is met if either the only technology that allows for utilising the ToA becomes obsolete or ceases to exist or a devastating quality problem of the ToA, e.g. a novel attack, was revealed, which cannot be fixed, and hence, all future work on the development of the ToA is likely being abandoned. The assessor has to validate whether one of these two conditions is met, and has to document his findings accordingly.

Given the above order from research to outdated, the achieved readiness level is implied by the highest threshold indicator that is met. For example, if the research→proof-of-concept threshold was met, but the proof-of-concept→pilot threshold was not met, the resulting readiness level of the ToA would be that of proof-of-concept.

Note this is particular important for outdated: if the threshold indicator to the outdated readiness level is met in any case, the resulting readiness level is always that of outdated. This may e.g. play a role if a devastating vulnerability is found in a ToA that e.g. still is of research level. Hence, the assessor must always evaluate each of the threshold indicators and document his findings.

Soft readiness indicator assessment

Unlike the threshold indicators for readiness, the soft indicators have no direct reflection in the methodology proposed. Their use is limited to being informative to the experts, so as to influence their decisions on the readiness assessment of the ToA.

The workflow for assessment of the soft readiness indicators is as follows. Initially, the assessor decides which soft readiness indicators are of potential relevance to the readiness assessment of the ToA, and documents his selection. Then, the assessor performs assessments of each of these soft indicators, trying to gather the correlated information from all available sources (e.g. Internet queries, questionnaires, library research, etc.). Again, all findings are to be documented. Then, the assessor provides his findings to all of the experts, who then can decide on their own on how to judge and consider these indicators in their readiness assessments. This way, consideration of soft indicators can be performed without the need for normalisation, cross-selection, and formalisation of the incorporation of measurable indicators into the overall process. However, if experts deviate from the implications obtained from the soft indicators, they may be asked by the assessor to comment on the reasons for such deviations, and these comments would then be documented as well. For instance, an expert may argue why she judges the ToA of research level, even though the soft indicator of lines of code with value 3256 implies there is an existing implementation, thus indicating a readiness level of proof-of-concept. A possible argument in this case could be that the expert thinks the implementation is not yet feature-complete with respect to the ToA, thus the number of lines of code does not suffice to imply proof-of-concept readiness. Hence, she can add a respective comment to her readiness assessment.

6.3.2 Indicators for privacy enhancement quality

We identified the following soft indicators to determine the quality of a particular ToA, and grouped them by the specific characteristic they contribute to.

Protection. 1) Documented protection levels and properties.

Trust assumptions. 1) Documented trust assumptions. 2) Described adversarial model. 3) Legal measures. 4) Organisational measures.

Side effects. 1) Documentation on known side effects.

Reliability. 1) Availability of stress test reports. 2) The number of (un)successful penetration tests. 3) Number of vulnerabilities discovered.

Performance efficiency. 1) Benchmarks or performance figures for storage, CPU power, bandwidth, latency and throughput.

Operability. See maintainability.

Maintainability. 1) Whether the system is modular in design. 2) Whether test suits exist. 3) Whether the ToA is open source. 4) Availability, extent and detail of documentation. 5) Whether an active developer community exists.

Transferability. 1) List of different software and/or hardware platforms the ToA has been ported to. 2) Evidence regarding the amount of work needed to port the ToA. 3) Whether the ToA uses general purpose programming languages and build environments, and standard libraries. 4) Availability and detail of instructions to port the ToA to other platforms.

Scope. 1) List of application domains the ToA is known to be applicable to. 2) Number of different products serving different markets that use the ToA.

6.4 Gathering expert opinions

In order to reasonably incorporate the expert's feedback into the assessment process, it is necessary to define a standardised process for collection and aggregation of expert opinions. As the expert inputs are needed for both the quality assessment and the readiness assessment, the most suitable approach consists in having all experts fill out a pre-defined questionnaire that covers all necessary inputs for these two aspects. This questionnaire is to be detailed next.

The experts are given access to the measurable indicators collected by the assessor. These measurable indicators form a baseline that the experts can use to base their expert opinion on. This is why the collection of measurable indicators needs to precede the process of gathering expert opinions.

6.4.1 Gathering readiness opinions

In terms of assessing readiness, the expert questionnaire in our approach allows for choosing one of the six possible readiness levels, as defined in Section 5.1. Thereby, the particular expert can input her opinion on which of the six readiness levels a given ToA has achieved. In case of doubt, we allow the expert to choose two adjacent readiness levels at once, in order to clearly illustrate that the expert thinks this ToA to be in transition from one level to the next. Beyond this, there is no expert input the readiness level assessment foreseen in our approach.

It should be stressed that the readiness of ToA should be based on the context provided in the ToA assessment itself. Therefore a more focused ToA that describes a very concrete PET or a concrete product will typically obtain a better, higher, readiness score.

The measurable readiness indicators (see Section 6.3.1) define a minimum overall readiness score that the experts can only increase, not decrease. That is, for each readiness level there is a threshold set of indicators such that, if the ToA meets this threshold set, the ToA is guaranteed to be assigned at least the associated readiness level. Experts are allowed to assign a higher score, but this needs to be motivated.[9] At the start of collecting the expert opinions, the assessor provides this initial readiness score to the experts. So, for example, if the indicators qualify a ToA as being at the pilot level, then the experts need to assess it at level pilot, product, or outdated.

Aggregation of the expert opinions on readiness is performed by the assessor. This task consists in determining the readiness level that the majority of experts have selected. If the majority happens to be two adjacent readiness levels, the corresponding transitional readiness level (e.g. pilot/product) is assigned.

[9] In case experts do not agree to the found threshold criteria, they can challenge it, too. Only in this case, lower scores could be chosen.

However, the assessor needs to additionally verify a certain level of consensus among all the experts. In order to reflect this, the following condition indicates a lack of consensus, which has to be dealt with before proceeding with the assessment process.

If more than one expert has picked a readiness level not adjacent to the level that received the majority of votes, consensus is not reached.

This may, e.g., be the case when two or more experts vote for research, but the majority voted for product. The assumption here is to allow for a single outlier, but to consider a deviation of more than one level by more than one expert as a non-homogeneous assessment, which needs to be resolved.

If such a situation of non-consensus is reached, the assessor needs to communicate with all the experts in the board, in order to determine the rationale behind the deviation of opinions. Potential reasons for such deviation may stem from lack of awareness (e.g. of an existing market product) or weak definitions of the ToA. In either case, the assessor has to determine and resolve the issue, and to re-perform the questionnaire until consensus is reached, or declare the assessment process to be unsuccessful due to sustainable lack of consensus.

6.4.2 Gathering quality opinions

Unlike readiness assessment, the expert's inputs to the quality assessment are more complex. This part of the questionnaire needs to address the manifold characteristics of expert feedback, which are quite challenging especially with respect to the privacy domain. Also, reaching consensus on a ToA's quality typically is not a trivial task.

Unlike readiness, there is no clear set of measurable quality indicators that by itself can be used to assign an initial score to each of the nine quality characteristics. So instead of providing an initial scoring of the quality characteristics based on these indicators, the assessor provides the experts with the values of the measurable indicators. The experts are required to use (and motivate their use) of the indicators provided to them when assigning their scores to the quality characteristics.

Utilising the nine quality characteristics defined in Section 5.2 as a basis, the questionnaire in our approach allows each expert to provide both a scale value (on the defined quality scale spanning from −− to ++) and a free text field for comments for each of the nine quality characteristics. Moreover, each expert is asked for an overarching total quality score, which also utilises the same scale. The rationale here is to allow experts to provide a general recommendation whether he/she considers the ToA's quality fair or not, despite the scores for the nine characteristics. This is somewhat in line with the approach taken for evaluation of scientific publications: the scores for the different parameters help the assessor to estimate these characteristics, but the overall score provides an overarching assessment that immediately reflects the primary tendency of the expert's opinion.

There is a straightforward way to compute the overall quality score given the score of the individual characteristics. This is in fact a weighted average of the individual scores s_c for characteristic c, interpreting scores as members of the set $\{-2, -1, 0, 1, 2\}$ with weight w_c computed as $quality = \sum_c w_c \times s_c$.

Different weights are assigned to the characteristics to reflect their different importance for the evaluation of a PET. In case the expert considers all characteristics as relevant, the weights are as follows: The characteristics protection and trust assumptions have weight 3/16. The characteristics side effects, reliability and performance efficiency have weight 2/16. The remaining characteristics operability, maintainability, transferability and scope have weight 1/16, see Table 1. Note for comparability these weights need to be fixed.

Characteristic c	**Weight** w_c	**Relevance yes/no?**
Protection	3/*x*	The relevance of each characteristic in the PET assessment is chosen by each expert.
Trust assumptions	3/*x*	
Side effects	2/*x*	
Reliability	2/*x*	
Performance efficiency	2/*x*	
Operability	1/*x*	
Maintainability	1/*x*	
Transferability	1/*x*	
Scope	1/*x*	
	Sum: 16/*x*	*x* as denominator is calculated by the sum of the numerators of the weights assigned to those characteristics the experts considers relevant.

Table 1. Weights for the overall quality score.

In case an expert chooses to omit a score for a characteristic because he argues that it is not relevant to the ToA, this has to be taken into account by the formula: For instance, if performance efficiency is not regarded as relevant, this characteristic is omitted and the remaining characteristics are multiplied by weights of *x*/14 instead of *x*/16 for $x \in \{1,2,3\}$.

This, admittedly arbitrary, distribution of weights roughly gives the more important characteristics a stronger influence on the overall quality score. In an assessment framework focused on privacy characteristics like protection and trust assumptions are most important, followed by the characteristics that correspond to performance issues.

This scoring follows a comply-or-explain approach as follows. The expert is supposed to use this formula to compute the overall score. Nevertheless the expert is allowed to adjust the overall score if he has sufficient reasons to do so. He is then however obliged to clearly explain this adjustment.

PET Maturity Assessment – Expert Questionnaire

Assessor: Dipl.-Inf. Marit Hansen	*Expert:* Dr.-Ing. Meiko Jensen
Target of Assessment:	*Pretty Bad Privacy (PBP)*

The Target of Assessment consists in the "Pretty Bad Privacy" toolsuite, as defined on the website http://pbp.tbd . It comprises of a client software product and supporting services, as described on the website, and of a server-side implementation of the PBPEnc and PBPDec protocols. […]

Readiness Assessment

◯ **idea** ◯ **research** ◯ **proof-of-concept** ◯ **pilot** ◯ **product** ◯ **outdated**

Comments on the Readiness Assessment:	

Quality Assessment

Overall Score:	--/-/0/+/++
Comments on the Quality Assessment:	

Quality Characteristics	*Score*	*Relevant?*	*Comment*
Protection	--/-/0/+/++	yes/no	
Trust Assumptions	--/-/0/+/++	yes/no	
Side Effects	--/-/0/+/++	yes/no	
Reliability	--/-/0/+/++	yes/no	
Performance Efficiency	--/-/0/+/++	yes/no	
Operability	--/-/0/+/++	yes/no	
Maintainability	--/-/0/+/++	yes/no	
Transferability	--/-/0/+/++	yes/no	
Scope	--/-/0/+/++	yes/no	

Figure 8. Example of a PET Maturity Assessment Expert Questionnaire.

A detailed example of such a questionnaire is shown in Figure 8.

With respect to aggregation of the quality assessment questionnaires filled out by the board of experts, the assessor again needs to perform the task of consensus validation. As with readiness, we propose a generic condition on quality level deviation, this time to be evaluated against each of the nine quality characteristics and on the overall quality score values.

If more than one expert has picked a quality score not adjacent to the score that received the majority of votes, consensus is not reached.

For example, if two or more experts chose a *protection score* of ++, whereas the majority voted for 0, consensus on this characteristic is not reached.

Unlike for readiness, reaching consensus for quality is complex, and thus we do not urge strict enforcement of harmonisation as for readiness. It is totally valid for the assessor to proceed with the assessment even in presence of non-consensus on at most four out of the nine quality characteristics. However, for protection and trust assumptions (the two characteristics with the largest weight), consensus is required.

If more characteristics are lacking consensus, the assessor should perform at least one round of negotiation among all board experts, trying to resolve the issues, and re-perform the questionnaire afterwards. These rounds will improve the written reasoning of the experts on explaining their scores. If still no consensus can be achieved on at least five of the characteristics, or on protection and trust assumptions, the assessor must cancel the assessment completely.

Also, if no consensus is reached for the overarching, total quality scores as determined from the set of expert's score values, this lack of consensus must be resolved. If no consensus can be reached on the overall quality score, this implies an insufficient information basis for an assessment, or an ambiguous definition of the ToA. In both cases, assigning a quality score, or even a PET maturity level to the ToA would be misleading, as it is not based on a sufficiently consented expert basis. Thus, if the assessor is not able to achieve consensus among the experts on the board with respect to their overall quality scores, she needs to cancel the assessment process without result.

6.4.3 A word on consensus

Both of the above descriptions utilise the concept of consensus, based on the two definitions given. We define consensus as follows.

Consensus is reached if at most one expert deviates from majority by two or more scale values, i.e. levels or score values.

Thus, as a direct consequence, resolving a lack of consensus can be performed by convincing some experts to change their scores, until consensus is reached. However, this does not necessarily imply that the assessor needs to convince one of the two deviating experts to follow the current majority. It is perfectly valid to have the majority of experts follow the deviating group, e.g. if the deviators manage to convince the other experts that a product for the ToA exists, thus justifying their readiness assessment of product instead of proof-of-concept.

In case of missing consensus, it is obviously not an option for the assessor to ignore expert's answers, or to expel experts from the board, in order to reach consensus this way. Any such attempt would invalidate and discredit the overall assessment and its results immediately. This can be enforced by requiring all communication be logged and kept on file, and make it widely known that any expert that is ever called upon to

evaluate a PET will check that his or her contributions are properly used and logged, and raise an alarm if not.

6.5 Combining indicators and opinions for the overall assessment

The measurable indicators and the expert opinions need to be combined to complete both the readiness assessment and the quality assessment. Once those have been determined, the overall maturity assessment can be completed. This process is described below.

6.5.1 Readiness assessment

Once the readiness indicators and the expert opinions have been collected the assessor can start compiling the Readiness Assessment Report.

The Readiness Assessment Report consists of the following parts.

- The final readiness score.
- The threshold readiness score according to the collected measurable indicators, with a summary of their values.
- The consented overall readiness score from the experts.

If there is unanimity on the readiness scores from all the experts, the assessor must assign this as the overall readiness score. If there is no overwhelming majority on a readiness score, the assessor must pick the intermediate readiness score that lies between both assigned scores. For example, if some experts assign score research and others assign the score pilot, then the assessor must assign the overall readiness score research/pilot. The assessor has to document the rationale for her decision in the assessment report.

6.5.2 Quality assessment

Once consensus on the quality characteristics and the overarching score value is reached, the assessor can compile the Quality Assessment Report. This report represents the result of the quality assessment part of our methodology (cf. Figure 7). Moreover, it is used to compile the quality input to the overall PET maturity level, as described in Section 6.5.3. The Quality Assessment Report of a PET maturity level assessment needs to be published alongside the PET maturity level achieved, in order to allow for subsequent investigation of the validity of the scores achieved.

The Quality Assessment Report consists of the following parts: First, the consented overall quality score is listed. If there is no unique majority (e.g. because two different quality scores received the same amount of votes from the experts), the assessor is allowed to pick one of these, according to the comments received from the experts. The assessor has to document the rationale for her decision in the assessment report, though.

Then, for each of the nine quality characteristics, the score value that received the majority of votes is listed (even if no consensus was reached). Moreover, all text comments given by all experts in the board are concatenated, and are listed alongside the quality characteristic they correspond to. This way, the rationale and motivations behind the quality scores of the nine characteristics can be assessed even for external observers that were not involved in the overall assessment process. Finally, the names and roles of all experts in the board and that of the assessor are listed and accompanied by a signature. If no consensus was reached on a characteristic, this is also recorded.

6.5.3 Maturity assessment

The last step in performing a full PET maturity assessment of the ToA consists in combining the results from the quality assessment part with the achieved readiness level. In our approach, this task narrows down to

aggregating the Quality Assessment Report's findings into a single quality indicator (on the quality scale described in Section 5.2), and attaching that quality indicator to the readiness level of the ToA. The combined result thus is a bipartite value anywhere in the range between idea–– to idea++ and outdated–– to outdated++, as illustrated in Figure 5.

6.6 Tool support

The illustrated assessment process can be supported by tools that provide guidance for the assessor and the experts throughout the entire assessment. A first step for developing such tools has been done within this project with the provision of a template of the expert questionnaire (see Appendix A) and the structured example of the PET Maturity Assessment Report (see Annex B:). On this basis, a tool could guide through the different steps in the process, make sure that the experts have dealt with all aspects mentioned in the questionnaire, and provide a structured documentation of the assessment. Also, tools could assist in the mechanism for finding a consensus. Further tool support could help in (semi-)automatically gathering indicators such as numbers of publications, citations, popular media references, research projects, curricula mentioning privacy enhancing technologies etc.

7. Pilot PET Assessment

In order to demonstrate feasibility of the methodology defined in the previous sections, we performed a pilot study that followed the methodology as close as possible, trying to detect issues of the process, feasibility problems, etc. In this section, we outline the approach for our study and present the results gathered from our methodology.

Beyond this study, we also performed a second pilot evaluation in a less controlled way, implemented as an open experiment among the participants of the 2015 IFIP Summer School on Privacy and Identity Management in August 2015. The results of this second evaluation are presented in this section as well.

The next section then focuses on evaluation of the first study on a meta-level, gathering the lessons learned from the study, and presenting the implications to the definition of our methodology and process.

7.1 The IRMACard pilot study

7.1.1 Choice of the PET to be evaluated

In preparation of the study, we identified the following key requirements towards the PET to be evaluated.

- Ideally, the PET in consideration should not be known to all of the experts. Hence, choosing one of the more well-known PETs was not an option.
- Ideally, the PET in consideration should not be kept in a confidential development container (such as in-house PETs developed at major companies), as this would probably have impacted on the availability of information on the PET. Thus, we chose an openly available PET.
- Ideally, the PET should be one of a relevant readiness level. Hence, we did not consider PETs that have a high probability of a readiness level of idea or outdated.

Based on these conditions, we chose a PET that fulfilled all of these conditions. Based on existing knowledge of the PET market within our project team, we eventually decided for the IRMACard project[10] as Target of Assessment for our pilot study.

7.1.2 Definition of the Target of Assessment

In order not to influence the readiness opinions of our experts up front, we decided not to refer to the IRMACard as a product, but merely took the approach of a generic description of the technology. Also, we intentionally implemented some marketing lingo in our definition, anticipating such to be present in other ToA definitions of the future as well. The resulting definition of the IRMACard Target of Assessment read as follows:

The IRMACard is a smartcard that has been developed in the IRMA project. IRMA is an acronym for “I Reveal My Attributes”, which already points to the core privacy enhancing technique used within IRMACards. The technique of Attribute-based Credentials, or ABCs for short, allows for provable attestation of selected attributes, potentially even derived attributes like age derived from a birthdate, of an identity. Given that the ABC solution is highly flexible on the selection of attributes to verify, it is no longer necessary to reveal irrelevant attributes in an authentication process. The IRMACard implements this ABC technology in an integrated, all-in-one solution, consisting of smartcards and verification devices. It is already in use as an authentication device in a printer kiosk for students. Here, students can use their IRMACard to authenticate as

[10] More information on this technology can be found at: https://www.irmacard.org/

members of a particular master program, and are allowed free printing services once this has been verified. Importantly, the IRMACard does authenticate the students as members of the specific programme, but does not reveal any other attributes. No name, no age, no identifier. Just the fact that this particular student is part of the specific master program. Thus, IRMACards provide an extensive degree of unlinkability in the context of authentication.

7.1.3 Selection and invitation of experts

The pilot study officially started with the assignment of the role of the study assessor, as defined in our process of Section 6. Then, following the anticipation of the real-world implementation of the process, the assessor chose the experts himself, without disclosure of their identity, neither among each other, nor towards other members of the project team. This was done by intention, in order not to affect the experts in their objectiveness.

Subsequently, the experts were invited to participate in the study via e-mail. Once the necessary number of experts was met (five experts for this study), and the board of experts was complete, the PET in consideration was disclosed to the experts. Hence, all experts had the same time to make themselves familiar with the ToA and judge it according to the scales of our methodology.

Given that this was the first implementation of our approach, we also added two more components that would probably not be necessary in a real-world implementation of our methodology. First, after starting the study period, we immediately arranged for a final, concluding telephone conference, not to discuss the PET or its assessment itself, but merely to collect the feedback of our experts on the meta level.

Second, we implemented the role of a study observer, whose assigned role was that of a supervisor of the study. If experts would have had feedback or issues with the implementation of the process itself prior to conclusion of the study period, they were able to contact the study observer directly, who then was empowered to decide whether and how to proceed or stop the study during its execution. This role was implemented separately from the role of the assessor, in order not to influence the actions of the assessor towards (parts of) the board of experts in any way.

During the study period, there were few contacts to the study observer, and none of them led to an active alteration of the study in progress.

7.1.4 Results of the study

After all experts have submitted their assessment forms, and after the assessor had performed his duties with respect to collection and assessment of measurable indicators (see Section 6.3), it turned out that the experts had reached consensus on both their quality and readiness assessments immediately. Thus, there was no need for conflict resolution, and the assessor proceeded with determining the final results of the study. As the last part of the study implementation, the assessor thus compiled the final assessment report Annex B:, and the pilot study concluded.

7.2 The TOR open experiment

Unlike the controlled study described above, which was performed under rather controlled conditions, and trying to reflect a close-to-real implementation of the evaluation process, we performed a second evaluation experiment for gathering feedback to our methodology as part of the 2015 IFIP Summer School on Privacy and Identity Management. Given the differences in nature to the controlled laboratory environment implemented in the previous study, the results of this experiment must be considered with caution, as they do not reflect the results of a reasonable, controlled, full execution of our assessment process. Nevertheless, given the impressive results, we provide the details and findings of this experiment here as well.

7.2.1 Choice of the PET to be evaluated

Given the audience attending the IFIP Summer School (which was a mixture of privacy experts and early researchers, Ph.D. students etc., also representing a mixture of scientific disciplines), we decided to take one of the most prominent PETs in the world: the TOR project.

7.2.2 Definition of the Target of Assessment

As we assumed that all attendants at least had heard of the TOR project, we decided to not affect their readiness assessment a priori. Thus, we just removed any descriptive text referring to the Target of Assessment, and simply stated the Target of Assessment to be “The Onion Router (TOR)”.

7.2.3 Selection of experts

Participation in this experiment was open to all attendants. Thus, the questionnaire was distributed openly to all participants, and no controls were implemented to prevent double submissions or unintended collaboration among participants. Moreover, expertise of the participants was not verified at all. Hence, the set of participants assumedly ranges from experienced scientists in the privacy area to Ph.D. students from the disciplines of law, computer science, and social sciences.

7.2.4 Results of the study

Unlike the previous study, there was neither an assessor nor a collection of measurable indicators. There was no resolution in case of lack of consensus, and there was no concluding results report. Hence, this section presents the most remarkable results of this experiment directly.

Figure 9 and Figure 10 show the results achieved in terms of readiness assessment and quality assessment. As can be seen, even though the number of participants (14) exceeds the assumed number of experts, consensus was still reached immediately for both readiness and quality. The resulting score that was reached was product+.

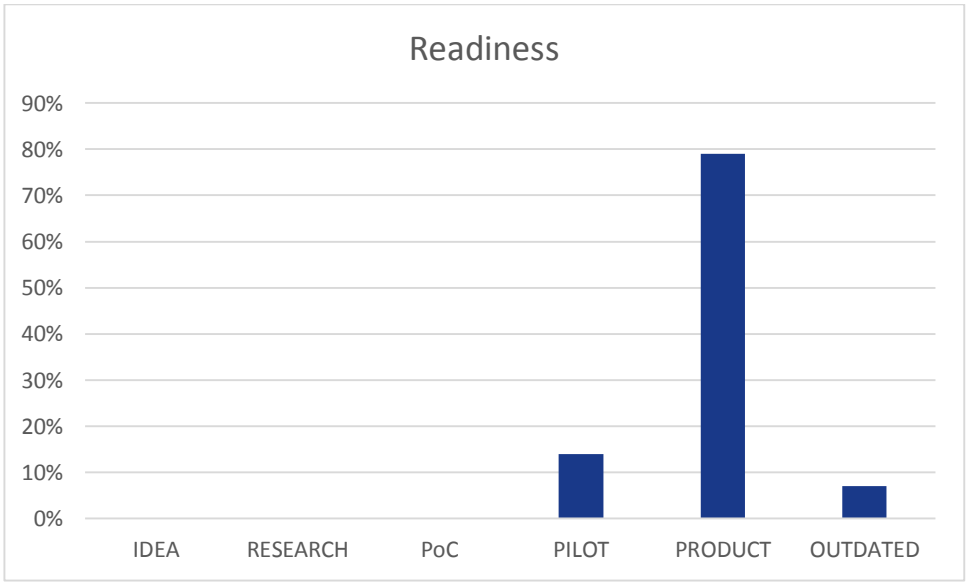


Figure 9. Readiness assessment results for the TOR experiment.

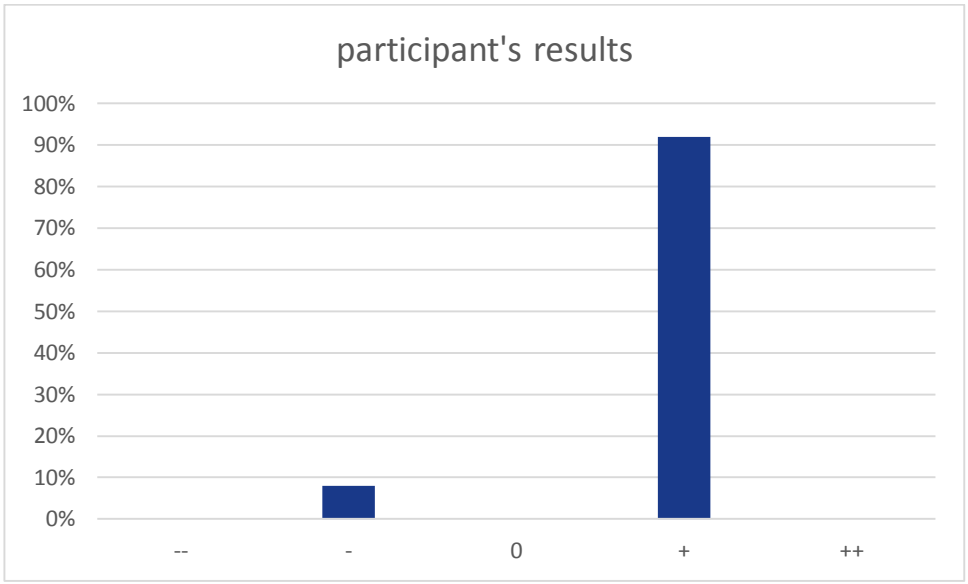


Figure 10. Quality assessment results for the TOR experiment.

8. Evaluation

In this section we evaluate the methodology concerning its adequacy, ease of use, effectiveness, and effort required to evaluate a PET. As input to this evaluation we use the experiences of the experts evaluating a PET in the pilot (described in Section 7).

8.1 Adequacy of the assessment methodology

The methodology is adequate, if the results are aimed at a large audience of software developers, policy makers and the like. The more it is applied consistently to a large set of different PETs, the more useful it becomes. It is important to make this target audience explicit in the methodology, and made clear to the experts involved in an evaluation.

The distinction between readiness and quality and scoring them separately was considered a good approach giving clear results. One expert wondered how the scores were combined to a single score (they are not, really), so this needs to be made clear in the documentation sent to experts.

One expert felt that the effect on the end user (i.e. whose privacy we are trying to protect) was underrepresented. For example whether the PET is usable, whether introducing the PET might introduce new risks to the user, or whether there is some over reliance on the fact that a PET protects the privacy more than it in fact can do. Also, whether an end user can comprehend the protection offered by the technology was felt missing.

8.2 Ease of use of the assessment methodology

The methodology is easy to use for an expert. The documentation that was sent was comprehensive enough for experts, but probably not as easy to understand for non-experts or industry people. Also, non-experts may find it hard to obtain all the necessary information to determine their scores. Perhaps the documentation package should be made more extensive.

It was not always clear what to use the comment field for, and how much text to provide. Some expert suggested using it to explain possible interpretation issues with certain characteristics. Lack of practical experience (especially for experts from the research domain) makes it hard to score characteristics like maintainability. Lack of information makes it hard to score reliability, performance and again maintainability.

8.3 Effectiveness of the scale: comprehensibility, comparability, scorability, reproducibility

Here we discuss the perceived effectiveness (as defined in Section 4.1) of the readiness and quality scales.

8.3.1 Readiness

The readiness score was considered for the most part effective. No ambiguity was perceived; it was easy to score. One expert suggested to also make explicit in the definition of the score how much effort would be required to advance to the next level.

There is a difference between a PET with only 4 research papers published, and one that had been studied in say 50 papers by different research teams. The first PET should still be classified as an ‘idea’. Some experts felt that it was not expressed anywhere in the scores what amount of attention (number of different products, for example) a PET had received.

It should be noted that it is not always easy to score the readiness of a PET. E.g.: Freenet has some proof-of-concepts, but which of the research papers do they implement (if any at all). Versions also influence this: one version of a PET may be still research, while an older version is already a product.

For industry people, a pilot is something that delivers useful information to decide if and how to build a product. So when evaluating products, make sure you have a good mix of academic and industry experts.

8.3.2 Quality

There were some questions one why the selected characteristics were chosen, and some were deemed less relevant (see below). The five-value scale was perceived adequate.[11] It is important that the experts also know how the final quality score is derived from the scores on the individual scores on each of the characteristics.

Even though the documentation was comprehensive, the quality was harder to score. One question was what to compare a PET to when scoring one particular characteristic. For example, do you compare a PET with other PETs when scoring protection, or with a broader scope. In the first case the score may be too harsh (low). It is important to create or establish a benchmark, or provide a couple of examples.

'Side effects' was a characteristic that was confusing to the experts. Maybe provide a couple of examples of the kind of side-effects we mean, either in the definition of 'side effects' or in the documentation package. Also, does ++ mean loads of side effects or good (i.e. low on side effects). This also holds for scoring trust assumptions.

For scope, the question is whether it is even relevant because even a PET with a limited scope may be very useful, more useful than a PET with a broader scope.

Availability (in terms of source code, producers, etc.) was a characteristic that was felt missing to assess the overall quality of a PET.

One way to address the lack of the user perspective would be to introduce a 'verifiability' characteristic to the quality score. Also one expert suggested adding a general 'usability' characteristic.

It was suggested to also add a catch-all characteristic.

Protection was considered quite course grained. It may be good to distinguish, e.g., integrity and other aspects separately. Clearly define the scope of what protection means: is it scored against just the advertised protection (i.e. does it deliver on its promises) or compared to protection offered by other products or compared to no protection at all. Also, what happens when a product gets hacked and then gets patched again: do you allow the score to jump up and down?

Finally in the feedback on the methodology the difference between scoring a product vs. scoring a research idea was pointed out. For example, transferability may be lower for actual products because of business reasons than for abstract research ideas that have not been implemented, yet. So a PET with a lower readiness level may have better scores for some characteristics than PETs with a higher readiness level.

[11] Although one expert predicted that a maximum score would never be given.

8.4 Effort required to perform the assessment

The assessment did not require much effort, and experts said it was less work than expected. It is helpful (when recruiting experts) to make clear that the evaluation does not involve a lot of work. Especially if experts know the technology.

9. Dissemination and Continuation

In the previous sections we have presented a methodology to assess the maturity of a privacy enhancing technology. We have applied our methodology to one particular PET to assess the applicability and validity of our approach in practice. This assessment showed that our approach is viable and delivers usable results, but clearly needs more work for applying it to arbitrary PETs and yielding reliable and comparable results that can be taken up by interested stakeholders. This then begs the question how to further develop our approach and what steps must be taken to ensure that also other PETs are evaluated for their maturity.

In the following, important aspects are outlined that have to be considered for that objective. Since there is currently no stakeholder that has assumed the task of maintaining an overview of PET maturity assessment, it is not clear whether the work will be continued. Therefore we refrain from showing a detailed roadmap for the progress of this work. Instead, we will briefly discuss the necessary steps: assessment of other PETs, establishment of a structured assessment process, maintenance of a PET maturity repository, analysis of utilisation venues, and dissemination.

9.1 Assessment of other PETs

We are convinced that assessment of other PETs is necessary to test and further develop the methodology and the defined process. For this undertaking, different kinds of PETs with varying complexity should be chosen to challenge the methodology and the assessment process. The selected PETs should be representative for a huge part of what currently the PET community is working at, both in research, in the open source community, and in the market. For the choice of PETs, the work presented at the PET Symposium and related conferences, references in the opinions of the Art. 29 Data Protection Party, European and international research projects, the Internet Privacy Engineering Network (IPEN), in standardisation, in the certification area with privacy seals, and in surveys or PET publications (such as [12]) should be taken into account.

Of course, on the basis of the proposed results any group could set up an assessor, invite experts, and conduct assessments. However, for learning from the assessments and refining the criteria, the scores, the scales, and the procedures, a coordinated approach would make sense to guarantee comparability and steady feedback from the exercises until the resulting assessment process can be regarded as sufficiently mature. At best, the assessor, or at least observers for the process, should stay within a dedicated consortium responsible for further developing the procedures. The experts, however, should not be the same for all assessments.

The European Commission should support this research line with an appropriate mechanism, e.g. through a network of excellence in the field.

9.2 Establishment of a structured assessment process

For reliable, reproducible and comparable results, the methodology must be turned into a more structured process. In particular, it is necessary to establish a method of generating and publishing the results and the reports. It is conceivable, and would be favourable, to develop tools that support a standardised step-by-step walk-through for the assessment of both readiness and quality of a PET, see Section 6.6. In case the lessons learnt from the assessment of further PETs results in changes of some aspects in the assessment process, it has to be checked how this would influence previous assessments and their results. In the phase of establishing a sufficiently mature process, it may become necessary to repeat earlier assessments, or at least partially adapt the previous evaluations and their outcomes. This, again, calls for a dedicated consortium with a clearly assigned responsibility as guardian and innovator for the structured assessment process.

Moreover, a repository of performed assessments has to be set up and maintained. The limits of such a repository should be clearly communicated: Certainly it is not sufficient for system design to pick and choose from a list of PETs on the basis of their maturity assessment results. As pointed out in [12], a naïve installation of a technical tool without sufficient understanding of the problem space, the solution approach, and potential side effects would most likely not work out. Instead, extensive expert guidance is necessary to prevent mistakes and choose the most appropriate solution.

ENISA should form a consortium that prototypes such an assessment tool. The consortium needs to involve all relevant stakeholders.

9.3 Maintenance of a PET maturity repository

For maintaining a PET maturity repository, the responsibility for the process and repository must be assumed by some organisation. This organisation should be independent from industry and not involved in development or provision of any of the PETs to be assessed. For the operation of this task, significant resources must be made available to that organisation. Here funding seems to be necessary to be independent from payments from PET developers.

Of course there could be more than one PET maturity repository, and more than one organisation being responsible. In this case, stakeholders should be enabled to compare the different PET maturity repositories and their results for specific PETs. This would require transparency of the criteria, the process, and people and organisations involved. In case of deviations of the criteria or the process, the different advantages or disadvantages should be evaluated for a better understanding and a comparability of results.

Also legal questions have to be considered when assuming the responsibility for a PET maturity repository. Manufacturers, researchers, or users of a PET may not agree with the results of the experts and send complaints. They may even try to legally stop the publication of an assessment if they believe it has been done in an unfair way or may have negative effects for them and their business. In any case, a thorough legal checking is highly recommended before starting a PET maturity repository.

The European Commission should mandate a supranational body to maintain a repository of best available techniques in the field of PETs. Without assuming a concrete structure for such a repository, the development and maintenance needs to be community driven, transparent, and independent form interests of a single stakeholder group.

9.4 Analysis of different utilisation venues of the assessment results

Another question that came up in many occasions throughout the project was that of utilisation purposes of the assessment results. The assessment process and result scales and scores have been developed in a way that they neither encourage nor discourage any specific way of utilisation. Hence, this flexibility allows using the results of an assessment in many different ways.

For instance, PETs that score on a readiness level of research and have a sufficient quality score as well may become of interest for funding agencies, trying to push their development into a proof-of-concept or even a pilot phase. Similarly, PETs that already are of pilot level may become of specific interest to investors and entrepreneurs in digital markets who may join the development of such PETs to advance them to product readiness. Researchers again may select their targets of interest in the range of PETs with sufficient quality levels, trying to increase knowledge on how they work and test their limits. Low quality scores of PETs, on the other hand, may cause companies to “eliminate” those from their existing products and services, thus advancing the state of PET utilisation towards higher quality ones in general.

Each of these potential venues may profit from the utilisation of our methodology. However, for each of them, it requires a lot more experience to define the correct path of utilisation of the scores and scales developed in this project. This, again, shows that further work is required for introducing a standardised and well-researched procedure for assessing the maturity of PETs. In particular, it would not be sufficient to rely only on computer scientists from the PET community, but additional disciplines (law, economics, politics, psychology) could be involved when analysing the utilisation venues and taking into account the differing incentives and interests of various stakeholders.

9.5 Dissemination

For disseminating the results of this project, but also the plans for further developing, testing, and verifying the assessment process it is necessary to spark a public discussion. Within the project's lifetime the project consortium has presented the interim results mainly at events in the international research domain, among others the 2015 IFIP Summer School on Privacy and Identity Management in August 2015 (http://www.ifip-summerschool.org/) and the Annual Privacy Forum 2015 in October 2015 (http://www.privacyforum.eu/) [19]. Moreover, a presentation was given at a workshop of the Internet Privacy Engineering Network (IPEN, https://secure.edps.europa.eu/EDPSWEB/edps/EDPS/IPEN).

As a next step, the dissemination should be extended towards Data Protection Authorities and their working groups, e.g. the Art. 29 Data Protection Working Group or similar associations on the European or national levels. Initiatives within standardisation bodies such as W3Cs Privacy Interest Group (PING, http://www.w3.org/Privacy/) or the groups working on Privacy Impact Assessments (e.g. in ISO/IEC) or on Privacy by Design (e.g. in the European Committee for Standardization (CEN)) should be addressed, too. In addition, certification bodies for privacy seals could be interested in combining their work with results from maturity assessment although this has to be thoroughly thought through to prevent potential conflicts of interest. Among the active players are certification initiatives from DPAs such as "Datenschutz-Gütesiegel" from ULD (https://www.datenschutzzentrum.de/guetesiegel/), CNIL Seal "Privacy Governance Procedures" (http://www.cnil.fr/), "ICO Privacy Seal" (https://ico.org.uk/), or the seal from the Federal Data Protection and Information Commissioner Switzerland (http://www.edoeb.admin.ch/), as well as the European Privacy Seal (https://www.european-privacy-seal.eu/), the work of Stiftung Datenschutz (http://www.stiftung-datenschutz.de/) or the "Data Protection Seal" as described in Art. 39 of the upcoming General Data Protection Regulation.

By focusing on how to disseminate the methodology itself to standardisation bodies and other relevant platforms, it may be possible to create a momentum that can generate the necessary resources in the future. At the very least it will lead to more attention being paid to privacy, and its assessment, in the standardisation process for privacy technologies itself.

The next years will show how the idea of "Privacy by Design" and technologies that support and enhance privacy and data protection will develop. One important cornerstone will be the European General Data Protection Regulation: not only "Data Protection by Design" will be demanded, but also instruments such as "Data Protection Impact Assessment" and a "European Data Protection Seal" will be introduced. One important aspect of a dissemination and evolvement strategy will be the cooperation with stakeholders active in the field such as Data Protection Authorities or research and industry groups: The considerations on readiness and quality of PETs will play a vital role when shaping and employing the legally demanded instruments.

10. Conclusions and Future Work

Assessing the maturity of PETs is not an easy task. However, the determination of both technology readiness and privacy enhancement quality will be of interest for many stakeholders. This deliverable has shown the first results considering PET maturity assessment.

Starting off from the initial observation that PET maturity comprises of PET readiness and PET quality, and that one should not investigate or measure one of these without the other, we developed and tested a full methodology to assess technology readiness, quality, and maturity of PETs in general, independent from their domain of application or type of utilisation. Based on inputs from both human experts and objectively measurable indicators, we defined an assessment process and methodology that allows for assessing and comparing PET readiness, quality, and maturity in an intuitive, easy-to-use way. The experiments and user studies we performed showed the feasibility and appropriateness of our methodology, and the expert feedback we received throughout the project duration validated our hypotheses and encouraged us to continue working on this methodology in the future.

Continuation of this work was often demanded, both from research and industry experts to whom we presented our results. However, continuity and sustainability of research on this highly challenging yet highly demanded direction of research depends on contributions by many. It demands experts to join the research community of PET maturity, as well as it demands financial support by organisations that profit from this work. If both could be raised in a sufficient quantity, the obvious next steps forward for the future work in this research domain would be as follows.

- Performing further pilot assessments for a wider variety of PETs to evolve and validate the scores, scales, methodology, and processes.
- Presenting the results for various stakeholders, e.g. Data Protection Authorities, researchers, industry groups, standardisation bodies, and collecting feedback for further improvement.
- Analysing incentives and obstacles for different potential solutions for a reliable and independent maintenance of a PET maturity assessment repository.
- Solving legal and financial issues.
- Developing tools that support the assessment by gathering information on indicators and guide the experts and the assessor through the assessment process.

In total, we conclude that the methodology presented in this document works, that it has found a lot of supporters among experts working in the field, and that it has the potential to initiate a new way of thinking about PETs and PET utilisation in digital markets for the decades to come.

Annex A: Assessment Form for Experts

For Readiness Assessment, please use the readiness scale below:

Idea. Lowest level of readiness. The PET has been proposed as an idea in an informal fashion, e.g. written as a blog post, discussed at a conference, described in a white paper or technical report.
Research. The PET is a serious object of rigorous scientific study. At least one (but preferably more) serious academic paper(s) have been published in the scientific literature, discussing the PET in detail and at least arguing its correctness and security and privacy properties.
Proof-of-concept. The PET has successfully been implemented, and can be tested for performance and other properties in practice. "Running code" is available.
Pilot. The PET is or has (recently) been used in some small or larger scale pilot applications with real users. The scope of application, and the user base may have been restricted (e.g. to power users, students, etc.).
Product. The highest readiness level. The PET has been incorporated in one or more generally available products that have been or are being used in practice by a significant number of users. The user group is not a priori restricted (by the developers).
Outdated. The PET is not used anymore, e.g., because the need for the PET has faded, because it is depending on another technology that is not maintained anymore, or because there are better PETs that have superseded that PET.

For Quality Assessment, please use a scale from -- (very poor) over 0 (moderate) to ++ (very good). Similar to paper reviews, please give scores for all of the listed criteria, with "Overall Score" being the most relevant one. If you consider a characteristic to be irrelevant, or if you feel unable to judge on that quality characteristic, you can indicate this by ticking "no" in the column "Relevant?" In that case, your scores for this characteristic will not be considered.

Protection. The degree of protection offered (in terms of for example unlinkability, transparency, and intervenability) to prevent privacy infringements while allowing access and normal functionality for authorised agents. Also depends on the type of threats and attacks against which the PET offers protection.
Trust assumptions. The number of components and/or agents that need to be trusted, and the nature and extent of trust that must assumed in order to use the PET. Also depends on whether these assumptions are legal, organisational, procedural, or technical.
Side effects. The extent in which the PET introduces (undesirable) side effects. Measured in terms of composability.
Reliability. The degree to which a system or component performs specified functions under specified conditions for a specified period of time. Measured in terms of fault tolerance, recoverability, and compliance. Also measured in terms of the number of vulnerabilities discovered.
Performance efficiency. The performance relative to the amount of resources used under stated conditions. Measured in terms of resource use (storage, CPU power, and bandwidth) and speed (latency and throughput).

Operability. The degree to which the product has attributes that enable it to be understood, and easily (and in particular securely) integrated into a larger system by a qualified system developer. Measured in terms of appropriateness, recognisability, learnability, technical accessibility, and compliance.
Maintainability. The degree of effectiveness and efficiency with which the product can be modified. Measured in terms of modularity, reusability, analysability, changeability, modification stability, and testability. Open source software typically scores high on this characteristic. Also, systems that have an active developer community, or that have official support, score high.
Transferability. The degree to which a system or component can be effectively and efficiently transferred from one hardware, software or other operational or usage environment to another. Measured in terms of portability and adaptability.
Scope. The number of different application domains the PET is applied in or is applicable to.

PET Maturity Assessment – Expert Questionnaire

Assessor:	*Expert:*
Target of Assessment:	*Pretty Bad Privacy (PBP)*
The Target of Assessment consists in the "Pretty Bad Privacy" toolsuite, as defined on the website http://pbp.tbd . I comprises of a client software product and supporting services, as described on the website, and of a server-side implementation of the PBPEnc and PBPDec protocols. […]	

Readiness Assessment

◯ idea ◯ research ◯ proof-of-concept ◯ pilot ◯ product ◯ outdated

Comments on the Readiness Assessment:	

Quality Assessment

Overall Score:	--/-/0/+/++
Comments on the Quality Assessment:	

Quality Characteristics	*Score*	*Relevant?*	*Comment*
Protection	--/-/0/+/++	yes/no	
Trust Assumptions	--/-/0/+/++	yes/no	
Side Effects	--/-/0/+/++	yes/no	
Reliability	--/-/0/+/++	yes/no	
Performance Efficiency	--/-/0/+/++	yes/no	
Operability	--/-/0/+/++	yes/no	
Maintainability	--/-/0/+/++	yes/no	
Transferability	--/-/0/+/++	yes/no	
Scope	--/-/0/+/++	yes/no	

Annex B: PET Maturity Assessment Report

PET Maturity Assessment Report

Target of Assessment	IRMACard
Assessor	Dr.-Ing. Meiko Jensen
Board of Experts	Prof. Dr. Claudia Diaz, Dipl.-Inf. Marit Hansen, Dr. Florian Kerschbaum, Dr. Gregory Neven, Prof. Dr. Thorsten Strufe
Start of Assessment Period	17.08.2015
End of Assessment Period	02.10.2015

Readiness Assessment

Measurable readiness indicators

Threshold RESEARCH: according to different sources, there exist at least three different published scientific papers that refer to IRMACard. Hence, I consider this threshold met.

Threshold PROOF-OF-CONCEPT: according to the project's website, there exists at least one implementation of an IRMACard. Hence, I consider this threshold met.

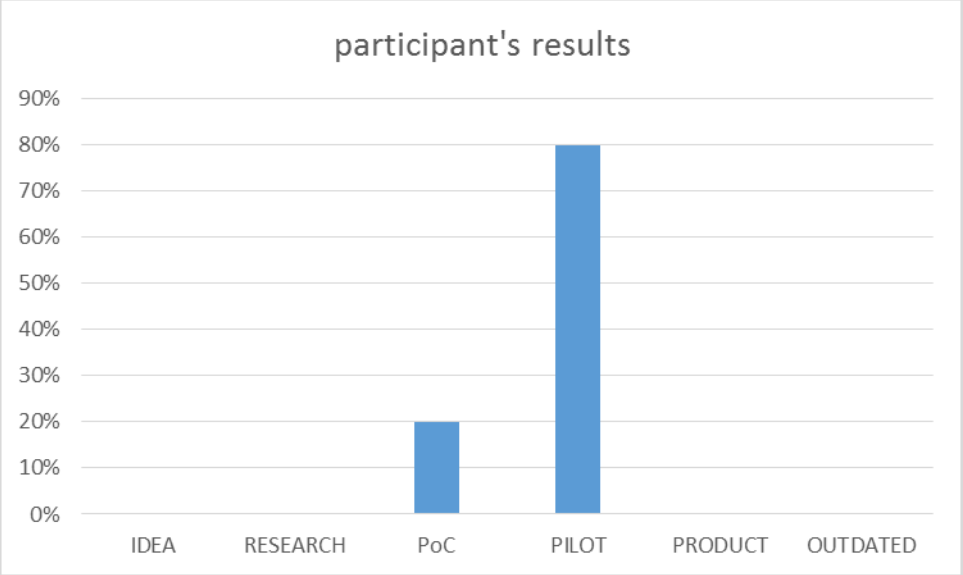
Threshold PILOT: according to the project's website, there exists at least one IRMACard pilot, for instance the IRMA printer kiosk. Hence, I consider this threshold met.

Threshold PRODUCT: according to the project website, it is feasible to apply for an IRMACard. However, to the best of my knowledge, there exists no commercially used product that utilises the IRMACard in a productive environment. Also, despite some research efforts, I was not able to gather any information on the market size of IRMACard technologies. Hence, I consider this threshold to be not met.

Threshold OUTDATED: To the best of my knowledge, there exists no official deprecation notice regarding the IRMACard. Moreover, I was not able to spot any report on a critical vulnerability (despite the trivial ones) concerning the IRMACard or attribute-based credentials (ABCs) in general. Hence, I consider this threshold to be not met.

Over all, I conclude that the IRMACard must be on the PILOT readiness level, according to the measurable indicators gathered.

Expert opinions



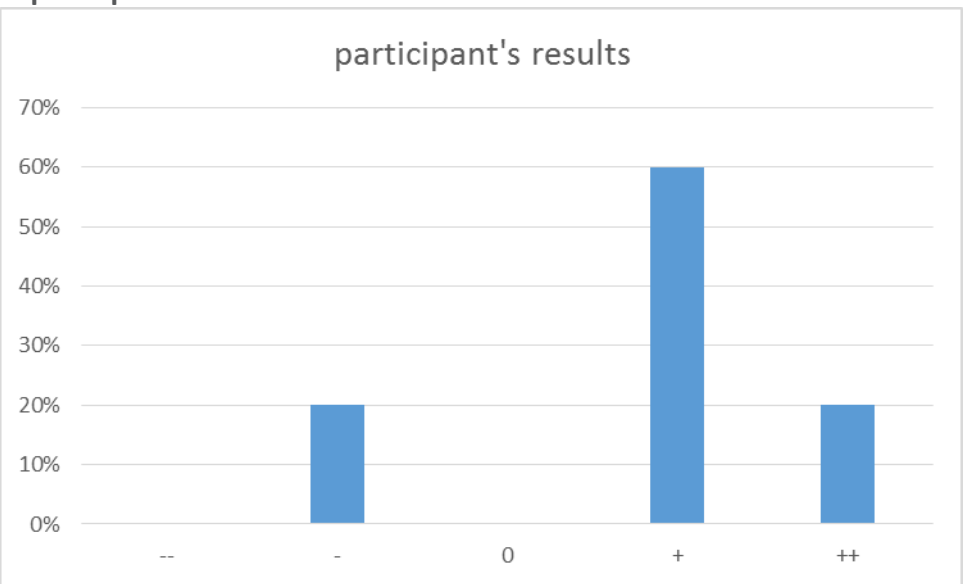
- Consensus was reached immediately.
- The final result of the experts' readiness assessment is PILOT.

Result. As both expert opinions and measurable indicators agree, the readiness assessment concludes with the resulting readiness level of

PILOT

Quality Assessment

Expert opinions



- Consensus was reached immediately.
- The final result of the experts' overall quality assessment is +.

Expert comments on the overall quality assessment

- "With all attribute-based credentials, the main problem is user-verifiability. It is not clear that the user may not be deceived into revealing more information than he intends. In many application scenarios simpler solutions achieve higher user verifiability. I am still looking for the case where attribute-based credentials are the fitting solution."
- "While not as feature-rich and flexible as the Privacy-ABC framework, the IRMA card is unrivaled as the best performing Identity Mixer implementation on smart cards to date."
- "I believe the IRMA card is a technical solution that provides the highest levels of privacy protection for identity management. The implementation seems to have a reasonable performance, it is implemented as open source, and presented together with extensive information for the general public. The solution provides a basic building block that can be useful for privacy-enhanced authentication in a variety of applications."

Results of assessment of the quality characteristics

Protection	++
Trust Assumptions	+
Side Effects	0
Reliability	+
Performance Efficiency	+
Operability	0
Maintainability	+
Transferability	0
Scope	+

Result

The quality assessment concludes with the resulting quality level of

+

Maturity Assessment

The PET maturity assessment concludes with the resulting level of

PILOT+

Annex C: Bibliography

[1] ISO/IEC 27004: Information technology – Security techniques – Information security management – Measurement, 2009.
[2] ISO/IEC 25010: Systems and software engineering – Systems and software quality requirements and evaluation (SQuaRE) – System and software quality models, 2011.
[3] ISO/IEC 15504-5: Information technology – Process assessment – Part 5: An exemplar software life cycle process assessment model, 2012.
[4] Article 29 Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques. 0829/14/EN, adopted on 10th April 2014.
[5] Article 29 Data Protection Working Party. Opinion 4/2007 on the Concept of Personal Data. 01248/07/EN, adopted on 20th June 2007.
[6] Article 29 Data Protection Working Party. Opinion 03/2014 on Personal Data Breach Notification. 693/14/EN, adopted on 25th March 2014.
[7] Assistant Secretary of Defense for Research and Engineering (ASD(R&E)). Technology Readiness Assessment (TRA) Guidance. Technical report, Department of Defense, USA, 2011.
[8] Abhaya Asthana and Jack Olivieri. Quantifying Software Reliability and Readiness. In
IEEE International Workshop Technical Committee on Communications Quality and Reliability – CQR 2009, pages 1–6, May 2009.
[9] James W. Bilbro. Systematic Assessment of the Program / Project Impacts of Technological Advancement and Insertion Revision A. Technical report, 2007.
[10] John J. Borking and Charles D. Raab. Laws, PETs and other Technologies for Privacy Protection. Journal of Information, Law & Technology (JILT), 1(1), 2001.
[11] Elizabeth Chew, Marianne Swanson, Kevin Stine, Nadya Bartol, Anthony Brown, and Will Robinson. Performance Measurement Guide for Information Security. NIST Special Publication 800-55 Revision 1, 2008.
[12] George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Métayer, Rodica Tirtea, and Stefan Schiffner. Privacy and Data Protection by Design – from policy to engineering. Technical report, ENISA, 2014.
[13] Claudia Díaz, Omer Tene, and Seda F. Gürses. Hero or Villain: The Data Controller in Privacy Law and Technologies. Ohio State Law Journal, 74(6):923–964, 2013.
[14] Director, Research Directorate (DRD), Office of the Director, Defense Research and Engineering (DDR&E). Technology Readiness Assessment (TRA) Deskbook. Technical report, Department of Defense, USA, 2009.
[15] European Commission. Privacy Enhancing Technologies (PETs) – the existing legal framework. MEMO/07/159, May 2007.
[16] European Commission. Horizon 2020 – Work Programme 2014-2015, Annex G. Technology readiness levels (TRL). European Commission Decision C (2014)4995 of 22 July 2014. Technical report, 2014.
[17] European Commission, Directorate-General for Research and Innovation. Innovation – How to convert research into commercial success story? Part 2: Analysis of commercial successes induced by innovation in the field of industrial technologies. Study carried out for the European Commission by PricewaterhouseCoopers EU Services EESV, Belgium. Technical report, 2013.
[18] Oleksandr Gordieiev, Vyacheslav Kharchenko, and Mario Fusani. Evolution of Software Quality Models: Green and Reliability Issues. In Proceedings of the 11th International Conference on ICT in Education, Research and Industrial Applications: Integration, Harmonization and Knowledge Transfer (ICTERI 2015), volume 1356, pages 432–445, 2015.
[19] Marit Hansen, Jaap-Henk Hoepman, and Meiko Jensen. Towards measuring maturity of privacy-enhancing technologies. In Proc. APF 2015, 2015.
[20] Marit Hansen, Meiko Jensen, and Martin Rost. Protection Goals for Engineering Privacy. In 2015 International Workshop on Privacy Engineering (IWPE), IEEE eXplore, 2015.
[21] Jaap-Henk Hoepman. Privacy Design Strategies (extended abstract). In ICT Systems Security and Privacy Protection – 29th IFIP TC 11 International Conference, SEC, pages 446–459, 2014.
[22] ISACA. COBIT 5 for Information Security. 2012.
[23] Edouard Kujawski. Analysis and Critique of the System Readiness Level. IEEE T. Systems, Man, and Cybernetics: Systems, 43(4):979–987, 2013.

[24] Zbigniew Kwecka, William Buchanan, Burkhard Schafer, and Judith Rauhofer. "I am Spartacus" — Privacy Enhancing Technologies, Collaborative Obfuscation and Privacy as a Public Good. Artificial Intelligence and Law, 22(2):113–139, 2014.
[25] John C. Mankins. Technology Readiness Assessments: A Retrospective. Acta Astronautica, 65:1216–1223, 2009.
[26] José P. Miguel, David Mauricio, and Glen Rodríguez. A Review of Software Quality Models for the Evaluation of Software Products. International Journal of Software Engineering & Applications (IJSEA), 5(6):31–54, 2014.
[27] National Aeronautics and Space Administration (NASA). NASA Systems Engineering Handbook. NASA/SP-2007-6105 Rev1. Technical report, NASA, 2007.
[28] William L. Nolte. Did I Ever Tell You About the Whale? Or Measuring Technology Maturity. Charlotte, North Carolina: Information Age Publishing, 2008.
[29] Alison L. Olechowski, Steven D. Eppinger, and Nitin Joglekar. Technology Readiness Levels at 40: A Study of State-of-the-Art Use, Challenges, and Opportunities. MIT Sloan Research Paper No. 5127-15. Technical report, MIT, 2015.
[30] Martin S. Olivier. A Layered Architecture for Privacy-Enhancing Technologies. South African Computer Journal, 31:53–61, 2003.
[31] Organisation for Economic Co-operation and Development (OECD). Inventory of Privacy-Enhancing Technologies (PETs). Report DSTI/ICCP/REG(2001)1/FINAL, working party on information security and privacy. Technical report, 2002.
[32] A. Parasuraman and Charles L. Colby. An Updated and Streamlined Technology Readiness Index TRI 2.0. Journal of Service Research, 18(1):59–74, 2015.
[33] Florin Paun. The Demand Readiness Level Scale as New Proposed Tool to Hybridise Market Pull with Technology Push Approaches in Technology Transfer Practices. In David B. Audretsch, Erik E. Lehmann, Albert N. Link, and Alexander Starnecker, editors, Technology Transfer in a Global Economy, volume 28 of International Studies in Entrepreneurship, pages 353–366. Springer US, 2012.
[34] Brian J. Sauser, Jose E. Ramirez-Marquez, Devanandham Henry, and Donald DiMarzio. A System Maturity Index for the Systems Engineering Life Cycle. International Journal of Industrial and Systems Engineering, 3(6):673–691, 2008.
[35] Brian J. Sauser, Dinesh Verma, Jose E. Ramirez-Marquez, and Ryan Gove. From TRL to SRL: The Concept of Systems Readiness Levels. In Proc. Conference on Systems Engineering Research, Los Angeles, CA, 2006.
[36] Siemens AG Corporate Technology. Security by Design with CMMI for Development, Version 1.3. An Application Guide for Improving Processes for Secure Products. Technical report, CMMI Institute, 2013.
[37] James D. Smith. An Alternative to Technology Readiness Levels for Non-Developmental Item (NDI) Software. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences – HICSS '05, 2005.
[38] Stanley S. Stevens. On the Theory of Scales of Measurement. Science, 103(2684):677–680, 1946.
[39] Daniel Corin Stig, Ulf Högman, and Dag Bergsjö. Assessment of Readiness for Internal Technology Transfer – A Case Study. In INCOSE International Symposium, volume 21, pages 898–912, 2011.
[40] Paul F. Velleman and Leland Wilkinson. Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. The American Statistician, 47(1):65–72, 1993.
[41] Stefan Wagner. Software Product Quality Control. Springer, 2013.

ENISA

European Union Agency for Network
and Information Security
Science and Technology Park of Crete (ITE)
Vassilika Vouton, 700 13, Heraklion, Greece

Athens Office

1 Vass. Sofias & Meg. Alexandrou
Marousi 151 24, Athens, Greece

Catalogue Number: TP-02-15-974-EN-N

PO Box 1309, 710 01 Heraklion, Greece
Tel: +30 28 14 40 9710
info@enisa.europa.eu
www.enisa.europa.eu

ISBN: 978-92-9204-151-9
DOI: 10.2824/614444